Project 2119 | June 2022

SJSU SAN JOSÉ STATE UNIVERSITY

MINETA TRANSPORTATION INSTITUTE

Modeling and Predicting Geospatial Teen Crash Frequency

Sonu Mathew, PhD
Srinivas S. Pulugurtha, PhD, PE, FASCE
Sarvani Duvvuri, PhD

Mineta Transportation Institute

Founded in 1991, the Mineta Transportation Institute (MTI), an organized research and training unit in partnership with the Lucas College and Graduate School of Business at San José State University (SJSU), increases mobility for all by improving the safety, efficiency, accessibility, and convenience of our nation's transportation system. Through research, education, workforce development, and technology transfer, we help create a connected world. MTI leads the Mineta Consortium for Transportation Mobility (MCTM) funded by the U.S. Department of Transportation and the California State University Transportation Consortium (CSUTC) funded by the State of California through Senate Bill 1. MTI focuses on three primary responsibilities:

Research

MTI conducts multi-disciplinary research focused on surface transportation that contributes to effective decision making. Research areas include: active transportation; planning and policy; security and counterterrorism; sustainable transportation and land use; transit and passenger rail; transportation engineering; transportation finance; transportation technology; and workforce and labor. MTI research publications undergo expert peer review to ensure the quality of the research.

Education and Workforce

To ensure the efficient movement of people and products, we must prepare a new cohort of transportation professionals who are ready to lead a more diverse, inclusive, and equitable transportation industry. To help achieve this, MTI sponsors a suite of workforce development and education opportunities. The Institute supports educational programs offered by the Lucas Graduate School of Business: a Master of Science in Transportation Management, plus graduate certificates that include High-Speed and Intercity Rail Management and Transportation Security Management. These flexible programs offer live online classes so that working transportation professionals can pursue an advanced degree regardless of their location.

Information and Technology Transfer

MTI utilizes a diverse array of dissemination methods and media to ensure research results reach those responsible for managing change. These methods include publication, seminars, workshops, websites, social media, webinars, and other technology transfer mechanisms. Additionally, MTI promotes the availability of completed research to professional organizations and works to integrate the research findings into the graduate education program. MTI's extensive collection of transportation-related publications is integrated into San José State University's world-class Martin Luther King, Jr. Library.

Report 2119

Modeling and Predicting Geospatial Teen Crash Frequency

Sonu Mathew, Ph.D.

Srinivas S. Pulugurtha, Ph.D., P.E., F.ASCE

Sarvani Duvvuri, Ph.D.

June 2022

A publication of the
Mineta Transportation Institute
Created by Congress in 1991

College of Business
San José State University
San José, CA 95192-0219

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. 2119 22-24	**2. Government Accession No.**	**3. Recipient's Catalog No.**
4. Title and Subtitle Modeling and Predicting Geospatial Teen Crash Frequency		**5. Report Date** June 2022 **6. Performing Organization Code**
7. Authors Sonu Mathew: 0000-0003-2263-2749 Srinivas S. Pulugurtha: 0000-0001-7392-7227 Sarvani Duvvuri: 0000-0001-9621-2726		**8. Performing Organization Report**
9. Performing Organization Name and Address Mineta Transportation Institute College of Business San José State University San José, CA 95192-0219		**10. Work Unit No.** **11. Contract or Grant No.**
12. Sponsoring Agency Name and Address U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology University Transportation Centers Program 1200 New Jersey Avenue, SE Washington, DC 20590		**13. Type of Report and Period Covered** **14. Sponsoring Agency Code**
15. Supplemental Notes		

16. Abstract

This research project 1) evaluates the effect of road network, demographic, and land use characteristics on road crashes involving teen drivers, and, 2) develops and compares the predictability of local and global regression models in estimating teen crash frequency. The team considered data for 201 spatially distributed road segments in Mecklenburg County, North Carolina, USA for the evaluation and obtained data related to teen crashes from the Highway Safety Information System (HSIS) database. The team extracted demographic and land use characteristics using two different buffer widths (0.25 miles and 0.5 miles) at each selected road segment, with the number of crashes on each road segment used as the dependent variable. The generalized linear models with negative binomial distribution (GLM-based NB model) as well as the geographically weighted negative binomial regression (GWNBR) and geographically weighted negative binomial regression model with global dispersion (GWNBRg) were developed and compared. This research relied on data for 147 geographically distributed road segments for modeling and data for 49 segments for validation. The annual average daily traffic (AADT), light commercial land use, light industrial land use, number of household units, and number of pupils enrolled in public or private high schools are significant explanatory variables influencing the teen crash frequency. Both methods have good predictive capabilities and can be used to estimate the teen crash frequency. However, the GWNBR and GWNBRg better capture the spatial dependency and spatial heterogeneity among road teen crashes and the associated risk factors.

17. Key Words Teen crash, spatial dependency, spatial heterogeneity, linear model, predictive	**18. Distribution Statement** No restrictions. This document is available to the public through The National Technical Information Service, Springfield, VA 22161.		
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 52	**22. Price**

Form DOT F 1700.7 (8-72)

DOI: 10.31979/mti.2022.2119

Mineta Transportation Institute
College of Business
San José State University
San José, CA 95192-0219

Tel: (408) 924-7560
Fax: (408) 924-7565
Email: mineta-institute@sjsu.edu

transweb.sjsu.edu/research/2119

ACKNOWLEDGMENTS

The authors thank the staff of the Highway Safety Information System (HSIS) for providing the crash data required for this research. The authors also thank the staff of the Charlotte Regional Transportation Planning Organization (CRTPO), the Charlotte Department of Transportation, and the North Carolina Department of Transportation for providing the demographic, land use, road, and traffic data needed for this research.

CONTENTS

Acknowledgements vi

List of Figures ix

List of Tables x

Executive Summary 1

1. Introduction 3

    1.1 Problem Statement 4

    1.2 Research Objectives 4

    1.3 Organization of the Report 5

2. Literature Review 6

    2.1 Factors Influencing Teen Crashes 6

    2.2 Geospatial Crash Frequency Modeling 7

    2.3 Limitations of Past Research 8

3. Study Segments, Data Collection, and Data Processing 10

    3.1 Selection of Study Segments 10

    3.2 Data Collection 11

        3.2.1 Crash Data 11

        3.2.2 Road and Traffic Volume Data 11

        3.2.3 Demographic Data 11

        3.2.4 Land Use Data 11

    3.3 Data Processing 12

        3.3.1 Crash Data 12

3.3.2 Road and Traffic Volume Data 12
3.3.3 Demographic Data 13
3.3.4 Land Use Data 15
4. Methodology 17
4.1 Descriptive Analysis of Data 17
4.2 Correlation Analysis 17
4.3 Develop Crash Estimation Models Using Global and Local Regression Approaches 17
4.4 Validate the Model 19
5. Descriptive Statistics and Correlation Analysis 20
5.1 Descriptive Analysis 20
5.2 Check the Correlation between Explanatory Variables 21
6. Crash Frequency Model by Buffer Width 27
6.1 Models for 0.25-mile Buffer Width Dataset 27
6.2 Models for 0.50-mile Buffer Width Dataset 30
6.3 Comparison of Crash Estimation Models by the Buffer Width 32
7. Conclusions 34
7.1 Limitations and Scope for Future Work 35
Bibliography 36
About the Authors 41

LIST OF FIGURES

Figure 1. Selected Study Segments 10

Figure 2. Spatial Distribution of Teen Crashes in the Study Area 12

Figure 3. Extracting Demographic Variables 14

Figure 4. Land Use: Spatial Distribution 15

Figure 5. Teen Crash Frequency Over the Years: 2015–2017 20

Figure 6. Spatial Distribution of Traffic Exposure Variable Coefficients from the GWNBR Model 33

LIST OF TABLES

Table 1. Selected Explanatory Variables from Road and Traffic Volume Data 13

Table 2. Selected Variables from TAZ-Level Demographic Data 15

Table 3. Selected Variables from Land Use Data 16

Table 4. Descriptive Statistics: Crash Frequency, Road Network, and Traffic Volume 21

Table 5. Descriptive Statistics: Demographic Characteristics and Land Use Characteristics (0.25-mile buffer width dataset) 22

Table 6. Descriptive Statistics: Demographic Characteristics and Land Use Characteristics (0.50-mile buffer width dataset) 24

Table 7. Correlation Matrix: Selected Variables (0.25-mile buffer width dataset) 25

Table 8. Correlation Matrix: Selected Variables (0.25-mile buffer width dataset) 26

Table 9. GLM-Based NB Model (0.25-mile buffer width dataset) 28

Table 10. GWNBR Model (0.25-mile buffer width dataset) 28

Table 11. $GWNBR_g$ Model (0.25-mile buffer width dataset) 29

Table 12. GLM-Based NB Model (0.50-mile buffer width dataset) 30

Table 13. GWNBR Model (0.50-mile buffer width dataset) 31

Table 14. $GWNBR_g$ Model (0.50-mile buffer width dataset) 31

Table 15. Model Performance Comparison 32

Executive Summary

Crashes involving teen drivers have received a great deal of attention from researchers and practitioners in the field of traffic safety and crash prevention. According to the North Carolina Department of Transportation's "2019 Crash Facts," teen drivers were involved in 51,061 car crashes during the year, resulting in 80 teen deaths and 11,776 teen injuries in the state.

The findings from past research indicate that factors such as inexperience, nighttime and weekend driving, distracted driving, speeding, alcohol usage, and drug/substance use contribute to teen crash frequency. The driving environment is another significant aspect that has an influence on teen crashes. For example, the driving challenges in an urban area and a rural area are entirely different for a new teen driver. Various types of land uses have different teen trip generation and attraction potential. Teen travel activity is higher at schools, commercial areas, recreational centers, and similar places. While such location-based features play a key role and can explain the spatial heterogeneity in teen crash frequency, studies exploring the relationship between teen crash frequency and driving environment are found to be very limited. Therefore, the objectives of this research are: (1) to evaluate the effect of road network, demographic, and land use characteristics on road crashes involving teen drivers, and (2) to develop geospatial models and compare the predictability of local and global regression models in estimating the teen crash frequency.

To achieve the aforementioned objectives, two different modeling methods to estimate teen crash frequency were examined. They are generalized linear models with a negative binomial distribution (GLM-based NB model) and geographically weighted negative binomial regression models (GWNBR and $GWNBR_g$). The evaluation process was carried out by selecting 201 spatially distributed road segments in Mecklenburg County, North Carolina. Demographic and land use characteristics were extracted within 0.25 miles and 0.5 miles of each road segment. The explanatory variables were screened by computing and comparing Pearson correlation coefficients.

The findings from this research indicate that AADT, light commercial, light industrial, and recreational land uses have an effect on the teen crash frequency. The multicollinearity between the AADT and other road network characteristics led to the exclusion of variables such as speed limit, access, and functional class type from the final model. Demographic variables such as population, total employment, number of households, and pupils enrolled in public or private high schools also have a significant effect on the teen crash frequency.

