5-2022

Network-Wide Pedestrian and Bicycle Crash Analysis with Statistical and Machine Learning Models in Utah

Md Rafiur Rahman
Utah State University

NETWORK-WIDE PEDESTRIAN AND BICYCLE CRASH ANALYSIS WITH

STATISTICAL AND MACHINE LEARNING MODELS IN UTAH

by

Md Rafiur Rahman

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Civil and Environmental Engineering

Approved:

Patrick Singleton, Ph.D. Major Professor	Ziqi Song, Ph.D. Committee Member
Michelle Mekker, Ph.D. Committee Member	Keunhyun Park, Ph.D. Committee Member

D. Richard Cutler, Ph.D.
Interim Vice Provost for Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

2022

ABSTRACT

NETWORK-WIDE PEDESTRIAN AND BICYCLE CRASH ANALYSIS WITH STATISTICAL AND MACHINE LEARNING MODELS IN UTAH

by

Md Rafiur Rahman, Master of Science

Utah State University, 2022

Major Professor: Dr. Patrick Singleton
Department: Civil and Environmental Engineering

The aim of this study was to investigate the impact of road network and environmental characteristics on pedestrian and bicycle crashes in Utah. Crash frequency modeling was undertaken to: (1) identify characteristics of segments and non-signalized intersections linked to significant differences in the number of non-motorized crashes and thus locate crash prone links in the road network in Utah, and (2) rank variables in terms of importance and explore possible non-linear associations of explanatory variables with crashes. This study uses innovative pedestrian and bicycle volume data as a measure of exposure on road segments and at unsignalized intersections. Pedestrian counts estimated from nearby signalized intersections were used for pedestrian exposure, and crowdsourced "Strava" app data was used for bicycle exposure. Using a spatial data joining process, this research created a feature-rich data source that included road geometry, traffic, and built environment characteristics for road segments and non-signalized intersections along with 10 years of pedestrian and bicycle crash information in Utah. Multiple negative binomial models investigated crashes at different spatial scales—12,204 segments and 4,555 intersections on state routes, and 46,497 segments and 50,737 intersections on state and federal aid routes—to account for different levels of data availability and completeness. Locations with high traffic volume, vertical grades, frequent transit stops, driveway density, and more legs at intersections tended to have more pedestrian and bicycle crashes. Greater residential and employment density, as well as a greater degree of low-income households and people of non-white race/ethnicity, were also associated with more crashes. Variable importance ranking illustrated that greater traffic volume, high bicycle and pedestrian volume, more driveways, and steeper vertical grades perhaps deserve additional attention. To investigate pedestrian and bicycle injury severity, logistic regression analysis was conducted and found that crashes occurring at mid-blocks are more severe as compared to intersections. High daily temperature was associated with greater severity. Human factors such as driving under the influence and distracted driving also increased severity in crashes. This study suggests potential countermeasures, policy implications, and the scope of future research for improving pedestrian and bicycle safety at segments and at non-signalized intersections.

(167 pages)

PUBLIC ABSTRACT

PEDESTRIAN AND BICYCLE CRASH ANALYSIS

Md Rafiur Rahman

Recent trends in crashes indicate a dramatic increase in both the number and share of pedestrian and bicyclist injuries and fatalities nationally and in many states. Crash frequency modeling was undertaken to identify crash prone characteristics of segments and non-signalized intersections and explore possible non-linear associations of explanatory variables with crashes. Crowdsourced "Strava" app data was used for bicycle volume, and pedestrian counts estimated from nearby signalized intersections were used as pedestrian volume. Multiple negative binomial models investigated crashes at different spatial scales to account for different levels of data availability and completeness. The models showed high traffic volume, steeper vertical grades on roads, frequent bus and rail stations, greater driveway density, more legs at intersections, streets with high large truck presence, greater residential and employment density, as a larger share of low-income households and non-white race/ethnicity groups are indicators of locations with more pedestrian and bicycle crashes. Crash severity model results showed that crashes occurring at mid-blocks and near vertical grades were more severe compared to crashes at intersections. High daily temperature, driving under influence, and distracted driving also increases injury severity in crashes. This study suggests potential countermeasures, policy implications, and the scope of future research for improving pedestrian and bicycle safety at segments and at non-signalized intersections.

ACKNOWLEDGMENTS

I want to express foremost appreciation for my advisor, Dr. Patrick Singleton from whom I have received endless encouragement, guidance, and generous support. I am indebted to him for the hands on learning and continual mentorship regarding all aspects of graduate study I was fortunate to receive. It has been a great pleasure to work with him.

I am extremely grateful to Dr. Ziqi Song, Dr. Michelle Mekker, and Dr. Keunhyun Park, my committee members, for their time and flexibility. Their coursework and research publications have guided my academic interest and curiosity that includes and goes beyond this thesis research. I want to gratefully acknowledge the contribution of the faculty members and staffs. Studying abroad has been quite an experience, and I feel very fortunate to be around amazing friends and fellow awesome graduate students through the non-linearity of this episode of life. Finally, I would like to thank my parents, lovely sisters and my very cute niece for believing in me!

Md Rafiur Rahman

CONTENTS

Page

ABSTRACT iii

PUBLIC ABSTRACT v

ACKNOWLEDGMENTS vi

LIST OF TABLES x

LIST OF FIGURES xiii

LIST OF SYMBOLS AND NOTATION xiv

1 Introduction 1

1.1 Problem Statement 1
1.2 Objectives 4
1.3 Organization of the Document 4

2 Literature Review 6

2.1 Risk Factors for Pedestrian Crashes 7

2.1.1 Road Geometry Characteristics 7
2.1.2 Intersection Characteristics 9
2.1.3 Traffic Characteristics 10
2.1.4 Built Environment and Community Characteristics 12
2.1.5 Weather Condition & Road Lighting 14

2.2 Risk Factors for Bicycle Crashes 14

2.2.1 Road Geometry Characteristics 15
2.2.2 Intersection Characteristics 17
2.2.3 Traffic Characteristics 19
2.2.4 Built Environment and Community Characteristics 21
2.2.5 Weather Condition and Road Lighting 23

2.3 ML Techniques Investigating Non-Linear Crash Associations 23
2.4 Limitations of Existing Research 26

3 Data Collection 28

3.1 Overview 28
3.2 Study Locations 28

3.2.1 Segments and Mid-block Locations 29
3.2.2 Non-signalized Intersections 31
3.3 Crash Data 32
3.3.1 Crash Data on Road Network 37
3.3.2 Individual Crash Characteristics 41
3.4 Exposure Data 43
3.4.1 Bicycle Exposure Data 43
3.4.2 Pedestrian Exposure Data 45
3.5 Roadway and Community Characteristics Data 49
3.6 Weather Data 53
3.6.1 Preparing Data for Weather Information 53
3.7 Data Preparation and Processing 56
3.7.1 Preparing Data for Segments and Mid-block Locations 56
3.7.2 Preparing Data for Non-signalized Intersections 62
3.8 Chapter Summary 67
4 Methodology 69
4.1 Overview 69
4.2 Count Data Models 69
4.3 Boosted Regression Tree Theory 72
4.3.1 Parameter optimization 76
4.3.2 Cross-validation 77
4.3.3 Relative importance of variables 77
4.3.4 Marginal effects of variables 78
4.3.5 Analysis Setup 79
4.4 Ordinal Data Models 79
4.5 Chapter Summary 81
5 Results 82
5.1 Overview 82
5.2 Pedestrian Crash Frequency Along Segments and at Mid-Block Locations 82
5.2.1 Negative Binomial Model Results 83
5.2.2 Boosted Decision Tree Model Results 87
5.3 Bicycle Crash Frequency Along Segments and at Mid-block Locations. 94
5.3.1 Negative Binomial Model Results 95
5.3.2 Boosted Decision Tree Model Results 98
5.4 Pedestrian Crash Frequency at Non-signalized Intersections 105

5.4.1 Negative Binomial Model Results 105
5.4.2 Boosted Decision Tree Model Results 107
5.5 Bicycle Crash Frequency at Non-signalized Intersections.....................113
5.5.1 Negative Binomial Model Results 114
5.5.2 Boosted decision tree model results 117
5.6 Pedestrian Crash Severity Results...123
5.7 Bicycle Crash Severity Results 124
5.8 Comparison between NB and DT models...126
5.9 Chapter Summary...129
6 Discussion 131
6.1 Overview 131
6.2 Key Findings 132
6.3 Contributions..139
6.4 Policy Implications...141
6.5 Limitations and Future Work 144
7 References 147

LIST OF TABLES

Page

Table 3-1 *Pedestrian Crashes at Mid-Block Locations along Segments (2010-2019)*... 38

Table 3-2 *Bicycle Crashes at Mid-Block Locations along Segments (2010-2019)* 38

Table 3-3 *Pedestrian Crashes at Non-Signalized Intersections (2010-2019)* 39

Table 3-4 *Bicycle Crashes at Non-Signalized Intersections (2010-2019)*...................... 40

Table 3-5 *Descriptive Statistics for Crashes on Segments & Non-Signalized Intersections (2010-2019)*... 41

Table 3-6 *Pedestrian & Bicycle Crash Characteristics (N= 6740)* 42

Table 3-7 *Descriptive Statistics of Bicycle Volume on Road Segments (2019)*.............. 45

Table 3-8 *Descriptive Statistics of Pedestrian Volume on Road Segments (2019)* 49

Table 3-9 *Data Types Investigated at Segments and at Non-Signalaized Intersections*. 52

Table 3-10 *Descriptive Statistics of Weather Data for Pedestrian Crashes* 54

Table 3-11 *Descriptive Statistics of Weather Data for Bicycle Crashes*......................... 55

Table 3-12 *Descriptive Statistics for Segments (State Routes Only): Categorical Variables*... 58

Table 3-13 *Descriptive Statistics for Segments (State Routes Only): Continuous Variables*... 60

Table 3-14 *Descriptive Statistics for Segments (State & Federal Aid Routes): Categorical Variables*.. 61

Table 3-15 *Descriptive Statistics for Segments (State & Federal Aid Routes): Continuous Variables*... 62

Table 3-16 *Descriptive Statistics for Non-Signalized Intersections at State Routes Only: Categorical Variables* 63

Table 3-17 *Descriptive Statistics for Unsignalized Intersections at State Routes Only: Continuous Variables* 65

Table 3-18 *Descriptive Statistics for Non-Signalized Intersections at State and Federal Aid Routes: Categorical Variables* 66

Table 3-19 *Descriptive Statistics for Non-Signalized Intersections at State and Federal Aid Routes: Continuous Variables* 67

Table 5-1 *NB Model for Pedestrian Crashes at Segments (State Routes, N = 4,979)* ... 84

Table 5-2 *NB Model for Pedestrian Crashes at Segments (State and Federal Aid Routes, N = 5,034)* 86

Table 5-3 *Variable Importance for Pedestrian Mid-Block Crashes (State Routes, N= 13,107)* 88

Table 5-4 *NB Model for Bicycle Crashes along Segments and Mid-Block Locations (State Routes, N = 11,910)* 95

Table 5-5 *NB Model for Bicycle Crashes Along Segments and at Mid-Block Locations (State and Federal Aid Routes, N = 11,865)* 97

Table 5-6 *Variable Importance for Bicycle Mid-Block Crashes* 99

Table 5-7 *NB Model for Pedestrian Crashes at Non-Sigalized Intersections (State Routes, N = 3,378)* 105

Table 5-8 *NB Model for Pedestrian Crashes at Non-Signalized Intersections (State and Federal Aid Routes, N = 4,738)* 106

Table 5-9 *Variable Importance for Pedestrian Non-Signalized Intersection Crashes*. 108

Table 5-10 *NB Model for Bicycle Crashes at Non-Signalized Intersections (State Routes, N = 3,192)* 114

Table 5-11 *NB Model for Bicycle Crashes at Non-Signalized Intersections (State and Federal Aid Routes, N = 14,430)* 116

Table 5-12 *Variable Importance for Bicycle Crashes at Non-Signalized Intersections*118

Table 5-13 *Pedestrian Crash Severity (N= 6740)* 123

Table 5-14 *Bicycle Crash Severity (N= 5,764)* 125

Table 5-15 *NB and DT Models Validation* 126

Table 5-16 *Significant Variables for Pedestrian and Bicycle Crashes* 129

LIST OF FIGURES

Page

Figure 3-1 *Spatial Join of Crashes on Segments* 36

Figure 3-2 *Spatial Join of Crashes on Intersections*... 37

Figure 5-1 *Marginal Effects for Pedestrian Mid-Block Crashes*.................................... 89

Figure 5-2 *Poisson Deviance for Bicycle Mid-Block Crashes*....................................... 94

Figure 5-3 *Marginal Effect for Bicycle Mid-Block Crashes*.. 100

Figure 5-4 *Poisson Deviance for Bicycle Mid-Block Crashes*...................................... 104

Figure 5-5 *Marginal Effects for Pedestrian Unsignalized Intersection Crash*............. 109

Figure 5-6 *Poisson Deviance for Pedestrian Intersection Crashes*.............................. 113

Figure 5-7 *Marginal Effects for Non-Signalized Intersection Bicycle Crash*.............. 119

Figure 5-8 *Poisson Deviance for intersection bicycle crash* 122

LIST OF SYMBOLS AND NOTATION

ATSPM	Automated Traffic Signal Performance Measures
ACS	American Community Survey
EPA	Environmental Protection Agency
NEMA	National Electrical Manufacturers Association
NOAA	National Oceanic and Atmosphere Administration.
UDOT	Utah Department of Transportation
USU	Utah State University

1 INTRODUCTION

1.1 Problem Statement

The number and share of pedestrian and bicyclist injuries and fatalities has been increasing, both in the US and in Utah. According to the National Highway Traffic Safety Administration, 6,205 pedestrians and 846 bicyclists were killed in road crashes in the United States, accounting for 17.2% and 2.3% of all traffic fatalities respectively (NHTSA, 2020). This was a substantial increase since 2010, when there were only 4,302 pedestrian and 623 bicycle crashes, representing 13% and 1.9% of fatalities respectively (NHTSA, 2018). Although Utah has seen decreases in bicyclist crashes and injuries in recent years, pedestrian crashes and injuries have increased. In 2019, 45 deaths (representing nearly 17% of all fatal crashes) and more than 900 injuries to people walking, as well as 6 deaths and nearly 500 crashes to people bicycling on Utah streets and highways, were reported (UDPS, 2020). These statistics highlight the need to focus on pedestrian and bicyclist safety.

To identify risk prone-locations on a road network, a crash frequency-based approach has been traditionally used that selects sites with a greater number of reported crashes and identifies site-specific safety issues (through diagnosis) using local condition, thus informing the selection of site-specific safety countermeasures and treatments (FHWA, 2016). However, crashes involving people walking and bicycling are (compared to motor vehicle crashes) highly dispersed, with many sites that have only a few crashes. Such highly dispersed crashes are difficult to address using site-based crash histories, and such site-based methods may not be able to address specific crash types (like bicycle & pedestrian crashes) due to their low frequencies (NCHRP Report 893, 2018).

An alternative and often complementary approach, the systemic approach, may be better suited for tackling many bicycle and pedestrian safety issues in a state like Utah. Rather than relying upon reported crashes to select and apply treatments to high-crash sites, a systemic approach to safety management instead first selects crash type(s) of interest and identifies geometric and operational risk factors across a network that are associated with those crash type(s), using crash data from a variety of sites and prior knowledge. Then, these risk factors guide the selection of sites with higher-risk characteristics, informed by but without having to rely upon site-specific crash histories or (in some cases) requiring exposure data. Systemic safety analysis is proactive, identifying potential improvements without waiting for crash histories and trends to develop (FHWA, 2016, NCHRP Report 893, 2018).

In addition to identifying factors that have a significant impact on pedestrian and bicycle safety, exploring the most important factors provide insights for prioritizing resources and develop safety management plans. Also, crash related variables often have a complex and non-linear association with crashes, so a closer look into these complex relationships may help to identify the ranges where they have the most influential effect on traffic crash occurrences. Application of machine learning techniques to investigate non-linearity in crashes is a recent development in traffic safety research, including methods that rank variables’ importance according to their contribution to crash frequency and illustrate the non-linear associations of explanatory variables with crashes.

These methods are relatively underexplored in the traffic safety context, and the network-wide analysis carried out in this study adds to that developing body of literature.

In order to perform a systemic safety analysis, network-wide traffic crash datasets were assembled for road segments to investigate mid-block crashes as well as crashes at non-signalized intersections. These datasets for state and federal aid routes contain detailed information regarding road geometry, traffic characteristics, bicycle and pedestrian exposure, and land use and community characteristics for the road network in Utah. This research has investigated two especially interesting exposure data sources. Pedestrian exposure data was estimated from Automated Traffic Signal Performance Measures (ATSPM) counters and their usefulness for analyzing crashes at mid-block locations and unsignalized intersections have been explored. Strava Metro app datasets were used for bicycle exposure. Investigation of these data sources provide evidence that they can be a viable application of exposure data that can account for non-motorized traffic exposure in crash studies.

Furthermore, no study has investigated state-wide road network and adjoining land use and community characteristics to investigate the impact of social equity on vulnerable road users such as pedestrian and bicycle safety. While crash studies to understand and model pedestrian and traffic crashes are not new, the use of unique data sources for the state of Utah and application of statistical and machine learning techniques are promising ways to meet this need to address bicycle and pedestrian safety issues and inform data-driven decision-making.

1.2 Objectives

The overarching objective of this study is to understand factors associated with pedestrian and bicycle crashes at mid-block locations and unsignalized intersections. This study aims to answer the following sub-objectives of the overall goal.

1. Identify and rank the common and most important factors affecting bicycle and pedestrian crash frequency and crash severity along mid-block road segments and at non-signalized intersections.
2. Test the applicability of machine learning models (boosted decision tree) and compare findings with statistical crash analysis models (negative binomial).
3. Determine associations with pedestrian crash frequency using more robust measures of pedestrian exposure from traffic signal ATSPMs and develop bicycle crash frequency models using crowd-sourced bicycle exposure data that has not been sufficiently studied in the literature.
4. Examine bicycle and pedestrian crash frequency disparities in Utah based on neighborhood sociodemographic characteristics.
5. Identify the impact of weather variables affecting pedestrian and bicycle crash severity.

1.3 Organization of the Document

This thesis contains six chapters. Chapter 1 contains a brief introduction and presents the motivation, scope, and objectives of the research program. Chapter 2 conducts a detailed literature review of previous research. Chapter 3 presents the data description, data collection, and assembly methods for this study. Chapter 4 discusses the methodology used to carry out the study presented in this thesis. Chapter 5 presents the data analysis results. This includes the results of statistical and machine learning models on pedestrian and bicycle crashes at mid-block locations unsignalized intersections. Estimated crash severity results for pedestrian and bicyclists are also discussed. Finally, Chapter 6 draws conclusions from this research, identifies key findings and contributions, and discusses future work on this topic

2 LITERATURE REVIEW

Crash analysis for pedestrians and bicyclists is a widely investigated topic in transportation safety research. Prior studies endeavoured to identify the contributing factors that increase or reduce the traffic risk to active transportation modes. This chapter reviews relevant literature to understand existing knowledge including factors affecting walking and bicycling safety conditions, and notes the recent development of data-driven machine learning techniques for crash prediction.

First, a systematic literature search using search terms such as “pedestrians” or “walking”, “bicycling”, “crash”, “frequency”, “severity”, “weather” was carried out on Google Scholar. The filtering of results involved reviewing relevant titles, reading abstracts, then finally reading papers in detail. Finally, only those relevant papers are reviewed in this section.

The organization of this section is as follows. First, it reviews which road geometry, traffic characteristics, and road network related factors are commonly found to affect pedestrian and bicycle crash safety at road segments and at unsignalized intersections. The following section surveys the recent methodological inventions including machine learning techniques that investigate non-linear relationships between these factors and pedestrian and bicycle crash frequencies. The final section discusses the limitations of existing research and identifies the scope of further studies.

2.1 Risk Factors for Pedestrian Crashes

The contributing factors for pedestrian safety regarding frequency and severity of crashes are traffic volumes and characteristics, roadway geometric conditions, intersection geometric conditions, built environment and community variables, weather, and lighting. Roadway geometry factors include lane width, number of turn lanes, shoulder, median, rumble strips, horizontal curvature, and vertical grades and their association with pedestrian crashes. Traffic variables include factors like traffic volume, pedestrian exposure, speed limit, traffic composition such as heavy truck percentage, and transit stops including bus stop and rail stations. Traffic safety factors related to community characteristics and land use information include residential density, employment density, household income, number of vehicles per household, and other demographic characteristics such as race/ethnicity. Additional risk factors like weather information (precipitation, snowfall, temperature), road light conditions, driving conditions, and vehicle characteristics are also common attributes that are linked with pedestrian crash risk.

2.1.1 Road Geometry Characteristics

Major arterials and roadways with more lanes are positively associated with increasing pedestrian crashes (Ukkusiri et al., 2012). Diogenes & Lindau (2010), Wang & Kockelman, (2013) found that streets with more lanes lead to increased pedestrian crashes at midblock locations and at intersections as pedestrians may require longer time to cross wide roadways. Abdel-Aty et al. (2007) informs that wide roads with more lanes near schools add to the risk of pedestrian crashes, and traffic calming measures such as road narrowing may improve pedestrian safety. However, crashes may commonly occur at urban streets with fewer number of lanes according to Zegeer et al., (2006).

Steep grades are negatively related to pedestrian crashes at mid-block locations, as found by Chen & Zhuo (2016). This may happen due to pedestrians' tendencies to avoid steep roads due to the possible physical strain. On the other hand, Poch and Mannering (1996) highlighted the uphill or downhill grades and an intersection approach may increase the likelihood of pedestrian crashes because of drivers' shortened sight distance and visibility.

Harkey et al. (2006) found that the presence of turn lanes indicated an increasing pedestrian crash frequency. Turning movements may create problems for crossing pedestrians at intersections, as drivers are focused on vehicles coming from the opposite direction. Schneider et al. (2010) found that intersections with many right turn only lanes are significantly associated with increased risk of pedestrian crash. Additionally, researchers found that tuning radius may have an impact on pedestrian crash occurrences as smaller turn radius may reduce pedestrian injury severity in case of a crash occurrence (Roudsari et al., 2007).

Frequent driveways can present additional conflict points between vehicles and pedestrians and thus create safety challenges. Schneider et al. (2004), Taquechel (2009), and Kim & Ulfarsson (2019) found that frequent driveways within 50 feet of intersections may lead to increasing numbers of pedestrian crashes, especially near downtown and commercial areas. Dumbaugh et al. (2013) found that pedestrians are particularly vulnerable at uncontrolled driveways.

2.1.2 Intersection Characteristics

More legs at intersections provide greater numbers of crossing stages for pedestrians. Prior studies found that intersections with more approaches are linked to increased pedestrian crash risk (Dumbaugh et al., 2013, Lee et al., 2017). Roundabouts are widely considered to be significantly safer for motor traffic; however, pedestrian safety benefits can be mixed. Stone et al., (2002) found roundabouts to be safer for pedestrian compared to traditional four-legged intersection. Inversely, Daniels et al. (2010) observed roundabouts to be a source of frequent pedestrian crashes considering the amount of traffic and pedestrian exposure. Generally, roundabouts without any pedestrian facility can be challenging for all and especially for people with a disability as pedestrians must find a gap in traffic flow before they start crossing. Pedestrian facilities such as well marked or raised crosswalk, high intensity activated crosswalk (HAWK) signals can mitigate some of these challenges.

Medians can provide refuge for pedestrian crossings and provide a comfortable crossing time, thus improving traffic safety for pedestrians at both signalized (Petritsch et al., 2005) and unsignalized intersections (Harwood et al., 2008, Schneider et al., 2010). Palamara & Broughton (2013) and Zegeer et al. (2006) reported fewer pedestrian crash at roads with raised medians. Medians can provide improved safety conditions for pedestrians at mid-block crossings as well. Baltes & Chu (2002) found that medians are associated with fewer crashes at mid-block locations.

2.1.3 Traffic Characteristics

Higher traffic volume results in a greater exposure to motor traffic for pedestrians and thus is likely to cause a greater number of collisions. Traffic volume is often found to be a critical factor associated with pedestrian crash frequency and injury crashes. Martin (2006), Loukaitou et al. (2007), Weir et al. (2009), and Palamara & Broughton (2013 reported that streets and intersections with higher traffic volumes are prone to see greater numbers of pedestrian crashes. In many cases, pedestrian crash rate increases at a higher rate at a lower traffic volume, as pointed out by Abdel-Aty et al. (2005). Pulugurtha & Sambhara (2011), and Siddiqui et al. (2012) also found that traffic volume is a common predictor of crash frequency at intersections, and an increase in traffic volume is associated with diminished pedestrian safety. Zegeer et al. (2006) found that high traffic volume at unsignalized intersections leads to increased fatal pedestrian crashes. However, a negative association between traffic volume and pedestrian crashes is also possible since that higher amount of traffic on roads may lead to reduced speed and thus create less risky situations for pedestrians, according to Chen & Zhuo (2016).

Similarly, the likelihood of pedestrian crashes increases when the pedestrian volume increases, because (collectively) pedestrians would have more exposure to traffic according to Schneider et al. (2004), and Pulugurtha & Sambhara (2011). Chen & Zhuo (2016) found that high volumes of crossing pedestrians at intersections increases the risk of more crashes. Pedestrian volume is also positively associated with an increasing number of pedestrian crashes at intersections and at mid-block crossing locations (Harwood et al., 2008; Petritsch et al., 2005).

However, findings from prior studies also showed a negative association of pedestrian crash rates with traffic volume, indicating that because of more pedestrians' presence on roads, they may be more visible to motorists (Chen & Zhuo, 2016; Singleton et al., 2020). These results are aligned with the pedestrian safety concept called "safety in numbers"—where an increase in pedestrian and bicycle volumes are possibly associated with increased caution among motorists when more people are walking or bicycling. As a result, pedestrian and bicycle crash rates (crashes per pedestrian or cyclist) decrease with increasing pedestrian and bicycle volume (Jacobsen, 2003; Elvik & Bjornskau, 2017).

Greater number of pedestrian crashes are found near high bus stop density. (Ukkusuri et al., 2012; Wang & Kockelman, 2013; Chen & Zhou, 2016; Abdel-Aty et al., 2007; Lee et al., 2005). Also, the location of bus stops at intersections (near side or far side) and stopped buses may prevent safe crossing sight distance for pedestrians. Schneider et al. (2004) and Diogenes & Lindau (2010) reported that the midblock crossings near public transit stops and other public transit system facilities may lead to high pedestrian crashes. Studies investigating both microscopic and macroscopic crash frequency modeling report that streets and intersections close to high public transit use locations are likely to experience more crashes (Amoh-gyimah et al., 2016; Abdel-Aty et al., 2011). Torbic et al. (2010) and Miranda-Moreno et al. (2011) found higher pedestrian crashes near intersections with frequent transit stops. However, Toranpour et al. (2018) found that mid-block pedestrian crashes are less severe near transit stations.

In most studies, the likelihood of pedestrian crashes increases with high vehicle speed limit (Lee et al., 2005; Zegeer et al., 2006; Chimba et al., 2014; Fridman et al, 2020). Senserrick et al. (2014) found that most pedestrian crashes occur not in low-speed urban areas but in high-speed rural areas. Severity of a pedestrian–vehicle crash increases substantially as the risks of fatal or injury crashes are much greater compared to low speeds (Sze & Wong, 2007; Wang & Kockelman, 2013; Olszewski et al., 2015; Doecke et al., 2018). A negative association with high-speed limit and pedestrian crashes are also found in literature, as pedestrians may avoid high speed streets (Moudon et al., 2011; Narayanamoorthy et al., 2013; Miranda-Moreno et al., 2011).

2.1.4 Built Environment and Community Characteristics

Pedestrian crashes may occur predominantly in mixed land use areas (Chen & Zhou, 2016). This may happen because of the fact that mixed land use areas are trip generators for pedestrian activities including neighborhoods, commercial buildings, schools, churches, playgrounds, etc. Pulugurtha & Sambhar (2011) found that primarily residential areas may see fewer pedestrian crashes due to the presence of low traffic volumes and low-speed roads near residential areas.

However, this finding is not uniform as residential area can be a source of greater pedestrian presence on roads and thus increase risks of more pedestrian crashes (Loukaitou-Sideris et al., 2007; Siddiqui et al., 2012). Ding et al. (2018) studied pedestrian crashes to investigate built environment effects and found that compact development areas with greater household density were related to increasing pedestrian crash rates.

Wier et al. (2009) estimated a macro-level pedestrian crash prediction model and found that employment density was positively associated with pedestrian crash frequencies using crash data from San Francisco. Nolan & Quddus (2004), Johnson et al. (2004) and Amoh-Gyimah (2016) found similar results, with an increased crash frequency near high residential and employment density. Miranda-Moreno et al. (2011) estimated a positive association of employment density and crash frequency in their models. They also introduced the number of jobs as a variable, and it had a similar effect as employment density.

A positive association between pedestrian crash occurrences and neighborhoods with low income as well as minority population has been observed by previous studies (Loukaitou-Sideris et al., 2007; Lyons et al., 2008; Cottrill & Thakuriah, 2010; Weir et al., 2009). Ukkusuri et al. (2011) showed a significant positive correlation between pedestrian crash frequencies and African American or Hispanic neighborhoods, as well as populations with lower educational attainment. Chimba et al. (2014) and Amoh-Gyimah (2016) found that higher rates of crashes were associated with lower household income. Torbic et al. (2010) and Ewing & Cervero (2011) also had similar findings regarding household income and pedestrian crashes. Jang et al. (2013), Chimba et al. (2014), and Loukaitou et al. (2007) reported that people who identify as Latino, Black and Hispanic were more likely to be involved in pedestrian-vehicle crashes.

Household vehicle ownership was negatively associated with pedestrian crashes (Chimba et al., 2014; Martin, 2006). It is possible because more vehicles in a household could be an indication of low pedestrian activity and less exposure to the risk of pedestrian collisions.

Additionally, studies determined that pedestrians' and drivers' alcohol and drug use is significant factor associated with pedestrian crashes.Driving under influence (DUI) and alcohol-impaired pedestrians were involved in more severe crashes (Zajac & Ivan, 2003; Chang, 2008).

2.1.5 Weather Condition & Road Lighting

Road lighting relates to drivers' sight and visibility directly and is thus a serious risk factor in pedestrian crashes. Fitzpatrick et al. (2014) indicated that 82% of the crashes in Texas from 2007 to 2011 occurred in dark conditions, almost half of which were at locations with no lighting. Senserrick et al. (2014) observed that pedestrians are more likely to be involved in crashes in poorer lighting conditions, particularly when crossing at a midblock location away from an intersection. However, Palamara & Broughton (2013) pointed out that most pedestrian crashes occurred during daylight hours in central business district areas. Lee et al. (2017) and Jang et al. (2013) included the weather factor in the models and found that rainy weather had a positive influence on both pedestrian crash frequency and severity level

2.2 Risk Factors for Bicycle Crashes

The contributing factors for bicycle safety are traffic volumes and characteristics, roadway geometric conditions, intersection geometric conditions, land use characteristics, demographic variables, weather condition and lighting. Roadway geometry factors consist of the presence of bike lanes, travel lane width, number of turn lanes, presence of shoulder, median, rumble strips, horizontal curvature, and vertical grades. Traffic variables include factors like traffic volume, bicyclist volume, speed limit, traffic composition such as heavy truck percentage, and transit stops including bus stop and rail station presence. Community characteristics and land use information include residential density, employment density, household income, vehicles per household, and additional demographic information. Other risk factors include driving conditions, type of vehicles involved in crash, weather information such as precipitation, snowfall, and temperature, as well as lighting conditions.

2.2.1 Road Geometry Characteristics

Bicycle crashes and overall bike safety condition relies heavily on road geometry. Bicycle infrastructures such as on road bike lanes, on road protected bike lanes, and separated cycle tracks are found to be effective in reducing bicycle injury crashes especially in roadways with lower traffic volume (AADT) (Park et al., 2015; Pedroso et al., 2016; Pucher & Buehler, 2016; Teschke et al., 2012). Abdel-Aty et al. (2011) analyzed bike crashes in Florida and found that safety conditions improved after the implementation of bike lanes. Although facilities such as bike lanes may encourage more cyclists on roadways, these facilities do not lead to more crashes (Harris et al., 2013; Reynolds et al., 2009). Moreover, the configuration of the nearby intersections and traffic volumes affects usefulness of bike lanes. Bike lane width can also affect cyclists' safety since bike lanes 4–8 feet in width are found to have improved safety effects and lower crash numbers (Park et al., 2015). This may be because motorists tend to view bike lanes with traditional lane width as another vehicle lane or parking area, according to Toole (2010). Reynolds, (2009) and Raihan et al., (2019) found the presence off-road bike paths improves the safety condition but cautions against mixed traffic of cyclists and pedestrians on sidewalks and multi-use trails as they may present a higher crash risk. However, the ‘dooring effect’ for bike lanes parallel to street parking can be a regular source of severe bike crashes (Schimek, 2018; Bhatia et al., 2016).

Travel lane width on roadways has a nonlinear effect on bicycle safety. Sadek et al. (2007) and Petritsch et al (2006) found that when bike lanes are present, drivers may be more aware of bicyclists in the bike lanes and drive more cautiously to avoid collisions in narrow lanes. In fact, for specific roadway conditions (such as the presence of shoulders), narrow lane width can provide better safety conditions for cyclists than roadways with wider lane widths (Gross et al., 2009). However, in mixed traffic without bicycle facilities, narrow roads may present more risk as vehicles pass cyclists closer (Walker, 2007). Moreover, crashes occurring at less busy local roads or multi-use paths can be more severe, as found by Cripton et al. (2015). Haleem and Abdel-Aty (2010) considered the effect of turning traffic on bike crashes and reported that the number of left-turn movements on both the major and minor approaches were significant factors that influenced bicycle risk.

