A SYSTEMIC PEDESTRIAN SAFETY PLANNING TOOL

FOR RURAL AND SMALL URBAN AREAS

by

Amir Jamali

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

of

Doctor of Philosophy

in

Civil Engineering

MONTANA STATE UNIVERSITY
Bozeman, Montana

July 2018

ACKNOWLEDGEMENTS

I would first like to express the deepest appreciation to my advisor Dr. Yiyi Wang, who has the attitude and the substance of a genius. She continually and convincingly conveyed a spirit of adventure regarding research and scholarship. I could not have imagined having a better advisor for my Ph.D. study. Without her guidance and persistent help this dissertation would not have been possible.

I would also like to thank my parents and elder brothers, whose love and guidance are with me in whatever I pursue. I am indebted to my parents for inculcating in me the dedication and discipline to do whatever I undertake well.

Finally, I am thankful of the FHWA UTC’s Small Urban and Rural Livability Center (SURLC) for the financial support.

TABLE OF CONTENTS

1. INTRODUCTION 1

Overview and Motivation 1
Systemic Safety Tools.....3
Study Objectives and Organization 6

2. LITERATURE REVIEW 8

Pedestrian Exposure Metrics.....8
Area-Based Approach.....11
Population (Density) 11
Walking Distance.....12
Walking Duration.....12
Number of Walk Trips 13
Point-Based Approach 14
Segment-Based Approach.....15
Distance-Based Approach.....15
Trip-Based Approach.....16
Crash Risk Factors 17
Pedestrian Characteristics 26
Driver 26
Roadway 27
Vehicle 28
Crash 28
Traffic 29
Socio-economic.....30
Built-Environment 30
Natural-Environment 31
Hotspot Locations Identification.....32
GIS Spatial Tools 33
Crash Rate Indicators 34
Empirical Bayesian (EB) 36
Project Prioritization 36
Before/After Analysis 37
Cost/Benefit Analysis 38
Multi-Criteria Decision-Making (MCDM) Method 39
Optimization Technique.....40

3. DATA SETS 42

NHTS Data Set 42

TABLE OF CONTENTS COUNTINUED

Crash Severity Data Set 45
Crash Frequency Data Set....................52

4. METHODOLOGIES 56

Initial Screening....................56
Pedestrian Exposure Estimation 59
Statistical Models....................59
Negative Binomial (NB) Model....................59
Zero-Inflated Negative Binomial (ZINB) Model 60
Finite Mixture of Negative Binomial Regression (FMNB) Model 61
Machine Learning Techniques....................62
K-Nearest Neighbors (K-NN) Technique....................62
Crash Risk Factors Identification....................63
Crash Severity Analysis....................64
Ordered Logit Regression (OR) Model 64
Ordered Probit Regression (OP) Model 65
Multinomial Logit (MNL) Model....................65
Multinomial Probit (MNP) Model....................66
Random Forest (RF) Technique....................66
Naïve Bayes Technique 67
Artificial Neural Network (ANN) Technique....................68
Crash Frequency Analysis 69
Hurdle Negative Binomial (HNB) Model....................69
Boosted Random Forest (BRF) Model 70
Hotspot Locations Identification....................72
Two-Step Floating Catchment Area (2SFCA) Method 72
Weight Matrix....................74
Severity Index 75
Countermeasure Selection 77
Project Prioritization....................77

5. RESULTS AND ANALYSIS....................80

Pedestrian Exposure Estimation 80
Statistical Models....................80
Machine Learning Techniques....................84
Crash Risk Factors Identification....................89
Crash Severity Analysis....................89
Multicollinearity 89
Statistical Models....................92
Machine Learning Techniques....................100

TABLE OF CONTENTS COUNTINUED

Crash Frequency Analysis 105
Multicollinearity 106
Statistical Models..107
Machine Learning Models 112
Hotspot Location Identification 114
Countermeasure Selection 119
Project Prioritization 121

6. CONCLUSION..130

Pedestrian Exposure Estimation 134
Crash Risk Factors Identification..137
Hotspot Locations Identification..140
Project Prioritization 141

REFERENCES CITED..143

APPENDIX A: Data Dictionary 165

LIST OF TABLES

Table Page

1. Pedestrian systemic safety process. 6

2. Summary of pedestrian exposure metrics and estimation methods. 9

3. Summary of previous research studies analyzed pedestrian crash severities. .. 18

4. Summary statistics of NHTS data (No. of observations = 58,227 households). 44

5. Severity levels definition (source: Texas CRIS data set)................................. 46

6. Descriptive statistics of crash severity factors. 47

7. Statistic summary of crash frequency data (No. of observations = 15,288 intersections). 54

8. Summary of methodologies in pedestrian systemic safety tool. 57

9. Pedestrian crashes and sites (where at least one pedestrian crash occurred) in Montana, Oregon, and Texas from 2011 to 2013. 58

10. Average comprehensive costs by injury severity... 77

11. NB, ZINB, and FMNB models of HB walking trips. 82

12. Walk trip frequency in rural and small urban areas. 85

13. Result summary of K-NN classification models. 87

14. GVIF measurement of crash severity factors... 90

15. GVIF measurement of crash severity factors after removing highly correlated factors... 91

16. The OL and MNL models of pedestrian crash severity. 93

17. The ordered Probit and multinomial Probit models of pedestrian crash severity. 95

LIST OF TABLES CONTINUED

Table Page

18. Result summary of OL, MNL, RF, Naïve Bayes, and ANN models (A=non-injury, B=injury, C=fatal). 103

19. GVIF measurement of crash frequency factors. 106

20. GVIF measurement of crash frequency factors after removing highly correlated factors. 106

21. The NB, ZINB, and Hurdle NB models of pedestrian crash frequency (# of observations = 15,288 intersections).. 109

22. Result Summary of NB, ZINB, HNB, RF, and BRF Models (A=0 crash, B=1 crash, C=2+ crashes)... 113

23. Proposed countermeasures for severity and risk factors................................. 120

24. Sensitivity analysis on cost constraint. 123

25. Selected countermeasures for hotspot locations. 125

26. Site characteristics data obtained for intersections... 166

LIST OF FIGURES

Figure Page

1. Three-layer neural network (Moghaddam et al., 2011). 68

2. Variable importance score plot from random forest technique. 101

3. Overall prediction accuracy percentage and weighted prediction accuracy percentage measures. 105

4. Hotspot locations map across Bozeman, MT. 117

5. Hotspot location - zone A (W College St. – Grant Chamberlain Dr.) 118

6. Hotspot location - zone B (E Main St. – Black Ave.).................................... 118

7. Hotspot location - zone C (N 19th Ave. – W Oak St.) 118

8. Optimization model's performance for different budget levels..................... 123

ABSTRACT

Rural areas bear a disproportionate number of pedestrian fatalities: the fatality rate per 100 million vehicle miles traveled is 2.5 times higher in rural areas than in urban areas. To measurably improve pedestrian safety, it is paramount to predict crash hot spots and apply cost-effective countermeasures. This dissertation work developed a new systemic pedestrian safety tool to enhance crash hotspot identification and safety project prioritization for rural and small urban areas. This new tool suggested a six-step systemic safety framework: (1) initial screening, which identifies what type of facilities are more prone to pedestrian crashes, (2) pedestrian exposure estimation, which provides an area-level exposure metric using National Household Travel Survey (NHTS) 2009, (3) crash risk factor identification, which identifies the factors that contribute to the occurrence and high severity levels of pedestrian crashes, (4) hotspot identification, which identifies the locations that are more likely to experience pedestrian crashes using two-step floating catchment area (2SFCA) method, (5) countermeasure selection, which provides candidate countermeasures through literature sources, and (6) project prioritization, which ranks safety projects through a mixed linear programming. This study incorporated three states' pedestrian crash data from 2011 to 2013: Texas, Oregon, and Montana. It was found that in rural and small urban areas pedestrian safety is associated negatively with male and elderly drivers, shoulder presence, bike lane presence, higher speed limit, number of lanes, wet surface, pedestrian exposure, hospital distance, population density, median income, share of industrial and commercial areas, and dark hours. In contrast, the pedestrian safety is associated positively with signal control, sidewalk and warning sign presence, median presence, icy and snowy surface, higher AADT, and high densely household areas. To validate the proposed hotspot and project prioritization methods, this study used the pedestrian crash data set from 2014 to 2016 in the City of Bozeman, a small urban area. According to findings, about 60 percent of crash locations fall on areas with a high crash risk index. Reasonable countermeasures were suggested for twenty intersections with highest crash risk index. It was found that budget of $100,000 is the optimal budget, where the crash risk index was reduced by 63 percent.

CHAPTER ONE

INTRODUCTION

Overview and Motivation

Walking has been recognized as an important transportation mode (especially for short distance trips) because it can improve physical and mental health while mitigating environmental issues associated with motorized travel (Gårder 2004; Islam and Jones 2014; Frank et al. 2006; Smith et al. 2008; Sallis et al. 2009; Warburton et al. 2006; Bauman et al. 2012). However, many individuals do not show interest in walking for their daily trips; in the United States, for example, 85 percent of residents do not walk for utilitarian purposes on a daily basis (Agrawal and Schimek 2007). One of the main reasons may be pedestrian safety, which remains a key issue and deters people from adopting this active travel mode. In 2015, a pedestrian was killed every 1.6 hours in the United States, and another was injured every 7.5 minutes in a traffic crash (NHTSA 2015). Pedestrian safety is even worse in rural and small urban areas. The fatality rate for pedestrians per 100 million vehicle miles traveled is 2.5 times higher in rural areas than in urban areas (1.88 in rural and 0.73 in urban)(NHTSA 2015).

For urban areas, numerous studies have investigated what traffic and roadway attributes contribute to pedestrian crashes and injury severities (J. K. Kim et al. 2008; Anne Vernez Moudon et al. 2011; Dai 2012; Abay 2013; Toran Pour et al. 2016). In addition, there has been considerable research to develop crash prediction models (Chang 2005; Miranda-Moreno et al. 2011; Wang and Kockelman 2013; Chen and Zhou 2016) and to create tools that identify countermeasures like the National Cooperative Highway Research Program (NCHRP) 500 Reports (Zegeer et al. 2004) and FHWA crash modification factor (CMF) Clearinghouse online tool. However, research studies on pedestrian safety that focus on rural and small urban areas have been sparse. This indicates a critical gap that needs to be bridged in order to fully understand what factors drive pedestrian crashes and how to improve pedestrian safety in rural and small urban settings.

There are several reasons why rural pedestrian safety needs to be studied in a different way than urban pedestrian safety. First, data details regarding traffic and land use are often unavailable for rural (and in some cases small urban) areas, which makes traffic operation and safety analysis more challenging (Jamali and Wang 2017). Moreover, rural areas have distinct roadway characteristics that cannot be captured from urban-based studies: for example, shoulders often substitute for sidewalks, and speed limits are usually higher in rural areas (Yan et al. 2012). Second, rural pedestrian crashes are infrequent and sporadic events compared to urban pedestrian crashes (Zajac and Ivan 2003), which challenges safety analysis such as estimating crash frequency, injury severity, and hotspot identification. Last but not least, travel behavior and activities are different in rural areas, which could change the level of pedestrian exposure to vehicular traffic (Hough et al., 2008; Millward and Spinney 2011). In sum, rural settings present unique challenges related to data availability, built environment, and travel behaviors, which require a context-sensitive study like this dissertation work.

Systemic Safety Tools

Site-specific analysis is a common approach in safety studies; it uses only crash frequency to identify hot spot locations for safety improvement. However, it overlooks the underlying risks of pedestrian-vehicle crashes that stem from the rare and random nature of these crashes. In rural and small urban areas, pedestrian crashes become more sporadic and scatter across a large area due to low exposure volume (pedestrian and vehicle volume). This approach is insufficient for rural areas, because although the percentage of fatalities is high, the crash density is typically low. This low crash density prevents the accurate identification of contributing factors and risky locations (Preston et al. 2013). In addition, existing research in the literature has not calibrated any safety performance functions (SPF) for rural and small urban areas.