The validation results indicate that the GWNBR model developed using a 0.25-mile buffer width dataset performed better than all the other models developed in this research. GWNBR can incorporate the effect of spatial variations in the data (by geographic location) while modeling the teen crash frequency. Overall, using the GWNBR model to locally estimate the teen crash frequency of a segment would improve the model fitting over the global NB regression model.

These findings are in line with the finding from the existing literature that GWNBR can better capture the spatial variations in data compared to global regression models.

The research outcomes can be used not only to estimate teen crash frequency by accounting for spatial variations in the explanatory variables but also to assist with planning, engineering, enforcement, and education activities. The research outcomes also imply that a higher emphasis should be placed on region-specific or localized crash prediction models.

1. Introduction

Road traffic crashes have emerged as a significant public health concern in this century. The fatalities and injuries from a road traffic crash affect the younger age group and their families more than any other age group. Hence, crashes involving teen drivers have received a great deal of attention from researchers and other practitioners in the field of traffic safety and crash prevention. According to the North Carolina Department of Transportation's "2019 Crash Facts," teen drivers were involved in 51,061 crashes during that year, resulting in 80 teen deaths and 11,776 teen injuries in the state (NCDOT 2020).

According to the Center for Disease Control and Prevention (CDC), inexperience, nighttime and weekend driving, not using seat belts, distracted driving, speeding, alcohol usage, and drug/substance usage are the major risk factors associated with teen crashes (CDC 2021). Teen drivers' lack of skills or lack of driving experience lead to severe crashes involving teen drivers (Hutchens et al. 2008).

Some researchers have shown that teen (or young) drivers are at the utmost risk of getting involved in crashes (Ma and Yan 2014; Zhou et al. 2015; Regev, Rolison, and Moutari 2018). Moreover, the risk is at a maximum within the first six months of their driving (Mayhew, Simpson, and Pak 2003). Over time, the crash risk may reduce as they gain more driving experience. On the contrary, some have researchers illustrated the higher risk associated with teen drivers during the transition from intermediate to full licensure (Curry et al. 2015; Li et al. 2018; Das et al. 2019). According to Voas and Kelley-Baker 2008, when the teens move away from parental control, under the influence of their peers, their exposure to risky driving may increase.

The driving environment is another factor that can influence teen crash frequency/severity. Various types of land uses have different trip generation and attraction potential (Pulugurtha, Duddu, and Kotagiri 2013; Mane and Pulugurtha 2020; Pulugurtha and Mathew 2021). Also, teen travel activity—as well as crash frequency—is higher at schools, commercial areas, recreational centers, etc. While such location-based features play a role in explaining the spatial heterogeneity of teen crashes, studies exploring the relationship between teen crashes and driving environment are found to be very limited.

The purpose of this research is to investigate the relationship between road network, demographic, and land use characteristics on teen crash frequency. The influence of location-specific indicators on teen crashes is difficult to capture from typical safety performance functions (SPFs) as they are based on global regression estimates (in which all data are used to develop a single model). The global regression estimates may not give accurate estimates at certain locations (Mathew and Pulugurtha 2021; Pulugurtha and Mathew 2021). For example, teen crash frequencies and other explanatory variables like traffic volume can vary significantly over similar road geometry within a study area. Methods based on Geographic Information System (GIS) methods, such as geographically weighted regression (GWR), can help generate localized SPFs (develop a model specific to the location based on data from its vicinity) by capturing the spatial variations in explanatory variables and thereby accurately estimate teen crash frequency. There will be a separate SPF developed at each road segment, representing the local relationships between teen crash frequency and other explanatory variables based on available data falling within the neighborhood of each target road segment.

This research focuses on the locally varying spatial association between teen crash frequency and road network, demographic, and land use characteristics using geographically weighted negative binomial regression models (GWNBR and $GWNBR_g$).

1.1 Problem Statement

The risk of getting involved in motor vehicle crashes is higher among teens compared to other age groups. Teens' driving behavior and safety can be influenced by policies, legislation, and other strategies. Better understanding the factors influencing teen crash frequency could, however, be useful in making such decisions.

At present, there are gaps in the research related to the factors influencing teen crash frequency and its estimation. Most of the previous studies have considered human factors, network characteristics, and environmental variables in the assessment process. Research studies related to the relationship between teen crashes and land use characteristics are found to be very limited. Besides, very few researchers in the past have assessed location-based information related to teen driver crashes. Most of the past researchers have used global models to assess the factors influencing teen crashes. In other words, the parameters in those models are fixed, and they neglect the spatial heterogeneity between the teen crashes and explanatory variables. To address this gap, this research employed negative binomial models based on geographically weighted regression (GWNBR and $GWNBR_g$) to model teen crash frequency.

1.2 Research Objectives

The objectives of this research are:

1. To evaluate the effect of road network, demographic, and land use characteristics on road crashes involving teen drivers, and,

2. To develop geospatial models and compare the predictive power of local and global regression models in estimating teen crash frequency.

1.3 Organization of the Report

The rest of the report is comprised of seven chapters. Chapter 2 summarizes literature related to the factors influencing teen crashes and the spatial analysis and modeling of road crashes. Chapter 3 presents the data collection and data processing methods adopted for this research. Chapter 4 provides a comprehensive framework to evaluate the effect of road network, demographic, and land use characteristics on teen crash frequency using various modeling approaches. Chapter 5 covers the descriptive statistics and correlation analysis. Chapter 6 summarizes the model development and comparison. Conclusions from this research and scope for future work are presented in Chapter 7.

2. Literature Review

This section synthesizes the previous research on the factors influencing teen crashes and the spatial analysis and modeling of road crashes. Both global and local models are considered, where global models are those in which all data are used simultaneously to develop a single model and local models involve the development of a model specific to the location.

2.1 Factors Influencing Teen Crashes

Vachal, Research Faculty, and Malchose (2009) analyzed North Dakota injury crash records of teen drivers to study the relationship between teen drivers' crash risk and the licensing age by considering driver, vehicle, and road factors. They evaluated the likelihood of teens getting involved in crashes when compared to experienced drivers. Dissanayanke and Amarsingha (2014) carried out a comparative assessment of crash causal factors among teen drivers and experienced drivers. Findings from their research indicated that factor such as speeding, failure to yield the right-of-way, disregarding traffic signs and signals, making improper turns or lane changes, aggressive driving, driving too slowly for the traffic, falling asleep, illness or fatigue, distracted driving, and not giving proper attention to driving have statistically significant effect on teen crash occurrence.

Das et al. (2019) studied the crash risk associated with teens under various licensing stages using the multivariate graphical method. They concluded that males with unrestricted licenses were the group with the highest risk, and they were involved in crashes more frequently than other drivers. Also, they pointed out the risky driving behavior of rural drivers and novice drivers with a learner's permit.

Traditionally, researchers have accounted for various environmental factors associated with teen crashes. Crash risks are higher at night compared to during the day (Voas and Kelley-Baker 2008). However, it is important to examine teen drivers' travel exposure intensity and other location indicators to understand the crash risk among teen drivers (Elander, West, and French 1993).

The variables affecting teen crashes may have a higher impact at some spatial locations and a smaller impact at some other locations. For example, teen crashes occurring at rural areas are more likely to lead to a fatal injury (Peek-Asa et al. 2010). Furthermore, crash frequency during school start times has been studied by many researchers Vorona et al. 2011; Vorona et al. 2014; Deka 2017; Foss, Smith, and O'Brien 2019). Past studies emphasize that spatial indicators of teen crashes need to be considered and evaluated. The spatial correlation assessment of socioeconomic and land use characteristics and teen crashes provides useful insights regarding the crash risks associated with certain location types.

Many researchers have illustrated the effect of spatial indicators on teen crashes and their frequency. The influence of factors such as alcohol consumption or drowsiness may be modulated by the driving environment (Goldstick et al. 2019); these factors and the driving environment are very likely to interact and produce increased crash risks at some point in time. A few research initiatives have also illustrated the effect of spatial factors such as socioeconomic characteristics (Aguero-Valverde and Jovanis 2006; Quddus 2008), traffic exposure (Li et al. 2018; Venkataraman, Ulfarsson, and Shankar 2013), and road characteristics (Quddus 2008; Guadamuz-Flores and Aguero-Valverde 2017; Kim, Svancara, and Kelley-Baker 2020) on teen crashes. Further, teen drivers are at greater risk on roads with grades/curves, intersections with no traffic signal, undivided roads, and unpaved roads (Kim, Svancara, and Kelley-Baker 2020).

Weast and Monfort (2021) evaluated the characteristics of vehicles driven by teens and adults killed in crashes from 2013 to 2017 and surmised that teens are more likely to be killed in older and smaller vehicles that are less often equipped with key safety features (Weast and Monfort 2021). Teen drivers are also more likely to be severely injured while driving sports utility vehicles and pickup trucks compared to passenger cars (Duddu, Kukkapalli, and Pulugurtha 2019).

Goldstick et al. (2019) assessed teen crash variations after implementing a Graduate Driver Licensing (GDL) program in Michigan. They examined the teen crash variations at the small-area level (alcohol outlets, movie theatres, and schools) after implementing GDL. They illustrated the spatial variation in the crash reduction. The crash reduction near alcohol outlets, school zones, and movie theatres after applying GDL was a good indicator of location-based assessment of teen crashes in the state of Michigan (Goldstick et al. 2019).

2.2 Geospatial Crash Frequency Modeling

To account for spatial non-stationarity or spatial heterogeneity across observations by allowing some or all parameters to vary (across space), researchers have put forth many modeling strategies. A few researchers have used random parameter models to address the unobserved heterogeneity in variables over space (Venkataraman, Ulfarsson, and Shankar 2013; Heydari et al. 2016). Some researchers have employed Bayesian spatial models for traffic safety analysis (Huang, Abdel-Aty, and Darwiche 2010; Darwiche 2009). The GWR method is also a proficient tool that can address the spatial heterogeneity in explanatory variables (Zhao and Park 2004; Du and Mulley 2006).

Pirdavani et al. (2013) employed the GWR method to explore the spatial variations and associations between the number of injury crashes and other explanatory variables. These included various indicators like traffic exposure, road network characteristics, and socioeconomic characteristics at the level of traffic analysis zones (TAZs). They showed the predictive accuracy of GWR models over traditional generalized linear models (GLMs) in crash prediction.

Hadayeghi, Shalaby, and Persaud (2010) studied the spatial variations and associations between the number of zonal collisions and potential transportation planning variables using the GWR method. They asserted the predictive performance of GWR over traditional GLMs.

Li et al. (2013) employed geographically weighted Poisson regression (GWPR) for county-level crash modeling in California. They compared the performance of GWR models with traditional GLMs. They illustrated the ability of GWPR models in capturing the spatially non-stationary relationships between crashes and predicting factors at the county level. Also, in that study, the GWPR method notably reduced the residual errors in predictions of fatal crashes over California counties.