Rumble strips are a proven countermeasure that reduces motor vehicle departure from traffic lanes. Findings in traffic safety literature suggest that rumble strips are effective in reducing bicycle crash rates as well (Garder, 1995; Elefteriadou et al., 2001; Spring, 2003; Torbic et al., 2003; O’Brian, 2009). Zebauers (2005) found that in a passing condition with cars, bicyclists get additional clearance on streets with rumble strips compared to streets where no rumble strips are present.

Wide shoulders are positively associated with cyclist safety. Greater shoulder width is associated with decreased crash rates and less severe injury (Klop & Khattak, 1999; Abdel-Rahim & Sonnen, 2012). Shared markings in the middle of the travel lane tend to encourage cyclists to adjust positions away from the curb and towards the center of the road, increasing their visibility. Wide curb lanes were found to have similar effects as bike lanes because cyclists can use the additional space on roadway to maneuver (Hunter et al., 1999; Metroplan Orlando, 2010).

The presence of horizontal and vertical curves are significant factors in frequent bicycle collisions (Pai, 2011; Elvik & Bjornskau, 2017) and severe injury crashes (Teschke et al., 2014; Klop & Khattak, 1999). Moore et al. (2011) found that crashes on curves or roadways with elevation cause more severe injuries for people bicycling. Kim et al. (2007) and Yang et al. (2011) found that vertical grades increase the likelihood of severe injuries once a crash has occurred.

2.2.2 Intersection Characteristics

Studies have found that intersection safety for bicyclists is influenced by vehicle volume, the presence of heavy vehicles, and speed limit on the major and minor roads (Dixon et al., 2012; Abdel-Aty et al., 2006). Intersections can pose greater safety risk to cyclists compared to mid-block locations (Kaplan & Prato, 2015; Romanov et al., 2012; Wei & Lovegrove, 2013). The configuration and design of intersections greatly influence bicycle crashes, as pointed out by Wang & Nihan (2004). Vandenbulcke et al. (2014) discussed that complicated rights-of-way at intersections, as well as larger and complicated designs at intersections, are a common source of danger for cyclists. The number of legs in an intersection is related to variations in crash frequency as studies show more legs at an intersection are related to an increase in bicycle crashes (Zen & Huang, 2004; Wang et al., 2017; Dumbaugh & Li, 2010; Strauss et al., 2013).

While bicycle paths and bike lanes are found to generally improve safety in busy urban areas, Prati et al. (2018) found that they might increase the risk of collisions at intersections. Kaplan & Prato (2015) posited that intersections and roundabouts in a bicycle network are the common source of crashes. However, Chen (2015) stated that closely spaced intersections may allow motor drivers and cyclists to be more alert to the surrounding traffic and reduce risk of bicycle crashes. According to literature, while roundabouts increase safety conditions for other types of road users, they may have an unfavorable effect on cyclist safety, sometimes leading to an increased risk of crashes (Poudel & Singleton, 2021; Hels & Orozova-Bekkevold, 2007; Daniels et al., 2008; Møller and Hels, 2008; Reynolds et al., 2009).

Wide medians at intersections are negatively associated with the likelihood of bicycle crashes. Medians and 2-way turn lanes often coincide at an intersection and they are associated with reduced bicycle crashes (Zegeer et al., 2006; Schepers et al., 2011). Wide medians are often found to be associated with low crash rates (Saha et al., 2015; Stamatiadis et al., 2009). On the other hand, Park et al. (2015) discussed that higher median width may be indirectly related to bike crashes, as wide medians are commonly present in roadways with multiple lanes and high-volume traffic.

The number of driveways near intersections influences the risk level of bicycles. Pulugurtha & Thakur (2015) suggested that reducing number of driveways to less than

50 per mile can reduce occurrences of bicycle crashes. Others (Gill, 2007; Davis & Hallenbeck, 2008; Shah et al., 2021) found that low driveway density especially near intersections and also along mid-block locations present lower risk to bicyclists.

2.2.3 Traffic Characteristics

Bicycle safety studies have recognized that road network traffic characteristics as such traffic volume (AADT), functional classifications of roadway are contributing factors for bicycle crashes.

Higher motor traffic volume causes an increased number of crashes as well as severe injury for cyclists at intersections (Vandenbulcke, 2014; Kim et al., 2007). Lee et al. (2017) found that traffic volume on volume on the major roads at an intersection had an especially larger impact on crashes. Nordback et al. (2014) found that crashes were equally sensitive to traffic volume. Dixon et al. (2012) presented that high traffic volume in an urban setting is associated with a greater number of bike crashes. At intersections, Haleem and Abdel-Aty (2010) considered the effect of turning traffic on bike crashes and found that the traffic volumes on the major and minor approach, as well as the distance to the nearest signalized intersection, were significant factors that influenced bicycle crash risk.

The presence of public transit stops such as bus stops, light rail, and commuter rails stations are often associated with greater numbers of bicycle crashes, since frequent bicycle activity may be centered around such stations (Pai, 2011; Morrison et al., 2019). Bus stops and bus transit stations can be a common source of bicycle collisions (Quddus, 2008;). Near intersections, the presence of bus stops attracts more bicyclists and can contribute to more bicyclist injuries, as found by Strauss et al. (2013)

Studies have found decreased crash rates along road segments where the speed limit is greater indicating that high speed roads tend to have fewer non-motorized travelers such as bicyclists (Morrison et al., 2019). Alternatively, a greater body of research has found that higher speed limits lead to worse injury severity levels (Haleem & Abdel-Aty, 2010; Kim et al., 2007; Cripton et al., 2015) and more frequent crash occurrences (Eluru 2008; Zahabi et al. 2011).

Bicyclists may also show avoidance behavior for streets with higher percentages of heavy trucks. Large vehicles' presence can significantly affect bicyclists' levels of comfort, and they tend to avoid streets with higher percentages of heavy trucks, as found by Pokorny & Pitera (2019). However, when large vehicles such as trucks are involved in a vehicle–bicycle crash, cyclists are more likely to be severely injured both at mid-block locations and at intersections (Moore et al., 2011; Yang et al., 2011; Walker, 2007; Boufous et al., 2012).

In crash analysis, bicycle exposure has been incorporated in the form of bicycle count from automated count stations or human counts (Guo et al., 2018; Prato et al., 2016), as well as bicycle trip estimation by fitness apps such as the STRAVA Metro app (Chen et al., 2020; Raihan et al., 2019; Sanders et al, 2017) based on the assumption that STRAVA counts are distributed evenly among total bicycle road users. However, accurate and complete bicycle exposure is difficult and time consuming to collect as manual collection of non-motorized volume can present significant error (Nordback et al., 2014; Chen et al., 2020; Roll, 2013), and bicycle app counts represent a low percentage of actual bike trips, as pointed out by Saha et al. (2018). Authors have attempted to account for bicycle exposure using macro-level spatial information such as density and find that bicyclists' higher exposure is associated with improved safety (Lee et al., 2017).

Studies have found that bicycle exposure estimated from STRAVA app is associated with increasing crash rates (Chen et al., 2020; Saad et al., 2019). Schepers et al. (2014) reported that unsignalized intersections with high bicycle volume were associated with higher numbers of crashes. However, there are also a number of studies that suggest a phenomenon of "safety-in-numbers" for bicyclists, since a greater number of cyclists on the road yields lower crash rates and contributes to an overall safety improvement for all road users, as discussed by Marshall & Garrick (2011). In a broader trend, safety-in-numbers effects are a common finding across different cities and regions, but the reasons behind it are incompletely known, and there are variations in the strength of the safety-in-numbers effect (Kroyer, 2015; Daniels et al., 2010; Elvik & Bjornskau, 2017).

2.2.4 Built Environment and Community Characteristics

Land use and community characteristics have an impact on bicycle crashes because depending on land use characteristics, road segments and intersections can face varying levels of bicycle trips and crashes. As expected, urban road network (high density) is related to an increased number of bicycle crashes, and rural road network (low density) is associated with fewer bicycle crashes (Zhai et al., 2019; Saha et al., 2018;

Nolan & Quddus 2004). Greater number of closely spaced intersections (intersection density) is also related to increased bicycle crashes (Wei & Lovegrove, 2013 Sando et al., 2011). Moreover, Pulugurtha & Thakur (2015) suggested that large spacing between unsignalized intersections may help to lower bicycle collisions.

Wier et al. (2009), Pulugurtha & Sambhara (2011), and Ukkusuri et al. (2011) showed that land use patterns such as residential and employment density could influence the occurrence of crashes for non-motorized users such as bicyclists. Guevara et al. (2004), Lee et al. (2015), and Loukaitou Sideris et al. (2007) found that population density was positively related to bicycle crash counts.

Among community characteristics, median household income is negatively associated with crash frequency, as streets and intersections in areas with more lower income households tend to face more crashes (Huang et al., 2017; Lee et al., 2014; Martinez & Veloz, 1996). The fact that low-income areas lag behind in bicycle and pedestrian safety is supported by numerous studies (Britt et al., 1998; Lyons et al., 2008; Siddiqui et al., 2012). Saha et al. (2018) found that fewer automobiles in are associated with increasing crash risk. This finding is similar to the study conducted by Loukaitou-Sideris et al. (2007). Siddiqui et al. (2012) found that neighborhoods with minority populations—including large proportions of Black or African American populations and Hispanic or Latino populations—are associated with increased likelihood of bicycle crashes.

Additional human factor found to contribute to high-risk levels is if the bicyclist was under the influence of alcohol or drugs (Boufous et al., 2012; Schepers & den

Brinker, 2011). Similarly, driving under influence (DUI) is an especially crucial factor that significantly affects severe injury crashes (Eluru et al, 2008; Noland & Quddus, 2004).

2.2.5 Weather Condition and Road Lighting

Inclement weather condition such as fog, snow, or rain are found to contribute to higher bike crashes (Mohan et al., 2006), Klopp & Khattak, (1999), Wanvik (2009) found that bad weather conditions increase the likelihood of increasing number of bike crashes; Kim et al (2007) found that and severe injuries are influenced by bad weather conditions.

Roadway lighting condition can have a substantial effect on bicycle safety at night (Chen 2015; Eluru et al., 2008; Bíl et al., 2010). Boufous et al. (2011) found that a lack of visibility due to darkness increased both crash frequency and fatality rates. Additionally, lack of roadway lighting is likely to cause higher severity crashes (Eluru et al., 2008)

2.3 ML Techniques Investigating Non-Linear Crash Associations

Data driven approaches in conjunction with machine learning are recent developments in non-motorized crash analysis (Luan et al., 2016). Machine learning techniques are non-parametric; a predefined functional form between the dependent and independent variables such as a natural log transformation is not required. Through loss function optimization, machine learning and data mining techniques can be utilized to analyze crash frequency and thus identify problematic links in a road network. The use of machine learning methods such as boosted trees, random forests, or other forms of ensembles of decision tree models are particularly useful because they can produce a prediction model in the form of an ensemble of weak (high variance) decision tree models. These models can reduce prediction error, and they are able to provide interpretable results such as the most contributing factors affecting crashes and non-linear marginal effects of explanatory variables on crashes (Saha et al., 2018; Abdel-Aty & Haleem, 2011). The non-linear and complex marginal effects of roadway geometry, traffic conditions, and land use characteristics are investigated through these models, and the effective ranges of these factors are easy to identify.

Ding et al. (2018) investigated the relationship between pedestrian crash frequency and road network characteristics, land use condition, and traffic characteristics with a Multiple Additive Poisson Regression Tree model. The authors reported effective range and often non-linear relationship between built environment variables and pedestrian crash frequency. Household density, commercial land use, and mixed land use were the most influential variables- accounting for 40% of all effects. Additionally, other variables such as speed limits below 25 mph and intersection density were found to assist in lowering pedestrian crash frequency.

One advantage of such models is that the determined relative contribution of factors in developing the models can help to find the priority of contributing factors (i.e., the ranking of explanatory variables). To quantify crash frequency on rural roads in Indiana, Karlaftis & Golias (2002) developed a hierarchical tree-based model to determine the relative contribution of explanatory variables. Kashani & Mohaymany (2011) applied regression trees to estimate crash characteristics and driving conditions affecting severe injury crashes. Alluri et al. (2012) used a random forest algorithm to find the most important variables in Highway Safety Manual for traffic crashes.

Pulugurtha & Sambhara (2011) investigated the non-linear associations between pedestrian crashes occurring at intersections and explanatory road characteristics. These characteristics included socio-economic characteristics, land use variables, and road geometry information. By estimating negative binomial models, the authors found that pedestrian crashes are affected by an increase in number of transit stops, the number of legs at an intersection, and the pedestrian volume.

Zhu (2021) has conducted pedestrian crash severity analysis with various data mining techniques and compared it to traditional logistic regression models. A data resampling method is applied to avoid the class imbalance issue, and several machine learning models—such as classification and regression tree (CART), gradient boosting (GB), random forest (RF), and artificial neural network (ANN) models—have been applied to predict the severity of crashes. The model results showed that adverse weather conditions such as light rain were associated with fatal and severe injury crashes, and unsignalized intersections with some type of traffic control such as stop signs were associated with low severity crashes.

Shirani-Bidabadi et al. (2020) utilized the multivariate adaptive regression splines (MARS) method to identify non-linear associations between bicycle crashes and crash predictor variables and developed safety performance functions (SPF) in Alabama. The presence of medians on major approaches and the presence of right turn lanes was found to reduce bicycle crashes at unsignalized intersections. Traffic volume (AADT) and bus stops near intersections contributed to increasing the predicted number of crashes.

Ensemble of decision tree models are more sensitive to outliers and effectively reduce high variance. This is particularly helpful in crash analysis, as there may be more zero crashes locations and only a few high crash frequency locations in a road network. These machine learning models are also advantageous for crash analysis because of their ability to rank explanatory variables according to their importance and illustrate their non-linear marginal effects on crash frequency.

2.4 Limitations of Existing Research

Many research studies on pedestrian and bicycle safety are limited by unavailability and incompleteness of exposure data. This is because pedestrian and bicycle exposures are inherently difficult to estimate due to the limitations of traditional data collection and travel demand models (Singleton, 2018; Lee & Sener, 2020). As discussed previously, manual counts are expensive and usually provide coverage of exposure over a small spatial scale. Automated counters for pedestrian and bicycle volumes are also set up targeting specific locations in mind such as streets near to trails. Bicycle exposure data provided by the Strava app is a source of big data, which can be utilized in estimating exposure data for larger spatial scales. While studies have investigated the use of this exposure method in smaller scale for bicycle crash estimations—for a few intersections (Wang et al., 2017, Sanders et al., 2017), a few segments (Raihan et al., 2019), and a few block group levels (Sener et al., 2021)—wide-scale estimation such as the road network of a region or state is very limited in the extant literature.

Moreover, estimating non-parametric machine learning models provide alternate ways to analyze crash frequencies, facilitate ranking of the most important variables, and identify ranges of explanatory variables having non-linear associations with crashes. While studies have explored the use of similar methods for finding important variables for motor crashes (Abdel-Aty & Haleem, 2011; Ahmed & Abdel-Aty, 2013), prioritizing Highway Safety Manual (HSM) variables affecting intersection and segment traffic crashes (Alluri et al., 2012; Saha et al., 2015), and classification methods to understand crash severity and crash types (Harb et al., 2009; Zhu, 2021), application of this recently developed technology is still underexplored in traffic safety research. To the author’s knowledge, no study has conducted network-wide pedestrian and bicycle crash analysis to rank factors’ importance and identify non-linear associations affecting these crash types. Identifying important factors can assist in selecting countermeasures, and detecting ranges of explanatory variables associated with non-motorized crashes provides a better understanding of the possible non-linear associations between roadway characteristics and pedestrian/bicycle crashes

3 DATA COLLECTION

3.1 Overview

This chapter discusses the data collection process including dataset sources and locations of the collected data, detailed descriptive statistics of the crash data, pedestrian and bicycle exposure, road geometry, traffic conditions, and weather data. The details of data assembly and processing are also summarized.

Pedestrian and bicycle crash data were collected from the Numetric website (Numetric, 2019). Detailed road geometry and traffic characteristics of state and federal aid roads in Utah were collected from the UDOT data portal (UDOT Open Data, 2019). Land use information for all block groups in the state of Utah were collected from the Environmental Protection Agency (EPA) Smart Location Database (Smart Location Database, 2014), and socio-demographic data at each block group level were collected from American Community Survey (ACS, 2017) 5-year survey. Weather information was used in the crash severity analysis, and detailed weather station data were collected from the National Centers for Environmental Information (NOAA, 2019). Detailed information about the datasets, relevant variables, and data collection procedures are described in the following sections.

3.2 Study Locations

A systemic safety analysis looks into detailed road conditions and traffic information about study locations to identify factors contributing to crash occurrences. Study locations were split into three types, depending on the roadway type and following common practices: segments and mid-block locations, non-signalized intersections, and signalized intersections. Signalized intersections have dedicated pedestrian and bicycle crossing time- aligned with directional green signal or shown in pedestrian signal heads if present. However, mid-block crossings and unsignalized intersections do not have a clearly defined pedestrian or bike crossing time, and these types of crashes are different as they are related to crossing pedestrian or bicyclist's judgement and yield behavior of motor drivers. This study has collected state routes and federal aid routes to investigate risk factors for pedestrian and bike crashes on unsignalized intersections and at mid-block locations. State routes were more complete in dataset containing more features – especially information regarding road geometry, and federal aid routes were more in numbers (better sample size). Crash data, exposure data for pedestrian and bicyclists, and road geometry data for study locations are collected and analyzed according to these two types of locations for state routes only and state and federal aid routes. Details about each of these types of study locations is described in the following subsections. (Signalized intersections were analyzed in a separate project and so are not included here.)

3.2.1 Segments and Mid-block Locations

Segments and mid-block locations include all locations where a crash occurred away from an intersection of two-or-more streets. To ensure relatively consistent spatial units between different datasets, segments were derived from the links in the "Road Centerlines" geodatabase, obtained in March 2020 from the Utah AGRC website (AGRC, n.d.).

Many of these segments represented local streets, not portions of the state or federal aid highway network. Segments on the state or federal aid highway network were identified by route numbers less than 999 (state routes) or less than 9999 (state and federal aid routes). 4,555 state only road segments and 46,497 state and federal aid route road segments were investigated in this study.

The crash dataset (2010–2019) collected from the Numetric website has been filtered to obtain only pedestrian and bicycle crashes that were not at any intersection. These crashes were identified as mid-block crashes and were matched to the nearest road segment if crashes are at least within 25m up to 50m distance. Segment ID and Unique ID were used as linked identifiers between both datasets.

Similarly, datasets with road geometry information collected from the UDOT data portal were spatially joined with the road centerline network now containing crash information. These datasets are: road centerlines geodatabase, lanes, shoulders, medians, rumble strips, traffic islands, barriers, and driveway shapefiles. Spatial joining of the road network information from various data sources was necessary, considering that the spatial shape lengths were not the same, and they did not always perfectly align at the exact location. On the first step, a spatial joining process was carried out to join these road geometry networks to the nearest road centerline network if the distance was less than 50m. The road centerline network dataset contained route numbers for roads on Utah as linked identifiers. So, on the next step, these road geometry datasets and same route numbers were matched to complete the joining process. The spatial joining process and relevant variables from these datasets are discussed on the following sections.

3.2.2 Non-signalized Intersections

Intersections include all locations where a crash occurred at or very close to an intersection of two-or-more streets. Intersections can be signalized (controlled by a traffic signal) or non-signalized—controlled by one-or-more stop signs or yield signs. This study investigated specifically unsignalized intersection crashes, as signalized intersection crash frequencies were analyzed in a different project. To ensure relatively consistent spatial units between different datasets, junctions were first derived from the nodes in the “Street Network” geodatabase, obtained in March 2020 from the Utah AGRC website (AGRC, n.d.). The links in the “Street Network” geodatabase were compared to those in the “Road Centerlines” geodatabase and were found to be almost identical.

Some of these junctions were signalized intersections, so junctions were allocated between signalized and non-signalized intersections using the following heuristic processes. All thresholds were determined through trial and error and visual inspection of maps.

- Signals were spatially joined to the nearest junction and were matched unless one of the following conditions were true:
    - The distance from the signal to the junction was greater than 75 m.
    - The distance from the signal to the junction was greater than 25 m and the signal was likely to not be at an intersection (e.g., HAWKs, streetcar, fire station, flasher, gantry, queue, or lab).

This study investigated only those intersections without any signal, and a closer look at the geography of road network showed many of the non-signalized intersections were located on local streets, not on the state or federal aid highway network. Junctions were linked to segments using the following heuristic processes. All thresholds were determined through trial and error and visual inspection of maps.

- Segments links were spatially intersected with shapes obtained from a 1 m buffer of junction nodes. In other words, all junctions and segments were matched if they were no more than 1 m away from each other.

Non-signalized intersections on the state or federal aid highway network were identified by any matched segments with route numbers less than 999 (state routes) or less than 9999 (state and federal aid routes). 4,555 intersections on state routes and 50,737 intersections on state and federal aid routes were included in this study.

3.3 Crash Data

Crash data for all study locations from 2010 through 2019 were obtained in August 2020 from the Utah Department of Transportation (UDOT) through the Numetric website. Each crash record contained information on temporal characteristics, spatial characteristics, contributing factors, crash severity, weather conditions, and crash participants. This information was extracted from police crash reports. No personally identifying information was included.

From the set of all crashes over the 10-year study period, bicycle and pedestrian crashes were extracted using the fields “bicycle involved” and “pedestrian involved.” Crashes were also segmented by severity, and fatal and serious injury crashes were extracted using the field “crash severity” and levels “fatal” and “suspected serious injury.” Next, crashes were assigned to segments or mid-block locations, non-signalized intersections, and signalized intersections using the following heuristic procedures. All thresholds were determined through trial and error and visual inspection of maps and crash records.

- Crashes with “false” for the field “intersection involved” were assumed to have occurred along segments or at mid-block locations.
    - These crashes were then spatially joined to the nearest segment with the same route number.
        - If the distance from the crash to the segment was less than or equal to 50 m, then the crash was assigned to that segment.
        - If not, then the crash proceeded to the following steps.
    - Remaining crashes were then spatially joined to the nearest segment, not matching on route number.
        - If the distance from the crash to the segment was less than or equal to 25 m, then the crash was assigned to that segment.
        - If not, the crash was discarded as being too far from a segment.
- Crashes with “true” for the field “intersection involved” were assumed to have occurred at non-signalized or signalized intersections.
    - These crashes were then spatially joined to the nearest junction and to the nearest signal.

    - If the junction was a signalized intersection, then the crash was assigned to that signal.
    - If not, then the crash proceeded to the following steps.
  - Remaining crashes with "signal" in the field "traffic control device" were assumed to have occurred at signalized intersections.
    - If the distance from the crash to the signal was less than or equal to 150 m (and no more than 75 m further away from the signal than from the junction), then the crash was assigned to that signal.
    - If not, for crashes with "ramp intersection with crossroad" in the field "roadway junction type" and if the distance from the crash to the signal was less than or equal to 300 m, then the crash was assigned to that signal.
    - If not, for crashes with "4-leg intersection" in the field "roadway junction type" and if the distance from the crash to the signal was less than or equal to 125 m, then the crash was assigned to that signal.
    - If not, then the crash proceeded to the following steps.
  - Remaining crashes without "signal" in the field "traffic control device" were assumed to have occurred at non-signalized intersections, with one exception:

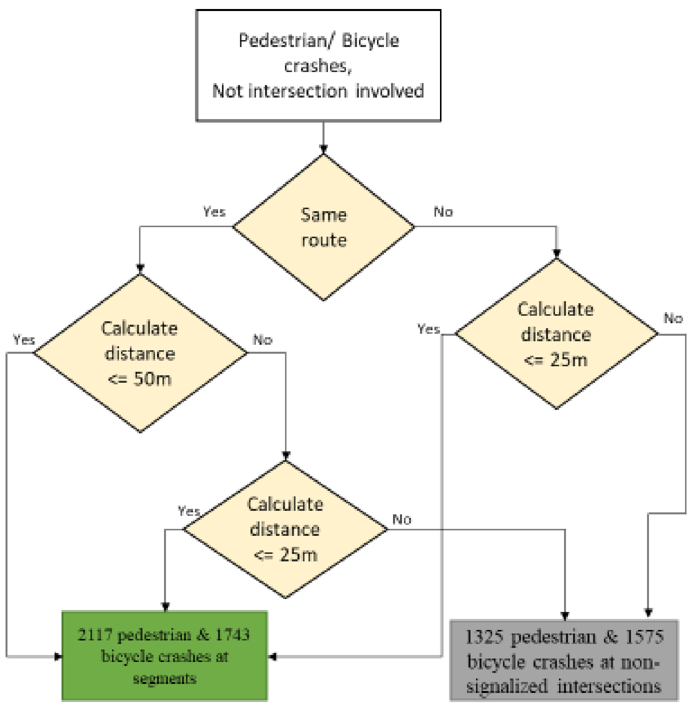
    - If the distance from the crash to the signal was less than or equal to 75 m (and no more than 25 m further away from the signal than from the junction), then the crash was assigned to that signal.
  - Remaining crashes were then spatially joined to the nearest junction (with 3+ legs, and with 2+ legs) and with the same route number, as well as to the nearest junction (with 3+ legs, and with 2+ legs), not matching on route number.
    - If the distance from the crash to the junction with 3+ legs and the same route number was less than or equal to 100 m, then the crash was assigned to that non-signalized intersection.
    - If not, if the distance from the crash to the junction with 2+ legs and the same route number was less than or equal to 100 m, then the crash was assigned to that non-signalized intersection.
    - If the distance from the crash to the junction with 3+ legs and the same route number was less than or equal to 200 m, then the crash was assigned to that non-signalized intersection.
    - If not, if the distance from the crash to the junction with 2+ legs and the same route number was less than or equal to 200 m, then the crash was assigned to that non-signalized intersection.
    - If the distance from the crash to the junction with 3+ legs and any route number was less than or equal to 100 m, then the crash was assigned to that non-signalized intersection.

- If not, if the distance from the crash to the junction with 2+ legs and any route number was less than or equal to 100 m, then the crash was assigned to that non-signalized intersection.
- If not, the crash was discarded as being too far from an intersection.

As shown in Figure 3-1, spatial join for crashes included route matching and distance calculation to filter crashes on segments.

Figure 3-1

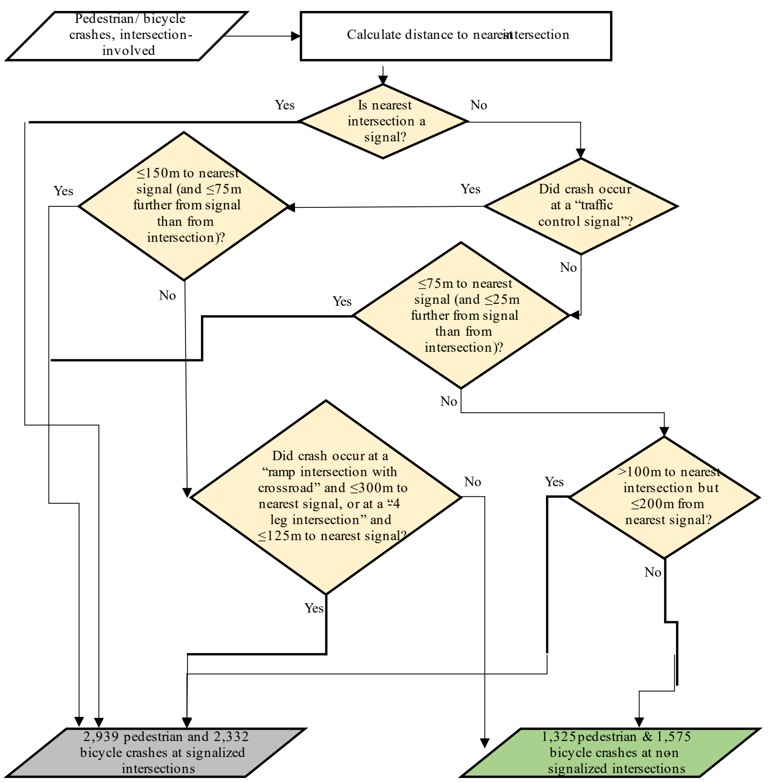
Spatial Join of Crashes on Segments



As shown in Figure 3-2, spatial join for crashes included filtering out segments and signalized intersections, route matching and distance calculation for non-signalized intersections.

Figure 3-2

Spatial Join of Crashes on Intersections



3.3.1 Crash Data on Road Network

Crash distributions in terms of severity at different spatial scales show that at mid-block locations or along segments, the majority of crashes are possible injury and minor injury: 35% and 24% (respectively) for pedestrian crashes on state and federal aid routes, and 37% and 25% for pedestrian crashes on state only routes. Fatal and serious injury crashes are 12% and 21% (respectively) for state and federal aid routes and 9% and 19% for state only routes. Table 3-1 shows descriptive statistics of pedestrian crashes.

Table 3-1

Pedestrian Crashes at Mid-Block Locations along Segments (2010-2019)

	State & Federal Aid Routes		State Routes Only	
Type of crash	#	%	#	%
Pedestrian crashes				
Fatal injury (k)	204	12%	110	9%
Serious injury (a)	404	21%	192	19%
Minor injury (b)	791	35%	316	37%
Possible injury (c)	540	24%	220	25%
No injury (o)	178	7%	69	8%
Total	2117	100%	907	100%

Among bicycle crashes at mid-block locations or along segments, the majority of crashes are possible injury and minor injury: 48% and 31% (respectively) for bicycle crashes on state and federal aid routes, and 49% and 32% for bicycle crashes on state only routes. Fatal and serious injury crashes are 1% and 20% (respectively) for state and federal aid routes and also 1% and 10% for state only routes. Table 3-2 shows descriptive statistics of bicycle crashes.

Table 3-2

Bicycle Crashes at Mid-Block Locations along Segments (2010-2019)

	State & Federal Aid Routes		State Routes Only	
Type of crash	#	%	#	%
Bicycle Crashes				
Fatal injury (K)	21	1%	9	1%
Serious injury (A)	180	10%	83	10%
Minor injury (B)	867	48%	394	49%
Possible injury (C)	547	31%	248	32%
No injury (O)	128	9%	73	7%
Total	1743	100%	807	100%

At intersections, the majority pedestrian crashes are possible injury and minor injury which are 50% and 29% for crashes on state and federal aid routes, and 50% and 34% for crashes on state only routes. Fatal and serious crashes are 2% and 11% for state and federal aid routes and also 1% and 10% intersections for state only routes. Table 3-3 shows descriptive statistics of these variables.

Table 3-3

Pedestrian Crashes at Non-Signalized Intersections (2010-2019)

	State & Federal Aid Routes		State Routes Only	
Type of crash	#	%	#	%
Pedestrian Crashes				
Fatal injury (K)	40	2%	1	1%
Serious injury (A)	186	11%	6	9%
Minor injury (B)	585	50%	28	50%
Possible injury (C)	412	29%	16	34%
No injury (O)	102	9%	5	6%
Total	1325	100%	56	100%

Similarly, among bicycle crashes at intersections, majority crashes are possible injury and minor injury which are 50% and 29% for bicycle crashes on state and federal aid routes, and 50% and 34% for bicycle crashes on state only routes. Fatal and serious crashes are 2% and 11% for state and federal aid routes and 1% and 9% for state only routes. Table 3-4 shows descriptive statistics of bicycle crashes.

Table 3-4

Bicycle Crashes at Non-Signalized Intersections (2010-2019)

	State & Federal Aid Routes		State Routes Only	
Type of crash	#	%	#	%
Bicycle Crashes				
Fatal injury (K)	13	2%	1	1%
Serious injury (A)	138	10%	6	10%
Minor injury (B)	782	54%	33	49%
Possible injury (C)	542	26%	16	32%
No injury (O)	103	8%	5	7%
Total	1575	100%	61	100%

In this study, the spatial joining process merged appropriate pedestrian and bicycle crashes with nearby segments and non-signalized intersections to calculate crash frequencies. For the crash frequency analysis at mid-block locations, 12,204 segments were found to be on state routes, with a mean of 0.06 pedestrian crashes and 0.07 bicycle crashes. 46,497 segments were found to be on state and federal aid routes, with means 0.05 pedestrian and 0.03 bicycle crashes per intersection. Larger standard deviations compared to means shows the overdispersion of the crash frequency data. 4,555 intersections were found to be on state routes, with mean 0.03 pedestrian and bicycle crashes. 50,737 intersections were found to be on state and federal aid routes with mean 0.02 pedestrian and bicycle crashes per intersection. Larger standard deviations compared to means shows the overdispersion of the crash frequency data. Descriptive statistics are shown in Table 3-5.