A systemic safety approach, which is a cost-effective solution, can be substituted for site-specific analysis. The systemic approach is a step by step process that begins by identifying focus crash types (e.g., pedestrian crashes) and associated risk factors, and then evaluates the entire road system with a set of criteria (i.e., primary risk factors) to identify high risk locations. In fact, this approach uses the risk factors and crash history to identify locations for potential safety improvements (Preston et al. 2013). It then recommends cost-effective measures for those candidate locations and concludes by prioritizing locations for implementation.

As an example, the U.S. Department of Transportation Federal Highway Administration (FHWA) adopted a systemic safety planning approach to identify roadway safety problems (Preston et al. 2013). It includes four main steps: (1) identify focus crash type and risk factors, (2) screen and prioritize candidate locations, (3) select countermeasures, and (4) prioritize projects. Step 1 determines the risk factors using descriptive statistics and the known characteristics identified from published research. Step 2 determines the list of locations that are candidates for safety improvements; locations are prioritized based on the number of risk factors present. Step 3 selects the most appropriate countermeasures for each risk factor based on expected crash reductions and estimated implementation and maintenance cost. Finally, step 4 prioritizes the projects by assessing and accounting for available funding, implementation time, amount of public outreach needed, and environmental constraints.

The American Association of State Highway and Transportation Officials (AASHTO) also developed a systemic safety tool, which involves six steps (AASHTO 2010): (1) network screening, (2) diagnosis, (3) countermeasure selection, (4) economic appraisal, (5) project prioritization, and (6) safety effectiveness evaluation. In the network screening step, the network is reviewed to identify type of facility being screened, which helps to determine the appropriate data needs and analytical methods. The diagnosis step itself consists of three levels including: (a) safety data review, which screens the crash dataset and identifies crash locations, severity levels, and environmental conditions to identify the crash patterns; (b) supporting documentation review, which reviews the literature to determine opportunities and constraints; and (c) field condition review, which involves visiting the sites to observe present transportation services and conditions. The economic appraisal step assesses the benefits of safety improvements as compared to the implementation cost through a B/C ratio or cost-effectiveness analysis. For project prioritization, the AASHTO tool recommends the use of three main methods: B/C ratio, cost-effectiveness analysis, and optimization. An optimization method such as linear programming is preferred due to its ability to account for factors like budget constraints. Finally, the evaluation step assesses how safety improvements may influence the crash frequency and severity at a specific location.

Finally, the Texas Department of Transportation (TxDOT) developed a four-step systemic safety tool (Walden et al. 2015). The first three steps are identifying target crash type and risk factors, screening candidate locations, and selecting countermeasures, which are similar to steps in the previous safety tools. However, in step four, a decision process is utilized for prioritizing the projects. The decision tree is created based on factors such as traffic volume, environment, and adjacent land use. For each candidate location, the decision tree is applied to identify the most appropriate countermeasure. Then, the projects are prioritized with consideration of funding, time, and amount of public outreach needed.

Based on this literature, this study adopts a six-step systemic safety planning tool to customize the process for pedestrian safety in small urban and rural settings, as shown in Table 1.

Table 1. Pedestrian systemic safety process.

Initial Screening	Reviewing the archived crash datasets to identify the type of sites (e.g., intersections, mid-block segments) or type of road facilities (e.g., highways, local roads) to be screened.
Pedestrian Exposure Estimation	Utilizing statistical and machine learning methods to estimate pedestrian exposure, which refers to a pedestrian's contact with vehicular traffic.
Risk Factor Identification	Utilizing statistical and machine learning methods to determine which factors pertain to or contribute to pedestrian injury levels or crash frequency.
Hotspot Location Identifcication	Screening the network for high-risk locations using the identified risk factors after controlling for pedestrian and traffic exposure.
Countermeasure Selection	Providing a comprehensive list of countermeasures to address the risk factors associated with pedestrian collisions.
Project Prioritization	Ranking safety improvement projects for hotspot locations through a mixed linear program.

Study Objectives and Organization

This study develops a systemic pedestrian safety planning tool to address multiple safety concerns in rural and small urban areas. The main tasks of this study are:

- to identify the hotspot locations and contributing factors associated with pedestrian injury severity levels, and
- to develop an optimization toolbox to screen the road networks for hotspots and to recommend countermeasures.

The proposed systemic safety planning tool aims to diagnose high-risk locations for pedestrian crashes in rural and small urban areas and to make effective and appropriate recommendations on safety measures. In addition, the tool is scalable with anticipated results pertinent to the livability goals at the federal and state level—ranging from measuring safety performance and identifying problems, to strategy development, and to integration with state-wide transportation planning.

The remainder of the dissertation is organized as follows: Chapter 2 (Literature Review) synthesizes the methodologies employed in pedestrian exposure estimation, crash risk factor identification, hotspot location identification, and project prioritization. Chapter 3 (Data Sets) presents the data used and provides summary statistics. Chapter 4 (Methodology) describes the proposed six-step systemic safety tool, with details on each specific step. Chapter 5 (Analysis and Results) reports and interprets estimation outputs for each specific step. Chapter 6 (Conclusions) summarizes the work's key contributions and suggestions for future research efforts.

CHAPTER TWO

LITERATURE REVIEW

This chapter provides a synthesis of research studies in the field of pedestrian crash safety. The literature is divided into four major categories including: (1) pedestrian exposure metrics, (2) crash risk factors, (3) hotspot identification techniques, and (4) project prioritization methodologies.

Pedestrian Exposure Metrics

Estimation models become an attractive option to infer pedestrian exposure in the absence of field-observed data (e.g., pedestrian counts). Pedestrian count data are sparse in part due to data collection challenges such as costs (Qin and Ivan 2001) and complexity of pedestrian behavior (e.g., multiple stops along a walk trip, usage of informal routes like trails or footpaths, and sporadic path changing) (Kuzmyak et al. 2014). Given the large volume of work on pedestrian exposure estimation, but extremely limited information that compares these methods with regard to accuracy, computation, and transferability (how well the methods can be applied to another area), it would be valuable to synthesize this information, as provided in this section of the literature review.

Five general types of metrics are used as proxies for pedestrian exposure (Table 2): area-based metrics that measure pedestrian exposure (e.g. population density and number of walk trips) across zones like municipalities (Kerr et al. 2013) or census tracts (Loukaitou-Sideris et al. 2007; Cottrill and Thakuriah 2010); point-based metrics that note the exposure at specific locations like pedestrian crossings (Silcock et al. 1996; Zegeer et al. 2005); segment-based metrics that measure the exposure along roadway links (Molino et al. 2012; Clifton et al. 2008); distance-based metrics that measure walk miles traveled on facilities (Molino et al. 2012); and trip-based metrics that exploit trip characteristics such as choice of crossing locations and space-time prisms (Yao et al. 2015; Lassarre et al. 2007).

Table 2. Summary of pedestrian exposure metrics and estimation methods.

Metrics	Papers	Estimation method (Performance)	Data Source
		Area-based	
Population (Density)	Loukaitou-Sideris et al. (2007)	-	U.S. Census (2000) in 860 census tracts in Los Angeles
	Wier et al. (2009)	-	U.S. Census (2000) in 176 census tracts in San Francisco
	Cottrill & Thakuriah (2010)	-	U.S. Census (2000) in 6 counties in the Chicago metro area
	Chakravarthy et al (2010)	-	U.S. Census (2000) in 577 census tracts in a California county
Walking Distance	Jonah & Engel (1983)	-	Interview of 956 respondents in Ottawa, Canada
	Wang & Kochelman (2013)	Weighted least squares model (-)	Austin Travel Survey (2005/2006) in 218 zones in Austin, Texas
	McAndrews et al. (2013)	-	Add-on NHTS (2001) sample, aggregated in the state of Wisconsin
	Haddak (2016)	-	French National Travel Survey (2007-2008), national aggregate

Table 2 Continued.

Walking Duration	Chu (2003)	- (Overestimation of crash risk due to missing walk trips and over-reported travel time)	NHTS (2001), national aggregate
	McAndrews et al. (2013)	-	Add-on NHTS (2001) sample, aggregated in the state of Wisconsin
	Haddak (2016)	-	French National Travel Survey (2007-2008), national aggregate
Number of trips	Beck et al. (2007)	-	NHTS (2001), national aggregate
	Kerr et al. (2013)	-	NHTS (2009) and U.S. Census 2010 in 553 municipalities in North Carolina
	Haddak (2016)	-	French National Travel Survey (2007-2008), national aggregate
	McAndrews et al. (2013)	-	Add-on NHTS (2001) sample, aggregated in the state of Wisconsin
Point-Based			
Pedestrian volume	Ivan et al. (2000)	Poisson linear regression model (Low model transferability and limited combinations of site features)	Pedestrian count data at 32 intersections in rural Connecticut
	Zegeer et al. (2005)	Adjustment factors (-)	Pedestrian count at 2,000 crosswalk sites across 30 cities in US
Interaction between pedestrian volume & traffic volume	Cameron (1982)	-	Pedestrian count data at 4 crosswalk sites in New South Wales, Australia
	Tobey et al. (1983)	-	Pedestrian count data at 1,357 crosswalk sites in Brooklyn, St. Louis, Seattle, St. Petersburg (Florida), Maryland, and Washington D.C.

Table 2 Continued.

	Lyon & Persaud (2002)	-	Pedestrian count data at 1,069 intersections in Toronto, Canada
	Geyer et al. (2006)	Space Syntax (Underestimation)	Pedestrian count data at 274 intersections in Oakland, California
	Miranda-Moreno et al. (2011)	Negative binomial regression model (-)	Pedestrian count data at 519 intersections in Montreal, Canada
Segment-Based			
Pedestrian volume	Jonsson (2005)	Adjustment factors (-)	Pedestrian count data at 393 links in seven Swedish cities
	Clifton et al. (2008)	Four-step travel demand model (walk trips to/from transit were not considered)	NHTS (2001) in 1,709 pedestrian analysis zones (street blocks) in Maryland
Distance-Based			
Annual walk Distance	Molino et al. (2012)	Adjustment factors (Over-estimation due to over-representation of intersections)	Pedestrian count data at 122 sites (intersections and roadway segments) in Washington D.C.
Trip-Based			
Space-Time method	Lam et al. (2013), Lam et al. (2014), Yao et al. (2105)	Shortest path algorithm & space-time prism (Small-area application; Pedestrians are assumed to choose only the shortest path)	Travel Survey of 924 elderly people in Hong Kong
Discrete-Choice	Lassarre et al. (2007)	Nested logit model (Good estimation accuracy when compared with travel survey data)	1,870 pedestrian crossing decisions in Florence, Italy and 1,793 pedestrian crossing decisions in Athens, Greece

Area-Based Approach

Population (Density). Some studies employed population or population density as a proxy for pedestrian exposure (Loukaitou-Sideris et al. 2007; Cottrill and Thakuriah

2010). It is hypothesized that areas with more residents are associated with more walking (Cottrill and Thakuriah 2010). While population (density) is updated frequently by the U.S. Census Bureau (Greene-Roesel et al. 2007), it does not represent the number of people who walk, nor does it represent the distance or duration walked (Greene-Roesel et al. 2007; Lee and Abdel-Aty 2005).

Walking Distance. Others used walking distance to represent pedestrian exposure (Wang and Kockelman 2013; Jonah and Engel 1983; McAndrews et al. 2013). Generally, there are two ways to derive walking distance: from travel survey data directly (e.g., annual kilometers travelled [McAndrews et al. 2013; Haddak 2016]), or from travel survey data via statistical models (Wang and Kockelman 2013). McAndrews et al. (2013) estimated total walking distance for the entire state of Wisconsin by weighting an add-on sample of the NHTS (2001). Wang and Kockelman (2013) utilized the weighted least squares (WLS) regression model to estimate walk-miles traveled across 217 zones in Austin, Texas, using the Austin Household Travel Survey (2005/2006) (an add-on program to the National Household Travel Survey [NHTS]). The sample-reported walk distances were scaled up by a ratio of zone population to zone sample size to reflect population-level walk distances across zones.