Gomes, Cunto, and da Silva (2017) employed geographically weighted negative binomial regression (GWNBR) to develop zonal-level safety performance models in Fortaleza, Brazil. They conducted a comparative assessment of global crash prediction models, GWPR, and GWNBR. Traffic exposure, road network characteristics, socioeconomic characteristics, and land use characteristics were considered as the explanatory variables in their modeling. Both the spatial methods outperformed the traditional method. However, the capability of GWNBR to address the spatial heterogeneity and the overdispersion of the data makes it more sophisticated than GWPR (Gomes, Cunto, and da Silva 2017). Silva and Rodrigues (2014) illustrated that GWNBR converges with GWPR when the overdispersion parameter reduces whereas GWNBR converges with the global negative binomial (NB) regression model when the data display a stationary pattern.

Liu, Khattak, and Wali (2017) assessed the variations in SPFs for predicting crash frequency over space using GWNBR. They illustrated the importance of developing location-based spatial models (which can vary over space) in developing SPFs. Also, the predictive accuracy of GWNBR was better than that of the global NB models.

2.3 Limitations of Past Research

Most of the previous research on teen crashes assessed human factors, network characteristics, and environmental variables. Research studies related to the relationship between teen crashes and land use characteristics are found to be very limited. Besides, very few researchers in the past have assessed location-based information related to teen driver crashes. Most past researchers have used global models to assess the factors influencing teen crashes. In other words, the parameters in those models are fixed, and they neglect the spatial heterogeneity between the teen crashes and explanatory variables.

The GWR model allows the parameters to vary over space to capture the spatially varying relationship between teen crashes and the explanatory variables. All the previous research studies on GWR-based crash prediction models have illustrated the predictive accuracy of those models over traditional methods. Among the GWR models, GWPR and GWNBR are best suited for crash analysis and prediction. The GWNBR model can also account for the overdispersion that is usually observed in the crash data. However, the multicollinearity among the local coefficients is one point that needs special attention in the modeling process. This research aims to address the aforementioned gaps.

3. Study Segments, Data Collection, and Data Processing

This chapter presents the study segments, data collection, and data processing methods used in this research. The descriptions of the chosen explanatory variables after data processing are also included in this chapter.

3.1 Selection of Study Segments

Mecklenburg County in North Carolina was chosen as the study area. The state-maintained road segments were identified for crash prediction modeling. A total of 201 road segments were identified based on the annual average daily traffic (AADT) data available for the study area. The selected road segments are geographically distributed and cover all area types in Mecklenburg County. Figure 1 illustrates the selected road links for this research.

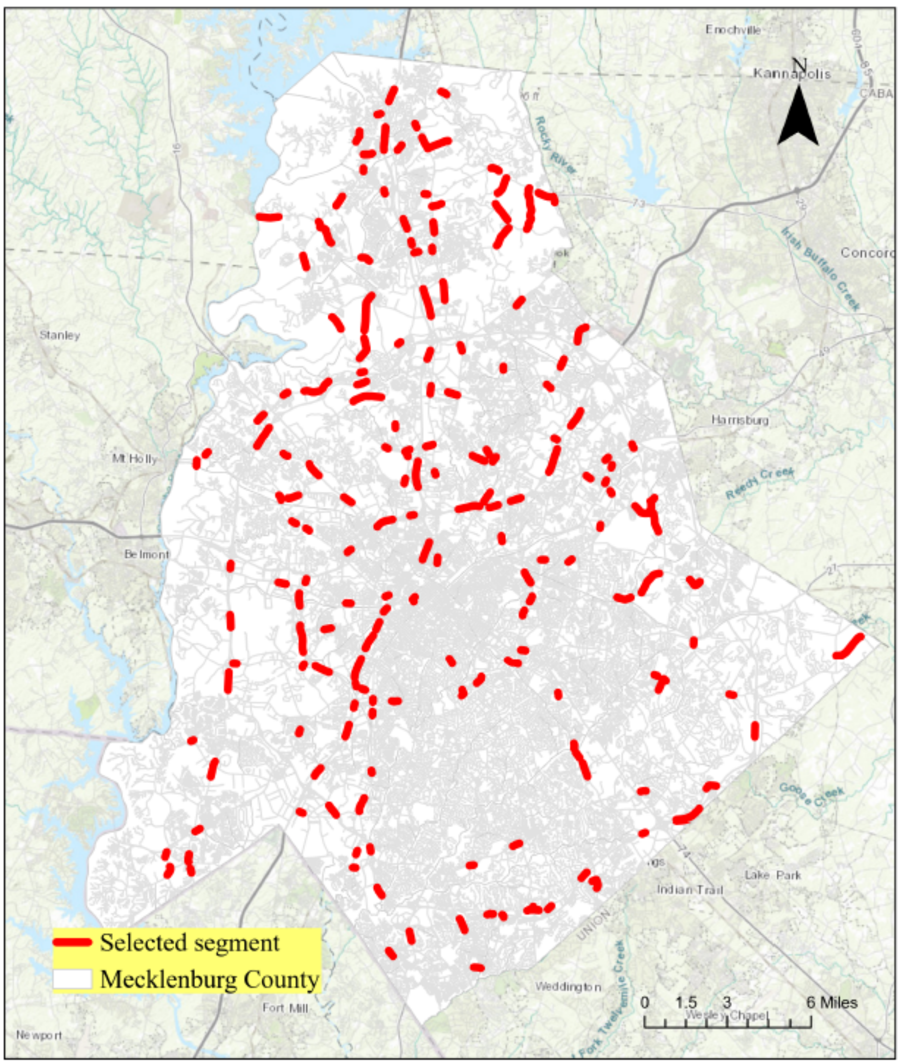

Figure 1. Selected Study Segments

3.2 Data Collection

Four sets of data were considered for the teen frequency modeling: crash data, road and traffic volume data, demographic data, and parcel-level land use data. Crash data for the years 2015 to 2017 were obtained from the Highway Safety Information System (HSIS). The demographic data for this research were obtained from the Charlotte Regional Transportation Planning Organization (CRTPO). The parcel-level land use development data were downloaded in geospatial (shapefile) format from the open mapping portal of Mecklenburg County.

3.2.1 Crash Data

The HSIS database consists of all the reported crashes on state-maintained facilities across the state of North Carolina. The raw database has four sub-files: vehicle data, occupant data, accident data, and road inventory data. As the purpose of the research is to model teen crash frequency, all pedestrian and bicycles crashes were removed from the database. Non-motorized crashes were removed from the database, as were crashes involving drivers aged over 20 years or younger than 15 years. The final resulting database consists of crash details involving drivers who are 15–19 years old (considered teen drivers in this research).

3.2.2 Road and Traffic Volume Data

The road-network-related information was obtained in a geospatial format—a digital file from the road inventory database of the NCDOT that describes a subset of characteristics of the state road network. The state road system consists of interstates, US and NC routes, secondary roads, ramps, and all non-state roads maintained in North Carolina. The traffic data include the observations associated with traffic count stations in Mecklenburg County between 2002 and 2017.

3.2.3 Demographic Data

The demographic data obtained from the Charlotte Regional Transportation Planning Organization (CRTPO) contained information at the TAZ level. There are 1,170 TAZs in the study area. The TAZ file was a TransCAD geospatial file consisting of variables such as population, number of households, total employment, number of pupils enrolled in public or private schools, number of pupils in public or private colleges and universities, and so on.

3.2.4 Land Use Data

The raw dataset consists of 115 different categories of land use. These parcel-level data include information related to the year of the structure and the heated area (living area of any kind of land use).

3.3 Data Processing

The data processing was carried out at various levels using software tools such as ArcGIS 10.6.2 and ArcGIS Pro. Each level of data processing is summarized in the following sections.

3.3.1 Crash Data

The sub-files (vehicle data, occupant data, accident data, and road inventory data) were combined into a single file using the a common field named case number (CRASH ID).The crash data obtained from the HSIS have milepost and road-related information. The linear referencing option in the ArcGIS was used for spatial referencing. This feature locates the point events along a route based on the milepost information. Teen crashes that are on the selected state-maintained roads in Mecklenburg County are shown in Figure 2.

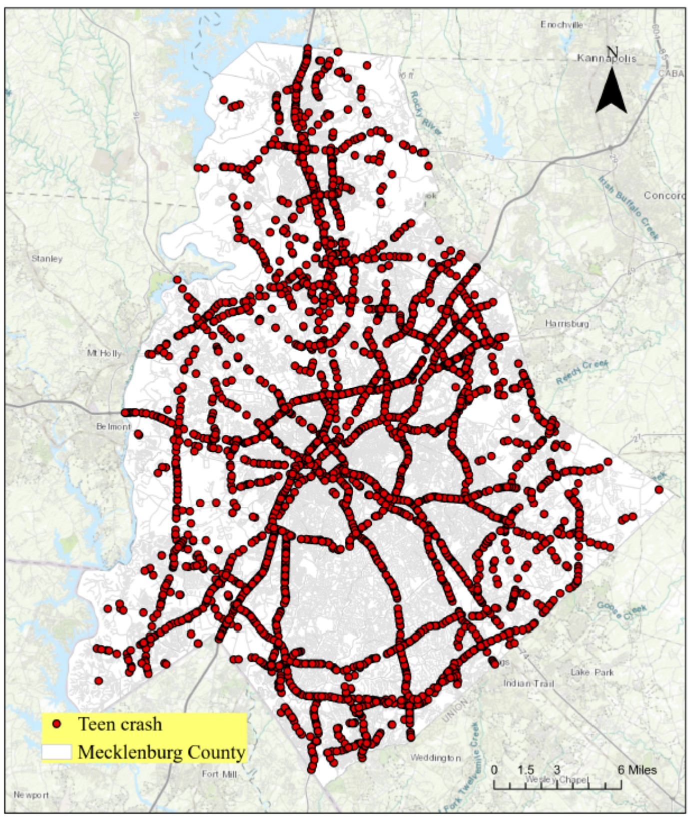

Figure 2. Spatial Distribution of Teen Crashes in the Study Area

3.3.2 Road and Traffic Volume Data

The traffic volume shapefile was overlaid with the road characteristics data obtained from NCDOT to identify the count-based AADT. The road-related variables extracted from the crash database and road characteristics shapefile are summarized in Table 1.

3.3.3 Demographic Data

The traffic volume and the number of crashes could increase with an increase in population. Hence, the number of households, population, total employment, number of pupils enrolled in public or private schools, and number of pupils in public or private colleges and universities were chosen as the potential variables from the demographic data for the study area. Based on the network buffers created around each road segment (0.25 miles and 0.5 miles), the weighted average values of demographic and socioeconomic variables are extracted as shown in Figure 3.