Table 3-5

Descriptive Statistics for Crashes on Segments & Non-Signalized Intersections (2010-2019)

	State & Federal Aid Routes		State Routes Only	
	Mean	Standard Deviation	Mean	Standard Deviation
Segments				
Pedestrian Crashes	0.05	0.26	0.06	0.31
Bicycle Crashes	0.03	0.23	0.07	0.34
Non-signalized Intersections				
Pedestrian Crashes	0.03	0.21	0.02	0.13
Bicycle Crashes	0.03	0.20	0.02	0.15

3.3.2 Individual Crash Characteristics

In order to conduct the pedestrian crash severity analysis, information linked to each crash was considered. Variable information—such as road network characteristics, human factors, vehicle characteristics, and crash characteristics—were available in the crash dataset for each of the 6,740 pedestrian crash observations. For the bicycle crash severity analysis, information such as road network characteristics, human factors, vehicle characteristics, and crash characteristics were available in the crash dataset for each of the 5,764 crash observations. Descriptive statistics for both of these crash characteristics are shown in Table 3-6.

Table 3-6

Pedestrian & Bicycle Crash Characteristics (N= 6740)

	Pedestrian crash characteristics, N= 6740		Bicycle crash characteristics, N= 5764	
	#	%	#	%
Crash severity				
No injury	436	6%	442	6%
Possible injury	2138	32%	1932	32%
Minor injury	2941	44%	2869	44%
Serious injury	947	14%	482	14%
Fatal injury	278	4%	39	4%
Lighting condition				
Lighted	3963	58%	3228	58%
Unlighted	2764	41%	2306	41%
Vehicle body type				
Small (Sedan, motorcycle)	3640	54%	3113	54%
Large (SUV, Pickup, Bus)	4100	46%	2651	46
Roadway surface condition				
Dry	5864	86%	4380	76%
Wet	886	14%	1384	24%
Crash involving				
Drivers at fault (Disregarding traffic control/ Distraction/ Drowsy Driving)	89	1%	230	4%
DUI	190	2%	330	5%
Work zone	492	19%	979	17%
Crash location				
Mid-block	2500	37%	1902	33%
Intersection	4240	63%	3862	67%
Horizontal curve (present)	206	0.5%	192	1%
Vertical grade (present)	716	10%	900	15%

3.4 Exposure Data

Pedestrian and bicycle exposure are some of the key variables in crash frequency modeling. While traffic volumes are common ways to represent exposure in road network, pedestrian and bicycle exposure helps to determine the level of walking and bicycling activity on road segments and at intersections. However, sufficiently complete pedestrian and bicycle volume data are often unavailable due to the difficulty involved in collecting these sorts of exposure data. This study has attempted to overcome this challenge by using pedestrian exposure data collected from traffic signals and bicycle exposure data from the STRAVA mobile application datasets in the crash analyses.

The pedestrian exposure data for non-signalized intersections were derived from direct demand models of annual average daily pedestrian volumes utilizing pedestrian traffic signal data and estimated crossing volumes from signalized intersections. The bicycle exposure data were collected from the STRAVA Edge dataset and the STRAVA Node dataset. These datasets contained average daily count of bicycle trips for road segments and intersections in Utah for the year 2019. More details about these datasets are presented in the following sections.

3.4.1 Bicycle Exposure Data

The STRAVA dataset created by STRAVA METRO (Strava Metro, 2019) is a large collection of aggregated and de-identified bicycle trip information. It is a recently developed source of trip data that is used by urban planners, engineers, and researchers to understand non-motorized mobility patterns. This recently emerged database has been used in transportation research—travel demand estimation (Roll, 2018), infrastructure evaluation (Skov-Peterson et al., 2017), and crash exposure (Sanders et al., 2017)—albeit in a limited manner. The data are collected from users' phone apps that help people keep track of their rides using GPS. The aggregated and de-identified dataset of bicycle counts in Utah for the year 2019 was used in this study. However, STRAVA app counts are a small sample, and may not be representative of all bicycle trips made in Utah.

In order to make sure bicycle trip data were assigned to the road segment network accurately, the trip information dataset and UDOT road network geodatabase collected from STRAVA were first joined together by Edge IDs. On the next step, the joined dataset and the road segments derived from the links in the "Road Centerlines" geodatabase were spatially joined. This two-step process allowed the joined dataset from STRAVA and the road network segments from the "Road Centerlines" geodatabase to be inspected and filtered according to the segment Unique IDs. This network of road segment information now contained bicycle trip counts, and this broad network was filtered to identify state and federal aid routes as well as state routes with bicycle trip information.

Aggregated STRAVA data were available for intersections as well. The bicycle trip information dataset and UDOT nodes geodatabase collected from STRAVA were first joined together. On the next step, the joined dataset and the street junction network derived from the links in the AGRC geodatabase were spatially joined to the nearest features. This two-step process was required because the datasets from the two different sources did not contain common identifiers. This network of intersections now contained bicycle trip counts, and this broad network was filtered to identify non-signalized intersections on state and federal aid routes as well as on state routes.

The total annual count of bicycle trips in the roll-up data were averaged over all days in the year 2019 to calculate the annual average daily bicyclist (AADB) volumes. In this analysis, the mean AADB on state route segments was 2.18 with a standard deviation of 4.46. AADB count on state route segments ranged from less than 1 to 87. On the state and federal aid route segments, the mean AADB was 2.43 with a standard deviation of 5.02. AADB on state and federal aid route segments ranged from 5 to 31,630. For intersections, the mean AADB at intersections on state route segments was 1.25 with a standard deviation of 4.32. AADB at intersections on state route segments ranged from less than 1 to 104. At intersections on the state and federal aid route segments, the mean AADB were 1.24 with a standard deviation of 4.00. AADB at intersections on state and federal aid route segments ranged from less than 1 to 105. All these values are for the year 2019. Table 3-7 shows these descriptive statistics.

Table 3-7

Descriptive Statistics of Bicycle Volume on Road Segments (2019)

Variable	State routes		State and federal aid routes	
	Mean	Standard deviation	Mean	Standard deviation
On segments	2.18	4.46	2.43	5.02
At intersections	1.25	4.32	1.24	4.00

3.4.2 Pedestrian Exposure Data

One unique aspect and contribution of this study is the use of novel and more complete pedestrian exposure data, which (as the literature review noted) is often missing from pedestrian safety studies. Pedestrian count from signals are a ubiquitous and relatively consistent source of data for the Utah road network. Hence, they were chosen as the exposure variable for this analysis. The pedestrian exposure data used here came from traffic signals, specifically derived from pedestrian activity events at signalized intersections that were recorded in high-resolution traffic signal controller logs (Sturdevant et al., 2012). When a traffic signal includes walk indications and pedestrian pushbuttons for detection, two relevant events can be recorded. First, pedestrian detection events occur whenever the push-button is pressed, which could happen multiple times per signal cycle. Second, a pedestrian call registered event is recorded the first time in a cycle (usually) that a push-button is pressed for a particular phase or crossing. Either (or both) of these events may be used as a proxy for pedestrian crossing volumes, which is the typical measure of pedestrian exposure, within a given time period.

Although pedestrian traffic signal data are not perfect measures of pedestrian volumes (Blanc et al., 2015; Kothuri et al., 2017), recent work in an earlier UDOT research project by (Singleton et al. 2020, Singleton & Runa, 2021) has demonstrated that such data can be used to predict pedestrian crossing volumes at signalized intersections with relative accuracy. Throughout 2019, more than 10,000 hours of videos of pedestrian crossing events were recorded at 90 signalized intersections throughout Utah, and more than 175,000 pedestrians were manually counted. These data were then compared to traffic signal push-button-based measures of pedestrian activity, using simple non-linear (quadratic and piece-wise linear) regression models predicting hourly pedestrian crossing volumes as a function of pedestrian signal activities. Over more than 22,500 hours of data, the correlation between observed and model-predicted hourly pedestrian crossing volumes was 0.84, with a mean absolute error of only 3.0 (Singleton et al., 2020). Overall, that research project demonstrated that pedestrian signal data can be used to estimate reasonably accurate pedestrian crossing volumes. For the purposes of this research project, these pedestrian signal data provide greater temporal and spatial coverage for measuring pedestrian exposure (more locations over longer time periods), thus improving the understanding of relationships between pedestrian crashes and pedestrian volumes.

Due to data and scale challenges with including pedestrians in regional travel demand forecasting models (Singleton et al., 2018), planners interested in facility-specific information on walking activity levels have instead turned to using direct demand models (Kuzmyak et al., 2014) which predict pedestrian volumes using observed counts and measures of the surrounding streetscape, land uses, built environment, and street network. Recent work by Singleton, et al. (2021) has developed direct demand models of annual average daily pedestrian volumes utilizing pedestrian traffic signal data and estimated crossing volumes from signalized intersections.

To obtain pedestrian exposure data for non-signalized intersections and for segments in this project, estimated pedestrian crossing volumes for all intersections from Singleton et al, (2021) were first obtained. Then, an iterative process was used to assign estimated pedestrian volumes to adjacent segments and junctions, using the following heuristic procedures. These procedures assume that: people make an average of one crossing per intersection (including people who cross two legs and those who turn a corner without crossing the street), people walk along a total of two legs (one from, one to) when passing through an intersection, and there are four legs at each intersection. Thus: # crossings ÷ (1 crossing / person) = # people at an intersection, and # people × (2 legs / person) × 4 legs = 50% × # crossings = # people on a segment.

- Step 1a: Transfer pedestrian volumes to segments from adjacent junctions.
- Step 1b: Calculate segment pedestrian volumes to be 50% of the average of pedestrian volumes from all adjacent junctions.
- Step 2a: Transfer pedestrian volumes to junctions from adjacent segments.
- Step 2b: Calculate junction pedestrian volumes to be 200% of the average of pedestrian volumes from all adjacent segments, or the originally estimated pedestrian volumes if available.
- Step 3: Repeat Steps 1 and 2 once more, and then repeat Steps 1 again.

Due to limitations in the signal data as well as the data used to develop the built environment regression models in Singleton et al., (2021), estimated pedestrian volumes were only available for most intersections located in the six most populous Utah counties: Salt Lake, Utah, Davis, Weber, Washington, and Cache. Thus, pedestrian exposure data was only available for non-signalized intersections and segments in these counties.

In this analysis, the average annual daily pedestrian (AADP) volume on state route segments was 44 with a standard deviation of 58. AADP on state route segments ranged from less than one to 616. On the state and federal aid route segments, the mean AADP was 42 with a standard deviation of 53. AADP on state and federal aid route segments ranged from less than one to 721. For intersections, the mean AADP at intersections on state route segments was 93 with a standard deviation of 125. AADP at intersections on state route segments ranged from less than one to 1,057. At intersections on the state and federal aid route segments, the mean AADP was 454 with a standard deviation of 161. AADP at intersections on state and federal aid route segments ranged from less than one to 1,635. All these values are for the year 2019. Descriptive statistics of pedestrian volume are shown in Table 3-8.

Table 3-8

Descriptive Statistics of Pedestrian Volume on Road Segments (2019)

Variable	State route		State & Federal aid route	
	Mean	Standard deviation	Mean	Standard deviation
On segments	44.34	57.91	42.37	52.80
At intersections	93.22	125.14	160.70	453.56

3.5 Roadway and Community Characteristics Data

Datasets with road geometry information collected from the UDOT data portal were spatially joined with the “Road centerline” network from the Utah AGRC (Automated Geographic Reference Center) to create one dataset containing information about each segment. These datasets included road geometry information such as the number and width of lanes, shoulders, medians, roadside and centerline rumble strips, traffic islands, barriers, driveway numbers, curb ramps, as well as traffic variables such as traffic volume, speed limit, bus stops, and light and commuter rail stations. These datasets were collected from the UDOT data portal (UDOT, 2019) as shapefiles. Community variables regarding employment density (jobs /acre), residential density

(Housing unit/ acre), household income (USD), vehicle occupancy, disabled populations (%), and demographic ethnicity information (%) were collected from American Community Survey's (ACS, 2017) 5-year survey data and the EPA (Environmental Protection Agency) Smart Location Database (EPA, 2019).

This study investigated 13,107 state routes and 46,497 state and federal routes. Although the sample containing state routes only provided fewer observations, the state only routes provided more information about roadway geometry and traffic variables. These variables included detailed road geometry information about lanes, shoulders, medians, rumble strips, barriers, sidewalks, traffic volume and composition, pedestrian volume, bicyclist volume, transit stop locations, speed limit, driveway count, driveway density (number of driveways per quarter mile), percent of vertical grade on road segments. There was also information about adjacent land use and neighborhood community characteristics.

On the other hand, state and federal aid routes provided a larger sample size but fewer variables regarding road geometry. These variables did not include the road geometry variables, but did contain traffic, pedestrian, and bicycle volumes, speed limit, driveway, roadway vertical grade information, and neighborhood community characteristics.

This study investigated 4,555 intersections on state routes and 50,737 intersections on state and federal routes. Although the sample containing state routes only provided fewer observations, the state only routes provided more information about roadway geometry and traffic variables. These variables included intersection geometry information about the number of legs, lanes, shoulders, medians, traffic volume and composition, pedestrian volume, bicyclist volume, transit stop locations, speed limit, driveway count, and percent of vertical grade on road segments. There was also information about adjacent land use and neighborhood community characteristics. On the other hand, state and federal aid routes provided a larger sample size but fewer variables regarding road geometry. These variables did not include the road geometry variables, but did contain traffic, pedestrian, and bicycle volume, speed limit, driveway, roadway vertical grade information, land use, and neighborhood community characteristics. Table 3-9 presents road geometry, traffic characteristics and community characteristics for segments and unsignalized intersections at a glance.

Table 3-9

Data Types Investigated at Segments and at Non-Signalaized Intersections

Data type	Segments	Non-signalized intersections
Roadway Geometry Characteristics	• Number of through lanes • Through lane width • Road segment length • Number of left turn lanes • Number of right turn lanes • Number of two-way turn lanes • Number of passing lanes • Number of acceleration lanes • Number of deceleration lanes • Number of transit lanes • Presence of bike lanes • Presence of shoulder • Shoulder type • Presence of painted island • Raised island • Median type • Rumble presence • Percent of vertical grade • Left barrier type • Center barrier type • Right barrier type • Island type • Percent of vertical grade	• Number of approaches at intersections, • Major and minor road lane width • Major and minor road total lanes, • Major and minor road through lanes • Major and minor road left turn lanes • Major and minor road right turn lanes • Major and minor road two-way lanes • Major and minor road with median • Major road with bike lane • Major and minor road with left shoulder • Major and minor road with right shoulder, • Major and minor road with painted island • Major and minor road with raised island • Major and minor road median width • Major and minor road island width • Major and minor road shoulder width • Percent of vertical grade
Roadway Traffic Characteristics	• Bus stations on road segment • Commuter rail station within quarter mile of road segment • Light rail station within quarter mile of road segment • Speed limit • Cartocode • One way street • Traffic volume (AADT), • Truck volume (%), • Driveway density on major roads • Driveway density on minor roads	• Bus stop • Commuter rail station • Light rail station • Maximum speed limit (mph) • Major and minor road traffic volume (AADT) • Major and minor road truck volume • Distance to nearest intersection • Distance to nearest traffic signal.
Neighborhood Community Characteristics	• Residential density (Housing unit/acre) • Employment density (Jobs/ acre) • Zero vehicle household • Jobs per household • Household income (USD) • Percentage of disabled population • Percentage of 'non-white' population including ethnic communities	• Residential density (Housing unit/acre) • Employment density (Jobs/ acre) • Zero vehicle household • Jobs per household • Household income (USD) • Percentage of disabled population • Percentage of 'non-white' population including ethnic communities

3.6 Weather Data

Weather data collected for each station in the six major counties in Utah were matched with individual pedestrian and bicycle crashes by matching date and location. It is to be noted that the six major counties (Cache, Davis, Salt Lake, Utah, Washington, and Weber) were selected to be included in the study considering the availability and completeness of datasets provided by weather stations located in those counties.

Weather stations provided data regarding daily maximum and minimum temperature (F), precipitation (inch), snowfall (inch), and snow depth (inch). However, not every station had information about all the selected variables, and the availability of data over time and for different weather variables were varied by stations as well. After trial and error, only the stations with more available observations for each weather variable were included in the analysis.

3.6.1 Preparing Data for Weather Information

First, pedestrian and bicycle crashes were spatially joined with the nearest weather stations if they were within 30 miles distance. Next, the dates of crash occurrences and the dates of weather station reported data were used to merge and create datasets that contained the crash characteristics as well as weather variables (daily averages) for the pedestrian and bicycle crash severity analysis. To ensure the completeness of observations, and to provide a large enough sample size, this process was repeated for different combinations of weather stations. Finalized datasets contained

6,740 pedestrian crashes and 5,764 bicycle crashes, with crash characteristics and weather information linked to each crash observation.

Next, data for each weather station in the six major counties in Utah were collected from the National Oceanic and Atmospheric Administration (NOAA) online data portal. Daily summaries of weather variables were collected between 2010-01-01 and 2019-12-31. This is the same date range as the pedestrian and bicycle crash data analyzed in this study. The selected variables—daily maximum and minimum temperature, precipitation, snowfall and snow width—were assumed to be most relevant to pedestrian and bicycle traffic crashes.

Table 3-10 presents the minimum, maximum, mean, standard deviation, skewness, and kurtosis for precipitation, snowfall, snow depth, min, and max temperatures observed for 6,740 pedestrian crashes. Among these observations, more than a few missing observations were found: 478 temperature observations, 549 precipitation observations, 968 snowfall, and 1,021 snow depth observations were missing. Missing data were removed before carrying on the statistical analysis.

Table 3-10

Descriptive Statistics of Weather Data for Pedestrian Crashes

Variable	Min	Max	Mean	Std. Dev.	Skewness	Kurtosis
Precipitation (in)	0	2.18	0.06	0.15	5.74	42.65
Snowfall (in)	0	15	0.14	0.75	8.57	88.55
Snow depth (in)	0	15	0.66	2.19	2.96	9.09
Min temperature (°F)	-15	84	41.98	15.45	-0.14	-1.00
Max temperature (°F)	4	105	63.73	21.42	-0.07	-1.25

To inspect the issue of multicollinearity, Pearson correlations were calculated among the five weather variables. Min/max temperatures were strongly correlated with each other (+0.96), and snow depth was moderately correlated with min/max temperatures (−0.42, −0.44) and snowfall (+0.41). Precipitation was very weakly correlated with min/max temperatures (−0.14, −0.15) and low-to-moderately correlated with snow depth (+0.12) and snowfall (+0.43).

Similar descriptive statistics were calculated for 5,764 bicycle crashes as well. Table 3-11 presents the minimum, maximum, mean, and standard deviation, skewness, and kurtosis for precipitation, snowfall, snow depth, min, and max temperatures. Among these observations, more than a few missing observations were found. 362 temperature observations, 459 precipitation observations, 1,020 snowfall, and 1,058 snow depth observations were missing. Missing data were removed before carrying on the statistical analysis.

Table 3-11

Descriptive Statistics of Weather Data for Bicycle Crashes

Variable	Min	Max	Mean	Std. Dev.	Skewness	Kurtosis
Precipitation (in)	0	1.97	0.04	0.13	6.74	39.65
Snowfall (in)	0	10.30	0.04	0.40	7.68	86.55
Snow depth (in)	0	19	0.20	1.24	3.02	9.19
Min temperature (°F)	-8	89	49.39	14.28	-0.15	-1.05
Max temperature (°F)	12	105	73.83	17.23	-0.07	-1.28

To inspect the issue of multicollinearity, Pearson correlations were calculated among the five weather variables. Min/max temperatures were strongly correlated with each other (+0.92) and snow depth was moderately correlated with min/max temperatures (−0.35, −0.32) and snowfall (+0.36). Precipitation was very weakly correlated with min/max temperatures (−0.16, −0.11) and with snow depth (+0.03) and snowfall (+0.28).

3.7 Data Preparation and Processing

3.7.1 Preparing Data for Segments and Mid-block Locations

Most attributes for data about roadway characteristics were provided in spatial files with line features, and the lines may or may not have matched perfectly to the spatial line features used to define segments. For example, some segments were shorter than, longer than, or overlapping the relevant links from the roadway data files. Therefore, a spatial matching process was needed. In most cases, roadway shapefile lines were first buffered (usually using a 5 m buffer), and then segments were spatially joined to the buffered roadway lines, only links with the same route number as the segment were retained, and relevant attributes were transferred over to segments.

In some cases, there were multiple matches of roadway characteristics to segments, such as in the case when a segment overlapped with two roadway shapefile lines, leading to multiple values for each attribute. In these situations, a data consolidation process was needed to obtain one value for each attribute. Depending on the attribute, one of four functions was applied to consolidate multiple values into one value:

- Longest distance: Attributes were tabulated according to their unique values, and the total link lengths of each attribute value were calculated. The value present for the longest total distance was retained. This was the most common function applied, especially for categorical or integer attributes.

- Distance-weighted average: The attribute values were multiplied by the link lengths and divided by the total lengths, yielding a distance weighted average attribute value. This was commonly applied to continuous numeric attributes that measured widths or heights.
- Maximum: The largest value among all values was retained. This was used only for a few categorical or integer attributes where the maximum was more relevant.
- Sum: The sum of all values was used. This was used only for continuous numeric attributes that measured lengths.

Table 3-12 represents the categorical variables measured for each road segment in the year 2019. After filtering for missing variables, categorical variables for 12,204 segments at state routes are represented in terms of frequency and percentage.

Table 3-12

Descriptive Statistics for Segments (State Routes Only): Categorical Variables

Variables		#	%
Cartocode	Local Road	134	1%
	State highway	9995	76%
	US highway	2978	23%
One Way Street	Present	338	2.50%
	Absent	46159	97.50%
Through lanes	1 lane	45	0.01%
	2 lanes	8400	65%
	3 lanes	324	2.50%
	4 lanes	2960	23%
	5 lanes	191	1.50%
	6 lanes	777	6.10%
	7 lanes	24	0.19%
	8 lanes	46	0.36%
Left turn lanes	No left turn	8279	64.80%
	1 lane	1863	14.60%
	2 lanes	2234	17.50%
	3 lanes	214	1.70%
	4 lanes	179	1.40%
Right Turn lanes	No right turn	9210	72%
	1 lane	2349	18.40%
	2 lanes	1189	9.30%
	3 lanes	21	0.20%
Two-way turn lane	Present	2459	19.30%
	Absent	10310	80.10%
Passing Lane	Present	283	2.20%
	Absent	12486	97.80%
Acceleration Lane	Present	449	3.50%
	Absent	12320	96.50%
Deceleration Lane:	Present	74	0.60%
Absent	Absent	12695	99.40%
Transit Lane	Present	10	0.05%
	Absent	12759	99.50%
Bike Lane	Present	666	5.30%
	Absent	12103	94.70%
Shoulder	Absent	11245	88%
	Present	1524	12%
Shoulder type:	Concrete Curb	3272	32.80%
	Gravel	7034	61.90%
	Barrier	179	1.60%
	Other	417	3.70%
Painted Island	Present	111	1%
	Absent	12658	99%
Raised Island	Present	27	0.03%
	Absent	12742	99.70%
Median Type	Undivided	8482	66.80%
	Painted	390	3.20%
	Raised/ Separate grade	1271	10%
	Two-way left turn lane	2543	20%

Rumble presence	Present	3004	23.20%
	Absent	9895	76.80%
Left barrier type	Present	1465	12%
	Absent	11642	88%
Center barrier type	Present	401	3%
	Absent	12706	97%
Right barrier type	Present	1739	13.20%
	Absent	11368	86.80%
Island type	No island	11807	90%
	Median island	1300	10%
Commuter rail station within 400m of road segment	0	46309	
	1	188	
Light rail station within 400m of road segment	0	45838	
	1	550	
	2	90	
	3	18	
	4	1	
Speed limit	25 mph or less	15784	4%
	30 mph	1971	4.20%
	35 mph	4229	9%
	40 mph	10738	23%
	45 mph	3521	7.50%
	50 mph	365	0.40%
	55 mph	3071	6.60%
	60 mph	14	0.01%
	65 mph	5868	12.60%
	70 mph	121	0.20%
	75 mph	759	1.63%

Table 3-13 summarizes the continuous variables collected for 4,555 number of road segments on state routes in the year 2019. The mean represents each segment's road geometry, traffic characteristics, and community neighborhood information. High standard deviation compared to mean represents greater variability.

Table 3-13

Descriptive Statistics for Segments (State Routes Only): Continuous Variables

Variable	Mean	Standard deviation
Percent of vertical grade	1.51	1.55
Traffic volume	10818	12769
Bus stops	0.18	0.45
Population density	0.63	1.14
Employment density	1.27	3.60
Zero vehicle household	1.90	4.34
Jobs per household	11.51	352.99
Disabled population	0.69	0.09
Non-white population	0.11	0.14

Table 3-14 represents the categorical variables measured for each road segments on all state and federal aid routes in the year 2019. After filtering for missing variables, categorical variables for 46,497 segments on state and federal aid routes are represented in terms of frequency and percentage.

Table 3-14

Descriptive Statistics for Segments (State & Federal Aid Routes): Categorical Variables

Variable		#	%
	Local Road	131	1.02%
Cartocode	State highway	9961	75.83%
	US highway	3015	23.24%
One Way Street	Present	2896	2.50%
	Absent	36460	97.50%
Number of commuter rail station within 400m of road segment	0	46309	91.27%
	1	188	0.37%
Number of light rail station within 400m of road segment	0	45838	90.34%
	1	550	1.08%
	2	90	0.18%
	3	18	0.04%
	4	1	0.00%
Speed limit	25 mph or less	15784	30.78%
	30 mph	1971	3.95%
	35 mph	4229	7.54%
	40 mph	10738	20.80%
	45 mph	3521	6.80%
	50 mph	365	1.30%
	55 mph	3071	5.75%
	60 mph	14	0.05%
	65 mph	5868	12.10%
	70 mph	121	0.05%
	75 mph	759	1.10%

Table 3-15 summarizes the continuous variables collected for 4,555 road segments on state and federal aid routes in the year 2019.

Table 3-15

Descriptive Statistics for Segments (State & Federal Aid Routes): Continuous Variables

Variable	Mean	Standard Deviation
Percent of vertical grade	1.46	1.52
Traffic volume	8906	16244
Bus stops	0.15	0.45
Population density	0.63	1.14
Employment density	1.63	3.60
Zero vehicle household	0.03	0.04
Jobs per household	46.80	804.44
disabled population	0.69	0.09
Non white population	0.11	0.14

3.7.2 Preparing Data for Non-signalized Intersections

Most attributes for data about roadway characteristics were provided in spatial files with line features, not point features. Therefore, most attributes for non-signalized intersections had to be transferred over and derived from attributes for the adjacent segments. (See Section 3.2.2for details on how junctions and segments were matched.) Since most intersections had at least three adjacent segments, leading to multiple values for each attribute, a data consolidation process was needed to obtain one value for each attribute. Depending on the attribute, one of four functions was applied to consolidate multiple values into one value:

- Mean: The arithmetic mean (or average) of all attribute values was calculated, and this single value was retained. This was used for most attributes, including integer and continuous numeric attributes like counts and widths.

- Maximum: The largest value among all values was retained. This was used only for a few categorical or integer attributes where the maximum was more relevant.

Table 3-16 summarizes the categorical variables collected for each unsignalized intersections (N = 4,555) on state routes in the year 2019. Major and minor leg information were collected separately and represented in terms of frequency and percentage.

Table 3-16

Descriptive Statistics for Non-Signalized Intersections at State Routes Only: Categorical Variables

Variables		#	%
Number of Legs	1	62	1.36%
	2	2243	49.24%
	3	1192	26.17%
	4	984	21.60%
	5 or more	74	1.76%
Major road 2way lanes	Present	769	16.88%
	Absent	3376	74.12%
	Not available	410	9.00%
Minor road 2way lanes	Present	38	0.83%
	Absent	341	7.49%
	Not available	4176	91.68%
Major road with median	Present	1	0.02%
	Absent	4144	90.98%
	Not available	410	9.00%
Minor road with median	Present	0	0.00%
	Absent	379	8.32%
	Not available	4176	91.68%
Major road with transit lane	Present	3	0.07%
	Absent	4142	90.93%
	Not available	410	9.00%
Minor road with transit lane	Present	0	0.00%
	Absent	379	8.32%
	Not available	4176	91.68%
Major road with bike lane	Present	282	6.19%
	Absent	3863	84.81%
	Not available	410	9.00%
Minor road with bike lane	Present	25	0.55%
	Absent	354	7.77%

	Not available	4176	91.68%
Major road with left shoulder	Present	620	13.61%
	Absent	3525	77.39%
	Not available	410	9.00%
Minor road with left shoulder	Present	28	0.61%
	Absent	351	7.71%
	Not available	4176	91.68%
Major road with right shoulder	Present	3709	81.43%
	Absent	436	9.57%
	Not available	410	9.00%
Minor road with right shoulder	Present	328	7.20%
	Absent	54	1.19%
	Not available	4173	91.61%
Major road with painted island	Present	115	2.52%
	Absent	4030	88.47%
	Not available	410	9.00%
Minor road with painted island	Present	11	0.24%
	Absent	368	8.08%
	Not available	4176	91.68%
Major road with raised island	Present	24	0.53%
	Absent	4121	90.47%
	Not available	410	9.00%
Minor road with raised island	Present	0	0.00%
	Absent	379	8.32%
	Not available	4176	91.68%
Bus stop	Present	226	4.96%
	Absent	4329	95.04%
Commuter rail station	Present	16	0.35%
	Absent	4539	4.61%
Light rail station	Present	71	1.56%
	Absent	4484	93.83%
Maximum speed limit (mph)	Below 25	224	4.92%
	30	57	1.25%
	35	228	5.01%
	40	1774	38.95%
	45	162	3.56%
	50	166	3.64%
	55	210	4.61%
	Over 60	1734	38.07%

Table 3-17 summarizes the continuous variables collected for 4,555 number of unsignalized intersections on state routes in the year 2019.

Table 3-17

Descriptive Statistics for Unsignalized Intersections at State Routes Only: Continuous Variables

Variable	Mean	S.D.
Major road lane width	11.94	1.12
Minor road lane width	12.20	1.70
Major road total lanes	2.98	1.58
Minor road total lanes	3.64	2.11
Major road through lanes	2.68	1.16
Minor road through lanes	2.84	1.26
Major road left turn lanes	0.5	0.80
Minor road left turn lanes	1.15	1.16
Major road right turn lanes	0.32	0.56
Minor road right turn lanes	0.64	0.73
Major road median width	8.96	35.2
Minor road median width	14.16	42.88
Grade %	1.91	1.68
Major road island width	1.3	4.32
Minor road island width	2.69	8.60
Major road shoulder width	4.86	3.22
Minor road shoulder width	4.40	3.28
Major road traffic volume (AADT)	14230	19541
Minor road traffic volume (AADT)	7283	9177
Major road truck volume	2671	3457
Minor road truck volume	2032	2489
Distance to nearest intersection	319	726
Distance to nearest traffic signal	11866	20546
Zero vehicle household	0.03	0.04
Residential density	0.64	1.11
Employment density	1.57	4.1
Jobs per household	62.5	955
Household income	64876	21965
Disabled population	0.69	0.08
Non-white population	0.10	0.12

Table 3-18 represents the categorical variables measured for each non-signalized intersection at state and federal aid routes in the year 2019. After filtering for missing variables, categorical variables for 50,737 non-signalized intersections at state and federal aid routes are represented in terms of frequency and percentage.

Table 3-18

Descriptive Statistics for Non-Signalized Intersections at State and Federal Aid Routes: Categorical Variables

Variables		#	%
N Legs	1	273	0.54%
	2	13369	26.35%
	3	28060	55.30%
	4	8765	17.28%
	5 or more	270	1%
Bus stop	Present	2768	5.46%
	Absent	47969	94.54%
Commuter rail station	Present	122	0.24%
	Absent	50615	5.22%
Light rail station	Present	507	1.00%
	Absent	50230	93.79%
Maximum speed limit (mph)	<25	15838	31.22%
	30	1932	3.81%
	35	4145	8.17%
	40	10375	20.45%
	45	4646	9.16%
	50	456	0.90%
	55	4262	8.40%
	60+	9083	17.90%

Table 3-19 represents the continuous variables measured for 50,737 non-signalized intersections at state and federal aid routes in the year 2019, represented in terms of mean and standard deviation.

Table 3-19

Descriptive Statistics for Non-Signalized Intersections at State and Federal Aid Routes: Continuous Variables

Variable	Mean	S.D
Major road traffic volume (AADT)	8242	15671
Minor road traffic volume (AADT)	5465	7414
Major road truck volume	2327	3748
Minor road truck volume	1972	2360
Pedestrian volume (AADP)	79	99
Bicycle volume	453	1460
Distance to nearest intersection	272	546
Distance to nearest traffic signal	15095	25111
Zero vehicle household	0.03	0.04
Residential density	0.64	1.11
Employment density	1.57	4.1
Jobs per household	62.5	955
Household income	64874	21965
Disabled population	0.70	0.08
Non-white population	0.11	0.14

3.8 Chapter Summary

This chapter summarized the data collection, assembly, and processing of crash data, roadway geometry and traffic characteristics, and exposure for pedestrian and bicycle volume data appropriate for road segments and non-signalized intersections on only state routes and on state and federal aid routes. Weather information during the same day and crash characteristics related to each pedestrian and bicycle crashes were also assembled.

First, road segments and intersections on state routes and state and federal aid routes were identified. Descriptive variables for pedestrian and bicycle crashes on segments and intersections were detailed. Individual crash characteristics were also presented. Then, a detailed discussion about pedestrian and bicycle exposure as well as descriptive statistics were discussed. This study used STRAVA bicycle volumes as a measure for bicycle trips on streets and intersections. Pedestrian volumes interpolated from ATSPM signal counts were used as pedestrian activity along roadways and intersections. Land use and community characteristics data collected from the ACS 2017 5-year survey were discussed. Daily temperature, precipitation, snowfall, and snow depth observed for pedestrian and bicycle crashes were merged with individual crash characteristics. Finally, this chapter concluded by describing the descriptive statistics of all the relevant road geometry variables investigated for crash analysis on mid-block locations along segments and at non-signalized intersections on only state routes and on state and federal aid routes.