Walking Duration. Critics of walking distance as a proxy for pedestrian exposure note that walking speed varies by age and gender, and hence, the distance walked does not reflect the level of exposure. For example, individuals who walk slowly are exposed to crash risk more than those who walk briskly (Lee and Abdel-Aty 2005).

Some researchers advocate for the use of trip duration because it accounts for different walking speeds. Trip duration is captured by the NHTS as self-reported travel time. Lee and Abdel-Aty (2005) claimed that survey respondents more accurately reflect on (and report) walk trip duration than they do with walk distance. In an attempt to estimate pedestrian fatality risk by age and gender, Chu (2003) used the self-reported travel time from the NHTS (2001) data as a proxy for pedestrian exposure and cited a potential pitfall as people tend to over-report trip duration for non-motorized modes in travel surveys (Chu 2003). In addition, the self-reported travel time as provided in NHTS data sets consists of time intervals during which a pedestrian is not exposed to vehicle traffic like walking upstairs (Molino et al. 2012).

Number of Walk Trips. Number of walk trips can also be used as an exposure metric for zones (Kerr et al. 2013; Beck et al. 2007; McAndrews et al. 2013; Haddak 2016). Trip frequency is regularly provided by travel survey data (Kerr et al. 2013; Beck et al. 2007) and can be used to develop statistical models based on zonal land use and socioeconomic attributes to make inferences on travel behavior (Greene-Roesel et al. 2013; Sabir et al. 2011; Montigny et al. 2012; Hatamzadeh et al. 2014).

In their safety analysis, Kerr et al. (2013) computed the number of walk trips by trip purpose across 553 municipalities in North Carolina. Four trip purposes were considered. Work trips were computed as the number of persons who walked to work, available from the American Community Survey [ACS] 5-year estimates in 2010. School trips were imputed by the percentage of children who walked to school from the Safe Routes to School 2010 Report and the number of children enrolled in school obtained from the US Census (2010). To compute the number of walk to/from college trips, they assumed that bike/walk-to-college trips shared the same mode split as bike/walk-to-work trips. Non-commute trips were estimated using the 2009 NHTS data.

Point-Based Approach

Point-based measurements note the intensity of pedestrian movements at point locations (e.g., intersections or mid-block locations). Zegeer et al. (2005) and Ivan et al. (2001) used pedestrian volume as an exposure term to analyze pedestrian safety at crossing locations. Ivan et al. (2001) estimated the weekly crossing pedestrian volume at rural intersections by controlling for site characteristics (e.g., sidewalk provision and traffic control type), median household income, area type (e.g., downtown area and residential area), and road attributes (e.g., number of lanes and lane width).

Raford and Regland (2004) used the Space Syntax technique to infer pedestrian volume at 730 intersections in Oakland, California. This technique used multivariate linear regression models to explain pedestrian volumes at observed locations while controlling for land use (e.g., population and employment density) and network configuration (including connectivity and accessibility). While this method utilizes readily available data from the U.S. Census (e.g., population and employment), it does require observed pedestrian counts and a well-defined pedestrian network. Raford and Ragland (2004) compared the predicted values with observed data across many unspecified sites in downtown Oakland. They found that observed pedestrian volumes were four times larger than the estimated values.

Segment-Based Approach

A segment-based approach is used to describe the exposure of pedestrians walking along streets. Clifton et al. (2008) developed a Model of Pedestrian Demand (MoPed) based on the four-step travel demand model (McNally 2007). The four steps include: (1) estimation of walk trips while controlling for land-use (e.g., percentage of commercial area), employment counts, and network features (e.g., street connectivity) across pedestrian analysis zones (PAZs); (2) trip distribution between origins and destinations using gravity models; (3) network assignment to allocate the trips between each origin-destination (O-D) pair onto road links via all-or-nothing method; and (4) mode split (which is not required because only one transportation mode [walking] is concerned). The MoPed model captures land use and network detail of the study area at fine geography (road segment), but can be difficult to transfer to a different region due to data constraints (Clifton et al. 2008).

Distance-based Approach

Molino et al. (2012) estimated annual walk distance using manual pedestrian count data from 122 sites that span eight facility types across the entire Washington, D.C. area during the fall of 2006 and the summer of 2007. The annual pedestrian exposure of each facility type was calculated by multiplying annual pedestrian volume (based on adjusted 15-minute pedestrian counts) with the average walking distance (approximated by the crossing distance, i.e., driveway width) for each facility type. As a result, annual walk distance was computed for intersections and road links totaling 80 million miles for the entire study area. To validate the method, they compared the estimated lump sum walk distance with the U.S. Census's QuickFacts data and found overestimation as a main drawback of this method, possibly caused by applying the same adjustment factors across all facility types and an overrepresentation of intersections in the sample.

Trip-Based Approach

Different from the approaches noted earlier, the trip-based approach further exploits trip characteristics in the estimation of pedestrian exposure, either by capturing the time lapse associated with a walk trip so that the amount of pedestrian exposure varies not only over space but also over different time intervals during a day (Yao et al. 2015; Lam et al. 2014; Lam et al. 2013) or by considering a pedestrian's choice of crossing locations when gauging the overall exposure to traffic during a trip (Lassarre et al. 2007). The unit of analysis for the trip-based approach can span point, segment, or area.

Lam et al. (2013) proposed a space-time (ST) approach to estimate the distance walked across 1,102 road segments. A spatiotemporal framework was developed to overlay pedestrian activities with crash occurrences to determine the amount of pedestrian exposure at the time of a crash. Pedestrian trajectories were inferred by the shortest path algorithm using the origin and destination data of 924 elderly people in the Hong Kong Travel Characteristics Survey (2002). Trip origins and destinations were recorded at street block level. Pedestrian exposure is defined for each ST slice as the product of the distance walked (proxied by the shortest path length) and crash frequency, summed over all pedestrian trips that occurred during that ST slice. The drawback of this approach is that it set pedestrian exposure to zero for segments where no crash was reported.

Lassarre et al. (2007) estimated pedestrian exposure for a generic crossing location along a pedestrian trip. The pedestrian exposure at a crossing location is defined as a multiplication of traffic volume, traffic speed, and number of traffic lanes, and the probability of that crossing being selected. They used the nested logit model to estimate the probability of each crossing opportunity being chosen while controlling for walk distance, traffic volume, and crossing distance.

Crash Risk Factors

This section reviews the existing litarature about factors that contribute to the occurrence of a specific level of injury severity in a traffic crash. Generally, there are nine categories of factors: pedestrian, motorized vehicle driver, motorized vehicle, roadway, crash, traffic, socio-economic, built environment, and natural environment characteristics. Table 3 summarizes the most frequently cited papers that were published after 2000, which have contributed to the pedestrian-vehicle crash analysis.

Table 3. Summary of previous research studies analyzed pedestrian crash severities.

Paper	Data Size	Significant Variables								
		Pedestrian	Driver	Roadway	Vehicle	Crash	Traffic	Socio-economic	Built Environment	Natural Environment
Descriptive Analysis										
Lefler & Gabler 2004	543	-*	-	-	type	impact speed	-	-	-	-
Oikawa & Matsui 2017	5,134	gender	-	-	type	impact speed	-	-	-	-
Linear Regression										
Noland & Quddus 2004	1,122,691	inebriety	inebriety	-	-	-	-	population, income	percent of local roads	-
Clifton & Kreamer-Fults 2007	1,513	-	-	-	-	-	-	population density, population, race	land use type, transit access, driveway presence, recreational facilities presence, commercial access	-
Binary Logistic Regression										
Roudsari et al. 2004	522	-	-	-	type	--	-	-	-	-

Table 3 Continued.

Ballesteros et al. 2004	16,838	-	-	-	type	impact speed	-	-	-	-
Roudsari et al. 2006	255	age	-	control devices	type	impact speed, pedestrian and vehicle movement	-	-	-	-
Sze & Wong 2007	73,746	age, gender	-	-	-	-	speed limit, volume	-	-	time of day
Poudel-Tandukar et al. 2007	1,557	traffic violation	-	-	-	-	-	-	-	-
Kim et al. 2008	2,275	age	age, gender, inebriety	control devices, traffic sign, functional class, sloped, curve, median type	type	speed-involved, location, vehicle movement, pedestrian and driver at fault		-	land use type	weather, time of day
Rosén & Sander 2009	2,127	age	-	-	-	impact speed	-	-	-	-

Table 3 Continued.

Ma et al. 2010	851	age, gender, income, education, family type	-	-	-	-	-	-	-	-
Moudon et al. 2011	711	age, gender, inebriety	-	-	-	# of pedestrian involved, location, vehicle movement	speed limit, volume	-	median home value, # of residential units, # of restaurants	-
Sarkar et al. 2011	4,976	age	-	control devices	type	pedestrian movement	-	-	-	weather
Dai 2012	8,403	age, gender, inebriety	gender, inebriety	lighting	-	-	-	-	setting	time of day, day of week, season
Tefft 2013	315	age	-	-	type	impact speed	-	-	-	-
Zhang et al. 2014	6,976	age	gender, inebriety, driving license	control devices, functional class, lighting	type	pedestrian at fault, vehicle movement	-	-	-	time of day, day of week
Bennet & Yiannakoulias 2015	199	age	-	control devices	-	location	flow	-	land use type	-

Table 3 Continued.

Verzosa & Miles 2016	7,628	age, gender	-	surface, # of lanes	type	-	-	-	land use type, transit access	time of day
Ordered Logit/Probit Regression										
Zajac & Ivan 2003	264	age, inebriety	inebriety	width	type	-	-	-	land use type	-
Lee & Abdel-Aty 2005	7,000	age, gender, inebriety	-	control devices, lighting	type	impact speed	-	-	setting	weather
Siddiqui et al. 2006	160,119	age, gender, inebriety	age, gender, inebriety	lighting	type	location	-	-	-	weather, time of day
Zahabi et al. 2011	5,820	-	-	functional class	type	location, vehicle movement	-	-	land use type, transit access, park presence	time of day
Tarko & Azam 2011	9,453	age, gender	-	-	type	location	-	-	setting	-
Khattak 2012	440	gender	-	traffic sign, # of lanes	-	impact speed	-	-	-	-

Table 3 Continued.

Obeng & Rokonuzzaman 2013	211	gender	-	sight obstruction	type	-	speed limit, volume	-	land use type, sidewalk presence	-
Jang et al. 2013	4,939	age, inebriety, cell phone use	-	-	type	vehicle movement	-	-	-	weather, time of day, day of week
Yu 2015	1,407	age	-	control devices, lighting	-	location, vehicle movement	speed limit	population density,	land use type, sidewalk density, intersection density, transit access, school and park usage	weather
Kim et al. 2017	137,470	age, gender	gender, inebriety	width	type	pedestrian movement	-	population density, percentage of elderly residents	# of doctors	weather, time of day
Generalized Order Probit/Logit										
Eluru et al. 2008	1,223	age, gender, inebriety	inebriety	control devices	type	vehicle movement	speed limit	-	-	weather, time of day
Clifton et al. 2009	4,500	age, gender, inebriety, cloth, traffic violation	-	-	type	-	-	-	pedestrian connectivity, transit access	weather
Latent Segmentation Based Ordered Logit										

Table 3 Continued.

Yasmin et al. 2014	7,354	age	-	lighting	type	-	-	--	-	weather, season
Partial Proportional Odds (PPO)										
Sasidharan & Menéndez 2014	12,766	age, gender	-	lighting	-	pedestrian movement	-	-	-	-
Pour-Rouholamin & Zhou 2016	19,361	age, cloth	age, inebriety	control devices, median type, # of lanes, lighting	type	location	-	population	setting	weather, time of day, season
Multinomial Logit (MNL)										
J. K. Kim et al. 2008	5,808	age	age, gender, inebriety	control devices, lighting, functional class, traffic sign, curved	type	speed-involved, pedestrian and vehicle movement, and pedestrian and vehicle at fault	-	-	land use type	weather, time of day
Tay et al. 2011	45,201	age, gender	age, gender, inebriety	width, functional class	type	location	speed limit	-	-	weather, time of day

Table 3 Continued.