Table 1. Selected Explanatory Variables from Road and Traffic Volume Data

Variable	**Description**
Annual average daily traffic (AADT)	Total volume of vehicle traffic during a year divided by 365 days
Segment length	Length of the segment (in miles)
Speed limit	Posted speed limit (in mph)
Number of lanes	Number of lanes in the selected segment
Direction	Traffic in both directions or only one direction
Functional class—Freeways	Principal arterial: interstate Principal arterial: freeways and expressways
Functional class—Major arterial	Principal arterial: other
Functional class—Minor arterial	Minor arterial
Functional class—Collector or local	Collector roads or local roads
Access—Full	Full access control
Access—Partial	Partial access control
Access—None	No access control

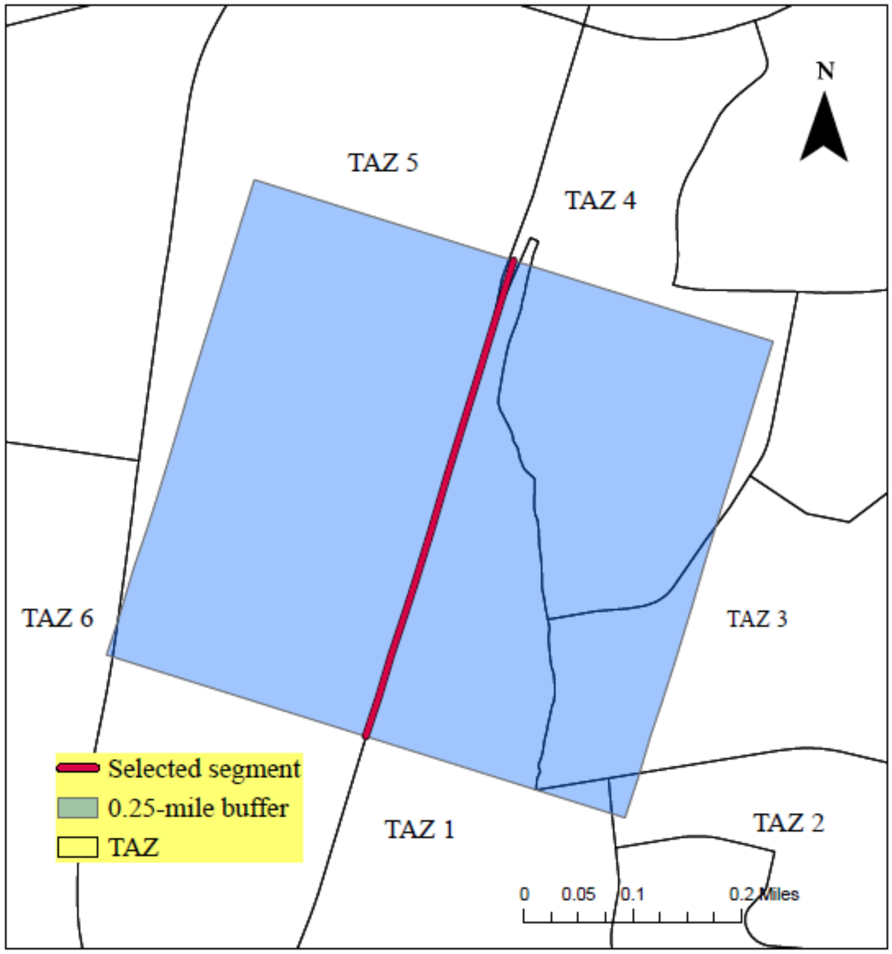

Figure 3. Extracting Demographic Variables

Figure 3 shows that the selected road segment buffer covers portions of six TAZs. The proportion of each TAZ area was determined by intersecting the buffer with each TAZ. The selected demographic variables mentioned in Table 2 are assumed to be uniform across the TAZ. Therefore, the proportion of TAZ area that falls in a buffer was considered to estimate the weighted average estimates of the demographic variables. For example, the population of buffer b (P_b) was estimated using Equation (1).

$$Population\ (P_b) = \sum_t \frac{A_{t,b}}{A_t} \times P_t \qquad (1)$$

where P_b= population of buffer b, $A_{t,b}$ = actual area of TAZ t in buffer b, A_t = area of TAZ t, and P_t = population of TAZ t.

A similar approach was adopted to estimate the other demographic variables in the study area. The descriptions of all the demographic variables considered in this research are included in Table 2.

Table 2. Selected Variables from TAZ-Level Demographic Data

Variable	Description
Population	Total population
Number of households	Number of households in the TAZ
Total employment	Total number of employees
Student—HS	Number of pupils enrolled in public or private high schools
Student—CU	Number of pupils in public or private colleges and universities

3.3.4 Land Use Data

The land use characteristics around the study segments were also identified using the buffer approach. Land use characteristics were captured within the 0.25-mile and 0.50-mile buffers generated around each selected road segment. As the crash data is for the years 2015-2017, the land use developments after the year 2017 was removed from the database. The land use distribution of a portion of the study area is shown in Figure 4.

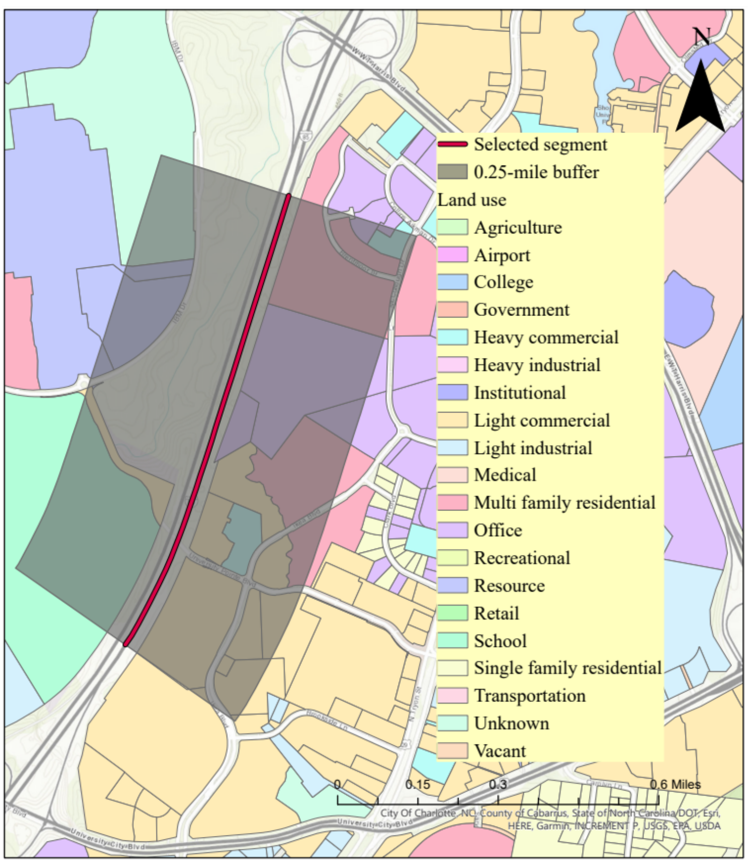

Figure 4. Land Use: Spatial Distribution

The descriptions of all the land-use-related variables are listed in Table 3.

Table 3. Selected Variables from Land Use Data

Variable	Description
Agriculture	Land use parcels such as farms, commercial forestry, pasture, tree farms, etc.
School	School parcels (public, municipal, or private)
College	College/university parcels (public-/private-owned institutions)
Government	Land use parcels owned by state or municipal authorities
Institutional	Parcels where services are provided for the community, such as daycare, church, etc.
Medical	Hospitals, pharmacy, and medical-based parcels
Light commercial	Constrained to community-based services such as fast-food centers, commercial stores (like laundry), service stations, etc.
Heavy commercial	Commercial land use parcels such as shopping mall, furniture stores, etc.
Light industrial	Light manufacturing-based industries and warehouse-based land use parcels
Heavy industrial	Industry-based land use parcels involving small manufacturing services, wastewater treatment plans, etc.
Single-family	Residential: fully detached, semi-detached, row house, or townhome
Multi-family	Residential: condominium houses, multi-dwelling residential units, apartment buildings, and mobile home parks
Office	Land use parcels mainly for administrative, office-related or business parks
Recreational	Land use parcels such as bowling alley, theatre, golf course, etc.
Resource	Resource land use parcels include wetlands, creeks, etc.
Retail	Parcels allocated for retail purposes; includes convenient/department store, supermarket, etc.
Transportation	Parcels such as trucking rest areas, right-of-way, or transportation/parking services
Unknown	Unknown parcels
Vacant	No land use category is allocated

4. Methodology

This chapter presents the methodology adopted in this research.

4.1 Descriptive Analysis of Data

A descriptive analysis was conducted to understand the influence of selected explanatory variables on the teen crash frequency. The minimum, mean, maximum, and standard deviation of each selected variable was computed and examined. The analysis was separately carried out for the 0.25-mile buffer dataset and the 0.50-mile buffer dataset.

4.2 Correlation Analysis

Pearson's correlation coefficients were computed to perform a correlation analysis. The Pearson correlation coefficient illustrates the strength of the linear relationship between two variables. If the correlation coefficient fell within a 95% confidence level, that variable was further considered in the assessment process. If two explanatory variables are correlated to each other, the variable with a higher Pearson correlation coefficient with teen crash frequency is chosen for modeling.

A positive value of the Pearson correlation coefficient indicates that teen crash frequency increases with an increase in the related explanatory variable (for example, the speed limit or the number of lanes), and a negative value indicates that teen crash frequency decreases with an increase in the related variable.

4.3 Develop Crash Estimation Models Using Global and Local Regression Approaches

The teen crash frequency was considered as the dependent variable. Road, demographic, and land use characteristics were considered as the explanatory variables. Separate models were developed for data extracted using the 0.25-mile buffer width and 0.50-mile buffer width.

Most previous studies have used global count-based regression models to estimate the crash frequency on any road segment—an approach which assumes that a single model can account for or adequately describe the entire study area. Generally, GLMs with log-normal distribution, Poisson distribution, and NB distribution are widely used for the same purpose. However, spatial heterogeneity is one issue that needs to be addressed in the model development process. GWR models can accommodate the non-stationarity spatial data to a good extent. This research, therefore, sought to evaluate the predictability of both NB-based GLM (Global) and GWNBR (local) crash prediction models.

The general form of the NB log-link distribution is shown in Equation (2).

$$Y = e^{\beta_0+\beta_1 X_1+ \beta_2 X_2+\cdots+\beta_n X_n} \quad (2)$$

Here, Y = estimated number of teen crashes per year, the $\beta_1,\ \beta_2, \ldots, \beta_n$ series = coefficients of the explanatory variables, and the $X_1, X_2, \ldots, X_n$ series = explanatory variables considered in the model development.

Similarly, the GWNBR model considers the log-link distribution at the local level. Equation (3) shows the general functional form of GWNBR.

$$Y = NB[t_j \exp\,(\textstyle\sum_k \beta_k(u_j, v_j) X j_k), \theta(u_j, v_j)] \quad (3)$$

Here, (u_j, v_j) are the location coordinates of data point j, NB represents the negative binomial distribution, t_j is an offset variable, θ is the overdispersion parameter, and β_k is the parameter related to the explanatory variable X_k.