4 METHODOLOGY

4.1 Overview

The objective of this study is to identify effects of roadway geometry, traffic characteristics, land use, and community characteristics of road segments and non-signalized intersections on pedestrian and bicycle crashes. This chapter presents a brief review of statistical models used in crash analysis. Among potential analysis methods, choosing an appropriate method is a crucial part of any research, and this study has examined crash analysis with statistical models and machine learning models. This chapter discusses the details of negative binomial models which are widely used in analyses of count data such as traffic crash frequencies. Then, various machine learning techniques and the selection of Gradient Boosting methods for crash analysis are discussed. Interpretation of these models requires a discussion regarding the relative importance of the investigated factors and the generation of partial dependence plots to understand non-linear marginal effects. Finally, the analysis setup for this study that is applied to explore pedestrian and bicycle crashes at mid-block locations and non-signalized intersections is discussed.

4.2 Count Data Models

Crash data consists of nonnegative integer values and is a common example of count data in transportation research. Applying ordinary least squares regression is not appropriate for crash data analysis since linear regression assumes normally distributed errors and estimates non-integer values and negative values for the predicted variable.

Crash data can follow a Poisson distribution and generalized linear models with Poisson error distributions such as Poisson regression model and negative binomial regression models are appropriate and widely used statistical models that estimate and predict crash data.

The Poisson regression model is perhaps the most popular model to estimate crash data. In a Poisson regression model, the probability of a road segment or intersection i having y_i crashes per year is given by:

$$P(y_i) = \frac{exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} \qquad \text{Eq: 4-1}$$

where y_i is a non-negative integer, $P(y_i)$ is the probability of a road segment or intersection i having y_i crashes within a time period, λ_i is the Poisson parameter for road segment or intersection i, which is equal to i's expected number of crashes within a time period $E[y_i]$. The Poisson parameter λ_i is estimated as a function of the explanatory variables.

The most common relationship between explanatory variables and the Poisson parameter is the log-linear model,

$$E[y_i] = \gamma_i = \exp(\beta X_i)\,, or\ \ln(\gamma) = \ (\beta X_i) \qquad \text{Eq: 4-2}$$

where X_i is a vector of explanatory variables and β is a vector of estimated parameters, One important note in crash prediction modeling is that the explanatory variable traffic volume (AADT), and similar exposure variables such as pedestrian volume (AADP) and bicycle volume (AADB), are often log transformed when used in the model.

One fundamental assumption while using the Poisson regression model is that the mean and variance of the dependent variable is assumed to be equal, E[y_i] = VAR[y_i].

The parameter vector is biased if this is the case, and the data are either underdispersed (E[y_i] > VAR[y_i]) or overdispersed (E[y_i] < VAR[y_i]). To account for under dispersion or overdispersion, the negative binomial model is rewritten by, for each observation i:

$$\gamma_i = \exp(\beta X_i + \varepsilon_i) \qquad \text{Eq: 4-3}$$

where $\exp(\varepsilon_i)$ is a Gamma-distributed disturbance term with mean 1 and variance α. This term allows the variance to differ from the mean:

$$VAR[y_i] = E[y_i]\big[1 + \alpha \cdot E[y_i]\big] \qquad \text{Eq: 4-4}$$

Thus, the probability that road segment or intersection i having y_i accidents per time period is given by:

$$P(y_i) = \frac{\Gamma\big((1 \,/\, \alpha) + y_i\big)}{\Gamma(1 \,/\, \alpha) y_i!} \left(\frac{1 \,/\, \alpha}{1 \,/\, \alpha + \lambda_i}\right)^{1/\alpha} \left(\frac{\lambda_i}{(1 \,/\, \alpha) + \lambda_i}\right)^{y_i} \qquad \text{Eq: 4-5}$$

where Γ(.) is a gamma function. This results in the likelihood function:

$$L(\lambda_i) = \Pi \; \frac{\Gamma\big((1 \,/\, \alpha) + y_i\big)}{\Gamma(1 \,/\, \alpha) y_i!} \left(\frac{1 \,/\, \alpha}{1 \,/\, \alpha + \lambda_i}\right)^{1/\alpha} \left(\frac{\lambda_i}{(1 \,/\, \alpha) + \lambda_i}\right)^{y_i} \qquad \text{Eq: 4-6}$$

Moreover, dispersion parameter theta, is $\theta = \frac{1}{\alpha}$.

Considering the Poisson or negative binomial regression, $\lambda_i = \exp(\beta x_i + \varepsilon_i)$, we can interpret coefficient values. The interpretation of an untransformed regression coefficient is that a unit change in an independent variable X yields a $100(e^{\beta} - 1)$ percentage change in the dependent variable. The interpretation of β for a dummy variable is different, as presence (1) of an independent dummy variable yields a $100(e^{\beta} - 1)$ percentage change in the dependent variable, compared to absence (0). For a log transformed independent variable such as traffic volume (AADT):

$$\ln(\lambda_i) = \beta_1 \cdot \ln(AADT) + \sum\nolimits_i \beta_i \cdot X_i + \varepsilon_i \qquad \text{Eq: 4-7}$$

Thus, a unit change in the independent variable AADT yields a β unit increase or decrease in the dependent variable.

Moreover, in the road segment crash prediction models, the effect of the length of road segments is restricted to be perfectly proportional, so that the length of each segment observation does not have an impact on the result. Taking a natural log of the road segment length variable and fixing the coefficient equal to one (offset) yields:

$$\ln(\lambda_i) = 1 \cdot \ln(Length) + \beta_1 \cdot \ln(AADT) + \sum\nolimits_i \beta_i \cdot X_i + \varepsilon_i \qquad \text{Eq: 4-8}$$

4.3 Boosted Regression Tree Theory

Decision tree models are used in both regression and classification analysis. With a continuous dependent variable such as a crash frequency, regression trees are developed, and with a categorical dependent variable such as crash severity, classification trees are used. A decision tree creates binary partitioning on each "node" that is represented by the risk factors in this study. The partitioning or splitting is binary because each node can only result in two split groups. This process is recursive until no new nodes (child nodes) can be developed due to homogeneity in the child nodes, or a user defined minimum number of nodes represented by "tree complexity" is achieved. Thus, the stoppage of splitting occurs when all possible threshold values for all explanatory variables (splitters) have been assessed to find the greatest improvement in the purity scores of the resultant nodes (Toran Pour et al, 2016).

However, a single tree is occasionally a weak classifier especially when high variance is present in the data which can be common in traffic crash data. Single-tree models normally use few variables to prevent data overfitting. This makes them an unstable method where a small data change may cause a large change in a tree (Chung, 2013). If several variables are included to control the effects of confounding factors in a single-tree model, the model usually results in high variance and low bias – the bias-variance trade-off (De'ath, 2007). To balance the bias and variance, the boosting technique is introduced. Boosting is used where a weak algorithm is run repeatedly to overcome the variance or bias. Boosted Decision Tree is an ensemble technique that tries to find a more accurate model by merging a number of trees in a sequential process. Boosting uses a forward, stage-wise procedure that only uses the results from the previous tree rather than from all other previously fitted trees. In this approach, after the first tree is fitted, the residuals are calculated and observations with high residual values are defined as poor fit observations. The computed classifiers are then combined in the final prediction.

This research assumes that crash counts follow a Poisson distribution,

$$p(k,\lambda) = \frac{e^{-\lambda}\lambda^{k}}{k!} \qquad \text{Eq: 4-9}$$

where k = 0,1,2,3,…. λ. In a non-linear conditional probability, we can use a non-linear function $S = f(x_i)$ Then, we can use a transfer function, $\Psi(s) = exp(s)$.

$$\lambda = exp(f(x)) \qquad \text{Eq: 4-10}$$

The probability of $y_i|x_i$ is,

$$P(y = y_i|x_i: f) = \frac{e^{-\exp(f(x_i))}[\exp(f(x_i))]^{y_i}}{y_i} \quad \text{Eq: 4-11}$$

Using log on both sides:

$$log\, p(y= y_i|x_i\,;\, f) = -\exp(f(x_i)) + y_i f(x_i) - \log(y_i!) \quad \text{Eq: 4-12}$$

The following function gives high likelihood to our observed data,

$$\sum_{i=1}^{N}[-\exp(f(x_i)) + y_i f(x_i) - \log(y_i!)] \quad \text{Eq: 4-13}$$

Minimizing the Poisson loss is equivalent to maximizing the likelihood of the data under the assumption that the target comes from a Poisson distribution. The loss function for Poisson regression takes the form of:

$$L(y, f(x)) = \frac{1}{N}\sum_{i\cdot=}^{N}\langle f(\mathrm{x_i}) - y_i\, log\, f(\mathrm{x_i})\rangle \quad \text{Eq: 4-14}$$

where f(x) is the predicted value.

Gradient boosted decision tree models are used for supervised learning problems, where training data containing multiple features are used to predict the target variable. This study has used training data containing features regarding roadway geometry, traffic characteristics, land use, and community characteristics, to predict crash counts in the regression model.

Friedman (2001) developed gradient boosting, an approximation technique that utilizes a greedy gradient descent to reach minimized loss. Let *x* be a vector of predictor variables and response variable *y* is estimated by function *f*(*x*). This function is expressed as a sum of basic functions $b(x;\, \gamma_m)$ (Hastie et al., 2001).

$$f(x) = \sum_m f_m(x) = \sum_m \beta_m b(x\,;\, \gamma_m) \qquad \text{Eq: 4-15}$$

where β_m (m = 1, 2, …., M) are the expansion coefficients and $b(x; y_m)$ are single regression trees with the parameter γ_m representing the split variables, their values at the splitting nodes, and the predicted values at the terminal nodes.

Considering these parameters, Dearth (2007) summarizes this process for loss function (e.g, deviance) $L\big(y, f(x)\big)$.

1. Initialize $f_0(x)$
2. For $I = 1\ to\ M$ (number of trees)
    a. For $I = 1\ to\ N$ (number of observations),

       Residual $r_{im} = -\left[\frac{\partial L\left(y_i f(x_i)\right)}{\partial f(x_i)}\right] f(x) = f_{m-1}(x)$
    b. Fit a regression tree to estimate $b(x\,;\, \gamma_m)$
    c. Estimate by minimizing $L\big(y, f_{m-1}(x_i) + \beta b(x\,;\, \gamma_m)\big)$
    d. Update $f_m(x) = f_{m-1}(x_i) + \beta_m b(x; \gamma_m)$
3. Calculate $f(x) = \Sigma_m f_m(x)$

The sequential tree building process in gradient descent model adds trees until all the observations fit. This can lead to overfitting of the model. Overfitting occurs when the model captures the training samples too well and fits the outliers or “noise” of the data. Overfitting results in a good accuracy for training data but shows poor accuracy on testing data and thus does not perform well in predicting the outcomes for new cases.

4.3.1 Parameter optimization

Avoiding overfitting and bias requires regularization of the decision tree models. The regularization process involves finding an optimized learning rate, tree complexity, and optimum number of trees to find the minimum loss function.

Learning rate ϵ is introduced in Step 2d when the algorithm updates the estimated function:

$$f_m(x) = f_{m-1}(x_i) + \epsilon\beta_m b(x; \gamma_m) \qquad \text{Eq: 4-16}$$

The learning rate is a value between 0 and 1 and implemented in this model by testing a set of values 0.05, 0.01, 0.005, 0.001, 0.0005. A smaller learning rate better minimizes the loss function; however, it takes a longer time to train the model. A lower value of the learning rate also requires more trees to be fitted to reach the minimum value of the loss function.

Tree complexity stands for the depth of interaction levels among predictor variables. A tree complexity of 1 creates two terminal nodes from one node for each tree. A tree complexity of 2 generates models with up to two-way interactions between variables (a maximum of two nodes in each branch), and so on (Hastie et al., 2009; Saha et al., 2016). To utilize the strength of boosted tree models, higher depths of interaction (i.e., higher levels of tree complexity) should be used in developing trees.

A lower learning rate requires more iterations in the boosting sequence. Studies have indicated that a 10-fold reduction in the learning rate requires an approximate 10-fold increase in iterations (De'ath, 2007) and have recommended at least 1,000 trees (Elith et al., 2008).

4.3.2 Cross-validation

To avoid overfitting of the gradient boosted decision tree model, K-fold cross validation is used in this study. In order to avoid overfitting, training data is further split into train samples and validation samples. A k-fold cross validation method is widely used for this purpose. In k-fold cross validation, the original sample is randomly partitioned into k mutually exclusive subsets. In each run, a single subset is retained as the validation data for testing the model, and the remaining k-1 subsets are used as training data. The cross-validation process is repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds are usually averaged to generate a single estimation. This process uses each observation for validation exactly once and all observations are used in training and validation.

A value of k=10 is very common in the field of machine learning. This study has used 10-fold cross validation which performs the fitting procedure a total of ten times, with each fit being performed on a training set consisting of 90% of the total training set selected at random, with the remaining 10% used as a hold-out set for validation.

To reduce overfitting and improve the run time of the model, stochastic gradient descent was applied, where trees were fitted by a random extraction of 50% of the training data without replacement at each iteration. The stopping criterion for a node splitting further was that terminal nodes must have at least 10 observations.

4.3.3 Relative importance of variables

The relative importance of features (variables) provides a score that indicates how valuable or useful each feature was to build the ensemble of boosted trees within the model. In a single decision tree model, the importance of variables are computed based on how many times the variable was used to split the nodes and the improvement from minimizing error because of the splits. In a boosted decision tree, the gain or influence of the variable is summed over the ensemble of decision trees and the average value of the summation is presented as the gain of the variable. Relative importance is calculated by taking each feature's contribution (the sum of squared improvements at all splits determined by the feature) for each tree in the model and finally averaged over all trees. This value is scaled by a sum of 100. A higher value of this metric when compared to another feature implies it is more important for generating a prediction. The “XGBoost” package in R produces two other measures of variables: cover and frequency. Cover represents the relative number of observations related to each variable. Frequency implies the percentage representing the relative number of times a particular feature occurs in the trees of the model.

4.3.4 Marginal effects of variables

Since boosted decision trees are an ensemble of many decision trees, the simple interpretation of a single tree is lost. However marginal effects of each variable in a boosted decision tree are investigated using partial dependency plots.

Marginal effects measure the effect of an explanatory variable on the response variable crash count after accounting for the average effects of all other variables. The magnitude of interaction effects also represents the amount of residual variance explanatory variables can explain (Elith et al., 2008). Partial dependency plots showing marginal effects of independent variables are shown in chapter 5.

4.3.5 Analysis Setup

In a Poisson regression, the Poisson deviance is equal to:

$$D = 2\sum_{i=1}^{n}[y_i \log\left(\frac{y_i}{\exp(x_i\beta)}\right) - (y_i - \exp(x_i\beta)] \qquad \text{Eq: 4-17}$$

Poisson deviance denotes the model’s goodness of fit measure. A well fitted Poisson regression model will have the observed values close to their predicted means. The smaller the deviance value is, the better the model fits the data.

Boosted decision trees with different tree complexities and shrinkage factors were employed to investigate the minimum Poisson deviance. In general, tree complexity level 1 had the highest Poisson deviance value compared to more complex tree level. Hence, it is decided that boosted decision trees with only two terminal nodes (tree complexity 1) does not provide for optimized results. Poisson deviance computed by higher tree levels (5,10) can fit higher interaction levels and thus capture the complex nature of crash frequency modeling.

Shrinkage factors 0.05, 0.01, 0.005, and 0.001 were used in optimizing boosted tree models with different tree complexities. Also, in general, better results were achieved by fitting more trees to the gradient boosted decision tree models. However, models with higher tree complexity had fitted fewer trees to reach optimization and converge to the minimum value of the Poisson deviance.

4.4 Ordinal Data Models

Crash severities are considered to be ordinal data. This type of data is categorical and represent the continuous aspect of these categories they are placed in an increasing order of severity from lowest to highest: no injury, possible injury, minor injury, serious injury and fatal.

Multinomial or nested logit or probit models can deal with the categorical nature of the dependent variable, but these models fail to account for the ordinal nature of the dependent variables. A more appropriate technique to model these data is the ordered probit or ordered logit models, which assume that there is some underlying continuous version of the ordinal/categorical dependent variable. Ordered logit models follow a standard logistic distribution that assumes a linear relationship between the independent variables and a latent (unobserved) dependent variable, and it calculates coefficients of the variables and threshold values. Observed dependent variable's value depends on whether they have crossed a particular threshold. Moreover, since severity of crashes has a specific order of arrangement, ordered logistic regression is appropriate.

In light of this, an ordered logit model was used in this study. The specification of an ordered logit model is as follows:

$$y_i^* = \beta' x_i + \varepsilon_i \qquad \text{Eq: 4-18}$$

where y_i^* is the predicted level of injury severity by a pedestrian i, β' is a vector of unknown parameters, x_i is a vector of explanatory variables, and ε_i is the random error term that follows a standard logistic distribution. The classification of observed injury severity is done based on the predicted injury using the following criteria:

$$y_i = \begin{cases} 0 \; if \; y_i^* \leq \mu_1 \; (no \; injury) \\ 1 \; if \; \mu_1 < y_i^* \leq \mu_2 \; (possible \; injury) \\ 2 \; if \; \mu_2 < y_i^* \leq \mu_3 \; (minor \; injury) \\ 3 \; if \; \mu_3 < y_i^* \leq \mu_4 \; (major \; injury) \\ 4 \; if \; \mu_4 < y_i^* \; (fatal) \end{cases} \qquad \text{Eq: 4-19}$$

where μ_1, μ_2, and μ_3 are the thresholds estimated by the model.

In ordered logistic models, parameter interpretation is based on the null hypothesis that there is no relationship between the selected independent and dependent variables. Here, the results are calculated at 95% confidence level. Since logit is a natural log of odds ratio, the coefficients of the model independent variables are calculated by exponentiating the odds ratio and calculating the change in 100 percent.

4.5 Chapter Summary

This chapter illustrated the two methods—a negative binomial statistical model and a machine learning technique called a Gradient Boosted Decision Tree model—to analyze pedestrian and bicycle crashes at mid-block locations and at unsignalized intersections. Interpretation of the results for both these variables were discussed. Optimization of parameters necessary to implement the machine learning model were discussed. The methods to demonstrate the results of the Boosted Tree models, relative importance of features (variables), and partial dependence plots to present the non-linear marginal effects were reviewed. Finally, analysis setup for this study that was applied to explore pedestrian and bicycle crashes at mid-block locations and intersections were discussed.

5 RESULTS

5.1 Overview

This chapter reports the results of the crash frequency models—negative binomial statistical models and gradient boosted decision tree models for road segments and non-signalized intersections—as well as discusses the pedestrian and bicycle crash severity analysis. Among the crash frequency models, the negative binomial models present the statistically significant variables affecting crashes. Gradient boosted models rank the most important variables associated with predicted crashes. Detailed results, including statistical metrics of negative binomial models and cross-validation results of the boosted decision tree models, are presented. For the crash frequency models, the impact of individual variables on the predicted bicycle and pedestrian crashes in terms of direction of association and non-linearity are discussed. Finally, results of the ordered logistic regression models implemented for pedestrian and bicycle crash severity are presented, described, and interpreted.

5.2 Pedestrian Crash Frequency Along Segments and at Mid-Block Locations

Pedestrian crashes along segments and at mid-block locations were estimated with negative binomial (NB) models and boosted decision tree models. These two methods investigated crashes occurring on two spatial scales: state routes, and state and federal aid routes.

5.2.1 Negative Binomial Model Results

This section presents the results for pedestrian crashes occurring at mid-block locations on road segments. In order to apply statistical crash frequency models, a series of considerations were undertaken. First, all models were estimated using all possible explanatory variables except the ones generating very large standard errors (possibly due to large numbers of missing observations). Second, the best fit model type was determined using tests for overdispersion and zero-inflation. In all cases, the data were significantly over-dispersed, indicating that NB models were better than Poisson models. Third, both forward and backwards elimination processes were used to add one variable at a time, and then remove variables that were not statistically significant from the model one-by-one, and elimination was stopped when all variables were at least marginally significant ($p < 0.10$). Table 5-1 shows the model results of pedestrian crashes along segments and at mid-block locations for state routes.

Table 5-1

NB Model for Pedestrian Crashes at Segments (State Routes, N = 4,979)

Variable	*B*	*SE*	*z*	*p*
(Intercept)	-15.07	0.75	-20.12	0.00
Natural log of (length)	1.00	-	-	-
Natural log of AADT	0.42	0.08	5.20	0.00
Natural log of Pedestrian volume	0.44	0.06	7.52	0.00
Number of left turn lanes	-0.07	0.04	-1.69	0.09
Driveway density on major road	0.04	0.01	6.37	0.00
Driveway density on minor road	0.01	0.00	5.04	0.00
Bus stops	0.14	0.04	3.50	0.00
Roadside barrier present	-0.37	0.15	-2.43	0.02
Classification Local road	0.23	0.10	-2.21	0.03
Zero vehicle household	1.94	1.01	1.91	0.06
Employment density	-0.02	0.01	-1.67	0.10
Disabled population (%)	2.10	0.53	3.92	0.00
Non-white population (%)	2.31	0.35	6.57	0.00

A negative binomial model was estimated to determine significant factors affecting pedestrian crashes at mid-block on state routes. The model had an AIC of 3,745.3, a null deviance of 3,213.6, and a residual deviance of 2,089.4 on 4,978 degrees of freedom. Pedestrian crashes are positively associated with busy streets with higher traffic volume, since a 1% increase in annual average daily traffic (AADT) would be expected to yield a 0.42% increase in crashes. Also, streets with greater pedestrian volume see higher numbers of midblock pedestrian crashes, as a 1% increase in pedestrian volume would be expected to yield a 0.44% increase in crashes. Pedestrian crashes occurring at mid-blocks are fewer on roads with more left turn lanes, as a 1% increase in the number of left turn lanes would be expected to yield a 7% decrease in crashes. The presence of frequent driveways on both major and minor roads contribute to more pedestrian crashes, as a 1% increase in driveway density on major roads yields a

4% increase in crashes and a 1% increase in driveway density on minor roads yields a 1% increase in pedestrian crashes. The presence of a nearby bus station is also associated with a higher number of pedestrian crashes, specifically a 15% increase in crashes. Compared to highways, local roads are associated with more pedestrian crashes (26% increase in crashes). Fewer pedestrian crashes occur at roadways where roadside barriers are present (a 31% decrease in crashes). Also, pedestrian crash frequency varied depending on some land use and community characteristics, as streets adjacent to neighborhoods with higher percentages of non-white populations, zero-vehicle households, and disabled population are associated with more pedestrian crashes. High employment density is associated with fewer pedestrian crashes.

A negative binomial model with an AIC of 3,807.6, a null deviance of 3,146.6, and a residual deviance of 2,098.2 on 5033 degrees of freedom was estimated to determine significant factors affecting pedestrian crashes at mid-block on state and federal aid routes. Table 5-2 shows the model results of pedestrian crashes along segments and at mid-block locations for state and federal aid routes.

Table 5-2

NB Model for Pedestrian Crashes at Segments (State and Federal Aid Routes, N = 5,034)

Variable	*B*	*SE*	*z*	*p*
(Intercept)	-14.71	0.75	-19.73	0.00
Natural log of (length)	1.00	-	-	-
Natural log of (AADT)	0.44	0.07	6.17	0.00
Natural log of (Pedestrian volume)	0.45	0.05	8.15	0.00
Truck volume (%)	0.96	0.53	1.81	0.07
Speed limit	-0.02	0.01	-2.46	0.01
Bus stop	0.17	0.04	4.07	0.00
One way street	-1.35	0.25	-5.31	0.00
Jobs per household	-0.01	0.00	-2.12	0.03
Household income	0.01	0.21	-2.12	0.00
Disabled population (%)	2.67	0.47	5.65	0.00
Non-White population (%)	2.61	0.38	6.78	0.00

Higher pedestrian crashes are positively associated with higher traffic volumes, since a 1% increase in annual average daily traffic (AADT) yields a 0.44% increase in crashes. Also, streets with greater pedestrian volume see higher number of midblock pedestrian crashes, as 1% increase in pedestrian volume yields a 0.45% increase in crashes. Pedestrian crashes occurring at mid-blocks are fewer on roads with high-speed limit. The presence of a nearby bus station is also associated with a higher number of pedestrian crashes, specifically a 18.5% increase in crashes. One-way streets are associated with lower numbers of pedestrian crashes (a 74% decrease in crashes). Fewer pedestrian crashes occur at roadways with higher percentages of trucks. Also, pedestrian crash frequency varies depending on some land use and community characteristics, as streets adjacent to higher percentages of non-white and disabled populations are associated with more pedestrian crashes. Higher numbers of jobs per household is associated with fewer pedestrian crashes. Streets adjacent to neighborhoods with higher household incomes are associated with fewer pedestrian crashes, although the effect is minimal.

5.2.2 Boosted Decision Tree Model Results

After testing various optimization parameters such as learning rate (0.10, 0.05, 0.01, 0.005), tree complexity (1, 5, 10, 15), and sub-sample ratio (0.3, 0.5, 0.8), the optimized result with a minimum value for the loss function was achieved for the crash frequency model. The loss function was Poisson log-likelihood, also known as Poisson deviance. Minimum Poisson deviance was found at a 0.01 learning rate and at a tree complexity of 10. Iteration number 827 was the iteration with the lowest test error (0.23) and thus was considered the optimal number of iterations.

Table 5-3 shows the influence of explanatory variables for pedestrian crashes at mid-block locations. Pedestrian volume is the most influential variable with a relative contribution of 19% to the model. Traffic volume contributes 16% to the model. Driveway density on minor roads contributes 7% to the model. Vertical grade, household income, driveway density on major roads, and non-white population percentage contributes 6% each to the model. The presence of nearby bus stops and residential density contribute 5% each to the model. In summary, these nine variables account for more than 75% of the total effect of the model. The remaining ten variables have very small influence (between 1% to less than 5%) on predicted crashes.

Table 5-3

Variable Importance for Pedestrian Mid-Block Crashes (State Routes, N= 13,107)

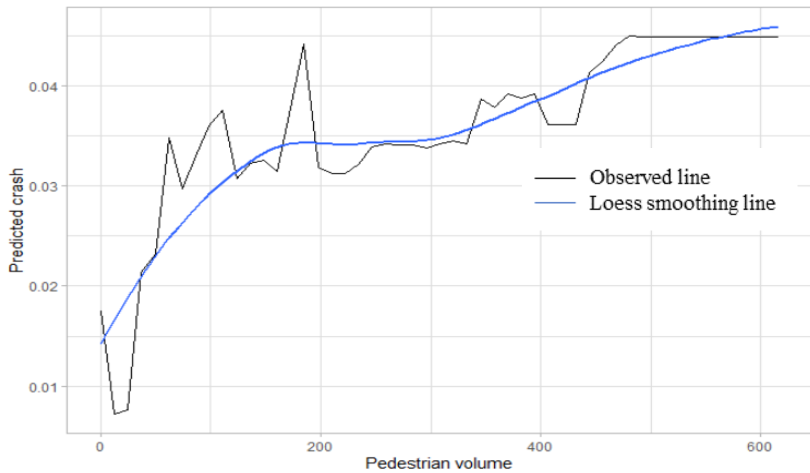
Feature	*Relative importance*	*Cumulative importance*	*Cover*	*Frequency*
Pedestrian volume	0.19	0.19	0.13	0.11
Traffic volume (AADT)	0.16	0.35	0.06	0.05
Driveway density on minor roads	0.07	0.42	0.12	0.11
Vertical grade (%)	0.06	0.48	0.11	0.10
Household income	0.06	0.54	0.06	0.07
Driveway density on major roads	0.06	0.60	0.08	0.06
Non-White population %	0.06	0.66	0.07	0.06
Bus stops	0.05	0.71	0.05	0.02
Residential density	0.05	0.76	0.05	0.05
Zero vehicle household	0.04	0.80	0.04	0.05
Jobs per household	0.04	0.84	0.03	0.05
Disabled population (%)	0.04	0.88	0.04	0.05
Truck volume (%)	0.03	0.91	0.05	0.05
Employment density	0.03	0.93	0.04	0.04
Number of Left turn lane	0.01	0.94	0.01	0.02
Number of Right turn lane	0.01	0.96	0.02	0.02
Two-way turn lanes	0.01	0.97	0.00	0.01
Lane width	0.01	0.98	0.01	0.01
Number of Through lane	0.01	0.99	0.00	0.01

Marginal effects of the explanatory variables on predicted crashes are illustrated by partial dependence plots. These plots show the relationship of an explanatory variable while all the other variables have an average effect on the model. Figure 5-1 present the marginal effects of the most influential variables with a minimum of 5% contribution to the models for pedestrian crashes at mid-block locations.