Rifaat et al. 2011	2,249	inebriety	age, gender, inebriety	control devices, median type, surface	-	location	-	-	street pattern	time of day, day of week, season
Kwigizile et al. 2011	13,106	age	careless driving	functional class, shoulder type, lighting	type	impact lane, location, vehicle movement	speed limit	-	land use type	weather, time of day, day of week
Kröyer 2014	8166	age	-	-	type	impact speed	-	-	-	-
Mixed Logit Model										
Kim et al. 2010	5,808	age	age, inebriety	control devices, lighting functional class, traffic sign	type	impact speed, location	-	-	land use type	weather, time of day
Aziz et al. 2013	7,354	age	-	control devices, # of lanes, surface, sloped, lighting	type	location, vehicle movement	speed limit	-	land use type	-
Abay 2013	4,952	age, gender, inebriety	age, inebriety, crime history	lighting	type	location, vehicle and pedestrian movement and at fault	speed limit	-	land use type	time of day

Table 3 Continued.

Haleem et al. 2015	7,630	age	-	control devices, lighting	type	pedestrian movement and at fault	speed limit, volume, % truck volume	-	-	weather
Clustering Regression										
Mohamed et al. 2013	5,820 and 6,896	age	age, inebriety	control devices, lighting, functional class, bike lane and metered parking presence	type	-	-	-	land use type, transit access	-
Sasidharan. et al. 2015	9,659	inebriety, traffic violation	Familiar with route	sight obstruction	type	location, pedestrian movement	speed limit	-	setting	time of day, day of week, season
Decision Tree										
Toran Pour et al. 2016	5,346	age, gender	-	# of lanes sloped, width	type	-	speed limit, volume	population density	transit access	-

Pedestrian Characteristics

Age, gender, and inebriety are the most common factors cited for pedestrian injury severity (Dai 2012; Moudon et al. 2005). Elderly pedestrians were more likely to sustain severe injuries than middle-aged or young pedestrians due to longer reaction time, physical weakness, and medical conditions (Moudon et al. 2011; Verzosa and Miles 2016; Clifton et al., 2009; Tay et al. 2011), while pedestrians younger than 15 years bore a higher fatality rate when compared to other population cohorts because they are less responsive to the risk and more susceptible to the impacts (Sarkar et al., 2011; Jang et al. 2013). Intoxicated pedestrians were prone to severe injuries in traffic crashes (Sasidharan et al., 2015; Jang et al. 2013; Eluru et al. 2008) due to their reckless behaviors and longer reaction time (Jang et al. 2013; Eluru et al. 2008). However, mixed results were found about the effect of pedestrian gender on severity levels. Some studies found that females were linked to an increase in fatality risk (Lee and Abdel-Aty 2005; Siddiqui et al.,2006; Tay et al. 2011), some studies pointed to the opposite (Dai 2012; Abay 2013), while others did not find any significant correlation between gender and severity (Moudon et al. 2011; Zhang et al., 2014; J. K. Kim et al. 2010). These mixed results are attributed to the difficulties of controlling for pedestrian behaviors under various conditions (Tay et al. 2011; Abay 2013).

Driver

Age, gender, and intoxication of the drivers also significantly affect the injury levels of pedestrians involved in traffic crashes. The drivers over the age of 65 were found to be associated with lower injury levels (J. K. Kim et al. 2008; Tay et al. 2011;

Mohamed et al. 2013) since older drivers typically drive more cautiously and mostly during daytime (Tay et al. 2011). In contrast, Rifaat et al. (2011) found that older drivers increased the probability of fatality risk of the pedestrian, due to longer reaction time. Although there is less existing research about driver gender than driver age, some studies found that male drivers were more likely to engage in crashes that caused severe pedestrian injuries (Kim et al., 2017; Dai 2012). In addition, intoxicated driving was a primary factor contributing to pedestrian injury and deaths (Kim et al., 2017; Pour-Rouholamin and Zhou 2016).

Roadway

The most significant factors noted for pedestrian injury are roadway functional class and lighting condition. Freeways and highways, which have higher average speed limits than local roadways, are associated with an increase in the probability of fatal injury (Yasmin et al., 2014; Tay et al. 2011; Kim et al. 2010). A number of studies have also found that absence of street lighting increased the probability of severe and fatal injuries (Siddiqui et al., 2006; Abay 2013; Arvin et al., 2017; Mohamed et al. 2013) .

Generally, the presence of traffic control devices (e.g., signal and stop sign) and traffic signs decrease the fatality risk, because drivers reduce vehicle speed and drive more cautiously when approaching the signs (Rifaat et al. 2011; Sarkar et al. 2011; Yu 2015). However, Kim et al. (2008) found that the presence of traffic signs is associated with higher levels of severe injuries and fatalities. Their findings suggest that both pedestrians and drivers are confused about who has the right-of-way at locations with a traffic sign (Kim et al. 2008).

Other roadway attributes such as sloped roadways (Kim et al. 2008; Aziz et al. 2013; Toran Pour et al. 2016), wider roadways (Tay et al. 2011; Toran Pour et al. 2016; Kim et al. 2017), multilane roadways (Aziz et al. 2013; Mohamed et al. 2013; Pour-Rouholamin and Zhou 2016), divided roadways (Rifaat et al. 2011; Pour-Rouholamin and Zhou 2016), and sight obstruction (Sasidharan et al. 2015) are associated with higher fatality risk. However, wet surfaces (Aziz, Ukkusuri, and Hasan 2013; Verzosa and Miles 2016), curved segments (Kim et al. 2008), bike lane and metered parking presence (Mohamed et al. 2013), and shoulders with a curb and gutter (Kwigizile et al. 2011) cause fewer fatal injuries.

Vehicle

Vehicle type emerged as the primary vehicle factor for pedestrian injury severities. Pedestrians are more likely to suffer severe injury and fatality in crashes that involve heavy vehicles (e.g., truck and bus) due to heavier mass and larger impact area (Lee and Abdel-Aty 2005; Kim et al., 2017; Zahabi et al. 2011; Eluru et al., 2008; Pour-Rouholamin and Zhou 2016). For example, Pour-Rouholamin and Zhou (2016) found that SUVs and buses increase the likelihood of serious injury to pedestrians by 8.6% and 22.9%, respectively.

Crash

Crash location (i.e., intersection versus midblock) is the most studied crash characteristic. Siddiqui et al. (2006), Zahabi et al. (2011), and Mohamed et al. (2013) found that pedestrians experienced less severe injuries at intersections because drivers traveled more slowly and more cautiously when approaching intersections, whereas Abay

(2013) concluded that intersection crashes led to more serious injuries since intersections represented a complex segment of traffic networks. Furthermore, some studies have investigated how vehicle movement (e.g., turning, forwarding, and backing) and pedestrian movement (e.g., walking along and crossing roadway) at time of crash impact pedestrian injury levels, but their findings were inconsistent (Roudsari et al. 2004; Kim et al., 2008; Moudon et al. 2011; Zhang et al., 2014; Yu 2015). For example, Yu (2015) suggested that pedestrians hit by turning vehicles were less likely to sustain fatal injuries than pedestrians hit by vehicles moving forward, but Kim et al. (2008) found that turning or backing vehicles caused more pedestrian fatalities than other types of vehicle movement did. In addition, Sze and Wong (2007), Tay et al. (2011), and Kim et al. (2017) found pedestrians who were crossing the roadway were more prone to fatal injuries as compared to pedestrians who were walking along the roadway, because crosswalks are points of conflict for pedestrians and vehicular traffic, and impact speed is also higher (Tay et al. 2011). In conclusion, the findings of the literature are inconsistent regarding crash attributes, which might be due to random nature of crash occurrence.

Traffic

Traffic volume is the most common traffic characteristic considered for pedestrian severity analysis. Since traffic volume data at the time of crash are often unavailable, the literature has used a series of proxies such as the average annual daily traffic (AADT) (Haleem et al., 2015; Moudon et al. 2011; Obeng and Rokonuzzaman 2013), average daily traffic (ADT) (Toran Pour et al. 2016), and congestion level (i.e., moderate versus severe [Sze and Wong 2007]). The majority of the studies concluded that higher traffic volume contributed to more serious injury of pedestrians (Toran Pour et al. 2016; Haleem et al., 2015; Bennet and Yiannakoulias 2015). However, Moudon et al. (2011) found that higher annual average daily traffic (AADT) tends to damper the risk of severe injury on state routes but increases the risk of severe injury along city streets, possibly due to lower speeds on state routes (which are often congested) and higher pedestrian exposure along city streets (Moudon et al. 2011).

Socio-economic

The literature also examined the effects of the socio-economic characteristics on the pedestrian severity analysis. Clifton and Kreamer-Fults (2007) and Yu (2015) indicated that population density is positively associated with severity levels because of higher pedestrian activity and more traffic volume. Conversely, Pour-Rouholamin and Zhou (2016), and Kim et al. (2017) found pedestrians are less likely to experience severe injuries in densely populated areas, possibly due to the lower traffic speed in such areas.

Built-Environment

The literature has considered three built environment attributes including transit access, land use (e.g., percentage of a certain land use type), and area setting (whether the area is rural, suburban, or urban).

With regard to transit accessibility, the literature found transit stop density (Zahabi et al. 2011), proximity to transit stops (Mohamed et al. 2013), and distance from crash location to the transit stops (Toran Pour et al. 2016) are associated positively with severe injuries and fatalities. These findings are expected as bus stops generate pedestrian movements and can increase the chance of conflict between pedestrians and vehicular traffic.

With regard to land use types, a high percentage of commercial land use was estimated to increase the risk of severe and fatal injuries, because the interaction between pedestrians and vehicles is more complex and pedestrians also might be more distracted in commercial areas. (Aziz et al., 2013; Clifton and Kreamer-Fults 2007). However, Yu's (2015) finding differed from previous studies, mainly because motorists usually drive at lower speeds in commercial areas.

With respect to area setting, crashes that occurred in rural areas appear to be more dangerous for pedestrians than those that occurred in urban areas (Lee and Abdel-Aty 2005; Sasidharan et al., 2015; Tarko and Azam 2011; Pour-Rouholamin and Zhou 2016). This might be due to higher vehicular traffic speed and fewer medical facilities in rural areas (Lee and Abdel-Aty 2005; Pour-Rouholamin and Zhou 2016).

Natural Environment

The literature has considered an array of variables such as weather conditions, time of the day, day of the week, and season (at the time of crash), but reached no consensus with regard to their effects on crash severity levels. The true linkage may rely on a wide range of confounding factors such as speed and driver behavior (Kim et al., 2017). For example, bad weather conditions (e.g., foggy, rainy, and snowy) were found to result in more severe and fatal injuries since low visibility affects drivers' ability to stop in time (Lee and Abdel-Aty 2005; Mohamed et al. 2013). However, the same variable appeared to be negatively associated with injury severity as drivers may opt to reduce speed and act more cautiously when weather conditions are poor (Yasmin et al., 2014; Yu 2015; Kim et al. 2008).

Hotspot Locations Identifcication

There are many studies focused on crash hotspot identification, but most of them targeted motor vehicle crashes. Therefore, this section summarizes the common hotspot identification methods regardless of pedestrian involvement.

The simplest approach is to rank sites based on their crash frequency (Da Costa et al. 2015), which assumes equal weights regardless of crash severity. An improved version is a severity index, which assigns larger weights to more severe crashes, hence prioritizing locations that have more potential for injuries and fatalities (Pulugurtha et al. 2007). However, the index overlooks key factors such as traffic exposure and spatial dependence (e.g., latent risks like visibility and alcohol consumption, which often trend in space and can affect neighboring zones).