Gaussian and bi-squared functions, discussed by Fotheringham et al. (2002), are used to assign weights (W_{ij}). The functional form of Gaussian and bi-squared functions, respectively, are provided in equations (4) and (5).

$$W_{ij} = exp\left[-0.5(d_{ij}/h)^2\right] \quad (4)$$

$$W_{ij} = \begin{cases} \left[1 - (d_{ij}/h)^2\right] & \text{if } d_{ij} \leq h \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

where d_{ij} is the Euclidean distance between the selected road segments (center to center), and h is the bandwidth.

Da Silva and Rodrigues (2014) proposed a GWNBR model with a global overdispersion parameter ($GWNBR_g$). In the $GWNBR_g$ model, the spatial variations mentioned in equation (3) are only allowed to $\beta_k(u_j, v_j)$, and the overdispersion parameter (θ) is used from the GLM-based NB model (da Silva and Rodrigues 2014). The modeling of teen crash frequency using GWNBR and $GWNBR_g$ was carried out with a set of a SAS® macro for Geographically Weighted Negative Binomial Regression (SAS/IML©macros) developed by da Silva and Rodrigues (2016).

The key criterion for selecting a distribution for modeling is to check the dispersion parameter. If the data turn out not to be overdispersed, the Poisson regression models are used because the Poisson model assumes the mean must be equal to the variance. NB distribution models are capable of modeling data with overdispersion. In the present study, the variance is greater than the mean for the considered dataset (the mean of the teen crash frequency is 3.43 whereas the variance is 10.37). Hence, the Poisson regression with the restrictive assumption of the equality of mean and variance may not be suitable for the model development. As a result, a GLM-based NB regression is suitable for this dataset.

The best fit was assessed using the Akaike Information Criterion (AIC) and corrected Akaike Information Criterion (AICc). The lower value of AIC and AICc indicate a better fit of the model. Overall, a total of six models were developed and compared.

4.4 Validate the Model

Data for 49 selected road segments were set aside for validation purposes. These links were randomly selected while ensuring that they represented a geographically/spatially distributed sample across the study area. Each of the developed models was validated using the mean absolute deviation and root mean squared error (RMSE). MAPE and RMSE are expressed as shown in equations (6) and (7).

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{C_j} - C_j\right| \tag{6}$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\widehat{C_j}-C_j)^2}{n}} \tag{7}$$

where n = the number of observations, $\widehat{C_j}$= the predicted number of teen crashes, and C_j= the observed number of teen crashes.

5. Descriptive Statistics and Correlation Analysis

This chapter presents the results obtained from the descriptive statistics and correlation analysis.

5.1 Descriptive Analysis

A total of 7,804 teen crashes that occurred in Mecklenburg County were considered for the model development and validation. Figure 5 shows the yearly variation in the number of crashes involving teen drivers in the study area. It is observed that the number of crashes involving teens has increased from 2015 to 2016. However, a reduction in the number of teen crashes was observed in the year 2017.

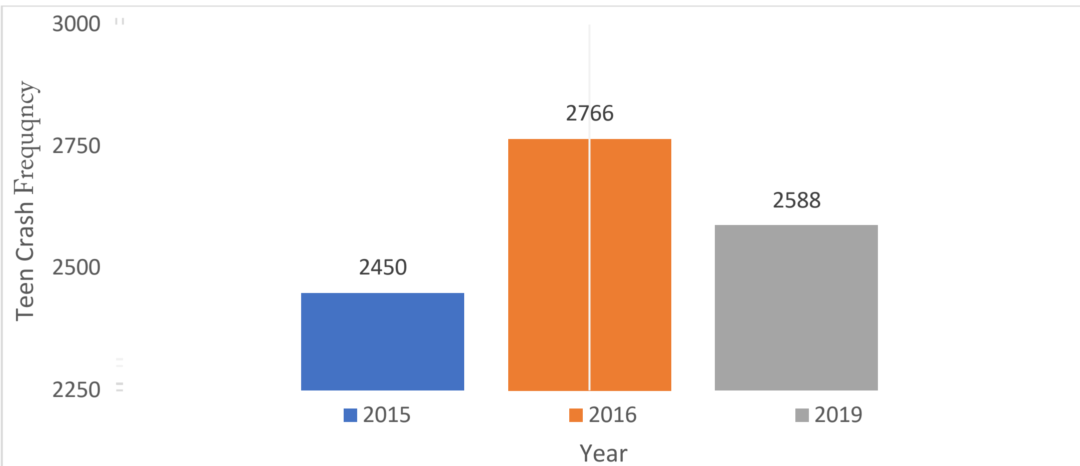

Figure 5. Teen Crash Frequency Over the Years: 2015–2017

Table 2 shows the descriptive statistics of crashes, road characteristics, and traffic volume/AADT variables considered in this research. The descriptive statistics consist of all 201 road segments considered for model development and validation. The selected road segments are distributed proportionally in different functional groups. However, most of the road segments in the state-maintained roads are in the urban area. Therefore, the area type indicator is not considered in the modeling process.

Table 4. Descriptive Statistics: Crash Frequency, Road Network, and Traffic Volume

Variable	Minimum	Mean	Maximum	Std. Deviation
Crash frequency	1	3.43	18	3.23
Annual average daily traffic (AADT)	70	33,197	181,000	38,621
Segment length	0.1	0.31	1.59	0.25
Speed limit	25	45.24	70	9.88
Number of lanes	2	3.57	10	1.75
Direction (one-way=1, bidirectional=2)	0	0.3	1	0.46
Functional class—Freeways	0	0.15	1	0.35
Functional class—Major arterial	0	0.24	1	0.43
Functional class—Minor arterial	0	0.38	1	0.49
Functional class—Collector or local	0	0.24	1	0.43
Access—Full	0	0.79	1	0.41
Access—partial	0	0.07	1	0.25
Access—None	0	0.15	1	0.35

From Table 4, one can notice that the teen crash frequency ranges from 1 to 18 in the selected road segments. The selected road segments are distributed proportionally in different functional groups. However, most of the road segments in the state-maintained roads are in the urban area. Therefore, the area type indicator is not considered in the modeling process.

The descriptive statics of demographic and land use characteristics for the 0.25-mile and 0.50-mile buffer width datasets are presented in Table 5 and Table 6, respectively.

5.2 Check the Correlation between Explanatory Variables

A Pearson correlation coefficient matrix was developed to examine the correlation between the explanatory variables. The analysis was carried out separately for each dataset (i.e., the 0.25-mile and 0.50-mile buffer width datasets) to examine and minimize the effect of multicollinearity. Table 7 summarizes the Pearson correlation coefficient analysis results of 0.25-mile buffer width dataset. The variables which have a statistically significant correlation (at a 95% confidence level) with the teen crash frequency are included in Table 7.

Road characteristics such as AADT, segment length, speed limit, number of lanes, freeway, and full access are positively correlated with the teen crash frequency. However, roads without any access are negatively correlated with the teen crash frequency. Land use characteristics such as light commercial, light industrial, and recreational areas are positively correlated with the teen crash frequency. Similarly, population, number of households, number of pupils enrolled in public or private high schools, and total employment are also positively correlated with the teen crash frequency.

Table 5. Descriptive Statistics: Demographic Characteristics and Land Use Characteristics (0.25-mile buffer width dataset)

Variable	Minimum	Mean	Maximum	Std. Deviation
Demographic characteristics (as multiples of one thousand)				
Population	0	0.23	1.27	0.20
Number of households	0	0.09	0.48	0.08
Total employment	0	0.22	2.70	0.31
# of pupils in public or private schools	0	0.02	0.53	0.07
# of pupils in public or private colleges and universities	0	0.04	2.45	0.26
Land use characteristics (in heated area per thousand square feet)				
Agriculture	0	0.03	3.44	0.26
College	0	0.01	1.07	0.08
Government	0	0.05	6.01	0.47
Heavy commercial	0	0.48	19.86	2.34
Heavy industrial	0	0	0.03	0
Institutional	0	0.46	23.4	2.18
Light commercial	0	2.71	28.54	5.1
Light industrial	0	3.34	57.34	9.26
Medical	0	0.03	3.83	0.28
Multi-family residential	0	2.2	42.6	5.39
Office	0	0.82	22.99	3.12
Recreational	0	0.04	2.66	0.28
Resource	0	0.16	16.31	1.37
Retail	0	0.06	3.27	0.35
School	0	0.24	12.76	1.22
Single-family residential	0	8.5	69.64	11.39
Unknown	0	1.79	80.1	10.23

Table 8 summarizes the Pearson correlation coefficient analysis results of the 0.50-mile buffer width dataset. Land use characteristics such as light commercial, light industrial, high commercial, and recreational areas are positively correlated with the teen crash frequency. Similarly, population, number of households, and total employment are also positively correlated with the teen crash frequency.

Some of the explanatory variables are highly correlated with each other. For example, the speed limit and the functional class are highly correlated with each other. Similarly, the number of lanes and the speed limit are also correlated with each other. Only one of the correlated variables was considered for modeling to reduce the effect of multicollinearity.

Table 6. Descriptive Statistics: Demographic Characteristics and Land Use Characteristics (0.50-mile buffer width dataset)

Variable	Minimum	Mean	Maximum	Std. Deviation
Demographic characteristics (as multiples of one thousand)				
Population	0	0.50	2.90	0.03
Number of households	0	0.19	1.14	0.01
Total employment	0	0.45	7.99	0.05
# of pupils in public or private schools	0	0.03	0.80	0.01
# of pupils in public or private colleges and universities	0	0.08	4.96	0.04
Land use characteristics (in heated area per thousand square feet)				
Agriculture	0	0.29	34.83	0.20
College	0	0.30	29.67	0.19
Government	0	0.59	60.11	0.33
Heavy commercial	0	9.01	303.74	2.62
Heavy industrial	0	0.01	0.57	0.00
Institutional	0	6.83	233.98	1.62
Light commercial	0	36.97	491.61	4.84
Light industrial	0	76.88	1810.02	14.34
Medical	0	0.32	38.36	0.19
Multi-family residential	0	36.63	528.45	5.51
Office	0	17.20	627.52	4.90
Recreational	0	0.52	26.89	0.20
Resource	0	3.18	181.17	1.51
Retail	0	0.81	66.76	0.41
School	0	3.79	132.19	1.06
Single-family residential	0	226.74	1279.73	17.97
Unknown	0	2.54	95.14	0.86

Table 7. Correlation Matrix: Selected Variables (0.25-mile buffer width dataset)

Variables	Crashes	AADT	One-way	Freeway	Fullacs	Noacs	Speedlim	# lanes	Pop	HH	Emp
Annual Average Daily Traffic (AADT)	0.41										
Direction (One-way)	0.27	0.70									
Functional class—Freeway (Freeway)	0.29	0.90	0.64								
Access—Full (Fullacs)	0.29	088	0.64	0.96							
Access—None (Noacs)	-0.26	-080	-0.77	-0.80	-0.80						
Speed limit (Speedlim)	0.16	0.68	0.59	0.79	0.79	-0.76					
Number of lanes (#lanes)	0.33	0.75	0.60	0.61	0.59	-0.60	0.41				
Population (Pop)	0.27	0.22	0.14	0.21	0.19	-0.18	0.17	0.20			
Number of households	0.29	0.22	0.14	0.20	0.18	-0.18	0.16	0.21	0.98		
Total Employment (Emp)	0.39	0.44	0.35	0.37	0.39	-0.39	0.25	0.54	0.23	0.27	
# of pupils in public or private schools (HS)	0.15							0.17			0.17
Light commercial land use (Lcom)	0.40		0.17					0.29			0.44
Light industrial land use (Lind)	0.16	0.24		0.17	0.19	-0.16		0.17			0.29
Recreational (Recr)	0.18								0.14	0.18	

Note: Blank cells indicate no statistically significant correlation at a 95% confidence level. Columns with no significant correlations were excluded.