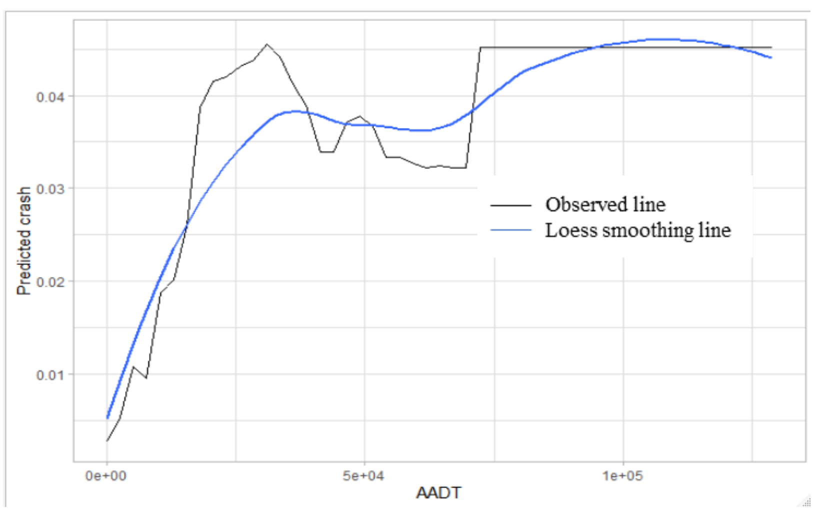
Figure 5-1

Marginal Effects for Pedestrian Mid-Block Crashes

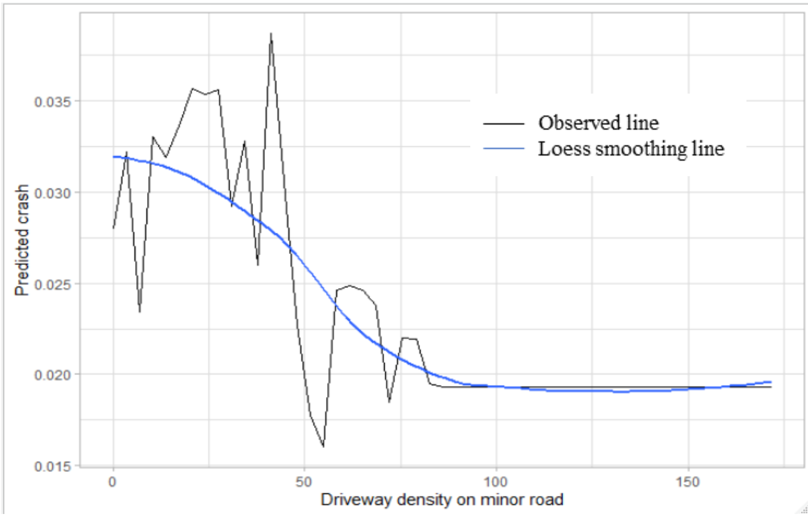
(a) Pedestrian Volume



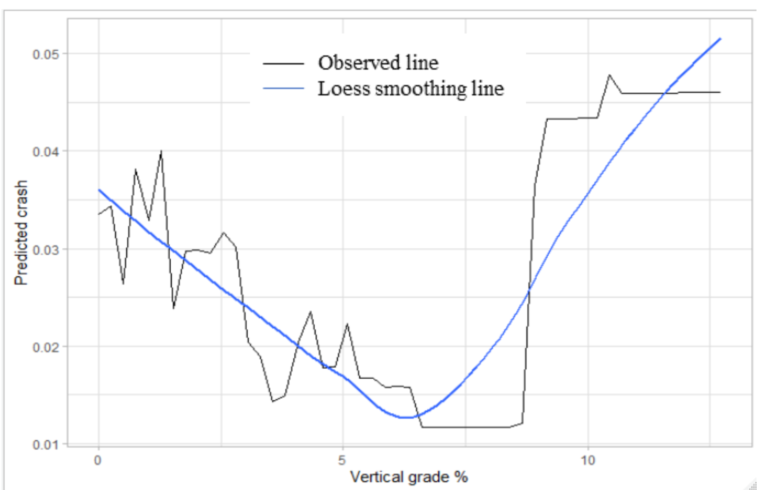
(b) Traffic Volume (AADT)



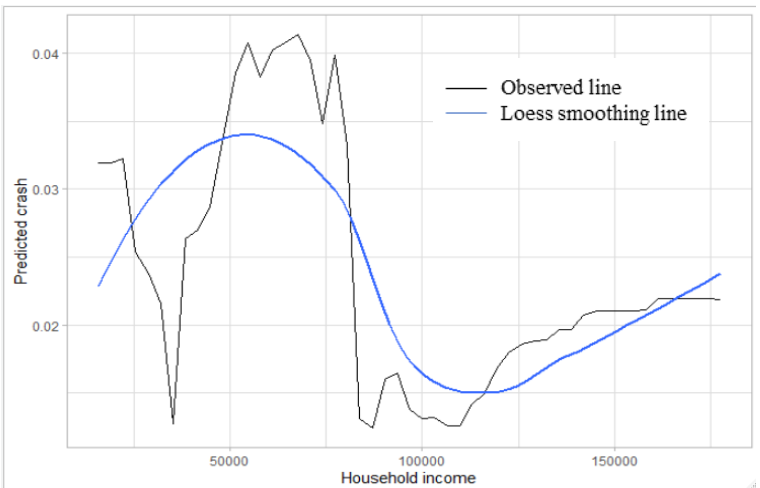
(c) Driveway Density on Minor Roads (# /Quarter Mile)



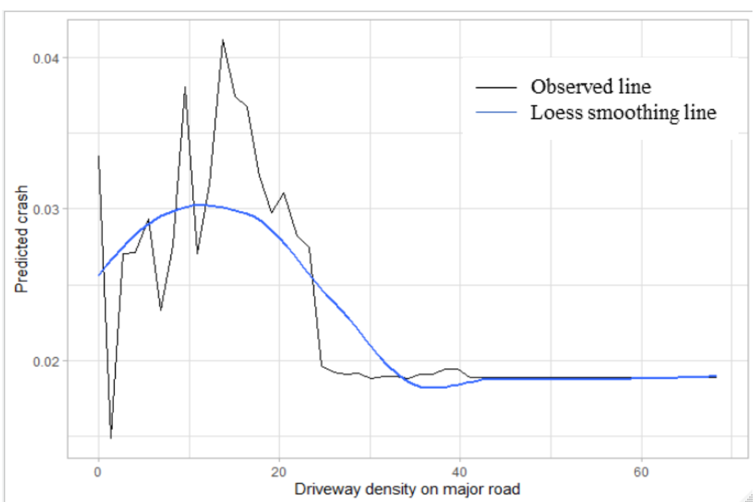
(d) Vertical Grade (%)



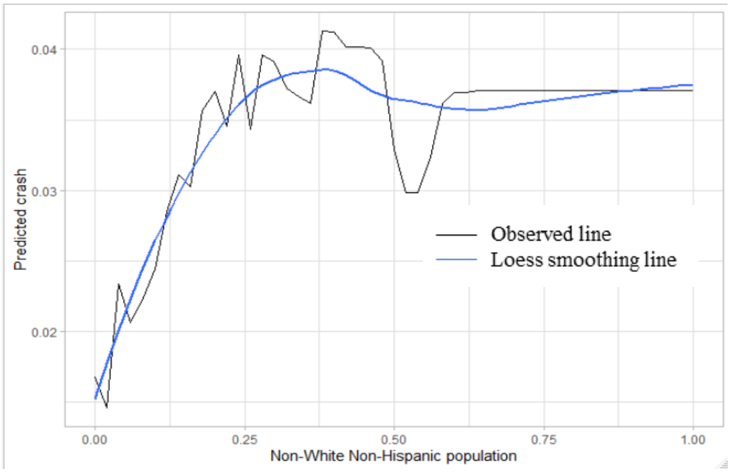
(e) Household Income ($)



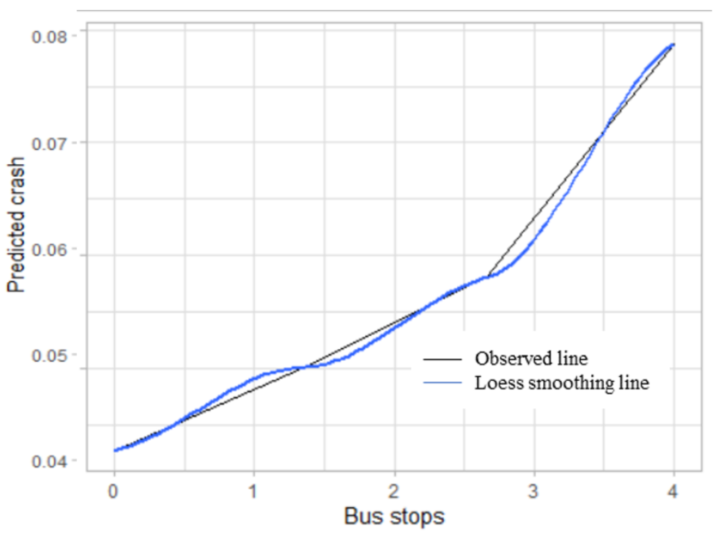
(f) Driveway Density on Major Roads (# /Quarter Mile)



(g) Non-White Non-Hispanic Population (%)



(h) Bus Stops

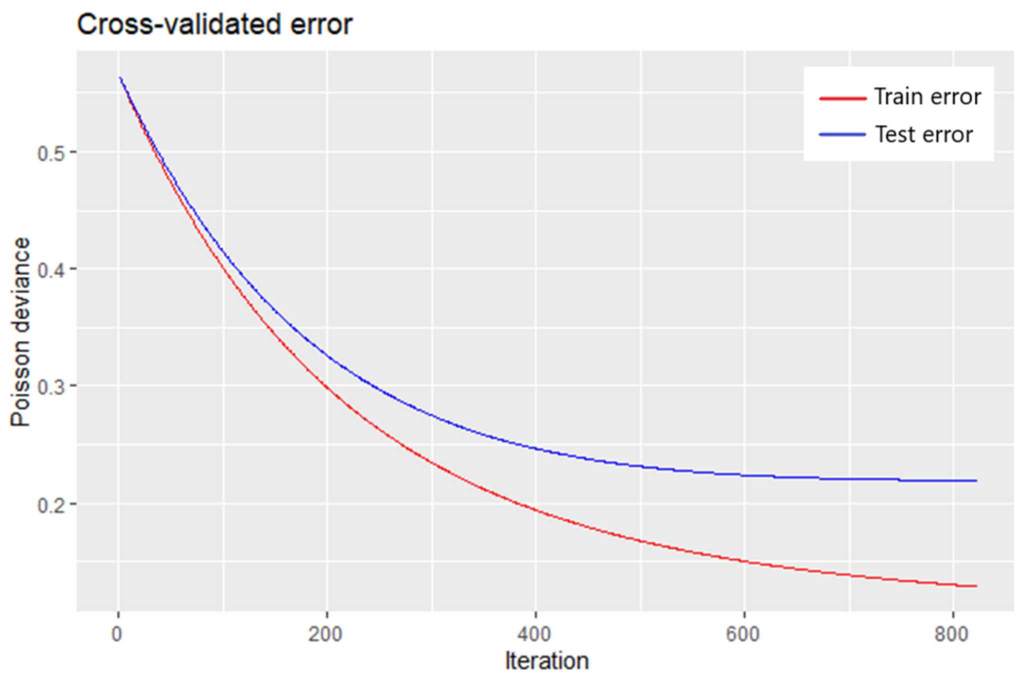


The marginal effect demonstrates a non-linear relationship with varying rate for predicted crashes at different values of pedestrian volume. Similarly, AADT would likely impact predicted crashes at a logarithmically decreasing rate, since there is a non-linear association between predicted crashes and AADT. Driveway density on minor roads is associated with fewer crashes, but only up to a certain threshold of about 80 driveways per mile. However, driveway density on major roads is associated with frequent pedestrian crashes when there are around 10 driveways per mile, but when driveways are denser on major roads, they are associated with a decreasing rate of pedestrian crashes, and the effect on predicted crashes plateaus after 40 driveway per mile. A nonlinear relationship among predicted pedestrian crashes and vertical grades are also presented as for vertical grades above 6%, crash frequency increases with increasing grade. However, for vertical grades below 6%, crash frequency decreases with increasing grade. Household income level shows a complex and non-linear relationship with pedestrian crashes, where neighborhoods with low-income households are generally associated with higher number of predicted crashes. Neighborhoods with more non-white race/ethnicity groups are associated with high predicted crash rate. The number of bus stops are generally associated with more frequent predicted crashes although at a varying rate.

Cross-validation of the fitted boosted tree shows acceptable predictive power of the model. Predictive performance was measured by implementing 10-fold cross validation at every iteration and computing train and test error indicated by Poisson deviance. 90% of the data (N = 12,204) was used for model fitting (train data) and 10% of the held-out data (test data) was used for validation in each iteration. Figure 5-2 shows the Poisson deviance obtained by the cross-validation procedure at each iteration. The validation process terminated when it reached the minimum test error and no further improvement was found in ten consecutive iterations.

Figure 5-2

Poisson Deviance for Bicycle Mid-Block Crashes



From Figure 5-2, it is clear that train data and test data performed similar, however train error was minimized the most compared to test data. From the results, the train error (0.11) and test error (0.23) were comparable. Since Poisson deviance measures how closely the model's predictions are to the observed outcomes, it may be used as the basis for a goodness of fit test of a boosted tree model.

5.3 Bicycle Crash Frequency Along Segments and at Mid-block Locations

Bicycle crashes along segments and at mid-block locations are estimated with negative binomial (NB) models and a boosted decision tree model. The following sections provide results on bicycle crash frequency occurring on state routes and state and federal aid routes.

5.3.1 Negative Binomial Model Results

This section presents the results for pedestrian and bicycle crashes occurring at mid-block locations and on road segments. Similar to the estimation process of pedestrian statistical models, over-dispersion and zero-inflation of the dataset were taken into consideration. Estimated negative binomial models showed a significant capability to account for over-dispersion. Using forward and backwards elimination, only statistically significant variables over 90% confidence interval were contained in the model. Table 5-4 shows the model results of bicycle crashes along segments and at mid-block locations for state routes.

Table 5-4

NB Model for Bicycle Crashes along Segments and Mid-Block Locations (State Routes, N = 11,910)

	B	*SE*	*z*	*p*
(Intercept)	-17.14	0.55	-31.11	0.00
Natural log of (length)	1.00	-	-	-
Natural log of (AADT)	0.80	0.06	13.48	0.00
Natural log of (Bicycle volume)	0.24	0.03	7.02	0.00
Number of left turn lanes	0.23	0.04	5.55	0.00
Vertical grade %	0.12	0.03	4.03	0.00
Bus stops	0.13	0.04	3.07	0.00
Commuter rail station	0.92	0.45	2.05	0.04
Presence of barrier	-0.56	0.15	-3.68	0.00
Presence of rumble strips	-1.65	0.27	-6.01	0.00
Driveway density on major roads	0.05	0.01	9.67	0.00
Driveway density on minor roads	0.02	0.00	9.08	0.00
Non-White population (%)	1.60	0.34	4.67	0.00
Residential density	0.11	0.03	3.67	0.00
Employment density	0.03	0.01	5.06	0.00

A negative binomial model with an AIC of 4,346.3, a null deviance of 5,463.6, and a residual deviance of 2,467.7 on 10,909 degrees of freedom was estimated to determine significant factors affecting bicycle crashes at mid-block locations on state routes. Bicycle crashes are positively associated with busy streets with higher traffic volume, since a 1% increase in annual average daily traffic (AADT) yields an 0.80% increase in crashes. Also, streets with greater bicycle volume see higher numbers of mid-block crashes, as a 1% increase in bicycle volume yields a 0.24% increase in crashes. Bicycle crashes occurring mid-block are higher on roads with more left turn lanes and with greater degree of vertical grades. The presence of nearby bus stations or commuter rail stations are also associated with higher numbers of bicycle crashes (14% and 150% respectively). The presence of frequent driveways in major and minor roads contribute to more bicycle crashes (5% and 2%). The presence of rumble strips and roadside barriers are associated with lower numbers of bicycle crashes. Also, bicycle crash frequency varied depending on some land use and community characteristics, as streets adjacent to areas with higher percentages of non-white populations, higher residential density, and higher employment density are associated with greater numbers of bicycle crashes.

A negative binomial model with an AIC of 4,367.6, a null deviance of 4,948.0, and a residual deviance of 2,401.3 on 11,864 degrees of freedom was estimated to determine significant factors affecting bicycle crashes at mid-block locations on state and federal aid routes. Table 5-5 shows the model results of bicycle crashes along segments and at mid-block locations for state and federal aid routes.

Table 5-5

NB Model for Bicycle Crashes Along Segments and at Mid-Block Locations (State and Federal Aid Routes, N = 11,865)

	B	*SE*	*z*	*p*
(Intercept)	-17.86	0.58	-30.74	0.00
Natural log of (length)	1.00	-	-	-
Natural log of (AADT)	0.98	0.06	17.08	0.00
Natural log of (Bicycle volume)	0.35	0.04	9.57	0.00
Heavy truck %	-0.84	0.51	-1.65	0.10
Bus stop	0.29	0.04	6.67	0.00
One-way streets	-3.25	0.35	-9.39	0.00
Residential density	0.22	0.03	7.01	0.00
Employment density	0.04	0.01	5.93	0.00
Household income	0.001	0.00	-2.90	0.00
Non-White population %	1.53	0.34	4.49	0.00

Higher bicycle crashes are positively associated with higher traffic volume, since a 1% increase in annual average daily traffic (AADT) yields a 0.98% increase in crashes. Also, streets with greater bicycle volume see higher numbers of midblock bicycle crashes, as a 1% increase in bicycle volume yields a 0.35% increase in crashes. Bicycle crashes occurring mid-block are fewer on roads with high truck volumes. The presence of a nearby bus station is also associated with higher numbers of bicycle crashes, specifically a 34% increase in crashes. One-way streets contribute to fewer bicycle crashes. Also, bicycle crash frequency varies depending on some land use and community characteristics, as streets adjacent to neighborhoods with higher percentages of people of non-white race/ethnicity, higher residential density, and higher employment density are associated with more bicycle crashes. Streets adjacent to areas with higher household incomes are associated with fewer bicycle crashes, although the effect is minimal.

5.3.2 Boosted Decision Tree Model Results

After testing various optimization parameters such as learning rate (0.10, 0.05, 0.01, 0.005), tree complexity (1, 5, 10, 15), and sub-sample ratio (0.3, 0.5, 0.8), the optimized result with a minimum value for the loss function was achieved for the crash frequency model. The loss function was Poisson log-likelihood (Poisson deviance). Minimum Poisson deviance was found at a 0.01 learning rate and at a tree complexity of 10. Iteration number 3,012 was the iteration with the lowest test error (0.22) and thus was considered the optimal number of iterations.

Table 5-6 shows the influence of explanatory variables for bicycle crashes at segments. As expected, traffic volume is the most influential variable with a relative contribution of 16% to the model. Employment density contributes 16% to the model. Driveway density on major and minor roads contributes 9% and 8% respectively. Bicycle volume contributes 7% to the model. Household income, non-white population percentage, residential density, and vertical grade contributes 5% each to the model. In summary, these nine variables account for more than 75% of the total effect of the model. The remaining variables have a very small influence (between 1% to less than 5%) on predicted bicycle crashes.

Table 5-6

Variable Importance for Bicycle Mid-Block Crashes

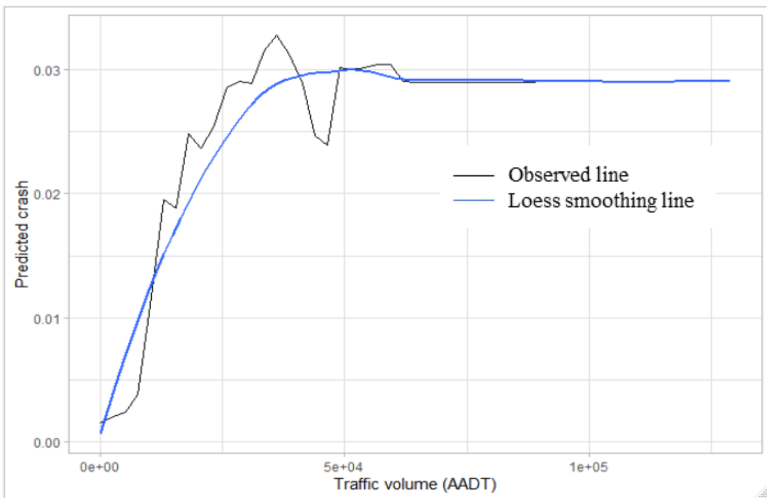
Variables	**Relative importance**	**Cumulative importance**	**Cover**	**Frequency**
Traffic volume (AADT)	0.16	0.16	0.07	0.05
Employment density	0.16	0.32	0.07	0.05
Driveway density on major roads	0.09	0.41	0.08	0.06
Driveway density on minor roads	0.08	0.49	0.12	0.12
Bicycle volume	0.07	0.56	0.09	0.11
Household income	0.05	0.62	0.06	0.07
Non-White population (%)	0.05	0.67	0.07	0.06
Residential density	0.05	0.72	0.06	0.06
Vertical grade (%)	0.05	0.77	0.09	0.09
Jobs per household	0.04	0.81	0.05	0.05
Truck volume (%)	0.04	0.85	0.06	0.05
Zero vehicle household	0.03	0.88	0.04	0.05
Disabled population (%)	0.03	0.91	0.03	0.05
Bus stops	0.02	0.93	0.01	0.02
Left turn lanes	0.02	0.95	0.02	0.03
Right turn lanes	0.01	0.96	0.01	0.02
Through lanes	0.01	0.97	0.01	0.01
Lane width	0.01	0.98	0.01	0.01

Figure 5-3 present the marginal effect of the most influential variables with a minimum of 5% contribution to the models for bicycle crashes at mid-block locations.

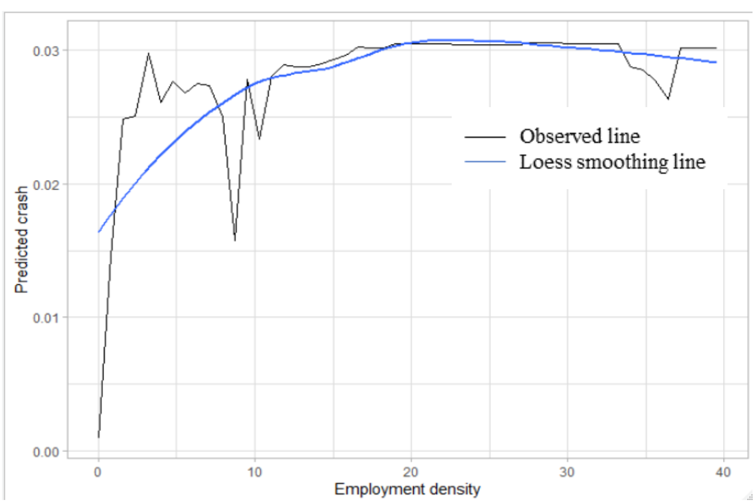
Figure 5-3

Marginal Effect for Bicycle Mid-Block Crashes

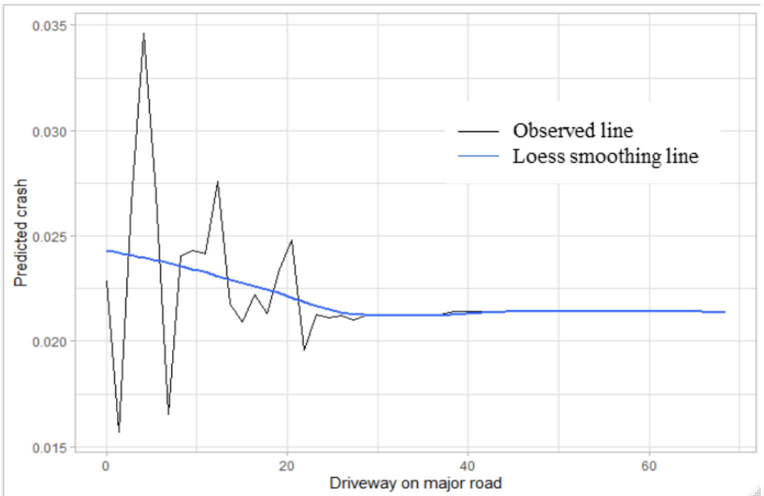
(a) Traffic Volume (AADT)



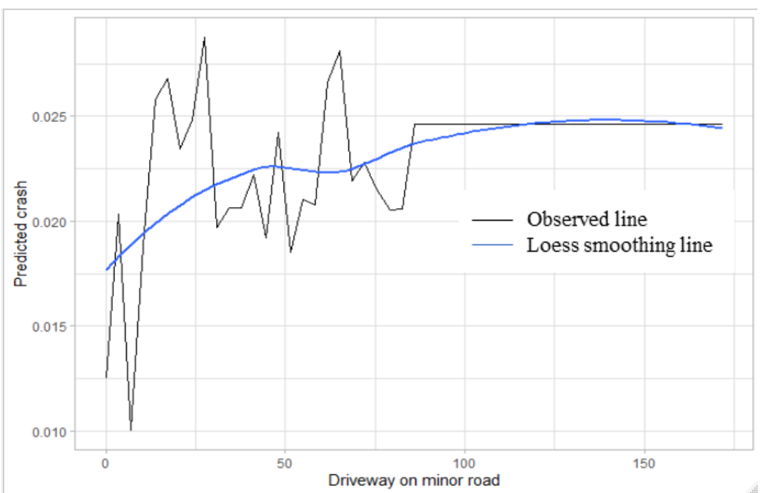
(b) Employment Density (Jobs /acre)



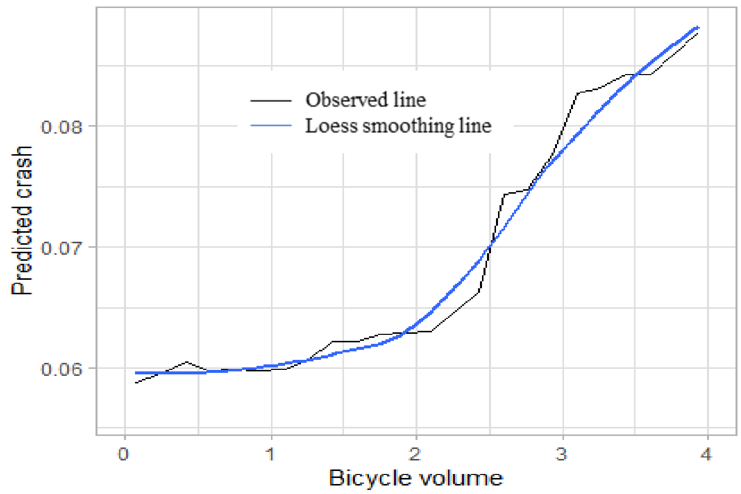
(c) Driveway Density on Major Road (# /Quarter Mile)



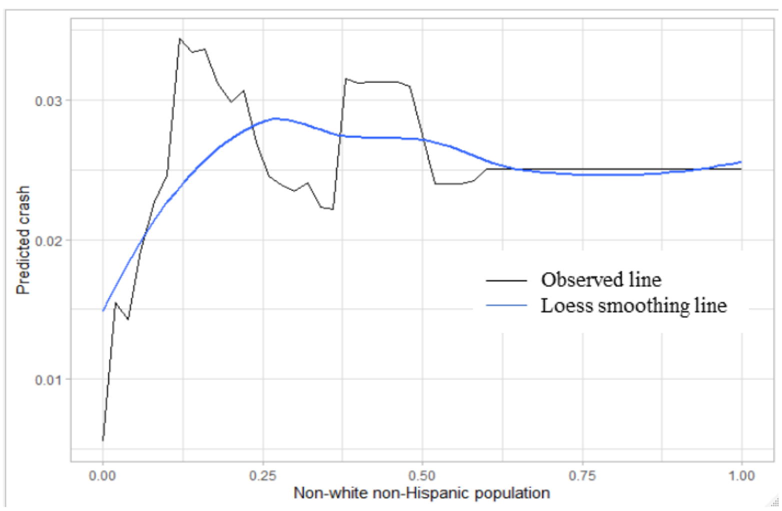
(d) Driveway Density on Major Road (# /Quarter Mile)



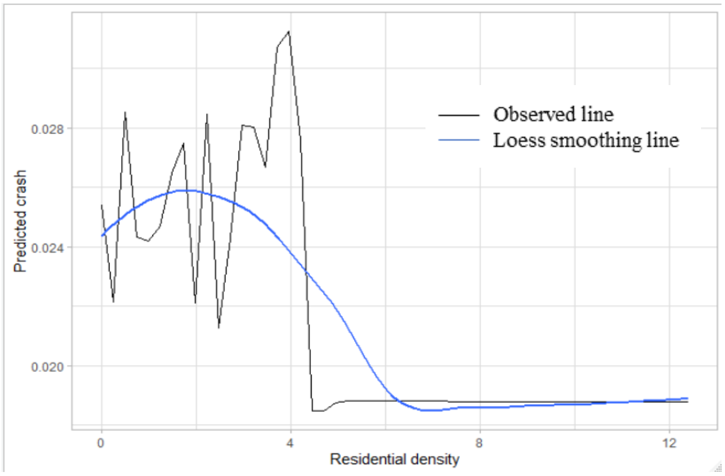
(e) Bicycle Volume



(f) Non-White Non-Hispanic Population (%)



(g) Residential Density (Housing Unit /Acre)

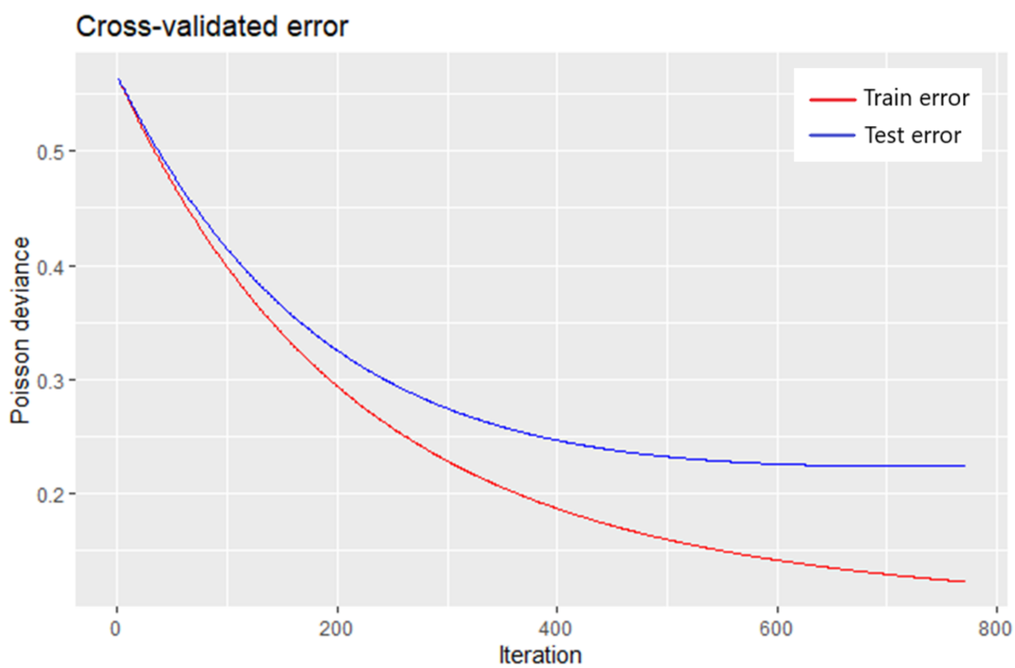


The marginal effect demonstrates a non-linear relationship with varying rate for predicted crashes for different values of traffic volume. After a certain threshold, the predicted crash is plateaued. Employment density is positively associated with more bicycle crashes. Driveway density on major roads is associated with fewer crashes, although the effect on predicted crash is limited. However, driveway density on minor roads is associated with more bicycle crashes. An exponential relationship among predicted crashes and bicycle volume shows rapid increase in predicted bicycle crashes with increasing volume. Household income level shows a complex and non-linear relationship with crashes, where low-income neighborhoods are generally associated with higher number of predicted crashes. Areas with more non-white population groups are associated with high predicted crash rate. Residential density is related to predicted crash in a nonlinear manner, as lower residential density areas see an increase in predicted crash compared to areas with high residential density.

Figure 5-4 shows Poisson deviance obtained by the cross-validation procedure at each iteration. The validation process terminated when it reached the minimum test error and no further improvement was found in ten consecutive iterations.

Figure 5-4

Poisson Deviance for Bicycle Mid-Block Crashes



From Figure 5-4 it can be noted that train data and test data performed similar, however train error was minimized the most compared to test data. Poisson deviance is in the same scale as the dependent variable (number of crash frequency). From the results, the train error (0.14) and test error (0.24) were comparable.

5.4 Pedestrian Crash Frequency at Non-signalized Intersections

In addition to the estimated crashes at road segments, this section provides estimated pedestrian crash frequency model results at non-signalized intersections estimated with negative binomial models and a boosted decision tree model.

5.4.1 Negative Binomial Model Results

This section presents the results for pedestrian crashes occurring at mid-block locations on road segments. Table 5-7 shows the model results of pedestrian crashes at non-signalized intersections on state routes.

Table 5-7

NB Model for Pedestrian Crashes at Non-Sigalized Intersections (State Routes, N = 3,378)

Variables	*B*	*SE*	*z*	*p*
(Intercept)	-11.00	1.67	-6.57	0.00
Natural log of (AADT on major approach)	0.81	0.19	4.28	0.00
Natural log of (Pedestrian volume)	-	-	-	-
Number of legs	1.39	0.19	7.16	0.00
Speed limit	-0.11	0.02	-4.71	0.00
Employment density	0.14	0.06	2.50	0.01
Jobs per household	-0.09	0.05	-1.80	0.07

Negative binomial model with AIC 406.1, null deviance 303.2 and residual deviance 172.4 on 3377 degrees of freedom was estimated to determine significant factors affecting pedestrian crashes at non-signalized intersections on state routes. Pedestrian crashes are positively associated with higher traffic volume (AADT) as a 1% increase in AADT would increase crashes by 0.81%. Intersections with a one increasing leg at an intersection is associated with 300% more pedestrian crashes. Higher speed limits on intersections are associated with fewer pedestrian crashes. Also, crash frequency varied depending on some land use and community characteristics, as intersections adjacent to high employment density are associated with more pedestrian crashes and higher jobs per household are associated with fewer crashes.

Negative binomial model with AIC 2211.9, null deviance 1493.1 and residual deviance 1088.2 on 4737 degrees of freedom was estimated to determine significant factors affecting pedestrian crashes at non-signalized intersections on state and federal aid routes. Table 5-8 shows the model results of pedestrian crashes at intersections on state and federal aid routes.

Table 5-8

NB Model for Pedestrian Crashes at Non-Signalized Intersections (State and Federal Aid Routes, N = 4,738)

Variable	*B*	*SE*	*z*	*p*
(Intercept)	-10.45	0.89	-11.74	0.00
Natural log of (Pedestrian volume)	0.70	0.07	9.82	0.00
Natural log of (AADT on major approach)	-	-	-	-
Truck volume %	0.22	0.14	1.61	0.10
Number of legs	1.02	0.08	11.70	0.00
Bus stops	0.32	0.09	3.46	0.00
Bike lanes	-0.47	0.24	-1.91	0.05

Pedestrian crashes are positively associated with greater pedestrian volume, as a 1% increase in pedestrian volume would be expected to yield a 0.70% increase in crashes. A higher percentage of truck volumes also yields in increasing number of crashes (25%). One increasing leg at an intersection is associated with more pedestrian crashes (177% increase in crash frequency). The presence of adjacent bus stops is associated with increasing numbers of bicycle crashes (38%). Interestingly, bike route presence on intersections is associated with fewer (37%) pedestrian crashes.

5.4.2 Boosted Decision Tree Model Results

After testing various optimization parameters such as learning rate (0.10, 0.05, 0.01, 0.005), tree complexity (1, 5, 10, 15) and sub-sample ratio (0.3, 0.5, 0.8), the optimized result with minimum value for loss function was achieved for the crash frequency model. The loss function was Poisson deviance, and the minimum Poisson deviance was found at a 0.01 learning rate and at a tree complexity of 10. Iteration number 972 was the iteration with the lowest test error (0.06) and thus was considered the optimal number of iterations.

Table 5-9 shows the influence of explanatory variables for pedestrian crashes at non-signalized intersections. Residential density is the most influential variable with a relative contribution of 27% to the model. Median width on the major leg at unsignalized intersection is the next important variable with a relative contribution of 15% to the model. The number of legs in an intersection contributes 15% to the model, and pedestrian volume (AADP) contributes 9% to the model. Vertical grade (%) and the distance to the nearest signalized intersection contribute 6% and 5% to the model. In summary, these six variables account for about 70% of the total effect of the model. The remaining variables have a very small influence (between 1% to less than 5%) on predicted crashes.