To control for the key factors mentioned above, previous studies used three main approaches, including GIS spatial tools (e.g., Getis-Ord Gi* and kernel density [Soltani and Askari 2017; Kuo et al. 2013]), crash rate indicators (Clifton et al. 2008; Kocatepe et al. 2017), and Empirical Bayesian (EB) approaches (Montella 2010; Elvik 2008). Spatial dependence can arise among neighboring locations because risk factors or conditions (e.g., inadequate sight distance or fog) often trend in space and therefore affect crash risk at neighboring locations. Crash rate indicators rank locations according to crash risk, which is defined as crash frequency per exposure unit. The EB approach accounts for expected number of crashes in addition to observed crash count, since crash frequency gives a biased estimate of the long-term crash counts as crash counts fluctuate within the observation period (Montella 2010).

GIS Spatial Tools

The Getis-Ord Gi* statistic is a very common tool to determine hotspot locations (Truong and Somenahalli 2011; Prasannakumar et al. 2011; Kingham et al. 2011; Soltani and Askari 2017). This method calculates a Gi statistic based on an attribute value (e.g., crash counts across zones) and a spatial weight matrix (for each zone, or spatial unit). A zone with a high Gi statistic means that the zone has a high concentration of crashes and is surrounded by other units with high values, denoting a statistically significant hotspot. Truong and Somenahalli (2011) used a severity index rather than absolute crash frequency to calculate a Gi statistic for pedestrian crash hot spot identification in Adelaide, Australia. Accordingly, values of 3, 1.8, 1.3, and 1 were applied to weigh fatal, serious injury, other injury, and property-damage only crashes, respectively. These values indicate that fatal, serious, and other injury crashes are 3, 1.8, and 1.3 times more important than property-damage only crashes.

The Kernel Density Estimation (KDE) is another widely-used GIS tool for hotspot identification (Kuo et al. 2013; Prasannakumar et al. 2011; Xie and Yan 2008). It provides a continuous surface map of risky locations and calculates the crash density around each given location according to the distances from the location to the crash points (Kuo et al. 2013). A spatial unit with a greater density estimate of larger than a given threshold is identified as a crash hotspot (Yu et al. 2014). Manepalli et al. (2011) compared KDE with the Getis-Ord Gi* method using crash data from seven years (2000-

2006) on I-630 in Arkansas. They found both approaches provided the same result, which is consistent with the findings from Flahaut et al. (2003). In addition, Yu et al. (2014) compared the different hotspot identification methods (e.g., crash frequency, crash rate, and KDE) using data from a 622.2 kilometer section on the A1 highway in the UK from 2001 to 2010 and found that the KDE outperformed the other methods.

Crash Rate Indicators

As noted earlier, GIS tools only consider spatial variation, but ignore other contributing factors like traffic characteristics (e.g., pedestrian volume [Clifton et al. 2008; Raford et al. 2003], traffic volume [Murat 2011]), and geometric and physical attributes (e.g., lane width, horizontal curve, and pavement type [Murat 2011]). In a study to define unsafe locations in a Maryland pedestrian network, Clifton et al. (2008) normalized crash counts with pedestrian volume, which was estimated by a Model of Pedestrian Demand (MoPed) (a four-step travel demand model tailored to pedestrian volume at link level). Murat (2011) examined an entropy approach that combined accident rate (proportion of crash count at a given location relative to the total crashes) with traffic volume, average speed, and geometric attributes to rank 11 crash intersections. The entropy value of the larger area (e.g., block group and specified buffer) was equal to the average entropy values of the hotspots that fall within that area. The locations with larger entropy values were considered as safer locations. The study indicated that the entropy approach result was statistically significant using the chi-square test.

Innovatively, Kocatepe et al. (2017) used the Gaussian-based two-step floating catchment area (2SFCA) method to estimate crash propensity while accounting for spatial heterogeneity and population (as a proxy for crash exposure). The 2SFCA method falls under the gravity models, which stem from accessibility principles that are used to measure how accessible a service is to a population center (Luo and Qi 2009). In particular, this method measures the availability of a health service to the population, and has numerous applications in the medical field (Luo and Qi 2009; Yang et al. 2006; Radke and Mu 2000).

Kocatepe et al. (2017) defined the crash proneness as a reverse version of accessibility, stating "the closer a specific area is to a crash, the more crash risk that area will have." First, they determined the crash hotspots using Getis-Ord Gi* statistics, based on 2008 through 2012 crash data from an area with mixed rural and urban development in the Tampa Bay, Florida region. Second, the 2SFCA method measured the crash proneness by a crash-to-weighted population ratio, which accounted for Euclidean distance to weigh population centers around hotspot locations. In this way, the 2SFCA considered both spatial heterogeneity and crash exposure simultaneously. However, Kocatepe et al. (2017) have overlooked some important issues. First, population does not necessarily represent traffic exposure, because not all residents in a block group walk or drive. Second, a block group is a large spatial unit for the crash proneness analysis because the influence area of crash locations is limited to neighboring blocks. Moreover, they used both Getis-Ord Gi* and 2SFCA together, which leads to a complex methodology.

Empirical Bayesian (EB)

The EB approach combines observed crash frequency with expected (model-predicted) crash frequency to more accurately identify hotspots. The EB method outperforms other methods thanks to its ability to mitigate random variation of crash occurrence (Qu and Meng 2014; Montella 2010; Cheng and Washington 2008; Elvik 2008). For instance, Montella (2010) compared the EB method with three alternative methods, including crash frequency, crash rate, and equivalent crash counts (converted into property damage only crashes), using data from a 135 kilometers long highway in Italy. Montella found that the EB approach outperformed the other methods based on four tests, including the site consistency test, the method consistency test, the total rank differences test, and the total score test. These tests mostly measure the methods' performance to recognize hotspot locations over repeated periods. One limitation associated with the EB appraoch is that it requires the calibration of a crash prediction model using high-quality crash data, often unavailable for rural and small urban cases.

Project Prioritization

After identifying candidate safety improvements, it is necessary to estimate the safety impacts and monetary aspects of those projects. The safety impact refers to effects of projects on crash frequency and severity levels, and monetary aspects include implementation and maintenance costs. The most effective countermeasures lead to the greatest reduction in crash frequency or severity level with the lowest costs. Generally, research studies in the literature have utilized four main approaches to select the most effective safety projects: (1) before/after analysis, (2) cost/benefit (C/B) analysis, (3) multi-criteria decision-making (MCDM) method, and (4) optimization technique.

Before/After Analysis

Before/after analysis compares safety measures before and after the implementation of given countermeasures to assess how countermeasures improve safety on the traffic network. After countermeasure implementation, if a significant increase in safety is experienced, it demonstrates the countermeasure's efficiency. In this regard, Chen et al., (2013) applied a two-group pretest–posttest design to evaluate the effectiveness of 13 safety countermeasures implemented in New York City from 1990 to 2008. The two-group pretest–posttest design, as its name suggests, considers two groups of locations: untreated and treated locations. The changes in crashes between pre- and post-treatment periods for two groups are compared. If a treated group experienced a greater reduction in crashes than an untreated group, the countermeasures would be considered as effective.

However, since pedestrian crashes are rare events and usually under-reported, Fitzpatrick et al. (2011) employed a before/after analysis that used behavioral surrogates rather than crash data to estimate the effectiveness of countermeasures. For example, for evaluating the effects of rectangular rapid-flashing beacons, they calculated the percentage of drivers who did and did not yield to pedestrians before and after the safety improvement. Behavioral surrogates are also able to measure the countermeasure effectiveness in relatively shorter periods, which prevent impacts from changes in environmental factors such as land use (Fitzpatrick et al., 2011). Despite the advantages of before/after analysis such as control over data collection, this method needs a statistically sufficient sample of sites and a long period of time to assess the countermeasures, which limit its application.

Cost/Benefit (C/B) Analysis

Some studies used cost/benefit analysis, which compares the associated costs and benefits (Al-Kaisy et al., 2017; Yang et al., 2016). There are three common C/B analysis methods: 1) net present value, which converts all costs and benefits of an improvement into a single present value over the life cycle, 2) the equivalent uniform annual return, which converts the net present value into a single annual value, and 3) cost/benefit ratio, which calculates benefits value over costs value such that improvements with benefit/cost ratio of greater than one are considered as economically feasible (Yang et al., 2016). The equivalent uniform annual return method is generally preferred over benefit/cost ratio due to the difficulty of defining cost and benefit for calculation of the benefit/cost ratio (Yang et al., 2016).

In contrast to experimental studies, C/B analysis is a quick way to evaluate the countermeasures' efficiency. Since safety improvements might exhibit different performance under different environmental conditions, C/B analysis offers researchers and planners more opportunities to determine the improvements' efficiency under different conditions. In this regard, Al-Kaisy et al. (2017) examined the economic feasibity of 27 proven safety countermeasures along rural low-volume roadways in Oregon. They found that the most economically feasible improvements on rural roadways are low-cost countermeasures, which was expected due to low traffic exposure on such roadways.

Generally, economic analysis is useful when costs and benefits are measurable in monetary terms. In addition, the economic results are sensitive to cost and benefit assumptions, which increase the uncertainty of the results (Greene-Roesel et al. 2013).

Multi-Criteria Decision-Making (MCDM) Method

Multi-criteria decision-making methods rank the projects based on a quantitative ranking value that is obtained through the evaluation of some criteria by a group of decision makers (Awasthi and Chauhan 2011). For example, Yu and Liu (2012) selected three criteria to prioritize highway safety improvements. They selected safety concerns, which referred to total reduction in crashes; economic concerns, which referred to costs associated with project implementation and service life; and social importance costs, which attempt to quantify the amount of traffic that can benefit from implementation of that seecific project.

The most commonly used MCDM method for crash safety problems is the Analytic Hierarchy Process (AHP) (Yi et al., 2011; Awasthi and Chauhan 2011; Yedla and Shrestha 2003). AHP is a multi-criteria decision-making technique developed by (Saaty 1987) and has applications in sustainable energy sources, environmental science, and transportation safety (Yi et al., 2011; Awasthi and Chauhan 2011; Yedla and Shrestha 2003). The AHP decomposes multiple complex decision problems into smaller and more managable sub-problems. Basically, it structures a multi-criteria poblem into three hierarchical levels: (1) top level or goal, which defines the ultimate problem objective, (2) criteria level, which identifies the considered concerns or factors to compare the alternatives, and (3) alternatives level, which represents the decision alternatives.

The MCDM techniques benefit from a hierarchical structure that facilitates understanding the problem and the relationship between individual criteria. However, they are not able to consider constraints such as budget constraints, which could be viewed as their limtation. In addition, the techniques require a subject matter expert's opinion to evaluate alternatives based on given criteria.

Optimization Technique

Another common approach for safety project prioritization is the optimization technique (Cook and Green 2000; Saha and Ksaibati 2016; Sadeghi and Mohammadzadeh Moghaddam 2016; Mishra et al. 2015), which maximizes the benefits to the traffic network, while accounting for some predefined constraints. Saha and Ksaibati (2016) formulated an integer linear program, where the objective function was to minimize the total number of predicted crashes, while accounting for the available budget. The number of predicted crashes was calculated by multiplying the number of crashes and the crash reduction factor (CRF). The methodology was implemented on 2444 miles of a county paved road network. This methodology minimized the overall expected crashes by selecting the best combination of safety improvement projects.

In another example, Mishra et al. (2015) applied an optimization program that maximized the weighted total number of crashes by severity level. The weights were equal to hospital/insurance costs associated with different severity levels. In addition to available budget limitations, they accounted for two more constraints: one that limited the maximum number of active projects at a given location, and another one that prevented implementing a project when a similar project is already active at that location.

Despite the advantages of optimization techniques, the results can be very sensitive to several parameters that are commonly used in formulation such as budget, costs, and CRFs. Therefore, it is necesssary to conduct a sensitivity analysis to obtain more robust results.

CHAPTER THREE

DATA SETS

This study collected data from three types of data sets: (1) the National Household Travel Survey (NHTS) 2009 data set, which was used for pedestrian exposure estimation, (2) a crash severity data set, which was used to identify severity factors, and (3) a crash frequency data set, which was used to estimate crash frequency and identify crash risk factors.