Table 8. Correlation Matrix: Selected Variables (0.25-mile buffer width dataset)

Variables	Crashes	AADT	One-way	Freeway	Fullacs	Noacs	Speedlim	# lanes	Pop	HH	Emp	Lcom
Annual Average Daily Traffic (AADT)	0.41											
Direction (One-way)	0.27	0.70										
Functional class—Freeway (Freeway)	0.29	0.90	0.64									
Access—Full (Fullacs)	0.29	088	0.64	0.96								
Access—None (Noacs)	-0.26	-080	-0.77	-0.80	-0.80							
Speed limit (Speedlim)	0.16	0.68	0.59	0.79	0.79	-0.76						
Number of lanes (#lanes)	0.33	0.75	0.60	0.61	0.59	-0.60	0.41					
Population (Pop)	0.33	0.28	0.16	0.25	0.23	-0.20	0.16	0.23				
Number of households	0.35	0.27	0.15	0.25	0.23	-0.20	0.16	0.25	0.98			
Total Employment (Emp)	0.42	0.48	0.33	0.39	0.40	-0.39	0.26	0.51	0.25	0.29		
Light commercial land use (Lcom)	0.36	0.17	0.18					0.35			0.56	
High commercial land use (Hcom)	0.25	0.26	0.17	0.18	0.18	-0.30					0.53	
Light industrial land use (Lind)	0.34	0.32	0.18	0.23	0.25	-0.24		0.22				
Recreational (Recr)	0.17											

Note: Blank cells indicate no statistically significant correlation at a 95% confidence level; Columns with no significant correlations were excluded.

6. Crash Frequency Model by Buffer Width

This chapter presents the results obtained from the developed global (GLM-based NB model) and local (GWNBR and $GWNBR_g$) models to examine the relationship between teen crash frequency and road, demographic, and land use characteristics. Separate models were developed for data extracted using a 0.25-mile buffer width and a 0.50-mile buffer width.

The number of crashes per year involving teen drivers in a segment is considered as the dependent variable. The GWNBR was used to capture the localized effect of the explanatory variables. This method allows the conditional relationship between the dependent variable and the different explanatory variables to vary across different spatial scales. In other words, teen crash frequency at a selected road segment within proximity to the target road segment has greater influence on the estimates of the regression coefficients than those areas far apart. The significant explanatory variables from the GLM-based NB model were used to develop the GWNBR models. The GWNBR builds a local regression equation for each feature in the dataset.

In GWNBR, it is necessary to decide an optimal bandwidth (number of neighbors for modeling) for model fitting. A high number of neighbors indicates a greater smoothing of the local coefficients, and a smaller number of neighbors indicates a greater degree of heterogeneity in local coefficients. The methodology explained in da Silva and Rodrigues (2016) were employed to determine the optimum bandwidth. The AIC statistics approach (minimizing AIC) was chosen to identify the optimum bandwidth. The lower value of AIC and AICc indicate a better fit of the model.

6.1 Models for 0.25-mile Buffer Width Dataset

The GLM-based NB regression model results from the 0.25-mile buffer width dataset are summarized in Table 9.

The sign of the coefficient indicates the role of each explanatory variable on the teen crash frequency. In the case of the GLM-based NB model, the AADT has a positive effect on the teen crash frequency. Similarly, the teen crash frequency is higher at locations with light commercial and light industrial land uses. The length and AADT are entered into the model as an offset to measure the exposure as a rate. The demographic variables such as number of households and number of pupils enrolled in public or private high schools have a positive effect on the teen crash frequency.

As mentioned earlier, all the variables identified from the GLM-based NB model (AADT, number of households, number of pupils in public or private schools, light commercial land use, and light industrial land use) were used for developing the GWNBR models. The GWNBR model results are summarized in Table 10.

Table 9. GLM-Based NB Model (0.25-mile buffer width dataset)

Parameter	Coeff. (B)	Std. Error	p-value
Intercept	0.672	0.11	<0.01
Segment length	-0.508	0.34	0.13
Log transformed AADT	0.006	0.01	<0.01
Number of households	2.003	0.70	<0.01
# of pupils in public or private schools	1.710	0.77	0.03
Light commercial land use	0.004	0.11	<0.01
Light industrial land use	0.001	<0.01	0.04
Overdispersion			0.200
AIC			644.75
AICc			645.77
RMSE			0.81
MAD			0.61

Table 10. GWNBR Model (0.25-mile buffer width dataset)

Parameter	β_{min} (Std. Error)	β_{25} (Std. Error)	β_{50} (Std. Error)	β_{mean} (Std. Error)	β_{75} (Std. Error)	β_{max} (Std. Error)
Intercept	0.608 (0.14)	0.629 (0.15)	0.686 (0.15)	0.681 (0.16)	0.733 (0.16)	0.739 (0.19)
Segment length	-0.676 (0.42)	-0.576 (0.45)	-0.539 (0.49)	-0.547 (0.49)	-0.507 (0.53)	-0.478 (0.63)
Log transformed AADT	0.007 (<0.01)	0.007 (<0.01)	0.007 (<0.01)	0.007 (<0.01)	0.007 (<0.01)	0.008 (<0.01)
Number of households	1.207 (0.92)	1.411 (0.93)	1.853 (0.98)	1.857 (0.97)	2.289 (1.01)	2.485 (1.12)
# of pupils in public or private schools	0.767 (0.97)	1.078 (1.03)	1.268 (1.08)	1.309 (1.13)	1.549 (1.20)	1.856 (1.54)
Light commercial land use	0.003 (<0.01)	0.004 (<0.01)	0.004 (<0.01)	0.004 (<0.01)	0.005 (<0.01)	0.005 (<0.01)
Light industrial land use	0.001 (<0.01)	0.001 (0.01)	0.001 (<0.01)	0.001 (<0.01)	0.001 (<0.01)	0.001 (<0.01)
Overdispersion parameter	0.141	0.186	0.198	0.195	0.209	0.221
AIC						645.75
AICc						647.22
RMSE						0.76
MAD						0.56

The minimum, 25th percentile, median, mean, 75th percentile, and maximum coefficient (β) are shown in Table 10. The GWNBR model shows that the average coefficients are close to the GLM-based NB regression coefficients. The signs of the coefficients are also consistent in both modeling approaches.

Regarding the AIC and AICc values, the GLM-based NB model and GWNBR model showed similar performance. While looking into the MAD and RMSE, the GWNBR model performed marginally better than the GLM-based NB model.

The $GWNBR_g$ model results are summarized in Table 11. Similar to the previous models, all the variables identified from the GLM-based NB models are considered in $GWNBR_g$. Also, a constant overdispersion parameter of 0.200 (obtained from the GLM-based NB model) is used in the model development process.

Table 11. $GWNBR_g$ Model (0.25-mile buffer width dataset)

Parameter	β_{min} (Std. Error)	β_{25} (Std. Error)	β_{50} (Std. Error)	β_{mean} (Std. Error)	β_{75} (Std. Error)	β_{max} (Std. Error)
Intercept	0.608 (0.14)	0.629 (0.15)	0.681 (0.16)	0.685 (0.15)	0.733 (0.16)	0.740 (0.21)
Segment length	-0.672 (0.43)	-0.578 (0.45)	-0.546 (0.50)	-0.535 (0.50)	-0.507 (0.53)	-0.478 (0.63)
Log transformed AADT	0.007 (<0.01)	0.007 (<0.01)	0.007 (<0.01)	0.007 (<0.01)	0.007 (<0.01)	0.008 (<0.01)
Number of households	1.218 (0.90)	1.400 (0.95)	1.854 (0.99)	1.845 (0.99)	2.302 (1.02)	2.476 (1.23)
# of pupils in public or private schools	0.767 (0.99)	1.082 (1.05)	1.311 (1.14)	1.27 (1.09)	1.551 (1.19)	1.857 (1.51)
Light commercial land use	0.004 (<0.01)	0.004 (<0.01)	0.004 (<0.01)	0.004 (<0.01)	0.005 (<0.01)	0.005 (<0.01)
Light industrial land use	0.001 (<0.01)	0.001 (<0.01)	0.001 (<0.01)	0.001 (<0.01)	0.001 (<0.01)	0.001 (<0.01)
Overdispersion parameter						0.200
AIC						650.93
AICc						653.74
RMSE						0.77
MAD						0.59

The minimum, 25th percentile, median, mean, 75th percentile, and maximum coefficient (β) are shown in Table 11. The $GWNBR_g$ model shows some coefficients close to the GWNBR model coefficients. For example, the coefficients of AADT are observed to be the same in both cases. On the contrary, the coefficients of number of households and light commercial land use are found to be different in both cases. The only difference between GWNBR and GWNBRg is in the estimation of the overdispersion parameter. The similarity or change in coefficients in some variables is due to the local overdispersion in data. Therefore, overdispersion as also an important factor in the modeling process in addition to the spatial nonstationarity.

While looking into the AIC and AICc values, GLM-based NB model and GWNBR performed better than GWNBRg. In the case of MAD and RMSE, $GWNBR_g$ performed slightly better than the GLM-based NB regression.

6.2 Models for 0.50-mile Buffer Width Dataset

The GLM-based NB regression model results for the 0.50-mile buffer width dataset are summarized in Table 12.

Table 12. GLM-Based NB Model (0.50-mile buffer width dataset)

Parameter	Coeff. (β)	Std. Error	p-value
Intercept	-1.689	0.67	0.01
Segment length	-0.048	0.31	0.88
Log transformed AADT	0.250	0.07	<0.01
Number of households	1.066	0.36	<0.01
Light commercial land use	0.003	<0.01	<0.01
Light industrial land use	0.001	<0.01	0.05
Overdispersion			0.219
AIC			649.88
AICc			650.67
RMSE			0.89
MAD			0.70

From Table 12, variables such as AADT, number of households, light commercial land use, and light industrial land use are significant explanatory variables influencing teen crash frequency.

The GWNBR and $GWNBR_g$ model results for 0.50-mile buffer width dataset are summarized in Table 13 and Table 14, respectively. Explanatory variables such as AADT, number of households, light commercial land use, and light industrial land uses have a positive effect on the teen crash frequency.