Table 5-9

Variable Importance for Pedestrian Non-Signalized Intersection Crashes

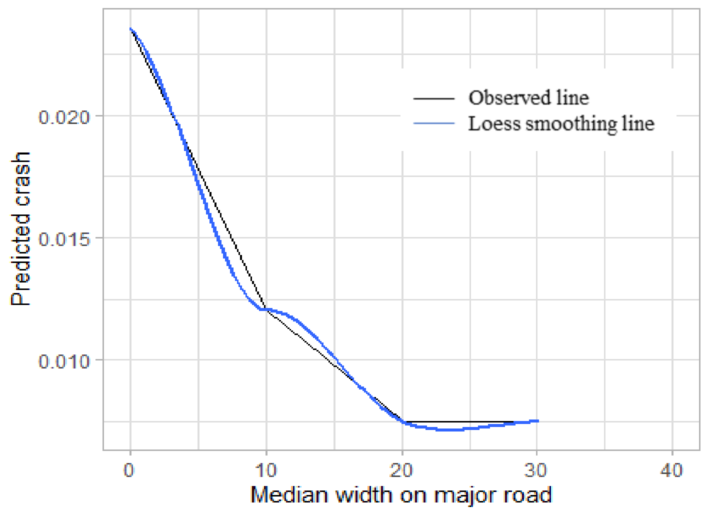
Feature	Relative importance	Cumulative importance	Cover	Freq
Residential density	0.27	0.27	0.06	0.12
Median width on major road	0.15	0.42	0.20	0.07
Number of legs	0.15	0.57	0.08	0.05
Pedestrian volume	0.09	0.66	0.10	0.05
Vertical grade (%)	0.06	0.71	0.02	0.08
Distance to nearest signal	0.05	0.77	0.05	0.09
Household income	0.03	0.80	0.02	0.06
Speed limit	0.03	0.83	0.03	0.08
Distance to nearest intersection	0.03	0.86	0.03	0.04
Disabled population (%)	0.02	0.88	0.02	0.05
Truck volume (%)	0.02	0.90	0.06	0.05
Non-White population (%)	0.02	0.92	0.01	0.03
Commuter rail station	0.02	0.93	0.11	0.01
Two-way turn lane on major road	0.02	0.94	0.01	0.01
Zero vehicle household	0.02	0.95	0.01	0.04
Right turn on major road	0.02	0.96	0.10	0.03
Jobs per household	0.01	0.97	0.00	0.03
Employment density	0.01	0.98	0.01	0.01
Lane width on major road	0.01	0.99	0.01	0.02

Figure 5-5 present the partial dependent plots of the most influential variables with a minimum of 5% contribution to the models for pedestrian crashes at non-signalized intersections.

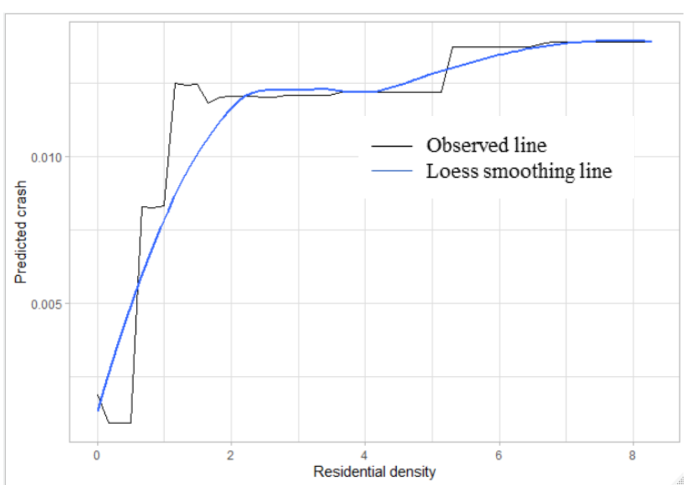
Figure 5-5

Marginal Effects for Pedestrian Unsignalized Intersection Crash

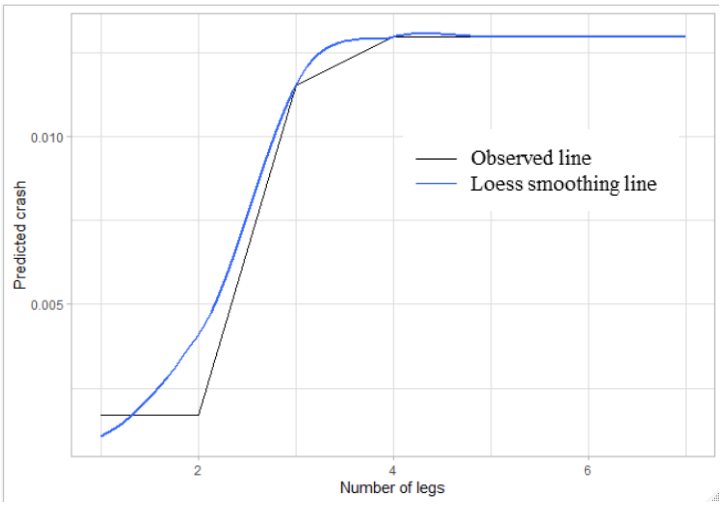
(a) Median Width On Major Roads (Ft)



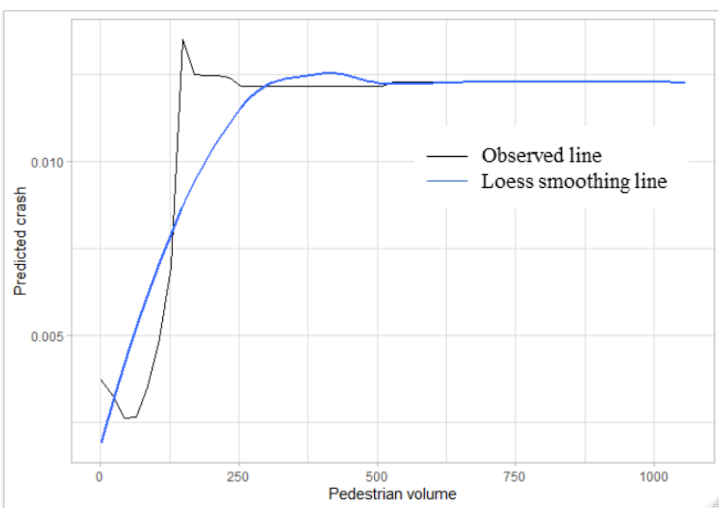
(b) Residential Density (Housing Unit /Acre)



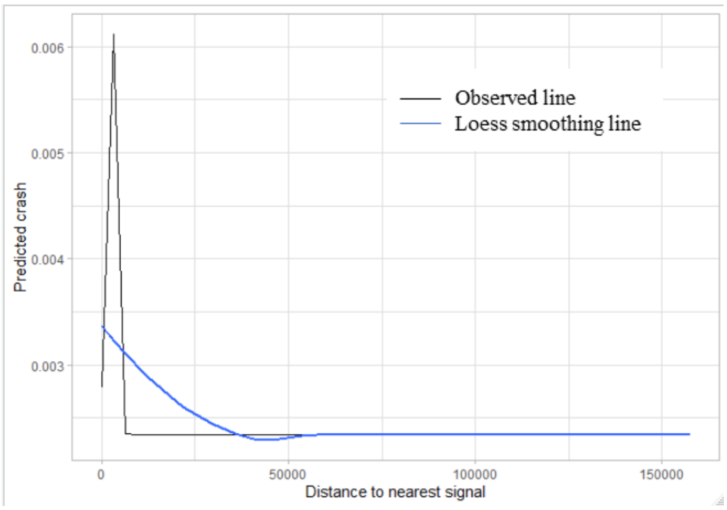
(c) Number of Legs



(d) Pedestrian Volume



(e) Distance to Nearest Signal (Ft)



(f) Vertical Grade (%)

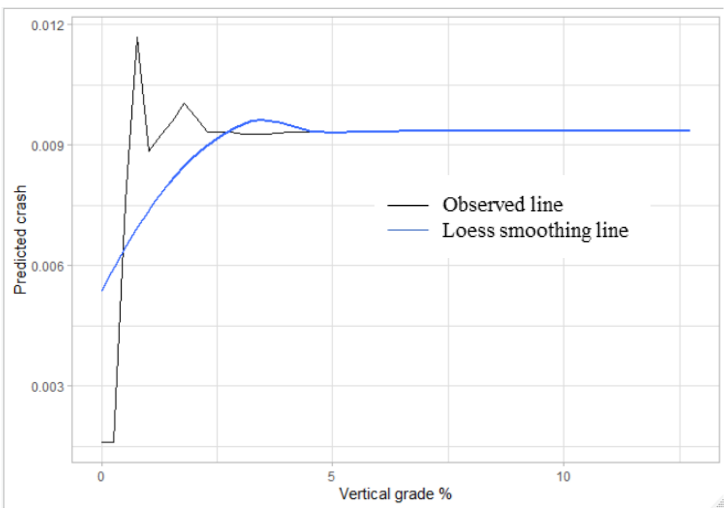
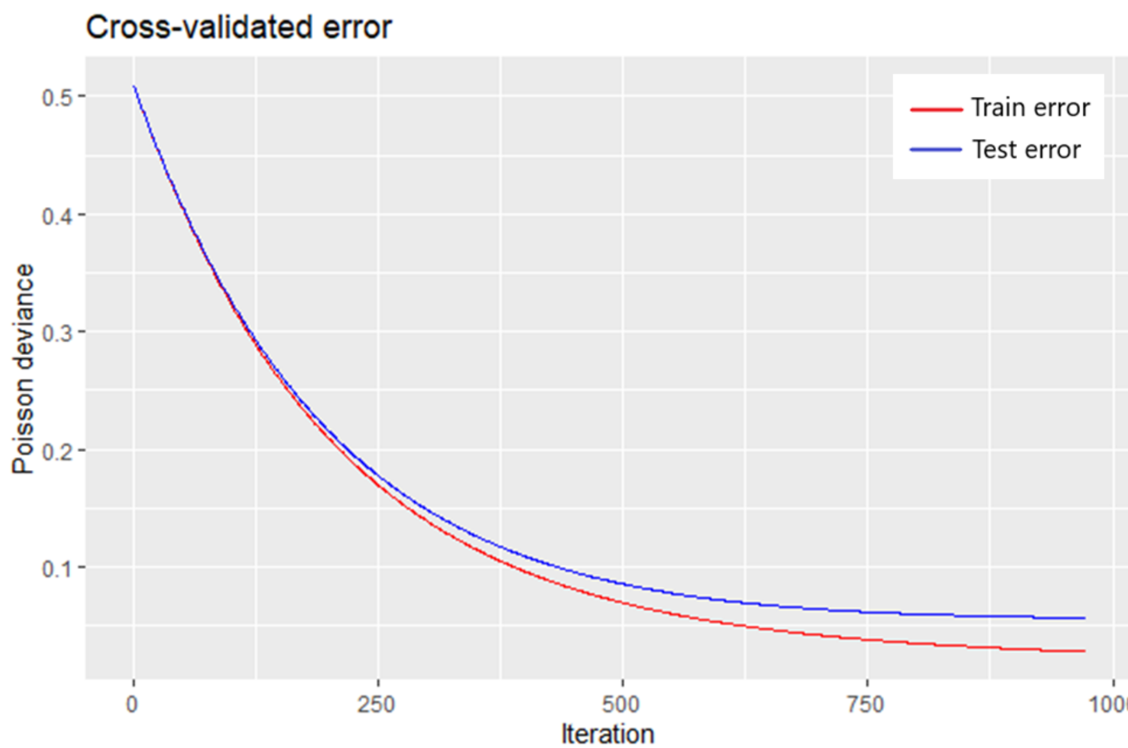


Figure 5-5 shows a non-linear relationship with predicted crashes for different values of residential density. The marginal effect of median width on major approach on predicted crashes are also shown. Generally, wide medians on the major leg at the non-signalized intersections improve pedestrian safety. Intersections with more legs are associated with greater numbers of predicted crashes for pedestrians. Pedestrian volume is associated with frequent predicted crashes. Distance to the nearest signalized intersection is inversely associated with predicted crashes, as non-signalized intersections closer to signalized intersections see more predicted crashes and non-signalized intersections situated further from signalized intersections see fewer predicted crashes. A nonlinear relationship among predicted pedestrian crashes and vertical grades is also presented as vertical grades (up to 5%) are associated with more crashes.

Predictive performance was measured by implementing 10-fold cross validation at every iteration and computing train and test error indicated by Poisson deviance. 90% of the data (N = 4,555) was used for model fitting (train data) and 10% of the held-out data (test data) was used for validation in each iteration. Figure 5-6 shows Poisson deviance obtained by cross-validation procedure at each iteration. The validation process terminated when it reached the minimum test error and no further improvement was found in ten consecutive iterations.

Figure 5-6

Poisson Deviance for Pedestrian Intersection Crashes



From Figure 5-6 it is clear that train data and test data performed similar, however train error was minimized the most compared to test data. From the results, the train error (0.03) and test error (0.06) were similar, indicating that predictive models were stable for new input data. Poisson deviance measures how closely the model's predictions are to the observed outcomes, and thus it may be used as the basis for a goodness of fit test of a boosted tree model.

5.5 Bicycle Crash Frequency at Non-signalized Intersections

Bicycle crashes along segments and at mid-block locations are estimated with negative binomial models and a boosted decision tree model. The following sections provide results on bicycle crash frequency occurring on state routes and state and federal aid routes.

5.5.1 Negative Binomial Model Results

Similar to the estimation process of pedestrian statistical models, over-dispersion and zero-inflation of the dataset were taken into consideration. Estimated negative binomial models showed significant capability to account for over-dispersion. Using forward and backwards elimination only statistically significant variables over 90% confidence interval were contained in the model. Table 5-10 presents negative binomial model results for bicycle crashes at intersections on state routes.

Table 5-10

NB Model for Bicycle Crashes at Non-Signalized Intersections (State Routes, N = 3,192)

	B	*SE*	*z*	*p*
(Intercept)	-15.95	1.91	-8.33	0.00
Natural log of (AADT)	1.42	0.32	4.44	0.002
Natural log of Bicycle volume	-	-	-	-
Heavy truck (%)	-0.47	0.28	-1.68	0.09
Two-way turn lane on major road	1.25	0.34	3.68	0.00
Number of legs	1.01	0.13	7.53	0.00
Median width on major road	-0.14	0.03	-4.02	0.00
Vertical grade (%)	-0.17	0.11	-1.64	0.10
Commuter rail station	2.29	0.68	3.35	0.00
Residential density	0.24	0.13	1.92	0.05

A negative binomial model with AIC 429, null deviance 416.2 and residual deviance 265.9 on 3191 degrees of freedom was estimated to determine significant factors affecting bicycle crashes at non-signalized intersections on state routes. Crashes are positively associated with busy intersections with higher traffic volume (AADT) on the major approach (1.42% increase in crash frequency). Intersections with one increasing leg are associated with more bicycle crashes (175% increase). Bicycle crashes occurring at non-signalized intersections are higher (additional 250% crashes) where two-way turn lanes are present on the major road. Steep vertical grades and wide medians on the major road on an intersection are associated with fewer bicycle crashes (31% and 13% respectively). The presence of nearby commuter rail stations is also associated with higher numbers of bicycle crashes. Also, bicycle crash frequency varied depending on some land use and community characteristics, as intersections adjacent to high residential density tend to see more bicycle crashes.

Negative binomial model with AIC 3159.6, null deviance 3130.2 and residual deviance 1788.9 on 14429 degrees of freedom was estimated to determine significant factors affecting bicycle crashes at non-signalized intersections on state and federal aid routes. Results are shown in Table 5-11.

Table 5-11

NB Model for Bicycle Crashes at Non-Signalized Intersections (State and Federal Aid Routes, N = 14,430)

	B	*SE*	*z*	*p*
(Intercept)	-11.37	0.79	-14.37	0.00
Natural log of (AADT)	1.31	0.12	11.15	0.00
Natural log of (Bicycle volume)	0.05	0.03	1.84	0.07
Truck volume (%)	-0.29	0.10	-2.83	0.00
Number of legs	0.80	0.07	11.42	0.00
Bus stops	0.36	0.08	4.66	0.00
Commuter rail station	0.86	0.46	1.88	0.06
Speed limit	-0.07	0.01	-11.56	0.00
Household income	0.00	0.00	-2.20	0.03
Residential density	0.09	0.04	2.27	0.02
Employment density	0.02	0.01	2.06	0.04
Disabled population (%)	-1.57	0.61	-2.56	0.01

Bicycle crashes are positively associated with busy intersections with higher traffic volume (AADT), with a 1% increase in AADT on major approach increases crash frequency by 1.31%. Also, intersections with greater bicycle volume see higher numbers of intersection crashes, as a 1% increase in bicycle volume increase crash frequency by 0.05%. Intersections with higher speed limits and high truck volume are associated with fewer bicycle crashes. One increasing leg at intersections is associated with more bicycle crashes (123% increase). The presence of nearby bus stations or commuter rail stations is also associated with higher numbers of bicycle crashes. Also, bicycle crash frequency varied depending on some land use and community characteristics. Intersections adjacent to areas with high residential density, high employment density, and high household income see more bicycle crashes, and a higher percentage of disabled population is associated with fewer crashes.

5.5.2 Boosted decision tree model results

After testing various optimization parameters such as learning rate (0.10, 0.05, 0.01, 0.005), tree complexity (1, 5, 10, 15) and sub-sample ratio (0.3, 0.5, 0.8), the optimized result with minimum value for loss function Poisson deviance was achieved for the crash frequency model. Minimum Poisson deviance was found at a 0.01 learning rate and at a tree complexity of 10. Iteration number 1,027 was the iteration with the lowest test error (0.07) and thus was considered the optimal number of iterations.

Table 5-12 shows the influence of explanatory variables for bicycle crashes at non-signalized intersections. Median width on the major leg on an intersection is the most influential variable with a relative contribution of 13% to the model. Residential density near the crash locations contributes similarly (13%) to the model. The number of legs in an intersection has a relative contribution of 12% to the model, and traffic volume (AADT) contributes 11% to the model. Distance to the nearest intersection is found to have an 8% contribution, and vertical grade percentage and household income level each contribute 6% to the model. In summary, these seven variables account for about 70% of the total effect of the model. The remaining variables have a very small influence (between 1% to less than 5%) on predicted crashes.

Table 5-12

Variable Importance for Bicycle Crashes at Non-Signalized Intersections

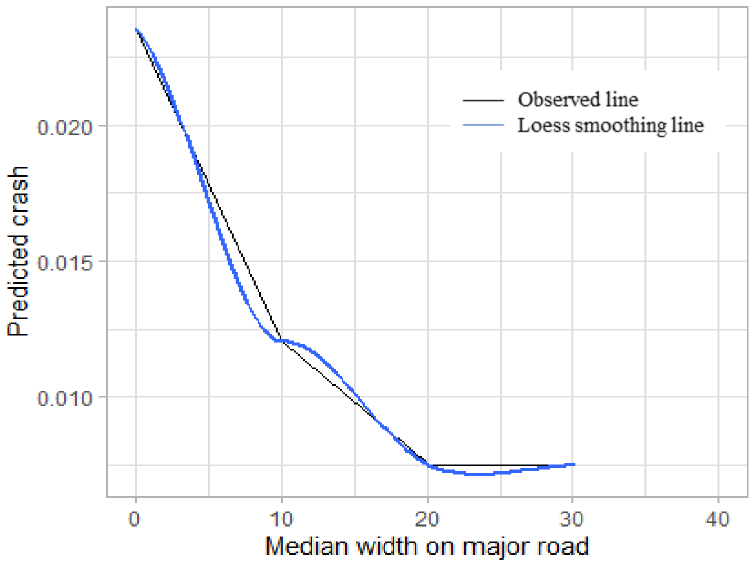
Feature	Relative importance	Cumulative Importance	Cover	Frequency
Median width on major road	0.13	0.13	0.06	0.11
Residential density	0.13	0.26	0.20	0.07
Number of legs	0.12	0.38	0.08	0.05
Traffic volume (AADT)	0.11	0.49	0.10	0.05
Distance to nearest intersection	0.08	0.57	0.02	0.08
Vertical grade (%)	0.06	0.63	0.02	0.06
Household income	0.06	0.69	0.05	0.08
Distance to nearest signal	0.04	0.73	0.03	0.06
Disabled population (%)	0.04	0.77	0.03	0.05
Truck volume (%)	0.03	0.80	0.02	0.04
Shoulder on major roads	0.03	0.83	0.06	0.05
Non-White population (%)	0.02	0.85	0.01	0.03
Commuter rail station	0.02	0.87	0.11	0.01
Right turn lane	0.02	0.89	0.10	0.04
Two-way turn lane on major road	0.02	0.91	0.01	0.01
Bicycle volume	0.02	0.93	0.01	0.04
Zero vehicle household	0.02	0.95	0.01	0.03
Employment density	0.01	0.96	0.01	0.02
Bus stop	0.01	0.97	0.01	0.01
Major road lane width	0.01	0.97	0.01	0.02
Left turn on major road	0.01	0.98	0.02	0.02
Jobs per household	0.01	0.99	0.00	0.02

Figure 5-7 present the partial dependent plots illustrating the most influential variables with a minimum of 5% contribution to the models for bicycle crashes at unsignalized intersections.

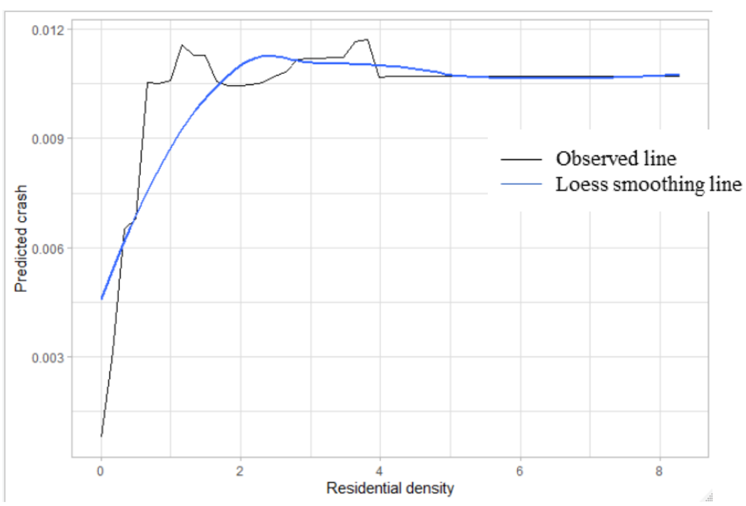
Figure 5-7

Marginal Effects for Non-Signalized Intersection Bicycle Crash

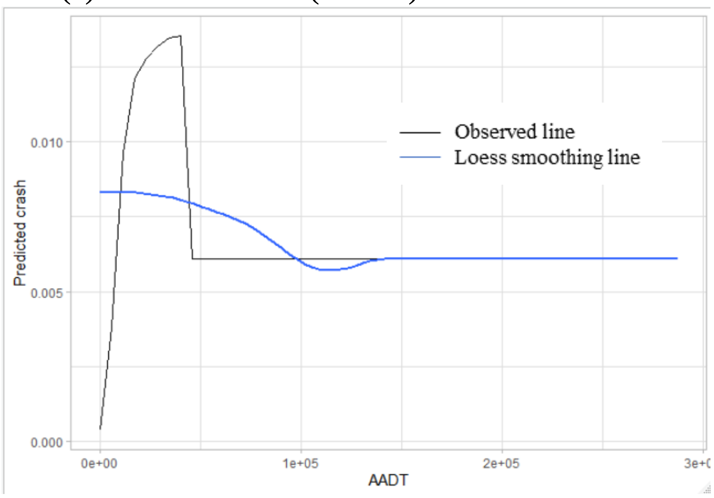
(a) Median width on major roads (ft)



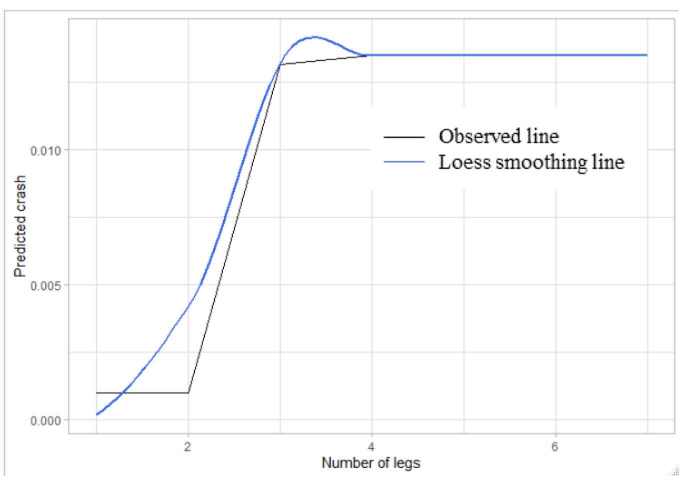
(b) Residential density (Housing unit /acre)



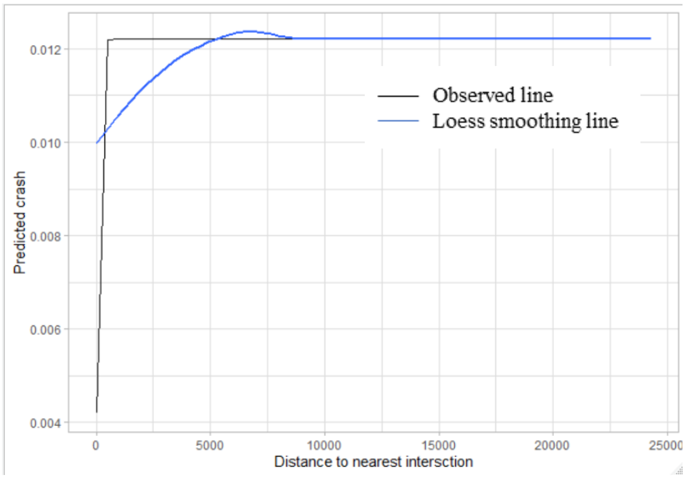
(c) Traffic volume (AADT)



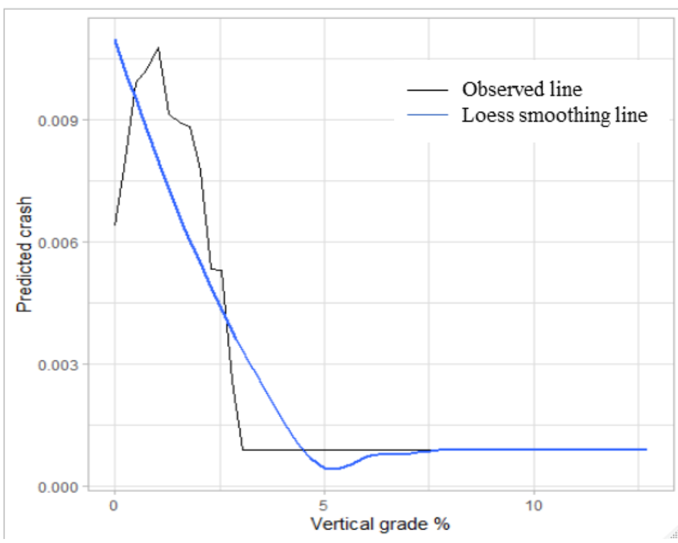
(d) Number of legs



(e) Distance to nearest intersection (ft)



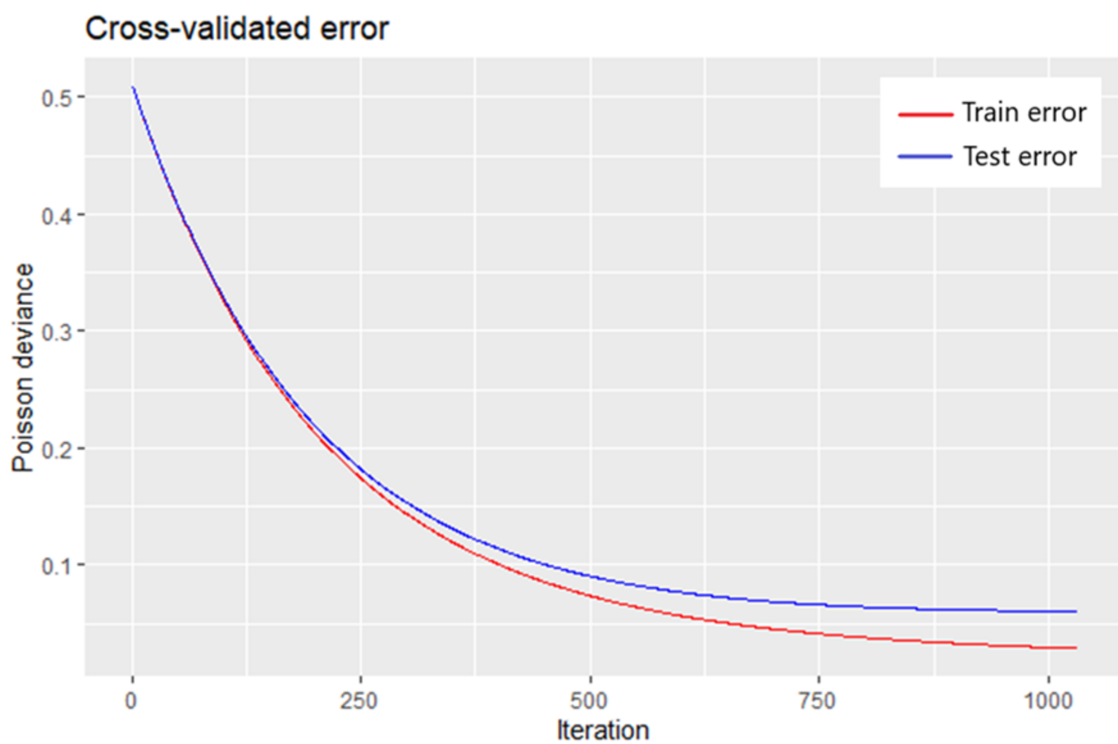
(f) Vertical grade (%)



The marginal effect of the explanatory variable median width on the major road at intersections on predicted crash frequency demonstrated a non-linear relationship. Similarly, predicted crashes generally increase with greater residential density. Predictably, traffic volume AADT has a positive and exponential impact on predicted crashes. Intersections with more legs are associated with greater numbers of predicted crashes for bicyclists. Distance to the nearest intersections is associated with predicted bicycle crashes, as non-signalized intersections closer to other intersections are predicted to see more bicycle crashes. Also, vertical grades (up to 5%) are associated with fewer bicycle crashes.

Figure 5-8

Poisson Deviance for Intersection Bicycle Crash



From Figure 5-8 train data and test data performed similar, however train error was minimized the most compared to test data. From the results, the train error (0.04) and test error (0.07) were similar.

5.6 Pedestrian Crash Severity Results

Table 5-13 provides the statistical results of the ordered logistic models investigating pedestrian crash severity. The model was fitted with a dataset containing 6,740 pedestrian crashes. The model had an overall good fit (McFadden's pseudo R^2 = 0.26).

Table 5-13

Pedestrian Crash Severity (N= 6740)

Variable	*B*	*SE*	*z*	*p*
Crash location: Intersection	-0.43	0.06	-7.42	0.00*
Horizontal curve (present)	-0.12	0.16	-0.73	0.46
Vertical grade (present)	0.25	0.09	2.74	0.01*
Speed limit	-0.01	0.00	-3.94	0.00*
Precipitation > 0.05 (inch)	-0.19	0.08	-2.28	0.02*
Snowfall (inch)	-0.01	0.04	-0.25	0.80
Snow width (inch)	-0.01	0.01	-0.89	0.37
Maximum Temperature > 90F	0.16	0.09	1.72	0.09~
Minimum temperature < 30F	0.01	0.07	0.20	0.84
Light condition: Dark	-0.61	0.06	-10.59	0.00*
Road surface condition: Wet	-0.13	0.09	-1.40	0.16
Vehicle size: large (SUV/ Pickup/ Van/ Large truck)	0.14	0.06	2.52	0.01*
Vehicle size unknown	-0.42	0.10	-4.15	0.00*
Crash involving DUI	1.29	0.20	6.49	0.00*
Crash involving driver at fault	0.23	0.08	2.65	0.01*
Crash involving drowsy driving	2.27	0.42	5.46	0.00*
Crash involving work zone	0.22	0.13	1.71	0.09

Threshold parameters are not included
*Statistical significance: * = p < 0.05, ~ = p < 0.10*

The results indicate that pedestrian crashes at mid-block locations resulted in more severe injury as they increase 35% chance of more severe injury compared to crashes occurring at intersection. Presence of vertical grades increase the chances of a more severe injury by 28%. Weather variables such as high temperature (over 90F) was associated with increased chances of severe crashes by 17%. Rainy weather indicated by daily precipitation over 0.05 inch was associated with 17% less severity pedestrian crashes. The absence of lighting at streets and at intersections greatly increased the odds of severe pedestrian crashes, as dark conditions increase the chances of severe pedestrian crashes by 46%. Involvement of a large vehicle—such as SUV, van, pickup truck, large truck, etc.—significantly increases pedestrian crash severity. Compared to smaller vehicles such as sedan and motorcycles, larger vehicles are linked with a 15% increase in the odds of a more severe injury. Results also indicate that human factors such as driving under influence (DUI), drowsy driving, and where drivers are at fault including distracted driving and disregard towards traffic control devices were detected to be associated with crash severity as they increase the chances of more severe crashes by 263%, 867% and 25% respectively.

5.7 Bicycle Crash Severity Results

Table 5-14 lists the estimation results of the ordered logistic regression model for bicycle crash severity. The model was fitted with 5,764 bicycle crash observations and had a good overall fit (McFadden's pseudo-R^2 = 0.38).