NHTS Data Set

The NHTS 2009 is a microscopic data set about daily travel patterns in the United States. This database includes trip purpose, travel time and date, travel distance, and transportation mode for 1,167,322 trips across the country. The NHTS data have been used for various purposes pertinent to non-motorized transportation planning. Planners and transportation engineers can use the data to analyze travel behavior, examine the relationship of socioeconomic factors and travel, and understand travel patterns change over time (NHTS, 2009).

The NHTS 2009 data were collected from March 2008 through May 2009 from a sample of 150,147 households in the entire country. The 2009 NHTS data set comprises four files, including a household file, a person file, a vehicle file, and a travel day trip file. The household file records unique data about households like number of vehicles, type of residence, household income, and education of household respondent. The person file includes data relating to each household’s respondent such as age, driver status, race and ethnicity, and miles driven. The vehicle file consists of data about each household's vehicle like annualized vehicle miles. The travel day trip file records data about each trip made by respondents on the randomly-assigned travel day like trip duration, travel distance, and trip purpose.

The National Household Travel Survey (NHTS 2009) has proved useful for analyzing trends of walking trips in both rural and urban settings (Carlson et al. 2015; Y. Yang et al. 2015). The NHTS 2009 used the 2000 Census Urban and Rural Classification and Urban Area Criteria (Census 2000). The NHTS (2009) provided data about 22,953 walk trips in rural areas and 8,770 trips in small urban areas. Among all of these trips, there were 1,418 walk-to-or-from transit trips and 30,305 walk-only trips (of which 23,220 walk trips were coded as home-based trips and 6,975 walk trips as non-home-based trips).

This study used 23,220 home-based walk-only trips recorded by the NHTS (2009) in rural and small urban areas. The research focused on home-based trips because home locations were discernable, at the block group level. This locational attribute is needed to relate household trip making to the household's neighborhood attributes (e.g., road network and land use). In addition, the study focused on walk-only trips because walk-to-or-from transit trips accounted for only four percent of the total walk trips that occurred in the study areas, where transit service is frequently unavailable.

Factors that have shown some impacts on the decision or the amount of walking include land use (e.g., "5D" variables [Wei et al. 2016]), road network attributes (e.g., street connectivity and sidewalk provision [Kuzmyak et al. 2014; Cao et al. 2009]), environment (e.g., soil slope, weather temperature, and lighting [Kuzmyak et al. 2014; Sabir et al. 2011; Montigny et al. 2012]), demographics (e.g., age, gender, and income [Hatamzadeh et al. 2014]), and perceptions and attitudes (e.g., innate preference for driving personal automobiles [Kuzmyak et al. 2014; Singh 2016]). Since the goal is to estimate walk trip frequency (over a one-day period) aggregated at the household level, person-specific (e.g., age and gender) and environmental factors (e.g., weather) are not (and cannot be) considered.

Consequently, covariates for the household trip frequency model include household-level variables (provided as part of the NHTS data set), and block-group-level socio-economic data from the U.S. Census American Community Survey (ACS] 5-year estimates in 2010. Land use data were captured from the U.S. Geological Survey (USGS) online resources (USGS 2016), which provide detailed area types including undeveloped (e.g., agricultural land, water, and forest land), commercial, residential, industrial, and mixed-use areas. Table 4 provides the summary statistics of these variables.

Table 4. Summary statistics of NHTS data (No. of observations = 58,227 households).

Variables		Avg.	Standard Deviation	Min.	Max.	Mode
Response Variable						
# of walk trips		0.40	1.35	0	21	0
Household Characteristics						
Household size		2.34	1.20	1	14	2
Vehicle count		2.25	1.23	0	27	2
Average age		55.18	18.13	11	92	61
# of workers		0.93	0.90	0	6	0
# of adults		1.91	0.65	1	8	2
Household income (dollar/year)*	Low (<=25,000)	0.24	0.43	-	-	-

Table 4 Continued.

	Median (>25,000 and <=75,000)	0.47	0.50	-	-	-
	High (>75,000)	0.29	0.45	-	-	-
Block Group Characteristics						
Population Density (per square mile)		869.92	2,147.15	50	30,000	300
Land use entropy		0.13	0.15	0	1.00	0
Share of land use by type	Undeveloped	0.84	0.26	0	1	0
	Commercial	0.02	0.07	0	1	0
	Industrial	0.01	0.03	0	0.70	0
	Mixed	0.01	0.04	0	0.96	0
	Residential	0.12	0.21	0	1	0
Setting*	Small Urban	0.25	0.43	-	-	-
	Rural	0.75	0.43	-	-	-
Regional effect*	West	0.08	0.27	-	-	-
	Southwest	0.18	0.38	-	-	-
	Midwest	0.12	0.33	-	-	-
	Northeast	0.16	0.37	-	-	-
	Southeast	0.46	0.50	-	-	-

Note: * indicates binary variables and average column shows the proportion of each variable level.

Crash Severity Data Set

This study incorporated three states' pedestrian crash data from 2011 to 2013: Texas, Oregon, and Montana. These three states were considered because each comes from a different region and each contains a large amount of rural land. As noted earlier, this study is concerned with pedestrian safety in rural and small urban areas. As a result, this study identified 180, 603, and 1,530 intersection-related pedestrian crashes that occurred in those areas of Montana, Oregon, and Texas, respectively.

This study categorized the crash severity levels into fatal, injury, and non-injury as shown in Table 5 because the Oregon crash dataset reported only three severity levels. The second column indicates the severity levels used by Texas CRIS dataset.

Table 5. Severity levels definition (source: Texas CRIS data set).

Severity Level	Original Severity Level	Description
Fatal	Fatal	Died due to injuries sustained from the crash, within 30 days of the crash.
Injury	Incapacitating Injury	Severe injury which prevents continuation of normal activities; includes broken or distorted limbs, internal injuries, crushed chest, etc.
	Non-Incapacitating Injury	Evident injury such as bruises, abrasions, or minor lacerations which do not incapacitate.
Non-injury	Possible Injury	Injury which is claimed, reported, or indicated by behavior, but without visible wounds; includes limping or complaint of pain.
	Not Injured	The person involved in crash did not sustain any injury.

Overall, this paper accounted for six characteristic categories as illustrated in Table 6. These variables were collected from multiple sources including the U.S. Census American Community Surveys (ACS), state Departments of Transportation crash data sets (DOT), U.S. Geological Survey (USGS), the state Departments of Health Services (DHS), Federal Highway Administration (FHWA), TIGER/Line, and Quarterly Workforce Indicators (QWI). Google Street Views (GSV) were also used to collect road and environment details (e.g., traffic control type, sidewalk presence, and pavement condition) that were unavailable from archival data (see Appendix A for the data dictionary that was used for GSV data collection). It is worthwhile to mention that some

Table 6. Descriptive statistics of crash severity factors.

Variable	Values	Source	Non-Injury	Injury	Fatal	Total
-	-	-	500(18) *	1321(72)	220(10)	2,041
Driver age	1=aged less than 24	DOT	62(34)	103(56)	20(11)	185
	2=aged 25-64		91(24)	255(67)	36(9)	382
	3=aged more than 65		280(24)	756(64)	146(12)	1,182
	0=unknown		67(23)	207(71)	18(6)	292
Driver gender	1=male	DOT	48(29)	96(58)	21(13)	165
	2=female		233(22)	688(65)	141(13)	1,062
	0=unknown		219(27)	537(66)	58(7)	814
Functional Class (Tiger/Line category)	1= highway-highway	GSV	19(31)	38(61)	5(8)	62
	2= highway-local		155(21)	484(67)	85(12)	724
	3= local-local		316(27)	760(64)	117(10)	1,193
	4= driveway		10(16)	39(63)	13(21)	62
Traffic control	1=Signalized	GSV	33(20)	99(61)	30(19)	162
	2=Four-way stop		131(27)	332(68)	23(5)	486
	3=Two-way stop		30(30)	65(65)	5(5)	100
	0=None		306(24)	825(64)	162(13)	1,293
Angle between intersecting streets	1= 90°	GSV	443(26)	1,118(64)	175(10)	1,736
	2= less than 90°		57(19)	203(67)	45(15)	305
Number of intersection legs	Leg=3	GSV	213(22)	595(63)	144(15)	952
	Leg=4		282(26)	721(67)	75(7)	1,078
	Leg>4		5(50)	5(50)	0(0)	10
Railroad crossing in the vicinity of the intersection	0= none	GSV	486(24)	1,295(65)	213(11)	1,994
	1= if railroad exists		14(30)	26(55)	7(15)	47
One-way or two-way operation	1= 1 way-1 way	GSV	2(13)	12(80)	1(7)	15
	2= 1 way-2 way		36(19)	134(71)	18(10)	188
	3= 2 way-2 way		462(25)	1,175(64)	201(11)	1,838
Pavement condition	1= if both approaches are paved	GSV	482(25)	1,280(65)	202(10)	1,964
	2= if only one of approaches is paved		14(22)	34(53)	16(25)	64
	3=other		4(31)	7(54)	2(15)	13
Shoulder type	0= if none of approaches has shoulder	GSV	398(27)	959(66)	96(7)	1453
	1= if one of approaches has shoulder		85(17)	295(60)	109(22)	489
	2= other		17(17)	67(68)	15(15)	99

Table 6 Continued.

Presence of sidewalk	0=if none of approaches has sidewalk	GSV	197(24)	470(58)	143(18)	810
	1= if one of approaches has sidewalk		99(26)	247(65)	35(9)	381
	2= other		204(24)	604(71)	42(5)	850
Presence of bicycle lane	0=if none of approaches has bicycle lane	GSV	489(26)	1,160(63)	204(11)	1853
	1= if one of approaches has bicycle lane		9(6)	135(85)	15(9)	159
	2= other		2(7)	26(90)	1(3)	29
Number of driveways within the intersection and its influence area.	Numerical	GSV	4.86 (3.52)	4.32 (3.32)	3.72 (3.22)	4.39 (3.37)
Presence of curb parking	0=if none of approaches has curb parking	GSV	276(23)	740(62)	176(15)	1192
	1= if one of approaches has curb parking		87(22)	283(71)	27(7)	397
	2= other		137(30)	298(66)	17(7)	452
Presence of lighting at the intersection	0=if none of approaches has street light	GSV	197(23)	524(61)	134(16)	855
	1= if one of approaches has street light		185(24)	512(67)	65(9)	762
	2= other		118(28)	285(67)	21(5)	424
Presence of curve	0=if none of approaches has horizontal curve	GSV	366(26)	927(65)	143(10)	1,436
	1= if one of approaches has horizontal curve		101(22)	313(67)	54(12)	468
	2= other		33(24)	81(59)	23(17)	137
Presence of curve	0=if none of approaches has vertical curve	GSV	427(25)	1105(65)	179(11)	1711
	1= if one of approaches has vertical curve		58(22)	172(66)	32(12)	262
	2= other		15(22)	44(65)	9(13)	68
Presence of advance warning signs such as crossroad, STOP ahead, or signal ahead	0=if none of approaches has sign	GSV	455(25)	1,172(64)	211(11)	1,838
	1= if one of approaches has sign		34(20)	127(75)	8(5)	169
	2= other		11(32)	22(65)	1(3)	34
Roadway surface condition	1=dry	DOT	460(26)	1,134(64)	188(11)	1,782
	2=wet		30(13)	173(74)	30(13)	233
	3=snow		3(25)	8(67)	1(8)	12
	4=ice		7(50)	6(43)	1(7)	14

Table 6 Continued.