While looking into the AIC and AICc values in Table 13 and Table 14, all the selected modeling approaches performed better. In the case of MAD and RMSE, the GWNBR and $GWNBR_g$ models performed better than the GLM-based NB regression.

Table 13. GWNBR Model (0.50-mile buffer width dataset)

Parameter	β_{min} (Std. Error)	β_{25} (Std. Error)	β_{50} (Std. Error)	β_{mean} (Std. Error)	β_{75} (Std. Error)	β_{max} (Std. Error)
Intercept	-2.155 (0.74)	-1.945 (0.77)	-1.783 (0.80)	-1.787 (0.82)	-1.637 (0.87)	-1.409 (1.01)
Segment length	-0.263 (0.35)	-0.162 (0.37)	-0.129 (0.38)	-0.136 (0.40)	-0.108 (0.42)	-0.062 (0.53)
Log transformed AADT	0.227 (0.08)	0.248 (0.08)	0.261 (0.08)	0.261 (0.08)	0.275 (0.09)	0.292 (0.10)
Number of households	0.872 (0.40)	1.02 (0.42)	1.101 (0.43)	1.100 (0.44)	1.187 (0.47)	1.298 (0.53)
Light commercial land use	0.002 (<0.01)	0.002 (<0.01)	0.003 (<0.01)	0.003 (<0.01)	0.003 (<0.01)	0.003 (<0.01)
Light industrial land use	0.001 (<0.01)	0.001 (<0.01)	0.001 (<0.01)	0.001 (<0.01)	0.001 (<0.01)	0.001 (<0.01)
Overdispersion parameter	0.188	0.212	0.219	0.219	0.228	0.241
AIC						652.79
AICc						654.26
RMSE						0.79
MAD						0.59

Table 14. $GWNBR_g$ Model (0.50-mile buffer width dataset)

Parameter	β_{min} (Std. Error)	β_{25} (Std. Error)	β_{50} (Std. Error)	β_{mean} (Std. Error)	β_{75} (Std. Error)	β_{max} (Std. Error)
Intercept	-2.144 (0.74)	-1.938 (0.77)	-1.784 (0.80)	-1.787 (0.82)	-1.642 (0.88)	-1.415 (1.01)
Segment length	-0.261 (0.35)	-0.167 (0.37)	-0.129 (0.38)	-0.136 (0.40)	-0.106 (0.42)	-0.064 (0.53)
Log transformed AADT	0.228 (0.08)	0.249 (0.08)	0.261 (0.08)	0.261 (0.08)	0.274 (0.09)	0.291 (0.10)
Number of households	0.882 (0.40)	1.023 (0.42)	1.102 (0.43)	1.101 (0.44)	1.181 (0.47)	1.292 (0.53)
Light commercial land use	0.002 (<0.01)	0.002 (<0.01)	0.003 (<0.01)	0.003 (<0.01)	0.003 (<0.01)	0.003 (<0.01)
Light industrial land use	0.001 (<0.01)	0.001 (<0.01)	0.001 (<0.01)	0.001 (<0.01)	0.001 (<0.01)	0.001 (<0.01)
Overdispersion parameter						0.219
AIC						653.13
AICc						654.60
RMSE						0.82
MAD						0.60

6.3 Comparison of Crash Estimation Models by the Buffer Width

When comparing with the 0.25-mile buffer width dataset model, the number of pupils in public or private schools is not found to be a significant variable influencing teen crash frequency. The goodness of fit and validation results for all the selected models are summarized in Table 15.

Table 15. Model Performance Comparison

Measure	0.25-mile buffer width dataset			0.50-mile buffer width dataset		
	GLM-based NB	GWNBR	$GWNBR_g$	GLM-based NB	GWNBR	$GWNBR_g$
AIC	644.75	645.75	650.93	649.99	652.79	653.13
AICc	645.77	647.22	653.74	650.67	654.26	654.60
RMSE	0.81	0.76	0.77	0.89	0.79	0.82
MAD	0.61	0.56	0.59	0.70	0.59	0.60

While looking into the goodness of fit measures and other validation measures in Table 15, the model developed using the 0.25-mile buffer width dataset performed better than the 0.50-mile buffer width dataset. The land use and demographic characteristics within a 0.25-mile buffer width of the selected road are better predictors of the teen crash frequency.

While comparing all the models, the GWNBR model using the 0.25-mile buffer width dataset performed best (RMSE = 0.76 and MAD = 0.56). Figure 5 shows the spatial variation in the coefficient of all the variables in the GWNBR model developed using the 0.25-mile buffer dataset. The coefficients have different ranges of effects on the teen crash frequency at various locations when using the GWNBR model. The traffic exposure variable has a different effect on teen crash frequency in the core urban area. However, the effect is comparatively lower on the northern side. A similar kind of localized effect is observed in all the selected explanatory variables. Such a localized effect indicates that GWNBR models could address the spatial variations in data by location than the GLM-based NB model.

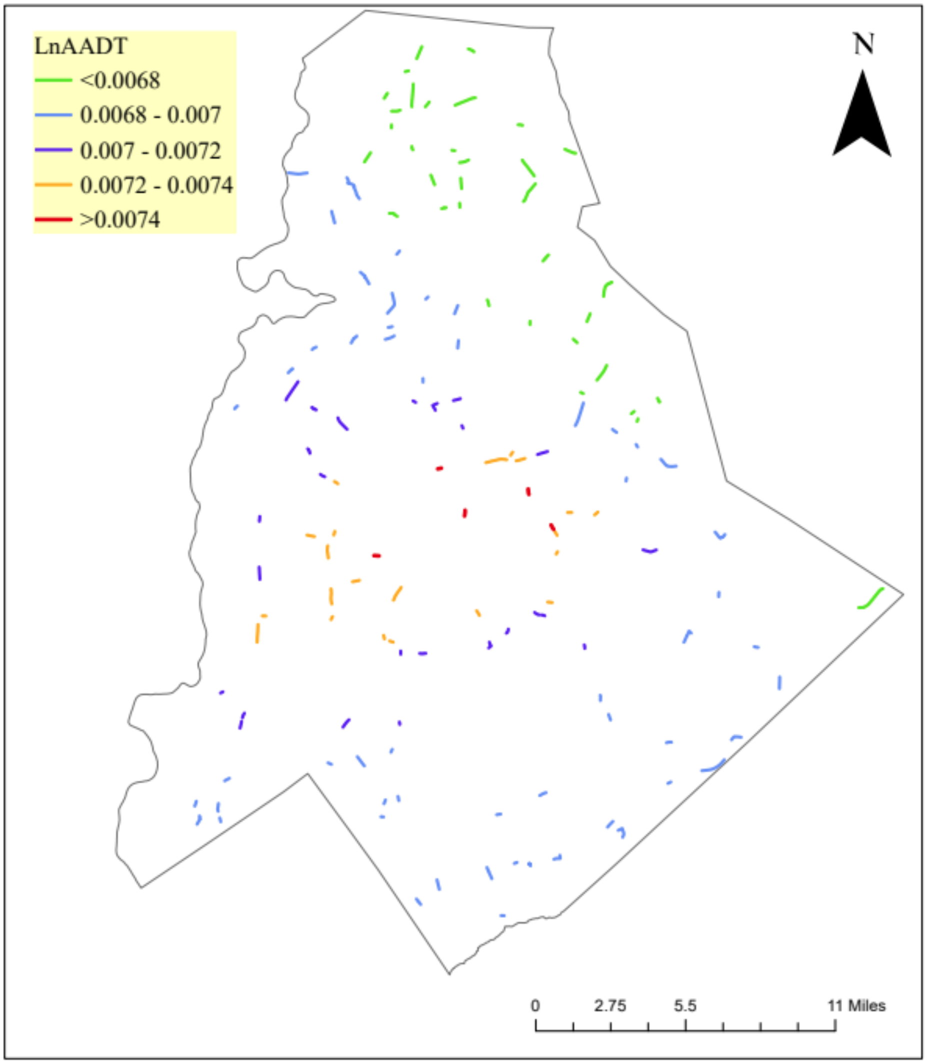

Figure 6. Spatial Distribution of Traffic Exposure Variable Coefficients from the GWNBR Model

7. Conclusions

This research focuses on assessing the effect of road network, demographic, and land use characteristics on the teen crash frequency. The evaluation process was carried out by selecting 201 spatially distributed road segments in the study area. Demographic and land use characteristics were extracted within 0.25 miles and 0.50 miles of each road segment. The explanatory variables were screened by computing and comparing Pearson correlation coefficients. Global GLM-based NB regression, GWNBR, and $GWNBR_g$ were then used to model and estimate link-level teen crash frequency, and the results were compared.

The AADT, light commercial, light industrial, and recreational land uses have an effect on the teen crash frequency. The multicollinearity between the AADT and other road network characteristics led to the exclusion of variables such as speed limit, access, and functional class type from the final model. Demographic variables such as population, total employment, number of households, and pupils enrolled in public or private high schools have a significant effect on the teen crash frequency. The GWNBR model developed using the 0.25-mile buffer width dataset performed better than all the other models developed.

The GLM-based NB regression model assumes that the variables are spatially stable. However, spatial heterogeneity is observed in most of the explanatory variables selected for modeling. The GWNBR model allows a conditional relationship between the teen crash frequency and different explanatory variables at each spatial location of interest. The goodness of fit and the model validation results indicated that the GWNBR model performed marginally better when compared to the global GLM-based NB regression model. GWNBR can incorporate the effect of spatial variations in the data, by geographic location, when modeling the teen crash frequency. Overall, using the GWNBR model to locally estimate the teen crash frequency of a segment will improve the model fitting over the global NB regression model.

The findings from this research can be used to understand the relative contributions of the various explanatory variables on teen crash frequency and, thereby, reduce the teen crash frequency by adopting effective solutions. Teen driver education schemes and enforcement can also be designed based on such findings. Based on the findings of this research, it is concluded that GWNBR is a promising method for crash frequency modeling, and we recommended it for use.

The crash data for the state-maintained roads were used in this research. Also, the intersection-level crashes are not separately considered/modeled in this research. Incorporating crash data for other routes/intersections may help examine the role of other specific land uses, local road types, intersection types, etc., on teen crash frequency. In addition, analysis using multiple buffer widths to capture the effect of spatial proximity on teen crash frequency merits an investigation.

7.1 Limitations and Scope for Future Work

Crash data for the state-maintained roads were used in this research. Incorporating crash data for other routes may help examine the role of other specific land uses and local road types on the teen crash frequency. In addition, analysis using multiple buffer widths to capture the influence of spatial proximity on teen crash frequency merits an investigation.

Bibliography

Aguero-Valverde, J., and P. P. Jovanis. "Spatial Analysis of Fatal and Injury Crashes in Pennsylvania." *Accident Analysis & Prevention* 38, no. 3 (2006): 618-625.