Table 5-14

Bicycle Crash Severity (N= 5,764)

	B	*SE*	*z*	*p*
Crash location: Intersection	-0.12	0.07	-1.66	0.10~
Horizontal Curve (present)	0.27	0.17	1.58	0.11
Vertical Grade (present)	0.54	0.09	5.84	0.00*
Speed Limit (mph)	0.43	0.10	4.31	0.00*
Precipitation > 0.05 inch	-0.13	0.11	-1.28	0.20
Snowfall (inch)	0.13	0.09	1.38	0.17
Snow width (inch)	0.01	0.03	0.43	0.67
Maximum Temperature > 90F	0.02	0.08	0.28	0.78
Minimum temperature < 30F	-0.34	0.11	-2.99	0.00*
Light Condition: Dark	0.21	0.08	2.65	0.01*
Road surface Condition: Wet	0.11	0.16	0.68	0.50
Vehicle Size: Large (SUV/ Pickup/ Van/ Large truck)	0.14	0.06	2.16	0.03*
Vehicle Size Unknown	0.08	0.14	0.57	0.57
Crashes involving DUI	1.55	0.43	3.63	0.00*
Crashes involving driver at fault	0.26	0.12	2.25	0.02*
Crashes involving work zone	0.04	0.17	0.25	0.80

Threshold parameters are not included
Statistical significance: * = $p < 0.05$, ~ = $p < 0.10$

The results indicate that bicycle crashes at mid-block locations resulted in more severe injuries as they increase chances of a more severe injury by 12%, compared to crashes occurring at intersections. Presence of vertical grades increases the chances of a more severe bicycle injury by 72%. High speed roadways are associated with highly severe bicycle crashes as well. Cold weather indicated by lower minimum temperature (below 30F) is associated with 28% increased chances of severe crashes. Finally, the absence of lighting at streets and at intersections greatly increases the odds of severe bicycle crashes, as they increase the chances of severe pedestrian crashes by 23%. Involvement of large significantly increases bicycle crash severity. Compared to smaller vehicles such as sedan and motorcycles, larger vehicles are linked to a 15% increase in the odds of a more severe injury. Results also indicate that human factors such as driving under influence (DUI), and where drivers are at fault including distracted driving and disregard towards traffic control devices were detected to be associated with severe bicycle crashes as they increase the chances of more severe crashes by 371% and 30% respectively.

5.8 Comparison between NB and DT models

Accuracy and predictive power of the negative binomial models and boosted decision tree models are compared through root mean squared error (RMSE), and Poisson deviance between observed and predicted values. Table 5-15 shows RMSE and Poisson deviance value for pedestrian and bicycle crash models.

Table 5-15

NB and DT Models Validation

	Pedestrian crash		Bicycle crash	
	Segments	Intersections	Segments	Intersections
RMSE for NB (train / test)	3.12 / 3.12	7.14 / 7.37	4.89 / 4.92	8.32 / 9.39
RMSE for BT (train / test)	0.23 / 0.31	0.03 / 0.17	0.21 / 0.31	0.06 / 0.14
Poisson deviance (10-fold) for NB	0.46 - 0.59	3.98 - 18.7	0.21 - 0.27	2.54 - 5.38
Poisson deviance (10-fold) for BT	0.23	0.06	0.18	0.07

For pedestrian crash models at segments, NB models show RMSE value of 3.12 in train data (70% of the sample) and 3.13 in test data (30% of the sample). Boosted tree models show RMSE value of 0.23 for train data and 0.31 for test data. In both models, RMSE value between train and test data are close, indicating stable models. Boosted decision tree models has lower RMSE value possibly because they have quite a few more variables in the model compared to the negative binomial models. At intersections, NB models show RMSE value of 7.14 in train data and 7.37 in test data. Boosted tree models show RMSE value of 0.03 and 0.17 for train and test data. Between train and test data RMSE value are close indicating relatively stable models. Boosted tree models has more variables in the model and thus show a lower RMSE value.

Additionally, a 10-fold cross validation model was applied to measure the Poisson deviance in the models. In a 10-fold validation process, in each iteration data is split between 9-folds as train data and 1-fold is for test data. Average Poisson deviance for all iteration is presented. For pedestrian segment crash, negative binomial models show Poisson deviance value ranging from 0.46 to 0.59. Boosted tree models show Poisson deviance value of 0.23 indicating a better fit model. For pedestrian intersection crashes, negative binomial models show Poisson deviance value ranging from 3.98 to 18.7. Boosted tree models show Poisson deviance value of 0.06 indicating a better fit model.

For bicycle crash models at segments, NB models show RMSE value of 4.89 in train data (70% of the sample) and 4.92 in test data (30% of the sample). Boosted tree models show RMSE value of 0.21 for train data and 0.31 for test data. In both models, RMSE value between train and test data are close, indicating stable models. Boosted decision tree models has lower RMSE value possibly because they have quite greater number of variables in the model compared to the negative binomial models. At intersections, NB models show RMSE value of 8.32 in train data and 9.39 in test data. Boosted tree models show RMSE value of 0.06 and 0.14 for train and test data. Between train and test data RMSE value are close indicating relatively stable models. Boosted tree models has more variables in the model and thus show a lower RMSE value.

Additionally, a 10-fold cross validation model with data split between 9-folds as train data and 1-fold is for test data for each iteration shows average Poisson deviance value for bicycle crashes. For pedestrian segment crash, negative binomial models show Poisson deviance value ranging from 0.21 to 0.27. Boosted tree models show Poisson deviance value of 0.18 indicating a better fit model. For bicycle intersection crashes, negative binomial models show Poisson deviance value ranging from 2.54 to 5.38.

Moreover, Table 5-16 shows all the significant variables in the crash models presented in this analysis.

Table 5-16

Significant Variables for Pedestrian and Bicycle Crashes

Variable	Negative binomial				Boosted Decision Tree				Severity	
	Segment		Intersection		Segment		Intersection		Segment/ Intersection	
	Ped	Bike	Ped	Bike	Ped	Bike	Ped	Bike	Ped	Bike
Traffic volume (AADT)	+	+	+	+	+	+	+	+	n/a	n/a
Pedestrian volume	+	n/a	n/a	n/a	+	n/a	+	n/a	n/a	n/a
Bicycle volume	n/a	+	n/a	n/a	n/a	+	n/a	n/a	n/a	n/a
Vertical grade	n/a	+	+	-	(+ to -)	n/a	+	(+ to -)	+	+
Number of legs at intersections	n/a	n/a	+	+	n/a	n/a	+	+	n/a	n/a
Left turn/ Two-way turn lanes	n/a	+	n/a	+	n/a	n/a	n/a	n/a	n/a	n/a
Median width at major leg	n/a	n/a	n/a	-	n/a	n/a	-	-	n/a	n/a
Transit stops	+	+	n/a	+	+	n/a	n/a	n/a	n/a	n/a
Density of driveways	+	+	n/a	n/a	-	n/a	+	(+ to -)	n/a	n/a
Residential or employment density	n/a	+	+	+	+	n/a	+	(+ to -)	n/a	n/a
Household income	n/a	n/a	n/a	n/a	(+ to -)	(+ to -)	n/a	(+ to -)	n/a	n/a
% Non-white population	+	n/a	+	+	+	+	n/a	n/a	n/a	n/a
% With a disability	+	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a

(+): Positive association with crashes
(-): Negative association with crashes

5.9 Chapter Summary

In this chapter, detailed results of negative binomial models and gradient boosted decision tree models investigating pedestrian and bicycle crashes at mid-block locations along road segments and at non-signalized intersections were presented. The negative binomial models identified statistically significant variables affecting pedestrian and bicycle crashes. The presence of bus stops, traffic volume, driveway density, employment density, pedestrian and bicycle exposure, left turn lanes, and the presence of roadside barriers were commonly found to be significant for mid-block pedestrian and bicycle crashes. The number of legs at intersections, traffic volume, pedestrian and bicycle exposure, vertical grade percentage, 2-way turn lanes on major legs, and residential and employment density were commonly found to be significant for pedestrian and bicycle crashes at non-signalized intersections. On the other hand, gradient boosted models identified the most important variables by explaining their relative contribution to the crash models. Traffic volume, driveway density, bike volume, pedestrian volume, household income level, and concentrations of minority demographic groups were some of the common important variables predicting pedestrian and bicycle crashes at mid-block locations. The number of legs at intersections, traffic volume, residential density, wide medians on the major leg at intersection, and vertical grade were often important predictors of pedestrian and bicycle crashes at non-signalized intersections. The non-linear effects of these variables on predicted pedestrian and bicycle crashes were also discussed. Finally, crash severity models presented road characteristics like vertical grades, mid-block crashes, weather conditions like temperature and precipitation, as well as human factors affecting injury severity once a crash has occurred.

6 DISCUSSION

6.1 Overview

This study investigated state-wide transportation networks in Utah to identify the roadway geometry, traffic characteristics, and land use and community characteristics of road segments and unsignalized intersections which affect pedestrian and bicycle crashes using both traditional statistical models (negative binomial) and machine learning techniques (boosted decision tree).

First, Chapter 1 provided the background and importance of a systemic analysis of pedestrian and bicycle crashes and mentioned the objectives of this research. Chapter 2 reviewed the literature to understand existing knowledge about the variables correlating to active transportation mode (walking and bicycling) safety. Next, the detailed data collection and assembly process were described in Chapter 3. Chapter 4 reported statistical and machine learning methods to identify factors affecting pedestrian and bicycle crash frequencies. Using the presented models, Chapter 5 reported the data analysis results.

This concluding section first highlights the major findings of this study. Next, it summarizes the major contributions of this work, and then policy implications are discussed. This section concludes by noting some limitations and potential future work.

6.2 Key Findings

The analyses presented in this study has been carried out to understand factors associated with pedestrian and bicycle crashes at mid-block locations and unsignalized intersections. It achieved five sub-objectives to fulfill the overall goal.

This study has identified and ranked several traffic, road geometry, land use and socio demographic characteristics to be common and most important sources of pedestrian and bicycle crash occurrences. Traffic volume (AADT), pedestrian and bicyclists volume on streets, vertical grades, greater number of bus stops, larger & complicated intersections, high residential or employment density, streets near low-income or minority populated neighborhoods were some of the common factors on Utah roadway directly affecting pedestrian and bicycle safety.

Analyzing crash data with traditional statistical models as well as with machine learning models has helped examine the transportation context and data availability around a network wide systemic safety study for pedestrian and bicyclists. Especially the machine learning models helped leverage the occasional incomplete data and imputed relationships between critical variables and crash occurrences as well as illustrated the non-linear association among variables. Results from these two approaches have confirmed many findings and helped draw sound and accurate conclusions.

Using pedestrian exposure from traffic signal ATSPMs, and crowd-sourced bicycle exposure data in the crash frequency analysis has statistically improved the models and provided better explanation of results. Safety-in-numbers effects have been confirmed for pedestrian and bicyclists on Utah road network: more walking and biking activities tend to improve overall traffic safety condition.

This study has identified a bicycle and pedestrian crash frequency disparity in Utah based on neighborhood sociodemographic characteristics. Neighborhoods with low-income residents or minority groups were related to greater number of pedestrian or bicycle crashes. Road safety conditions must improve equally for everyone regardless of their socio-economic standing.

The impact of weather on pedestrian and bicycle safety were interesting and complex. While adverse weather conditions can certainly create challenging conditions for bicyclists and pedestrians (dark road conditions posed possibilities for high severity crashes), pedestrian and bicycle crashes were found to carry greater level of injury on bright and sunny days. Continuous interventions and education programs should be carried on in Utah to ensure zero fatalities for pedestrian and bicyclists in all weather conditions.

This study identified significant risk factors affecting frequency and severity in pedestrian and bicycle crashes at mid-block locations along segments and at unsignalized intersections in Utah. Significant factors affecting pedestrian crash frequencies at 4,979 road segments and 3,378 intersections on state only routes, and 11,910 road segments and 20,381 intersections on state and federal aid routes, were identified. Bicycle crash frequencies at 11,910 (state only) and 11,865 (state and federal aid routes) road segments and at 3,192 unsignalized intersections on state only routes and 20,381 unsignalized intersections on state and federal aid only routes were analyzed to determine factors that affect crash frequency. Overall, the findings from negative binomial crash analysis models showed statistically significant factors and their magnitude and direction; and boosted decision tree models revealed the marginal effects and often non-linear associations between pedestrian and bicyclist crashes and roadway geometry, traffic characteristics, and land use and community characteristics. The following paragraphs summarize the key findings of this study.

Some road geometry characteristics illustrated significant impact on pedestrian and bicycle crash frequency. Streets with left turn lanes saw fewer pedestrian crashes but more bicycle crashes. Greater number of turn lanes can pose risk for crossing bicyclists and pedestrians as turning drivers are focused on the vehicles coming from opposite direction (Chen, 2009). High vertical grades were associated with more bicycle crashes. Boosted tree models illustrated some non-linear associations that were not captured by the statistical model. Streets and intersections with over 6% vertical grades saw more pedestrian crashes while grades below 6% were not related to more crashes. This finding suggests 6% or above vertical grade may provide increasing challenge for motorists to detect pedestrians at mid-bock locations. Predicted occurrence of increase in bicycle crashes with any degree of vertical grades present are also found. This may occur because of the fact that riding bicycles along vertical grades is a physically demanding task that may distract them from riding safely along with the fact that vertical grades obstruct vision for motor vehicle drivers (Chen & Zhuo, 2016). From these results, the safety issue of car drivers' visibility may be worse for bicyclists compared to pedestrians. The fact that estimating a bicyclist's speed or position can be challenging while driving can also be a reason behind this finding.

Roadside barriers were moderately associated with fewer pedestrian crashes. For bicycle crashes, roadside barriers and rumble strips along segments were significantly associated with lower crash occurrences. Wide medians on the major roads at unsignalized intersections are also generally associated with fewer pedestrian crashes. These findings suggest that presence of barrier and median devices assist in separating motor vehicle traffic from people walking and bicycling. Moreover, wide medians may provide a refuge for pedestrians while crossing and thus improve the safety conditions at intersections (Palamara & Broughton, 2013),.

Several intersection-related characteristics were linked to pedestrian and bicycle crash frequencies. The number of approaches at unsignalized intersections had a major impact on both pedestrian and bicycle crashes, since complex intersections with more approaches saw more pedestrian and bicycle crashes. It is an anticipated outcome because larger and complicated intersections can provide more conflict between motor traffic and pedestrian and bicyclists (Schneider et al., 2010). Moreover, intersections with 2 way turn lanes at the major road also saw greater number of bicycle crashes. This may occur because motor drivers may be too focused to find a gap by looking at the oncoming traffic and not looking for bicyclists at intersections or on driveways (Dumbaugh et al., 2013).

Presence of frequent bus stops near segments and intersections were related to more pedestrian and bicycle crashes since they promote more walking and biking activity in these locations. Transit stops are also a common place where non-motorized modes such as pedestrians and bicyclists interact at a greater extent with other non-motorized and motorized modes (Miranda-Moreno et al., 2011). In the case of both pedestrian and bicycle crashes, statistical model results showed positive associations between driveway density and crash occurrences, although at a moderate to low level. This finding may imply that driveway density created more conflict points between motorized modes and non-motorized modes such as walking and bicycling (Kim & Ulfarsson, 2019). The marginal effects from the boosted tree models revealed that the association between pedestrian or bicycle crashes and driveway density changed at a varying rate, and this relationship is quite complex. On major roads, up to 10 driveway per mile is associated with high pedestrian crashes, whereas 11 to 40 driveways per mile are associated with fewer pedestrian crashes. Driveway density up to 25 per mile were associated with fewer bicycle crashes on major roads. An interesting phenomenon was noticed for driveway density on minor roads as they were generally associated with fewer pedestrian crashes and more bicycle crashes. In all cases, the effect of driveway density was neutralized beyond a certain point.

An interesting finding from this study was the association between the presence of bike lanes and pedestrian crashes. Based on the statistical test results, the addition of bike lanes on at least two approaches at an unsignalized intersection could reduce pedestrian crashes by over 30%. This finding aligns with the similar conclusion drawn by a recent study that analyzed pedestrian crash factors at signalized intersections in Utah (Singleton et al., 2021). In some respects, bike lanes shorten the portion of the crossing distance at unsignalized intersections where pedestrians are exposed to motor vehicles while crossing the road. Bike lanes may also provide better visibility between people walking and driving, as well as a place for cars to wait before turning which can be helpful for detecting pedestrians on the road. Also, the presence of bike lanes could indicate other complete streets treatments, such as traffic calming devices, that have also been shown to improve pedestrian safety (LaPlante & McCann, 2008).

Employment density from the statistical models was significantly associated with more mid-block bicycle crashes. Residential density was associated with high pedestrian crashes. Unsignalized intersections near high residential and high employment density had seen greater number of walking and bicycling crash occurrences. This may happen due to increasing conflicts between different travel modes at these locations. Residential land use may encourage frequent pedestrian and bicycling activity especially by the nearby residents (Siddiqui et al., 2012). Employment density also typically see more walking and bicycling activity (Loukaitou Sideris et al., 2007).

When estimating crashes at a broader spatial scale with both state and federal aid routes, model results revealed that streets with higher truck volume percentage were associated with more pedestrian crashes and fewer bicycle crashes. Due to the large size of heavy trucks, there may be additional blind spots that are worsening the safety condition for pedestrian and bicyclists (Narayanamoorthy et al., 2013).

Generally, traffic volume was significantly associated with pedestrian and bicycle crashes. A 1% increase in traffic volume yielded 0.42% to 0.44% more crashes along mid-block locations and 0.81% to 1.42% at unsignalized intersections. A closer look at the non-linear relationship illustrated by the boosted tree models shows that traffic volume has critical impact on both pedestrian and bicycle crashes, especially because crashes increase significantly with higher traffic volume (Nordback et al., 2014). Similarly, at both road segments and at unsignalized intersections, high pedestrian and bicycle volumes are associated with high pedestrian crashes (Siddiqui et a., 2012). However, in all cases, a 1% increase in people walking and bicycling resulted in less than a 1% increase in crashes. This is strong evidence of the "safety-in-numbers" effect (Singleton et al., 2020). This less than proportional increase in crashes illustrates that as more pedestrians and bicyclists are on the road, drivers possibly adjust their driving behavior and thus improve the safety condition of roadways.

Statistical models identify low-income households as a significant variable related to more crashes for streets and unsignalized intersections; although the association was weak (less than 1% increase in crashes). However, the marginal effect of this variable shows a complex non-linear relationship between household income level and estimated pedestrian crashes. Low-income household locations were generally associated with higher pedestrian and bicycle crash frequency (Chimba et al., 2014). Additionally, road segments and intersections around racial minority groups (non-white population including Hispanic, African American and other demographic groups) were found to experience increasing number of pedestrian and bicycle crash frequency. This may happen because neighborhoods with primarily low-income residents or minority groups often lack access to better pedestrian and bicycle facilities (Ukkusuri et al., 2011).

Regarding crash severity, pedestrian and bicycle crashes that occurred in darkness exhibited a greater probability of high severity. Darkness significantly decreases the visibility of drivers and pedestrians, which in turn increases reaction times and braking distance. Visibility improvement design such as reflectors can help drivers in detecting pedestrian and bicycle movements. Street lights should also improve pedestrian and bicycle safety condition on roads (Kim et al., 2007). The findings also show that vertical grades are associated with greater severity in crashes, which again reiterates the importance of visibility in pedestrian and bicycle safety. Moreover, high daily temperature was associated with severe crashes. This may happen due to decreased driver and pedestrian or bicyclists' awareness on bright sunny days. This finding is aligned with those of previous studies, which found that pedestrians and drivers can be less patient and are more likely to violate traffic rules in high temperature (Zhai et al., 2019; Naik et al., 2016). This study found that rainy weather was associated with lower severity crashes. This may happen due to drivers' risk mitigation behavior during heavy precipitation.

Overall findings identify and illustrate significant factors that affect pedestrian or bicycle safety at mid-block locations or unsignalized intersections. Also, the general direction and the non-linear relationship of the variables considered are consistent with the previous body of literature.

6.3 Contributions

This research has implemented a machine learning technique—boosted decision tree model—that possesses several advantages in crash analysis. First, this method provides satisfactory prediction power by utilizing a training set and a test set to evaluate crash frequency models, and at the same time does not lose interpretability of the results (i.e., not a black box method). This helps to rank the most important variables contributing to the development on the model and thus possess the largest impact on crash frequency. Ranking of the explanatory variables associated with crash frequency assists in prioritizing resources and in selecting safety countermeasures for sites. To date, limited research has applied similar methods to study vulnerable road user (pedestrian and bicyclist) crash frequency. Second, machine learning approaches can handle big data and this is especially important since in the near future, developing new technologies can be a source of big data. For example, newer sources of collision data such as sensors and smartphone apps are becoming common which can better report minor collisions (Aichinger, 2016). When processing large amounts of data, computational efficiency is a major challenge. Third, machine learning techniques are more sensitive to outliers in the sample and capture the interactions among variables. This can be helpful in pedestrian and bicycle crash frequency analysis as these crash data contains large number of zeros and only a few high crash locations. Fourth, discrete variables with many categories with possibility of multicollinearity are more properly handled by machine learning techniques in contrast to conventional regression models (Prati et al., 2017). Fifth, although elasticities can be computed for explanatory variables and used for evaluation of significant variables in statistical models, the decision tree approach provides an alternative way to rank factors. Understanding the non-linear effects assist in determining cut off points for roadway and traffic characteristics that affect pedestrian and bicycle crashes to greater extent and thus need to be addressed first.

Furthermore, the systemic safety analysis of crashes investigated using negative binomial models can identify the locations with characteristics which contribute to frequent pedestrian and bicycle crashes (Monsere et al., 2016; Preston et al., 2013). This system-wide, data-driven study is a better approach to locate intersections and streets with elevated risk and provide evidence to support for transportation planning, policy recommendations, and road safety programs.

A network-wide study has provided the opportunity to examine the relationship of social equity and traffic safety in Utah. More specifically, a closer inspection of the association between walking and bicycling crashes with low-income, zero-vehicle households and racial minority groups has provided evidence that more work is required to eradicate the traffic safety discrepancy among different population groups and provide a safer transportation system for all.

This research study has also produced a rich and robust dataset with sufficient samples that included state and federal aid routes. This dataset can readily be used in further research, data-driven policymaking, and the development of road safety programs in the state of Utah.

6.4 Policy Implications

For transportation engineers, planners, and policymakers, this research provides several statistically founded implications and recommendations to improve pedestrian and bicycle safety through data-driven decision-making and appropriate engineering, education, and enforcement measures.

For instance, the results indicate that visibility is an issue for car drivers that are harming bicyclists' safety more as compared to pedestrians. Especially at unsignalized crossings, or at mid-block locations where vertical grades are present, greater caution is to be expected of motorists. This suggests a need for state- and national-level campaigns focused on promoting driver awareness and greater caution regarding sharing roads with bicyclists. Infrastructure-related measures such as bike lanes, improved signs, and markings may be warranted. Increasing the amount of walking and bicycling activity on streets and intersections may create more conflicts with motor traffic, but at a decreasing rate (per person). Introducing road diet measures, and speed setting based on neighborhood context and engineering judgement to provide better pedestrian and bicyclist safety, are likely to improve traffic safety conditions. Transportation planning and design policy "Complete Streets" have illustrated the added safety benefit for pedestrians and bicycles in various cities and towns nationally, after accommodating planning components in road design (Complete Streets, 2015). Stringent enforcement programs including detection of driving under the influence of alcohol or drugs can also reduce some of these safety challenges.

Pedestrians and bicyclists also may need to be more proactively safe on the road. Bicycle and pedestrian safety education programs can provide information regarding good practices like crossing the street carefully by stopping, looking, and listening to surroundings, safely negotiating turns and intersections, and demonstrating traffic awareness, being predictable to let drivers predict accurately on shared roads, etc. The National Highway Traffic Safety Administration (NHTSA) carries out a number of pedestrian safety programs and bicycle safety programs aimed at education for motor vehicle drivers, bicyclists and pedestrians to promote safe behavior when sharing the roadway.

Roadway and pedestrian access design has an important impact on non-motorized road users' safety. At intersections with major roadways with high traffic volumes, pedestrians face a difficult, high-speed environment. Better pedestrian-friendly design such as wide medians for refuge while crossing, better crosswalk visibility, sidewalk continuity, and connectivity can reduce crash occurrences.

Unsignalized intersections can create confusion and present a safety hazard to motor vehicle drivers and pedestrian and bicyclists. A tendency to cross streets at mid-block locations without marked facilities may create unexpected conflict situations. Pedestrian and bicyclist interactions with motor traffic in darkness, and in adverse weather conditions, often create traffic safety issues. To address these concerns, campaigns to increase visibility of all road users can have a significant impact on traffic safety.

Increased walking and bicycling activity near transit stops needs to be taken into account when planning for these facilities. Visible and marked crossing facilities, safe boarding locations, low-stress bicycle routes, providing sufficient signs, and markings cautioning drivers regarding non-motorized users' presence must be prioritized.

Interestingly, community characteristics such as low-income residential areas, households with no vehicle, and residential areas with racial minority groups including Hispanic and African American communities are found to be more vulnerable towards pedestrian and bicyclist crashes. Transportation policymaking and interventions sensitive to social equity issues should investigate these findings and strive to provide better and safer traffic conditions for all.

Practitioners should strive to apply the principle of no fatality to operate a safe and sustainable transportation system. UDOT's Zero Fatality program promotes practices and education programs to eliminate all traffic fatalities and serious injuries. In order to promote an active and healthy lifestyle, in view of the ongoing multimodal planning, road safety programs, and infrastructure development, continual research on pedestrian and bicycle crashes based on local context are urgently needed.

6.5 Limitations and Future Work

A few limitations are present in this study, especially regarding data quality and availability. The dataset used in this study was created with spatial information regarding road networks in Utah. A larger and more complete dataset may have yielded different results, especially if the unobserved features were correlated with factors affecting bicycle and pedestrian crash frequency. Also, this study has investigated 10 years of crash occurrences, but explanatory variables were observed at a single point in time. Due to this limitation in temporally varying data, possible changes in road network characteristics over time may not have been captured.

While the application of the STRAVA fitness app offered a proxy measure of exposure at a network-wide scale, these data may not represent total bicycle activity on segments and intersections to a satisfactory extent. Raw counts of STRAVA data captures about 5% bicycle activity (Lee & Sener, 2019), and the difference between total bicycle activity and app counts, and potential under- and over-representation of samples in a crash study, is a major concern, and adjustments to this crowdsourced data with respect to population distribution and field observations may be required (Saad et al., 2019). Better representation of STRAVA data or other similar big-data sources can be achieved by cross-use of such data with other count sources such as automated counter at selected locations. This method can also provide a validation for the use of app data for bicycle volumes in Utah. Aggregating app counts to macro level zones such as the Census block level, and estimating high frequency and low frequency zones, might be a useful data engineering technique to make the best of this limited data source.

The availability of a more complete dataset with detailed information and features would have strengthened this study. For instance, information regarding different types of bicycle facilities could provide important insights considering that all bike facilities are not effective at a similar level (Morrison et al., 2019; Wall et al., 2016). Increasing the availability of further datasets such as the presence of roadside parking, the type of unsignalized intersections (four-way stop, two-way stop, or no stop sign present) can provide better context to this crash analysis study.

Pedestrian and bicycle crash analysis at a state-wide scale cannot easily capture the variation in rural and urban areas, in more populous counties compared to less populous counties, and pedestrian and bicycle activity around metro areas compared to recreational trails. Site characteristics or land use characteristics may account for crashes in such diverse locations. Because statistical crash prediction methods like negative binomial models do not account for spatial autocorrelation (Siddiqui et al., 2012), building separate models at various spatial scales (such as for each county) to account for spatial correlation in future studies can be a significant improvement.

Finally, while boosted tree models perform very well when handling large amounts of data, they require the optimization of several parameters (such as learning rate, tree complexity, number of trees, subsample for observations and features, etc.) for determining the best model with improved predictive performance. Additionally, while these methods predict crash occurrences to a great extent of accuracy, they do not provide a determining confidence interval and significance of difference between relative contributions.

This study has attempted to model the crash severity with the same datasets (detailed traffic, road geometry, land use and community characteristics) used in crash frequency analysis. It would provide meaningful and comparable implications; however, after filtering for weather station data, it was not possible to get a sufficient sample size of observations for pedestrian and bicycle crashes. A network wide pedestrian and bicycle safety study conducted with traditional statistical models and machine learning models helps to analyze the various factors and data associated with it, and provides understanding to design effective countermeasure and policy making for a safer transportation system.

7 REFERENCES

AASHTO (2010). Highway Safety Manual. American Association of State Highway and Transportation Officials (AASHTO), Washington, DC.

Abdel-Aty, M., & J. Keller (2005). Exploring the Overall and Specific Crash Severity Levels at Signalized Intersections. *Accident Analysis & Prevention*, Vol. 37 No. 3, pp. 417-425.

Abdel-Aty, M., Chundi, S. S., & Lee, C. (2007). Geo-spatial and log-linear analysis of pedestrian and bicyclist crashes involving school-aged children. Journal of safety research, 38(5), 571-579.

Abdel-Aty, M., & Haleem, K. (2011). Analyzing angle crashes at unsignalized intersections using machine learning techniques. Accident Analysis & Prevention, 43(1), 461-470. https://doi.org/10.1016/j.aap.2010.10.002Abdel-Rahim, A., & Sonnen, J. (2012). Potential safety effects of lane width and shoulder width on two-lane rural state highways in Idaho. Idaho Transportation Department. https://rosap.ntl.bts.gov/view/dot/25160

Al-Ghamdi, A. S. (2002). Using Logistic Regression to Estimate the Influence of Accident Factors on Accident Severity. *Accident Analysis & Prevention*, Vol. 34 No. 6, pp. 729-741

Alluri, P., & Ogle, J. (2012). Effects of state-specific SPFs, AADT estimations, and overdispersion parameters on crash predictions using SafetyAnalyst (No. 12-4332).

Amoh-Gyimah, R., Saberi, M., & Sarvi, M. (2016). Macroscopic modeling of pedestrian and bicycle crashes: A cross-comparison of estimation methods. Accident Analysis & Prevention, 93, 147-159. https://doi.org/10.1016/j.aap.2016.05.001

Andersson A. L., & Bunketorp. O. (2002). Cycling and Alcohol. *Injury*, Vol. 33 No. 6, pp. 467-471.

Baltes, M. R., & Chu, X. (2002). Pedestrian level of service for midblock street crossings. Transportation Research Record: Journal of the Transportation Research Board, 1818(1), 125-133. https://doi.org/10.3141/1818-19

Berdica, K. (2002). An Introduction to Road Vulnerability: What Has Been Done, Is Done and Should be Done. *Transport Policy*, Vol. 9 No. 2, pp. 117-127.

Bhatia, D., Richmond, S. A., Loo, C. J., Rothman, L., Macarthur, C., & Howard, A. (2016). Examining the impact of cycle lanes on cyclist-motor vehicle collisions in the city of Toronto. Journal of Transport & Health, 3(4), 523-528. https://doi.org/10.1016/j.jth.2016.04.002

Bíl, M., Bílová, M. & Müller, I. (2010). Critical Factors in Fatal Collisions of Adult Cyclists with Automobiles. *Accident Analysis & Prevention*, Vol. 42 No. 6, 2010, pp. 1632-1636.

Boufous, S., de Rome, L., Senserrick, T., & Ivers, R. (2012). Risk Factors for Severe Injury in Cyclists Involved in Traffic Crashes in Victoria, Australia. *Accident Analysis & Prevention*, Vol. 49, pp. 404-409.

Britt, J., Silver, I., & Rivara, F. P. (1998). Bicycle helmet promotion among low income preschool children. Injury Prevention, 4(4), 280-283. http://doi.org/10.1136/ip.4.4.280

Carter, D. L., Hunter, W.W., Zegeer, C.V., Stewart, J.R. & Huang. H.F., (2006) Pedestrian and Bicyclist Intersection Safety Indices: Final Report. Publication FHWA-HRT-06-125. FHWA, U.S. Department of Transportation.

Chen, C., Wang, H., Roll, J., Nordback, K., & Wang, Y. (2020). Using bicycle app data to develop Safety Performance Functions (SPFs) for bicyclists at intersections: A generic framework. Transportation Research Part A: Policy and Practice, 132, 1034-1052. https://doi.org/10.1016/j.tra.2019.12.034

Chen, L., Chen, C., Srinivasan, R., McKnight, C. E., Ewing, R., & Roe, M. (2012). Evaluating the safety effects of bicycle lanes in New York City. American Journal of Public Health, 102(6), 1120-1127. https://doi.org/10.2105/AJPH.2011.300319

Chen, P. (2015). Built environment factors in explaining the automobile-involved bicycle crash frequencies: A spatial statistic approach. Safety Science, 79, 336-343. https://doi.org/10.1016/j.ssci.2015.06.016

Chen, P., & Zhou, J. (2016). Effects of the built environment on automobile-involved pedestrian crash frequency and risk. Journal of Transport & Health, 3(4), 448-456. https://doi.org/10.1016/j.jth.2016.06.008

Chimba, D., Emaasit, D., Cherry, C.R., & Pannell, Z., (2014). Patterning Demographic and Socioeconomic Characteristics Affecting Pedestrian and Bicycle Crash Frequency. No. 14-0600. Presented at Transportation Research Board 93rd Annual Meeting, Washington D.C.

Cottrill, C. D., & Thakuriah, P. V. (2010). Evaluating pedestrian crashes in areas with high low-income or minority populations. Accident Analysis & Prevention, 42(6), 1718-1728. https://doi.org/10.1016/j.aap.2010.04.012

Cripton, P. A., Shen, H., Brubacher, J. R., Chipman, M., Friedman, S. M., Harris, M. A., ... & Teschke, K. (2015). Severity of urban cycling injuries and the relationship with personal, trip, route and crash characteristics: Analyses using four severity metrics. BMJ Open, 5(1), e006654. http://doi.org/10.1136/bmjopen-2014-006654

Daniels, S., Nuyts, E., & Wets, G. (2008). The effects of roundabouts on traffic safety for bicyclists: an observational study. Accident Analysis & Prevention, 40(2), 518-526. https://doi.org/10.1016/j.aap.2007.07.016

Daniels, S., Brijs, T., Nuyts, E., & Wets, G. (2010). Explaining variation in safety performance of roundabouts. *Accident Analysis & Prevention*, *42*(2), 393-402.