Number of through lanes on first approach	=0	GSV	4(33)	7(58)	1(8)	12
	=1		352(26)	862(64)	142(11)	1,356
	=2		127(21)	403(67)	69(12)	599
	>2		17(23)	49(66)	8(11)	74
Number of through lanes on second approach	=0	GSV	7(15)	34(74)	5(11)	46
	=1		414(24)	1,101(65)	181(11)	1,696
	=2		66(26)	160(63)	28(11)	254
	>2		13(29)	26(58)	6(13)	45
Number of right lanes on first approach	=0	GSV	467(25)	1,221(65)	205(11)	1,893
	=1		31(21)	100(69)	15(10)	146
	=2		2(100)	0(0)	0(0)	2
Number of right lanes on second approach	=0	GSV	459(25)	1,205(65)	205(11)	1,869
	=1		41(24)	112(67)	15(9)	168
	=2		0(0)	4(100)	0(0)	4
Number of left lanes on first approach	=0	GSV	355(25)	916(64)	160(11)	1,431
	=1		142(24)	399(67)	58(10)	599
	=2		3(27)	6(55)	2(18)	11
Number of left lanes on second approach	=0	GSV	424(25)	1,090(64)	194(11)	1,708
	=1		72(22)	225(70)	25(8)	322
	=2		4(36)	6(55)	1(9)	11
Configuration of right-turn lane on first approach	0= none	GSV	463(25)	1,203(64)	204(11)	1,870
	1= right-turn lane only		29(23)	87(67)	13(10)	129
	2= channelizing island only		5(21)	17(71)	2(8)	24
	3= right turn lane and channelizing island		3(17)	14(78)	1(6)	18
Configuration of right-turn lane on second approach	0= none	GSV	450(24)	1,193(65)	199(11)	1,842
	1= right-turn lane only		29(22)	92(70)	10(8)	131
	2= channelizing island only		11(31)	17(49)	7(20)	35
	3= right turn lane and channelizing island		10(30)	19(58)	4(12)	33
Configuration of left-turn (LT) lane on first approach	0= none	GSV	352(25)	915(64)	161(11)	1,428
	1= painted		131(24)	366(66)	55(10)	552
	2 = curbed		15(29)	32(63)	4(8)	51
	3= prohibited		2(20)	8(80)	0(0)	10

Table 6 Continued.

Configuration of left-turn (LT) lane on second approach	0= none
	1= painted
	2 = curbed
	3= prohibited
		GSV	419(25)
			60(22)
			16(28)
			5(71)
				1,090(64)
				193(71)
				36(63)
				2(29)
					195(11)
					20(7)
					5(9)
					0(0)
						1,704
						273
						57
						7

Average AADT of roadways in 0.5-mile buffer	Numerical	FHWA	9,535 (9,623)	9,316 (7,725)	10,642 (10,104)	9,512 (8,516)
Average daily pedestrian trips on intersected approaches	Numerical		539 (832)	671 (1,265)	342 (944)	604 (1,148)

Area type indicator	1= small urban
	2= rural
		ACS	318(25)
			182(23)
				853(68)
				168(60)
					87(7)
					133(17)
						1,258
						783

Population density of block group	Numerical	ACS	1,702 (1,756)	1,870 (2,100)	1,069 (1,707)	1,742 (1,998)
Proportion of male residents in block group	Numerical	ACS	0.49(0.06)	0.49(0.06)	0.49(0.06)	0.49(0.06)
Proportion of residents younger than 24 years	Numerical	ACS	0.37(0.11)	0.37(0.11)	0.35(0.09)	0.37(0.11)
Proportion of residents aged 25 to 54	Numerical	ACS	0.39(0.09)	0.39(0.08)	0.35(0.09)	0.39(0.08)
Proportion of residents older than 55 years	Numerical	ACS	0.24(0.12)	0.24(0.12)	0.25(0.11)	0.24(0.12)
Median income of households available in block group	Numerical	ACS	52,881 (27,464)	49,743 (25,592)	54,368 (22,714)	51,011 (25,835)
Household density of block group	Numerical	ACS	12,061 (40,913)	3,184 (19,830)	3,158 (26,385)	5,359 (27,462)
Employment density of county	Numerical	QWI	127(274)	113(260)	156(335)	121(273)
Proportion of undeveloped area in 0.5-mile buffer	Numerical	USGS	0.46 (0.37)	0.45 (0.37)	0.67 (0.34)	0.48 (0.37)
Proportion of commercial area in 0.5-mile buffer	Numerical	USGS	0.14 (0.15)	0.14 (0.16)	0.09 (0.15)	0.13 (0.16)
Proportion of industrial area in 0.5-mile buffer	Numerical	USGS	0.01 (0.04)	0.02 (0.06)	0.01 (0.06)	0.02 (0.06)

Table 6 Continued.

Proportion of residential area in 0.5-mile buffer	Numerical	USGS	0.04 (0.08)	0.03 (0.07)	0.02 (0.04)	0.03 (0.07)
Proportion of mixed area in 0.5-mile buffer	Numerical	USGS	0.34 (0.27)	0.35 (0.27)	0.21 (0.25)	0.33 (0.27)
Proportion of agricultural area in 0.5-mile buffer	Numerical	USGS	0.01 (0.04)	0.01 (0.04)	0.01 (0.06)	0.01 (0.04)
Total centerline mile in 0.5-mile buffer	Numerical	TIGER/Line	13.17 (5.43)	12.25 (5.13)	8.93 (4.99)	12.12 (5.32)
Total number of intersections in 0.5-mile buffer	Numerical	TIGER/Line	91.68 (51.72)	87.17 (52.90)	54.29 (46.11)	84.73 (53.02)
Distance to nearest hospital (km)	Numerical	DOH	37.5 (48.4)	23.5 (33.5)	27.4 (34.8)	27.3 (38.3)
# of alcohol establishments (bars and liquor stores) in 0.5-mile buffer	Numerical	GSV	11(59)	28(93)	7(28)	22(82)
Day of week	1=weekday	DOT	385(25)	1,035(66)	150(10)	1,570
	2=weekend		115(24)	286(61)	70(15)	4,71
Season	1=spring	DOT	124(25)	313(64)	52(11)	789
	2=summer		98(25)	258(65)	39(10)	395
	3=fall		149(25)	372(63)	73(12)	594
	4=winter		129(23)	378(67)	56(10)	563
Time of day	1=20:00-5:59	DOT	124(20)	378(62)	111(18)	613
	2=06:00- 09:59		72(24)	191(65)	32(11)	295
	3=10:00- 15:59		143(26)	371(69)	26(5)	540
	4=16:00- 19:59		161(27)	381(64)	51(9)	593
Weather condition	1=clear	DOT	412(27)	962(63)	156(10)	1,530
	2=cloudy		65(20)	220(68)	38(12)	323
	3=foggy or smoke		4(15)	16(62)	6(23)	26
	4=rain		16(11)	115(78)	17(12)	148
	5=snow or sleet		3(21)	8(57)	3(21)	14

* This table displays frequencies (percentage in parentheses) for categorical variables and mean (standard deviation in parenthesis) for numerical variables.

variables that were identified as significant factors in previous studies were not explored due to data unavailability or missing information. These variables include: pedestrian age and gender, vehicle movement, roadway grade, and impact speed.

Crash Frequency Data Set

This study used a three-year pedestrian crash data set from the state of Texas (2011-2013) to estimate crash frequency in rural and small urban areas. Crash data sets from Oregon and Montana were excluded, because many roadway and environment factors (e.g., traffic control type, shoulder type, median type) were missing in their crash data sets. Besides, it was infeasible to use Google Street View to collect those factors for all intersections that are sparsely located in the study areas (which would amount to 480,725 intersections in total across Oregon and Montana).

However, the crash data set included only those intersections where at least one crash occurred from 2011 to 2013, and the attributes corresponding to intersections with no crashes were not present. To reduce the resulting bias, this study also used a crash dataset for 2014 to capture information for intersections that did not experience any crashes from 2011 to 2013.

In total, this study accounted for 15,288 intersections in the rural and small urban areas of Texas, among which 1,444 intesections have experienced one or more pedestrian crashes during the study period (2011 to 2013). Table 7 presents the summary statistics for the dependent and independent variables.

Four categories of independent variables were considered to quantify the relationship between contributing factors and pedestrian crash frequency:

- **Socio-economic characteristics:** the U.S. Census American Community Surveys (ACS) were used to obtain the block group-level population density, average age, median income, and household density.
- **Built-environment characteristics:** a half-mile buffer around intersections was used to calculate the U.S. Geological Survey (USGS) land use type shares, number of intersections, and roadway centerline mileage. Texas Department of Health Services (DHS) data were also used to measure the distance of the intersections to the nearest hospital or emergency medical service.
- **Traffic attributes:** the crash data set has reported the AADT of the intersection leg at which crash occurred. If AADT data were missing, this study calculated the average AADT (reported by the FHWA) of the roadways at half-mile buffers around the intersections. If there was still no data available, this study used the average AADT of other intersections in the dataset based on road type and number of lanes. Additionally, the pedestrian exposure at the block group level was obtained using the K-NN technique, which will be explained in a later section (see page 56). The pedestrian exposure of each intersection is equal to the pedestrian exposure of the block group level where that intersection is located.
- **Roadway characteristics:** the crash dataset has reported the roadway characteristics of the intersection leg at which crash occurred. These charactersitics contain speed limit, number of lanes, roadtype, traffic control type, surface type, curve presence, shoulder type, and median type.

Table 7. Statistic summary of crash frequency data (No. of observations = 15,288 intersections).

Variable		Mean	Std. Dev.	Min.	Median	Max.
Response Variables						
Total crash frequency		0.10	0.32	0	0	10
Fatal crash frequency		0.01	0.12	0	0	2
Injury crash frequency		0.06	0.25	0	0	7
Variable		**Mean**	**Std. Dev.**	**Min.**	**Median**	**Max.**
Non-injury crash frequency		0.03	0.18	0	0	3
Independent Variables						
Socio-economic Characteristics (block-group level)						
Population density		526.2	1011.3	0.3	142.3	25328
Average age		37.3	5.9	18.4	37.1	66.4
Median income		55478.8	25012.4	8828	50417	98359
Household density		182.5	343.4	0.66	49.0	7492.1
Built-Environment Characteristics (0.5-mile buffer)						
Land use share	Undeveloped	0.79	0.29	0	0.81	1
	Commercial	0.04	0.09	0	0.01	1
	Industrial	0.01	0.06	0	0	1
	Mixed	0.02	0.05	0	0	0.88
	Residential	0.14	0.22	0	0.07	1
Centerline miles		9.4	4.9	1.35	8.21	36.6
# of intersections		41.6	42.8	1	25	327
Traffic Characteristics						
Avg. AADT		8863.9	9388.0	63	6725	39,504
Ped. exposure (# of walk trips)		684.4	1059.5	22	311.5	2268
Roadway Characteristics						
Speed limit		39.2	15.3	5	45	60
# of lanes (two-directions)		2.8	1.1	2	2	6
Road Type*	Two-lanes, two-way	0.45	0.49	-	-	-
	4 or more lanes, divided	0.15	0.35	-	-	-
	4 or more lanes, undivided	0.14	0.33	-	-	-
	other	0.26	0.43	-	-	-
Traffic control*	Signal	0.12	0.32	-	-	-
	2- or 4-way stop sign	0.37	0.48	-	-	-
	Yield sign or flashing light	0.06	0.23	-	-	-
	None	0.45	0.50	-	-	-
Surface type*	Low type surface-treated	0.20	0.38	-	-	-
	High type flexible	0.80	0.42	-	-	-

Table 7 Continued.

Curve*	Straight	0.90	0.30	-	-	-
	Curved	0.10	0.30	-	-	-
Shoulder presence*	Present	0.59	0.50	-	-	-
	None	0.41	0.48	-	-	-
Median presence*	Present	0.15	0.10	-	-	-
	None	0.85	0.40	-	-	-

Note: * indicates binary variables and average column shows the proportion of each variable level.

CHAPTER FOUR

METHODOLOGIES

This study adopts a six-step systemic safety planning tool to customize the process for addressing pedestrian safety in small urban and rural settings as shown in Table 8. The following sections describe the methodologies that are used in each step in more detail.

Initial Screening

The mechanisms underlying pedestrian-vehicle collisions differ between intersections and mid-block segments (Aziz et al., 2013; Bennet and Yiannakoulias 2015; Abay 2013). At-grade intersections have impacts on driver decisions (e.g., speed reduction), vehicle queues, and vehicle movements. These impacts might be extended to the upstream and downstream areas of the intersections. Therefore, it is judicious to identify whether a pedestrian crash happened at an intersection (or its influence area) or a mid-block location. This study defined two crash types: midblock crashes (crashes that happened more than 250 feet from the centers of intersections) and intersection or intersection-related crashes (crashes that occurred within the influence areas of intersections, i.e., within 250 feet of the intersection centers). The cut-off value, 250 feet, was used because most of the intersection-related crashes occurred within 250 feet of the intersections on rural two-lane highways (Harwood et al. 2000).