CDC. Teen Drivers: Get the Facts. Center of Disease Control and Prevention, National Center for Injury Prevention and Control, 2021. https://www.cdc.gov/transportationsafety/teen_drivers/teendrivers_factsheet.html. Last accessed December 8, 2021.

Curry, A. E., M. R. Pfeiffer, D. R. Durbin, and M. R. Elliott. "Young Driver Crash Rates by Licensing Age, Driving Experience, and License Phase." *Accident Analysis & Prevention* 80 (2015): 243-250.

Da Silva, A. R., and T. C. V. Rodrigues. "Geographically Weighted Negative Binomial Regression—Incorporating Overdispersion." *Statistics and Computing* 24, no. 5 (2014): 769-783.

Darwiche, A. A GIS Safety Study and a County-Level Spatial Analysis of Crashes in the State of Florida. Electronic Theses and Dissertations, University of Central Florida, 2009.

Das, S., L. Minjares-Kyle, L. Wu, and R. H. Henk. "Understanding Crash Potential Associated with Teen Driving: Survey Analysis using Multivariate Graphical Method." *Journal of Safety Research* 70 (2019): 213-222.

Deka, D. "Impacts of Standardizing School Start Time on Children and Household Workers–An Examination with NHTS Data." *Journal of Transport Geography* 59 (2017): 40-48.

Dissanayake, S., and N. Amarasingha. Safety Experience of Young Drivers in Comparison to Experienced Drivers. In T&DI Congress 2014: Planes, Trains, and Automobiles (pp. 344-355).

Du, H., and C. Mulley. "Relationship between Transport Accessibility and Land Value: Local Model Approach with Geographically Weighted Regression." *Transportation Research Record* 1977, no. 1 (2006): 197-205.

Duddu, V. R., V. M. Kukkapalli, and S. S. Pulugurtha. "Crash Risk Factors Associated with Injury Severity of Teen Drivers." *IATSS Research* 43, no. 1 (2019): 37-43.

Elander, J., R. West, and D. French. "Behavioral Correlates of Individual Differences in Road-Traffic Crash Risk: An Examination of Methods and Findings." *Psychological Bulletin* 113, no. 2 (1993): 279.

Foss, R. D., R. L. Smith, and N. P. O'Brien. "School Start Times and Teenage Driver Motor Vehicle Crashes." *Accident Analysis & Prevention* 126 (2019): 54-63.

Fotheringham, A. S., C. Brunsdon, and M. Charlton. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley & Sons, 2002.

Goldstick, J. E., P. M. Carter, F. Almani, S. J. Brines, and J. T. Shope. "Spatial Variation in Teens' Crash Rate Reduction Following the Implementation of a Graduated Driver Licensing Program in Michigan." *Accident Analysis & Prevention* 125 (2009): 20-28.

Gomes, M. J. T. L., F. Cunto, and A. R. da Silva. "Geographically Weighted Negative Binomial Regression Applied to Zonal Level Safety Performance Models." *Accident Analysis & Prevention* 106 (2017): 254-261.

Guadamuz-Flores, R., and J. Aguero-Valverde. "Bayesian Spatial Models of Crash Frequency at Highway–Railway Crossings." *Transportation Research Record* 2608 no. 1 (2017): 27-35.

Hadayeghi, A., A. S. Shalaby, and B. N. Persaud. "Development of Planning Level Transportation Safety Tools using Geographically Weighted Poisson Regression." *Accident Analysis & Prevention* 42, no. 2 (2010): 676-688.

Heydari, S., L. Fu, D. Lord, and B. K. Mallick. "Multilevel Dirichlet Process Mixture Analysis of Railway Grade Crossing Crash Data." *Analytic Methods in Accident Research* 9 (2016): 27-43.

Huang, H., M. A. Abdel-Aty, and A. L. Darwiche. "County-Level Crash Risk Analysis in Florida: Bayesian Spatial Modeling." *Transportation Research Record* 2148, no. 1 (2010): 27-37.

Hutchens, L., T. M. Senserrick, P. E. Jamieson, D. Romer, and F. K. Winston. "Teen Driver Crash Risk and Associations with Smoking and Drowsy Driving." *Accident Analysis & Prevention* 40, no. 3 (2008): 869-876.

Kim, W., A. M. Svancara, and T. Kelley-Baker. "Understanding the Impact of Road Design Characteristic on Teen Driver's Fatality." *Traffic Injury Prevention* 21, no. 5 (2020): 313-318.

Li, Q., F. Guo, I. Kim, S. G. Klauer, and B. G. Simons-Morton. "A Bayesian Finite Mixture Change-Point Model for Assessing the Risk of Novice Teenage Drivers." *Journal of Applied Statistics* 45, no. 4 (2018): 604-625.

Li, Z., W. Wang, P. Liu, J. M. Bigham, and D. R. Ragland. "Using Geographically Weighted Poisson Regression for County-Level Crash Modeling in California." *Safety Science* 58 (2013): 89-97.

Liu, J., A. J. Khattak, and B. Wali. "Do Safety Performance Functions Used for Predicting Crash Frequency Vary Across Space? Applying Geographically Weighted Regressions to Account for Spatial Heterogeneity." *Accident Analysis & Prevention* 109 (2017): 132-142.

Ma, L., and X. Yan. "Examining the Nonparametric Effect of Drivers' Age in Rear-End Accidents through an Additive Logistic Regression Model." *Accident Analysis & Prevention* 67 (2014): 129-136.

Mane, A. S., and S. S. Pulugurtha. "Influence of Proximal Land Use and Network Characteristics on Link Travel Time." Journal of Urban Planning and Development 146, no. 3 (2020): 04020028.

Mathew, S., and S. S. Pulugurtha. "A Comparative Assessment of Geospatial and Statistical Methods to Estimate Local Road Annual Average Daily Traffic." ASCE Journal of Transportation Engineering, Part A: Systems 147, no. 7 (2021): 04021035.

Mayhew, D. R., H. M. Simpson, and A. Pak. "Changes in Collision Rates Among Novice Drivers During the First Months of Driving." *Accident Analysis & Prevention* 35, no. 5 (2003): 683-691.

NCDOT. North Carolina 2019 Traffic Crash Facts, North Carolina Department of Transportation, 2020. https://connect.ncdot.gov/business/DMV/CrashFactsDocuments/2019%20Crash%20Facts.pdf. Last accessed June 25, 2021.

Peek-Asa, C., C. Britton, T. Young, M. Pawlovich, and S. Falb. "Teenage Driver Crash Incidence and Factors Influencing Crash Injury by Rurality." *Journal of Safety Research* 41, no. 6 (2010): 487-492.

Pirdavani, A., T. Brijs, T. Bellemans, and G. Wets. "Spatial Analysis of Fatal and Injury Crashes in Flanders, Belgium: Application of Geographically Weighted Regression Technique. In The 92th annual meeting of transportation research board, Washington, DC, 2013.

Pulugurtha, S. S., and S. Mathew. "Modeling AADT on Local Functionally Classified Roads Using Land Use, Road Density, and Nearest Nonlocal Road Data." Journal of Transport Geography 93 (2021): 103071.

Pulugurtha, S. S., V. R. Duddu, and Y. Kotagiri. "Traffic Analysis Zone Level Crash Estimation Models Based on Land Use Characteristics." *Accident Analysis & Prevention* 50 (2013): 678-687.

Quddus, M. A. "Modelling Area-Wide Count Outcomes with Spatial Correlation and Heterogeneity: An Analysis of London Crash Data." *Accident Analysis & Prevention* 40, no. 4 (2008): 1486-1497.

Regev, S., J. J. Rolison, and S. Moutari. "Crash Risk by Driver Age, Gender, and Time of Day Using a New Exposure Methodology." *Journal of Safety Research* 66 (2018): 131-140.

Silva, A. R., and C. V. Rodrigues, C. V. A SAS® Macro for Geographically Weighted Negative Binomial Regression, 2016. https://support.sas.com/resources/papers/proceedings16/8000-2016.pdf. Last accessed June 25, 2021.

Vachal, K., Research Faculty, and D. Malchose. "What Can We Learn About North Dakota's Youngest Drivers From Their Crashes?" *Accident Analysis & Prevention* 41, no. 3 (2009): 617-623.

Venkataraman, N., G. F. Ulfarsson, and V. N. Shankar. "Random Parameter Models of Interstate Crash Frequencies by Severity, Number of Vehicles Involved, Collision and Location Type." *Accident Analysis & Prevention* 59 (2013): 309-318.

Voas, R., and T. Kelley-Baker. "Licensing Teenagers: Nontraffic Risks and Benefits in the Transition to Driving Status." *Traffic Injury Prevention* 9, no. 2 (2008): 89-97.

Vorona, R. D., M. Szklo-Coxe, R. Lamichhane, J. C. Ware, A. McNallen, and D. Leszczyszyn. "Adolescent Crash Rates and School Start Times in Two Central Virginia Counties, 2009-2011: A Follow-Up Study to a Southeastern Virginia Study, 2007-2008." *Journal of Clinical Sleep Medicine* 10, no. 11 (2014): 1169-1177.

Vorona, R. D., M. Szklo-Coxe, A. Wu, M. Dubik, Y. Zhao, and J. C. Ware. "Dissimilar Teen Crash Rates in Two Neighboring Southeastern Virginia Cities with Different High School Start Times." *Journal of Clinical Sleep Medicine* 7, no. 2 (2011): 145-151.

Weast, R. A., and S. S. Monfort. "Characteristics of Vehicles Driven by Teens and Adults Killed in Crashes, 2013–2017." *Journal of Safety Research* 77 (2021): 263-267.

Zhao, F., and N. Park. "Using Geographically Weighted Regression Models to Estimate Annual Average Daily Traffic." *Transportation Research Record* 1879, no. 1 (2004): 99-107.

Zhou, H., J. Zhao, M. Pour-Rouholami, and P. A. Tobias. “Statistical Characteristics of Wrong-Way Driving Crashes on Illinois Freeways.” *Traffic Injury Prevention* 16, no. 8 (2015): 760-767.

About the Authors

Sonu Mathew received his Ph.D. degree in Infrastructure and Environmental Systems from the University of North Carolina at Charlotte, NC. He is a Postdoctoral Researcher with the Infrastructure, Design, Environment, and Sustainability (IDEAS) Center at the University of North Carolina at Charlotte. His current research interests include traffic safety, traffic flow modeling and simulation, ITS, connected and automated vehicles, and spatial modeling.

Dr. Srinivas S. Pulugurtha, P.E., F.ASCE is currently working as a Professor & Research Director of the Department of Civil and Environmental Engineering at the University of North Carolina at Charlotte. He is also the Director of the IDEAS Center.

Sarvani Duvvuri received her Ph.D. degree in Infrastructure and Environmental Systems from the University of North Carolina at Charlotte. She is a Postdoctoral Researcher with the IDEAS Center. Her current research interests include traffic operations and safety, traffic simulation, and connected and automated vehicles.