Davis, K. D., & Hallenbeck, M. E. (2008). Evaluation of engineering treatments and pedestrian and motorist behavior on major arterials in Washington State. Washington Department of Transportation. https://rosap.ntl.bts.gov/view/dot/16612

De Geus, B., Vandenbulcke, G., Panis, L. I., Thomas, I., Degraeuwe, B., Cumps, E., & Meeusen, R., (2012). A Prospective Cohort Study on Minor Accidents Involving Commuter Cyclists in Belgium. *Accident Analysis & Prevention,* Vol. 45, pp. 683-693.

Ding, C., Chen, P., & Jiao, J. (2018). Non-linear effects of the built environment on automobile-involved pedestrian crash frequency: a machine learning approach. Accident Analysis & Prevention, 112, 116-126. https://doi.org/10.1016/j.aap.2017.12.026

Diogenes, M., & Lindau. L., (2010). Evaluation of Pedestrian Safety at Midblock Crossings, Porto Alegre, Brazil. *Transportation Research Record: Journal of the Transportation Research Board.* No. 2193, TRB, National Research Council, Washington D.C., pp. 37-43.

Dixon, K., Avelar, R., Brown, L., Mecham, M., & Van Schalkwyk, I. (2012). Quantifying Safety Performance of Driveways on State Highways Publication FHWA-OR-RD-13-02, FHWA, U.S Department of Transportation.

Doecke, S. D., Kloeden, C. N., Dutschke, J. K., & Baldock, M. R. (2018). Safe speed limits for a safe system: The relationship between speed limit and fatal crash rate for different crash types. Traffic Injury Prevention, 19(4), 404-408. https://doi.org/10.1080/15389588.2017.1422601

Dumbaugh, E., & Li, W. (2010). Designing for the safety of pedestrians, cyclists, and motorists in urban environments. Journal of the American Planning Association, 77(1), 69-88. https://doi.org/10.1080/01944363.2011.536101

Dumbaugh, E., Li, W., & Joh, K. (2013). The built environment and the incidence of pedestrian and cyclist crashes. Urban Design International, 18(3), 217-228. https://doi.org/10.1057/udi.2013.2

Elefteriadou, L., Torbic, D., El-Gindy, M., Stoffels, S., & Adolini, M. (2001). Rumble strips for roads with narrow or non-existent shoulders. Pennsylvania Department of Transportation. https://trid.trb.org/view/706708

Eluru, N., Bhat, C. R., & Hensher, D. A. (2008). A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accident Analysis & Prevention*, *40*(3), 1033-1054.

Elvik, R., & Bjørnskau, T. (2017). Safety-in-numbers: A systematic review and meta-analysis of evidence. Safety Science, 92, 274-282. https://doi.org/10.1016/j.ssci.2015.07.017

Ewing, R., & Cervero, R. (2001). Travel and the built environment: a synthesis. Transportation Research Record: Journal of the Transportation Research Board, (1780), 87-114.

Fitzpatrick, K., Iragavarapu, V., Brewer, M., Lord, D., Hudson, J. G., Avelar, R., & Robertson, J. (2014). *Characteristics of Texas pedestrian crashes and evaluation of driver yielding at pedestrian treatments* (No. FHWA/TX-13/0-6702-1). College Station: Texas A&M Transportation Institute, Texas A&M University System.

Fridman, L., Ling, R., Rothman, L., Cloutier, M. S., Macarthur, C., Hagel, B., & Howard, A. (2020). Effect of reducing the posted speed limit to 30 km per hour on pedestrian motor vehicle collisions in Toronto, Canada: A quasi experimental, pre-post study. BMC Public Health, 20(1), 1-8. https://doi.org/10.1186/s12889-019-8139-5

Furth, P. G., Mekuria, M. C., & Nixon, H. (2016). Network connectivity for low-stress bicycling. *Transportation research record*, *2587*(1), 41-49.

Garder, P. (1995). Rumble strips or not along wide shoulders designated for bicycle traffic? Transportation Research Record: Journal of the Transportation Research Board, 1502, 1-7. https://trid.trb.org/view/453065

Gill Jr., M. P. (2007). Analysis of bicycle-related and pedestrian-related roadway crashes. Colorado Department of Transportation. https://trid.trb.org/view/842584

Greibe, P. (2003). Accident Prediction Models for Urban Roads. In Accident Analysis & Prevention, Vol. 35 No. 2, pp. 273-285.

Gross, F., Jovanis, P. P., & Eccles, K. (2009). Safety effectiveness of lane and shoulder width combinations on rural, two-lane, undivided roads. Transportation Research Record: Journal of the Transportation Research Board, 2103(1), 42-49. https://doi.org/10.3141/2103-06

Guo, Y., Li, Z., Wu, Y., & Xu, C. (2018). Evaluating factors affecting electric bike users' registration of license plate in China using Bayesian approach. Transportation Research Part F: Traffic Psychology and Behaviour, 59, 212-221. https://doi.org/10.1016/j.trf.2018.09.008

Haleem, K., & Abdel-Aty, M. (2010). Examining traffic crash injury severity at unsignalized intersections. *Journal of safety research*, *41*(4), 347-357.

Harkey, D. L., Tsai, S., Thomas, L., & Hunter, W. W. (2006). Pedestrian and bicycle crash analysis tool (PBCAT): Version 2.0 application manual. Federal Highway Administration. https://rosap.ntl.bts.gov/view/dot/38707

Harris, M. A., Reynolds, C. C., Winters, M., Cripton, P. A., Shen, H., Chipman, M. L., ... & Teschke, K. (2013). Comparing the effects of infrastructure on bicycling injury at intersections and non-intersections using a case–crossover design. Injury Prevention, 19(5), 303-310. http://doi.org/10.1136/injuryprev-2012-040561

Harwood, D. W., Gilmore, D. K., Torbic, D. J., Souleyrette, R., & Hans, Z. N. (2008). United States Road Assessment Program: Pilot Program: Phase II.

Hels, T., & Orozova-Bekkevold, I. (2007). The effect of roundabout design features on cyclist accident rate. Accident Analysis & Prevention, 39(2), 300-307. https://doi.org/10.1016/j.aap.2006.07.008

Huang, H., Zhou, H., Wang, J., Chang, F., & Ma, M. (2017). A multivariate spatial model of crash frequency by transportation modes for urban intersections. Analytic Methods in Accident Research, 14, 10-21. https://doi.org/10.1016/j.amar.2017.01.001

Hunter, W., Stewart, J., & Stutts, J. (1999). Study of bicycle lanes versus wide curb lanes. Transportation Research Record: Journal of the Transportation Research Board, 1674(1), 70-77. https://doi.org/10.3141/1674-10

Jang, K., Park, S. H., Kang, S., Song, K. H., Kang, S., & Chung, S. (2013). Evaluation of pedestrian safety: pedestrian crash hot spots and risk factors for injury severity. *Transportation research record*, *2393*(1), 104-116.

Jacobsen, P. L. (2003). Safety in numbers: more walkers and bicyclists, safer walking and bicycling. Injury Prevention, 21(4), 271-275. http://doi.org/10.1136/ip.9.3.205rep

Johnson, E., Geyer, J. A., Rai, N., & Ragland, D. R. (2004). Low income childhood pedestrian injury: Understanding the disparate risk. UC Berkeley. https://escholarship.org/uc/item/68q8v836

Kaplan, S., & Prato, C. G. (2015). A spatial analysis of land use and network effects on frequency and severity of cyclist–motorist crashes in the Copenhagen region. Traffic Injury Prevention, 16(7), 724-731. https://doi.org/10.1080/15389588.2014.1003818

Karlaftis, M. G., & Golias, I. (2002). Effects of road geometry and traffic volumes on rural roadway accident rates. Accident Analysis & Prevention, 34(3), 357-365. https://doi.org/10.1016/S0001-4575(01)00033-1

Kashani, A. T., & Mohaymany, A. S. (2011). Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Safety Science, 49(10), 1314-1320. https://doi.org/10.1016/j.ssci.2011.04.019

Kim, J. K., Kim, S., Ulfarsson, G. F., & Porrello, L. A. (2007). Bicyclist injury severities in bicycle–motor vehicle accidents. *Accident Analysis & Prevention*, *39*(2), 238-251.

Kim, S., & Ulfarsson, G. F. (2019). Traffic safety in an aging society: analysis of older pedestrian crashes. *Journal of transportation safety & security*, *11*(3), 323-332.

Klop, J. R., & Khattak, A. J. (1999). Factors influencing bicycle crash severity on two-lane, undivided roadways in North Carolina. Transportation Research Record: Journal of the Transportation Research Board, 1674(1), 78-85. https://doi.org/10.3141/1674-11

Kröyer, H. R. (2015). The Relation Between Speed Environment, Age and Injury Outcome for Bicyclists Struck by a Motorized Vehicle – a Comparison with Pedestrians. In Accident Analysis & Prevention, Vol. 76, 2015, pp. 57-63.

Ladron de Guevara, F., Washington, S. P., & Oh, J. (2004). Forecasting crashes at the planning level: simultaneous negative binomial crash model applied in Tucson, Arizona. Transportation Research Record, 1897(1), 191-199.

Lee, C., & Abdel-Aty, M. (2005). Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida. *Accident Analysis & Prevention*, *37*(4), 775-786.

Lee, J., Abdel-Aty, M., & Cai, Q. (2017). Intersection crash prediction modeling with macro-level data from various geographic units. Accident Analysis & Prevention, 102, 213-226. https://doi.org/10.1016/j.aap.2017.03.009

Lee, J., Abdel-Aty, M., & Jiang, X. (2015). Multivariate crash modeling for motor vehicle and non-motorized modes at the macroscopic level. Accident Analysis & Prevention, 78, 146-154. https://doi.org/10.1016/j.aap.2015.03.003

Loukaitou-Sideris, A., Liggett, R., & Sung, H. G. (2007). Death on the crosswalk: A study of pedestrian-automobile collisions in Los Angeles. *Journal of Planning Education and Research*, *26*(3), 338-351.

Lowry, M.B., & Cool, S., (2015). Dangerous Situation Topology and Analysis for Comparing Bicycle Infrastructure Improvement Plans. Paper. Presented at the 94th Annual Meeting of the Transportation Research Board. Paper 15-5989.

Luan, S., Li, M., Li, X., & Ma, X. (2020). Effects of built environment on bicycle wrong Way riding behavior: A data-driven approach. Accident Analysis & Prevention, 144, 105613.

Lyons, R. A., Towner, E., Christie, N., Kendrick, D., Jones, S. J., Hayes, M., ... & Phillips, C. (2008). The advocacy in action study a cluster randomized controlled trial to reduce pedestrian injuries in deprived communities. Injury Prevention, 14(2), e1-e1. http://doi.org/10.1136/ip.2007.017632

Maring, W., & Van Schagen, I. (1990). Age dependence of attitudes and knowledge in cyclists. *Accident Analysis & Prevention*, *22*(2), 127-136.

Marshall, W. E., & Garrick, N. W. (2011). Evidence on why bike-friendly cities are safer for all road users. Environmental Practice, 13(1), 16-27. https://doi.org/10.1017/S1466046610000566

Martin, A. (2006). *Factors influencing pedestrian safety: a literature review* (No. PPR241). Wokingham, Berks: TRL.

Martinez, R., & Veloz, R. A. (1996). A challenge in injury prevention: The Hispanic population. Academic Emergency Medicine, 3(3), 194-197. https://doi.org/10.1111/j.1553-2712.1996.tb03418.x

McDaniel, S., Lowry, M. B., & Dixon, M. (2014). Using origin–destination centrality to estimate directional bicycle volumes. *Transportation Research Record, 2430*(1), 12-19.

McMahon, P., Duncan, C., Stewart, J., Zegeer, C., & Khattak, A. (1999). Analysis of factors contributing to" walking along roadway" crashes. *Transportation Research Record*, *1674*(1), 41-48.
McMahon, P., Duncan, C., Stewart, J., Zegeer, C., & Khattak, A. (1999). Analysis of factors contributing to" walking along roadway" crashes. *Transportation Research Record*, *1674*(1), 41-48.
Metroplan Orlando. (2010). Orlando area bicyclist crash study: A role-based approach to crash countermeasures. Metroplan Orlando. https://www.pedbikeinfo.org/resources/resources_details.cfm?id=3695

Miranda-Moreno, L. F., Morency, P., & El-Geneidy, A. M. (2011). The link between built environment, pedestrian activity and pedestrian–vehicle collision occurrence at signalized intersections. *Accident Analysis & Prevention*, *43*(5), 1624-1634.

Mohan, D., Khayesi, M., Tiwari, G., & Nafukho, F. M. (2006). *Road traffic injury prevention training manual*. World Health Organization.

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

Møller, M., & Hels, T. (2008). Cyclists’ perception of risk in roundabouts. Accident Analysis & Prevention, 40(3), 1055-1062. https://doi.org/10.1016/j.aap.2007.10.013

Morrison, C. N., Thompson, J., Kondo, M. C., & Beck, B. (2019). On-road bicycle lane types, roadway characteristics, and risks for bicycle crashes. Accident Analysis & Prevention, 123, 123-131. https://doi.org/10.1016/j.aap.2018.11.017

National Oceanic and Atmospheric Administration (2019). https://www.noaa.gov

Moudon, A. V., Lin, L., Jiao, J., Hurvitz, P., & Reeves, P. (2011). The risk of pedestrian injury and fatality in collisions with motor vehicles, a social ecological study of state routes and city streets in King County, Washington. Accident Analysis & Prevention, 43(1), 11-24. https://doi.org/10.1016/j.aap.2009.12.008

Narayanamoorthy, S., Paleti, R., & Bhat, C. R. (2013). On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. Transportation Research Part B: Methodological, 55, 245-264. https://doi.org/10.1016/j.trb.2013.07.004

Noland, R. B., & Quddus, M. A. (2004). Analysis of pedestrian and bicycle casualties with regional panel data. *Transportation research record*, *1897*(1), 28-33.

Nordback, K., Marshall, W. E., & Janson, B. N. (2014). Bicyclist safety performance functions for a US city. *Accident Analysis & Prevention*, *65*, 114-122.

Numetric. (2019). Numetric. Utah Department of Transportation. https://udot.numetric.com/signinO’Brien, S. W., Findley, D. J., Jackson, K. N., Holzem, A., & Vosburgh, E. (2014). Rumble strip gaps for high speed bicycles on downgrades. North Carolina Department of Transportation. https://rosap.ntl.bts.gov/view/dot/28335

Oh, J., Jun, J., Kim, E., & Kim, M. (2008). *Assessing critical factors associated with bicycle collisions at urban signalized intersections* (No. 08-0832).

Olszewski, P., Szagała, P., Wolański, M., & Zielińska, A. (2015). Pedestrian fatality risk in accidents at unsignalized zebra crosswalks in Poland. Accident Analysis & Prevention, 84, 83-91. https://doi.org/10.1016/j.aap.2015.08.008

Olkkonen, S., & Honkanen, R. (1990). The role of alcohol in nonfatal bicycle injuries. *Accident Analysis & Prevention*, *22*(1), 89-96.

Pai, C. W. (2011). Overtaking, rear-end, and door crashes involving bicycles: An empirical investigation. *Accident Analysis & Prevention*, *43*(3), 1228-1235.

Palamara, P., & Broughton, M. (2013). An investigation of pedestrian crashes at traffic intersections in the Perth Central Business. *regulation*, *5*, 169-175.

Park, J., Abdel-Aty, M., Lee, J., & Lee, C. (2015). Developing crash modification functions to assess safety effects of adding bike lanes for urban arterials with different roadway and socio-economic characteristics. Accident Analysis & Prevention, 74, 179-191. https://doi.org/10.1016/j.aap.2014.10.024

Pedroso, F. E., Angriman, F., Bellows, A. L., & Taylor, K. (2016). Bicycle use and cyclist safety following Boston's bicycle infrastructure expansion, 2009–2012. American Journal of Public Health, 106(12), 2171-2177. https://doi.org/10.2105/AJPH.2016.303454

Petritsch, T. A., Landis, B. W., McLeod, P. S., Huang, H. F., Challa, S., & Guttenplan, M. (2005). Level-of-service model for pedestrians at signalized intersections. Transportation Research Record: Journal of the Transportation Research Board, 1939(1), 54-62. https://doi.org/10.1177/0361198105193900107

Petritsch, T. A., Landis, B. W., Huang, H. F., & Challa, S. (2006). Sidepath safety model: bicycle sidepath design factors affecting crash rates. *Transportation research record*, *1982*(1), 194-201.

Poch, M., & Mannering, F. (1996). Negative binomial analysis of intersection-accident frequencies. *Journal of transportation engineering*, *122*(2), 105-113.

Pokorny, P., & Pitera, K. (2019). Truck-bicycle safety: an overview of methods of study, risk factors and research needs. European Transport Research Review, 11(1), 1-14. https://doi.org/10.1186/s12544-019-0371-7

Poudel, N., & Singleton, P. A. (2021). Bicycle safety at roundabouts: A systematic literature review. Transport Reviews, 41(5), 617-642. https://doi.org/10.1080/01441647.2021.1877207

Prati, G., Marín Puchades, V., De Angelis, M., Fraboni, F., & Pietrantoni, L. (2018). Factors contributing to bicycle–motorised vehicle collisions: A systematic literature review. Transport Reviews, 38(2), 184-208. https://doi.org/10.1080/01441647.2017.1314391

Prato, C. G., Kaplan, S., Rasmussen, T. K., & Hels, T. (2016). Infrastructure and spatial effects on the frequency of cyclist-motorist collisions in the Copenhagen Region. Journal of Transportation Safety & Security, 8(4), 346-360. https://doi.org/10.1080/19439962.2015.1055414

Preston, H., Storm, R., Dowds, J. B., Wemple, B., Hill, C., & Systematics, C. (2013). *Systemic safety project selection tool* (No. FHWA-SA-13-019). United States. Federal Highway Administration. Office of Safety.

Pucher, J., & Buehler, R. (2016). Safer cycling through improved infrastructure. American Journal of Public Health, 106(12), 2089-2091. https://doi.org/10.2105/AJPH.2016.303507

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

Pulugurtha, S. S., & Thakur, V. (2015). Evaluating the effectiveness of on-street bicycle lane and assessing risk to bicyclists in Charlotte, North Carolina. Accident Analysis & Prevention, 76, 34-41. https://doi.org/10.1016/j.aap.2014.12.020

Quddus, M. A. (2008). Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data. Accident Analysis & Prevention, 40(4), 1486-1497. https://doi.org/10.1016/j.aap.2008.03.009

Raihan, M. A., Alluri, P., Wu, W., & Gan, A. (2019). Estimation of bicycle crash modification factors (CMFs) on urban facilities using zero inflated negative binomial models. Accident Analysis & Prevention, 123, 303-313. https://doi.org/10.1016/j.aap.2018.12.009

Reynolds, C. C., Harris, M. A., Teschke, K., Cripton, P. A., & Winters, M. (2009). The impact of transportation infrastructure on bicycling injuries and crashes: A review of the literature. Environmental Health, 8(1), 1-19. https://doi.org/10.1186/1476-069X-8-47

Rodgers, G. B. (1995). Bicyclist deaths and fatality risk patterns. *Accident Analysis & Prevention*, *27*(2), 215-223.

Romanow, N. T., Couperthwaite, A. B., McCormack, G. R., Nettel-Aguirre, A., Rowe, B. H., & Hagel, B. E. (2012). Environmental determinants of bicycling injuries in Alberta, Canada. Journal of Environmental and Public Health, 2012, 487681. https://doi.org/10.1155/2012/487681

Roll, J. (2018). Bicycle count data: What is it good for? A study of bicycle travel activity in Central Lane Metropolitan Planning Organization. Oregon Department of Transportation. https://rosap.ntl.bts.gov/view/dot/36255

Roudsari, B. S., Mock, C. N., Kaufman, R., Grossman, D., Henary, B. Y., & Crandall, J. (2004). Pedestrian crashes: Higher injury severity and mortality rate for light truck vehicles compared with passenger vehicles. Injury Prevention, 10(3), 154-158. http://doi.org/10.1136/ip.2003.003814

Saad, M., Abdel-Aty, M., Lee, J., & Cai, Q. (2019). Bicycle safety analysis at intersections from crowdsourced data. Transportation Research Record: Journal of the Transportation Research Board, 2673(4), 1-14. https://doi.org/10.1177/0361198119836764

Sadek, A. W., Dickason, A. S., & Kaplan, J. (2007). Effectiveness of green, high-visibility bike lane and crossing treatment. Presented at the 86th Annual Meeting of the Transportation Research Board, Washington, DC. https://trid.trb.org/view/801418

Saha, D., Alluri, P., & Gan, A. (2015). Prioritizing Highway Safety Manual's crash prediction variables using boosted regression trees. Accident Analysis & Prevention, 79, 133-144. https://doi.org/10.1016/j.aap.2015.03.011

Saha, D., Alluri, P., Gan, A., & Wu, W. (2018). Spatial analysis of macro-level bicycle crashes using the class of conditional autoregressive models. Accident Analysis & Prevention, 118, 166-177. https://doi.org/10.1016/j.aap.2018.02.014

Sanders, R. L., Frackelton, A., Gardner, S., Schneider, R., & Hintze, M. (2017). Ballpark method for estimating pedestrian and bicyclist exposure in Seattle, Washington: Potential option for resource-constrained cities in an age of big data. Transportation Research Record: Journal of the Transportation Research Board, 2605(1), 32-44. https://doi.org/10.3141/2605-03

Sando, T., & Moses, R. (2011). Operational and safety impacts of restriping inside lanes of urban multilane curbed roadways to 11 feet or less to create wider outside curb lanes for bicyclists. Florida Department of Transportation. https://rosap.ntl.bts.gov/view/dot/23404

Schepers, J. P., Kroeze, P. A., Sweers, W., & Wüst, J. C. (2011). Road factors and bicycle–motor vehicle crashes at unsignalized priority intersections. *Accident Analysis & Prevention*, *43*(3), 853-861.

Schepers, P., & den Brinker, B. (2011). What do cyclists need to see to avoid single-bicycle crashes?. *Ergonomics*, *54*(4), 315-327.

Schepers, P., Hagenzieker, M., Methorst, R., van Wee, B., & Wegman, F. (2014). A conceptual framework for road safety and mobility applied to cycling safety. *Accident Analysis & Prevention*, *62*, 331-340.

Schimek, P. (2018). Bike lanes next to on-street parallel parking. Accident Analysis & Prevention, 120, 74-82. https://doi.org/10.1016/j.aap.2018.08.002

Schneider, R. J., Diogenes, M. C., Arnold, L. S., Attaset, V., Griswold, J., & Ragland, D. R. (2010). Association between roadway intersection characteristics and pedestrian crash risk in Alameda County, California. *Transportation Research Record*, *2198*(1), 41-51.

Schneider, R. J., Ryznar, R. M., & Khattak, A. J. (2004). An accident waiting to happen: a spatial approach to proactive pedestrian planning. *Accident Analysis & Prevention*, *36*(2), 193-211.

Sener, I. N., Lee, K., Hudson, J. G., Martin, M., & Dai, B. (2021). The challenge of safe and active transportation: Macrolevel examination of pedestrian and bicycle crashes in the Austin District. Journal of Transportation Safety & Security, 13(5), 525-551.

Senserrick, T., Boufous, S., De Rome, L., Ivers, R., & Stevenson, M. (2014). Detailed analysis of pedestrian casualty collisions in Victoria, Australia. *Traffic injury prevention*, *15*(sup1), S197-S205.

Shah, N. R., Aryal, S., Wen, Y., & Cherry, C. R. (2021). Comparison of motor vehicle-involved e-scooter and bicycle crashes using standardized crash typology. Journal of Safety Research, 77, 217-228. https://doi.org/10.1016/j.jsr.2021.03.005

Shirani-Bidabadi, N., Mallipaddi, N., Haleem, K., & Anderson, M. (2020). Developing bicycle-vehicle crash-specific safety performance functions in Alabama using different techniques. Accident Analysis & Prevention, 146, 105735. https://doi.org/10.1016/j.aap.2020.105735

Siddiqui, C., Abdel-Aty, M., & Choi, K. (2012). Macroscopic spatial analysis of pedestrian and bicycle crashes. Accident Analysis & Prevention, 45, 382-391. https://doi.org/10.1016/j.aap.2011.08.003

Singleton, P. A., Totten, J. C., Orrego-Oñate, J. P., Schneider, R. J., & Clifton, K. J. (2018). Making strides: state of the practice of pedestrian forecasting in regional travel models. Transportation Research Record: Journal of the Transportation Research Board, 2672(35), 58-68. https://doi.org/10.1177/0361198118773555

Singleton, P. A., & Runa, F. (2021). Pedestrian traffic signal data accurately estimates pedestrian crossing volumes. Transportation Research Record: Journal of the Transportation Research Board, 2675(6), 429-440. https://doi.org/10.1177/0361198121994126

Singleton, P. A., Runa, F., & Humagain, P. (2020). Utilizing archived traffic signal performance measures for pedestrian planning and analysis. Utah Department of Transportation. https://rosap.ntl.bts.gov/view/dot/54924

Singleton, P. A., Mekker, M., & Islam, A. (2021). Safety in numbers? Developing improved safety predictive methods for pdestrian crashes at signalized intersections in Utah using push button-based measures of exposure. Utah Department of Transportation. https://rosap.ntl.bts.gov/view/dot/56362

Singleton, P. A., Park, K., & Lee, D. H. (2021). Utilizing ATSPM data for pedestrian planning and analysis – Phase II: Extending pedestrian volume estimation capabilities to unsignalized intersections. Utah Department of Transportation.

Spring, G. S. (2003). Shoulder rumble strips in Missouri. Missouri Department of Transportation. https://rosap.ntl.bts.gov/view/dot/37983

Stone, J. R., Chae, K., & Pillalamarri, S. (2002). *The effects of roundabouts on pedestrian safety*. Southeastern Transportation Center.

STRAVA. About Us. Available on: https://www.STRAVA.com/about. Retrieved on 2020-11-2.

STRAVA. Global Heatmap. Available on: http://labs.STRAVA.com/heatmap/#5/-103.99658/38.44498/blue/bike. Retrieved on 2020-11-2.

Strauss, J., Miranda-Moreno, L. F., & Morency, P. (2013). Cyclist activity and injury risk analysis at signalized intersections: A Bayesian modelling approach. Accident Analysis & Prevention, 59, 9-17. https://doi.org/10.1016/j.aap.2013.04.037

Stamatiadis, N., Pigman, J. G., Sacksteder, J., Ruff, W., & Lord, D. (2009). Impact of shoulder width and median width on safety. National Cooperative Highway Research Program. https://doi.org/10.17226/14252

Summala, H., Pasanen, E., Räsänen, M., & Sievänen, J. (1996). Bicycle accidents and drivers' visual search at left and right turns. *Accident Analysis & Prevention*, *28*(2), 147-153.

Sze, N. N., & Wong, S. C. (2007). Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. Accident Analysis & Prevention, 39(6), 1267-1278. https://doi.org/10.1016/j.aap.2007.03.017

Taquechel, E. (2009). A spatial analysis of the relationship between pedestrian crash events and features of the built environment in downtown Atlanta (master's thesis). Georgia State University. https://scholarworks.gsu.edu/iph_theses/117

Team, R. C. (2013). R: A language and environment for statistical computing. Teschke, K., Harris, M. A., Reynolds, C. C., Winters, M., Babul, S., Chipman, M., ... & Cripton, P. A. (2012). Route infrastructure and the risk of injuries to bicyclists: a case-crossover study. American Journal of Public Health, 102(12), 2336-2343. https://doi.org/10.2105/AJPH.2012.300762

Toole, J. (2010). Update of the AASHTO Guide for the Planning, Design, and Operation of Pedestrian Facilities. Transportation Research Board. https://onlinepubs.trb.org/onlinepubs/nchrp/docs/NCHRP20-07(263)_FR.pdf

Tin, S. T., Woodward, A., & Ameratunga, S. (2013). Incidence, risk, and protective factors of bicycle crashes: Findings from a prospective cohort study in New Zealand. *Preventive medicine*, *57*(3), 152-161.

Toran Pour, A., Moridpour, S., Tay, R., & Rajabifard, A. (2018). Influence of pedestrian age and gender on spatial and temporal distribution of pedestrian crashes. Traffic injury prevention, 19(1), 81-87.

Torbic, D., El-Gindy, M., Elefteriadou, L., & Jiang, Z. (2003). Bicycle-friendly shoulder rumble strips. International Journal of Vehicle Design, 33(4), 440-466. https://www.inderscienceonline.com/doi/abs/10.1504/IJVD.2003.003575

Torbic, D. J., Harwood, D. W., Bokenkroger, C. D., Srinivasan, R., Carter, D., Zegeer, C. V., & Lyon, C. (2010). Pedestrian safety prediction methodology for urban signalized intersections. *Transportation research record*, *2198*(1), 65-74.

Ukkusuri, S., Hasan, S., & Aziz, H. A. (2011). Random parameter model used to explain effects of built-environment characteristics on pedestrian crash frequency. Transportation Research Record: Journal of the Transportation Research Board, 2237(1), 98-106. https://doi.org/10.3141/2237-11

Ukkusuri, S., Miranda-Moreno, L. F., Ramadurai, G., & Isa-Tavarez, J. (2012). The role of built environment on pedestrian crash frequency. Safety Science, 50(4), 1141-1151. https://doi.org/10.1016/j.ssci.2011.09.012

Utah Department of Transportation (UDOT). (2016). Available on: https://gis.utah.gov/data/transportation/roads-system/ Access on 2020-10-12

UDOT Data Portal (2019). https://data-uplan.opendata.arcgis.com/

United States Environment Protection Agency (EPA), 2019. https://www.epa.gov/smartgrowth/smart-location-mapping

Vandenbulcke, G., Thomas, I., & Panis, L. I. (2014). Predicting cycling accident risk in Brussels: a spatial case–control approach. *Accident Analysis & Prevention*, *62*, 341-357.

Walker, I. (2007). Drivers overtaking bicyclists: Objective data on the effects of riding position, helmet use, vehicle type and apparent gender. *Accident Analysis & Prevention*, *39*(2), 417-425.

Walker, S. H., & Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, *54*(1-2), 167-179.

Wang, X., Yang, J., Lee, C., Ji, Z., & You, S. (2016). Macro-level safety analysis of pedestrian crashes in Shanghai, China. Accident Analysis & Prevention, 96, 12-21. https://doi.org/10.1016/j.aap.2016.07.028

Wang, J., Huang, H., & Zeng, Q. (2017). The effect of zonal factors in estimating crash risks by transportation modes: Motor vehicle, bicycle and pedestrian. Accident Analysis & Prevention, 98, 223-231. https://doi.org/10.1016/j.aap.2016.10.018

Wang, Y., & Kockelman, K. M. (2013). A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. Accident Analysis & Prevention, 60, 71-84. https://doi.org/10.1016/j.aap.2013.07.030

Wang, Y., & Nihan, N. L. (2004). Estimating the risk of collisions between bicycles and motor vehicles at signalized intersections. *Accident Analysis & Prevention*, *36*(3), 313-321.

Wanvik, P. O. (2009). Effects of road lighting: An analysis based on Dutch accident statistics 1987–2006. Accident Analysis & Prevention, 41(1), 123-128. https://doi.org/10.1016/j.aap.2008.10.003

Wei, V. F., & Lovegrove, G. (2012). Sustainable road safety: A new (?) neighbourhood road pattern that saves VRU lives. Accident Analysis & Prevention, 44(1), 140-148.

Wier, M., Weintraub, J., Humphreys, E. H., Seto, E., & Bhatia, R. (2009). An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. *Accident Analysis & Prevention*, *41*(1), 137-145.

Yang, X., Huan, M., Abdel-Aty, M., Peng, Y., & Gao, Z. (2015). A hazard-based duration model for analyzing crossing behavior of cyclists and electric bike riders at signalized intersections. Accident Analysis & Prevention, 74, 33-41. https://doi.org/10.1016/j.aap.2014.10.014

Zahabi, S. A. H., Strauss, J., Manaugh, K., & Miranda-Moreno, L. F. (2011). Estimating potential effect of speed limits, built environment, and other factors on severity of pedestrian and cyclist injuries in crashes. Transportation Research Record: Journal of the Transportation Research Board, 2247(1), 81-90. https://doi.org/10.3141/2247-10

Zajac, S. S., & Ivan, J. N. (2003). Factors influencing injury severity of motor vehicle–crossing pedestrian crashes in rural Connecticut. Accident Analysis & Prevention, 35(3), 369-379.

Zebauers, V. (2005). Centerline rumble strips help cars and bicyclists. In ITE 2005 Annual Meeting and Exhibit Compendium of Technical Papers. Institute of Transportation Engineers. https://trid.trb.org/view/762422

Zegeer, C. V., Richard Stewart, J., Huang, H., & Lagerwey, P. (2001). Safety effects of marked versus unmarked crosswalks at uncontrolled locations: analysis of pedestrian crashes in 30 cities. *Transportation research record*, *1773*(1), 56-68.

Zegeer, C. V., Carter, D. L., Hunter, W. W., Stewart, J. R., Huang, H., Do, A., & Sandt, L. (2006). Index for assessing pedestrian safety at intersections. *Transportation research record*, *1982*(1), 76-83.

Zhai, X., Huang, H., Sze, N. N., Song, Z., & Hon, K. K. (2019). Diagnostic analysis of the effects of weather condition on pedestrian crash severity. Accident Analysis & Prevention, 122, 318-324.

Zhu, S. (2021). Analysis of the severity of vehicle-bicycle crashes with data mining techniques. Journal of Safety Research, 76, 218-227. https://doi.org/10.1016/j.jsr.2020.11.011