Table 9 summarizes the pedestrian crashes by severity and by locations (intersection or midblock). As these statistics show, 72% of pedestrian crashes in the data

Table 8. Summary of methodologies in pedestrian systemic safety tool.

Initial Screening

- **Data:** crash data sets from three states (Montana, Oregon, and Texas)
- **Approach:** trend analysis

Pedestrian Exposure Estimation

- **Data:** NHTS 2009 across rural and small urban areas in USA
- **Selected Approach:** finite mixture among statsitical models and K-Nearest Neighbor model among machine learning techniques
- **# of Obs.** =56,500 households
- **Response variable** = houshold walk trip frequency
- **Covariates** = household size, vehicle count, income, # of adults, # of workers, population density, land use shares, area setting, regional effect

Risk Factors Identification

1- Crash Severity Analysis:

-**Data:** crash datasets from three states of Montana, Oregon, Texas

-**Selected Approach:** Random Forest

-**# of Obs**. = 2,200 crashes

-**Response variables** = crash severity levels: 1- non-injury, 2-injury, 3-fatal

-**Covariates** = driver, roadway, traffic, bult-environment, socioeconomic, natural environment characteristics

2-Crash Frequency Analysis

-**Data:** crash datasets from Texas

-**Selected Approach:** Hurdle Negative Binomial

-**# of Obs**. = 15,288 intersections

-**Response variables** = total crash fequency

Candidate (Hotspot) Locations Identification

Approach: Two-Step Floating Catchment Area (2SFCA) method

Countermeasure Selection

- **Approach:** literature sources

Project Prioritization

- **Approach:** mixed linear program

occurred within the influence areas of intersections. Therefore, this study focused on intersection (or related) crashes. Statistics reveal rural areas are more likely to have fatal crashes; 24 and 9 percent of pedestrian crashes were recorded as fatal crashes in rural and small urban areas, respectively. However, small urban areas reported more injury and non-injury crashes than rural areas, so that 57 and 67 percent of pedestrian crashes were.

Table 9. Pedestrian crashes and sites (where at least one pedestrian crash occurred) in Montana, Oregon, and Texas from 2011 to 2013.

State		**Pedestrian crashes**									**Sites**		
		Sub-Total	Intersection-related				Midblock				Sub-Total	Inter-section	Mid-block
			Total	Fatal	Injury	Non-injury	Total	Fatal	Injury	Non-injury			
Montana	Urban Cluster	123 (45)	108 (88)	5 (5)	39 (36)	64 (59)	15 (12)	2 (13)	10 (67)	3 (20)	106 (42)	91 (86)	15 (14)
	Rural	151 (55)	72 (48)	10 (14)	38 (53)	24 (33)	79 (52)	16 (20)	47 (59)	16 (20)	146 (58)	72 (49)	74 (51)
	Total	274	180 (66)	15 (8)	77 (43)	88 (49)	94 (34)	18 (19)	57 (61)	19 (20)	252	163 (65)	89 (35)
Oregon	Urban Cluster	536 (79)	507 (95)	41 (8)	466 (92)	0 (0)	29 (5)	3 (10)	26 (90)	0 (0)	492 (78)	466 (95)	26 (5)
	Rural	143 (21)	96 (67)	23 (24)	72 (75)	1 (1)	47 (33)	9 (19)	38 (81)	0 (0)	139 (22)	94 (68)	45 (32)
	Total	679	603 (89)	64 (11)	538 (89)	1 (0)	76 (11)	12 (16)	64 (84)	0 (0)	631	560 (89)	71 (11)
Texas	Urban Cluster	852 (38)	756 (89)	65 (9)	422 (56)	269 (36)	96 (11)	18 (19)	56 (58)	22 (23)	802 (37)	706 (88)	96 (12)
	Rural	1,390 (62)	774 (56)	148 (19)	438 (57)	188 (24)	616 (44)	190 (31)	320 (52)	106 (17)	1344 (63)	741 (55)	603 (45)
	Total	2242	1530 (68)	213 (14)	860 (56)	457 (30)	712 (32)	208 (29)	376 (53)	128 (18)	2146	1447 (67)	699 (33)
Sum		3,195	2313 (72)	292 (13)	1475 (64)	546 (24)	882 (28)	238 (27)	497 (56)	147 (17)	3,029	2,179 (72)	859 (28)

recorded as injury crashes in rural and small urban areas, and 20 and 24 percent of pedestrian crashes were recorded as crashes without injury in rural and small urban areas

Pedestrian Exposure Estimation

This study demonstrates an area-based approach to estimate pedestrian exposure (number of walk trips) for rural and small urban areas using household-level NHTS data (2009). Disaggregate approaches (e.g., point- or segment-based) appear to be incompatible with the NHTS data for the study purpose, because the NHTS reports home locations at block group level and work (or other destination) locations at census tract level, obscuring any attempt to estimate pedestrian paths within traffic networks.

Statistical Models

A host of statistical models have been proposed to analyze count data that are over-dispersed and/or contain many zeros. Miranda-Moreno (2006), Zou et al. (2013), and Park et al. (2010) used a finite mixture of negative binomial regression (FMNB) model to estimate crash frequency. They found that FMNB outperformed the negative binomial in the case with excess zeros. Hence, this study employed this model for the estimation of pedestrian exposure, with a comparative look at conventional methods including the negative binomial (NB) and zero-inflated negative binomial (ZINB) models.

Negative Binomial (NB) Model. Generalized Linear Regression modeling (GLM) such as Poisson and negative binomial (NB) models have been widely employed to analyze count data. The NB regression method permits over-dispersion (the conditional variance is greater than the conditional mean of the response variable), which was found to exist in the data set. The negative binomial model is formulated as,

$$y_i \sim Poisson\ (\mu_i) \tag{1}$$

where y_i is the number of daily walk trips made by household i and μ_i represents the expected walk trip frequency of household i,

$$\mu_i = (E_i)^{\alpha} \cdot e^{X\beta+\varepsilon} \tag{2}$$

where E_i is an exposure term for household i, which is proxied by household size (assuming that larger households offer more opportunities for making more [walk] trips); X is a vector of independent variables, of which the effects are described by a vector of coefficients, β; α denotes any non-linear relationship between household size and the average trip frequency; and ε is a Gamma distributed error term.

Zero-Inflated Negative Binomial (ZINB) Model. A Zero-inflated model is an alternative to model count data containing an excess number of zeros. This model assumes the zeros originate from two types of sources: first, households that do not take walking into account as a transportation mode; second, households that make walk trips ranging from zero to many (New et al. 2016). The expected walk trip frequency is calculated as follows,

$$\begin{aligned} E(y = n) = P(&first\ type\ of\ households) * 0 \\ &+ P(second\ type\ of\ household) \\ &* E(y = n|second\ type\ of\ household) \end{aligned} \tag{3}$$

and the negative binomial probability density function (PDF) is structured as,

$$PDF(y;p,n) = \frac{(y_i + n - 1)!}{y_i!\,(n-1)!}\, p_i^n (1 - p_i)^{y_i} \tag{4}$$

where y is the number of walk trips $(0,1,2,\ldots)$ and p represents the probability of n walk trips.

Finite Mixture of Negative Binomial Regression (FMNB) Model. The FMNB model assumes the data set arises from a finite number of components (two or more) with unknown proportions, which permits accounting for heterogeneity without imposing strong distributional assumptions on the mixing variable (Park et al. 2010). In fact, the FMNB captures the additional heterogeneity within components not captured by covariates. In addition, the finite mixture model allows the components to have different regression coefficients. The finite mixture model is structured as follows (Park et al. 2010),

$$\begin{aligned} f_Y(y_i|X_i,\Theta) &= \sum_{k=1}^{N} w_k NB(\mu_{i,k},\phi_k) \\ &= \sum_{k=1}^{N} w_k \left[\frac{\Gamma(y_i+\phi_k)}{\Gamma(y_i+1)\Gamma(\phi_k)} \left(\frac{\mu_{i,k}}{\mu_{i,k}+\phi_k}\right)^{y_i} \left(\frac{\phi_k}{\mu_{i,k}+\phi_k}\right)^{\phi_k}\right] \end{aligned} \tag{5}$$

$$E(y_i|X_i,\Theta) = \sum_{k=1}^{N} w_k\, \mu_{i,k} \tag{6}$$

$$Var(y_i|X_i,\Theta) = E(y_i|X_i,\Theta) + (\sum_{k=1}^{N} w_k\, {\mu_{i,k}}^2 \left(1 + \frac{1}{\phi_k}\right) - E(y_i|X_i,\Theta)^2) \tag{7}$$

$$\mu_{i,k} = \left(exposure_{i,k}\right)^{\alpha} \cdot e^{(X\beta)_{i,k}+\varepsilon} \tag{8}$$

$$\Theta = \{(\beta_1,\ldots,\beta_N),(\phi_1,\ldots,\phi_N),(w_1,\ldots,w_N)\} \tag{9}$$

where f_Y, E, and Var are probability density function, mean, and variance, respectively; w_k indicates the weight of component k (1, 2, …, N); $X_i$and β are vector of covariates and regression coefficients; ϕ represents the dispersion parameter of the negative binomial distribution; and Θ is the vector of all unknown parameters. The number of components for the finite mixture model should be selected such that the goodness-of-fit measures are optimized (Park et al. 2014). Therefore, a series of models with different numbers of components are applied, and the model with best goodness-of-fit measurement would be selected.

Machine Learning Technique

Since most of the sampled households in NHTS reported zero walk trips, the data set suffers from excess zeros and over-dispersion issues. This will have a negative impact on the statistical model's application. In addition, the statistical models have assumed a predefined underlying relationship between response and explanatory variables (Chang and Chen 2005), which also limit their application. Addressing this issue, this study examines several machine learning techniques (i.e., Random Forest, Support Vector Machine, Decision Tree, Naive Bayes, Artificial Neural Network, and K-Nearest Neighbor), which are well-known tools for prediction and classification, to estimate the number of walk trips per household. However, only the K-Nearest Neighbor technique is explained here due to its prominent results. The other techniques are described later in this document.

K-Nearest Neighbors (K-NN) Technique. The K-NN is a non-parametric technique for classification and regression (Raj et al. 2016). In classification, prediction of an object equates to the prevailing class among its K- nearest neighbors in the feature space (Zheng and Su 2014). In regression, the predicted value of an object is the average of the values of its K-nearest neighbors. The K-nearest neighbors are selected based on their straight-line (Euclidean) distance from the object in feature space.

Crash Risk Factors Identification

The literature has analyzed pedestrian safety largely through the lens of crash frequency (Miranda-Moreno et al., 2011; Wang and Kockelman, 2013; Chen and Zhou, 2016) or crash severity (Sasidharan et al., 2015; Toran Pour et al., 2016; Kim et al., 2017) or both together. The first approach, univariate crash frequency analysis, estimates the expected crash count at an aggregate level. The second approach, crash severity analysis, predicts the probability of different severity levels once a crash happens. However, univariate crash frequency models neglect the correlation among different severity levels. Hence, some studies employed a third approach, crash frequency by severity level analysis, which explores total crash frequency and crash proportion by crash severity levels.

Crash frequency can be used to estimate the crash risk for hotspot identification and the crash reduction benefits associated with countermeasures. Moreover, crash severity plays a key role in health outcomes, treatment costs, and long-term repercussions for pedestrians (Dai 2012). Therefore, this study explores both approaches (i.e., crash frequency and crash severity) to compare their performance and identify contributing factors associated with pedestrian–vehicle injury severity levels and crash frequency in rural and small urban areas.

Crash Severity Analysis

As previously mentioned, a wide range of methodologies has been applied for crash severity analysis. These methods range from simple methods like descriptive analysis to more complicated statistical models, to machine learning techniques. However, there is no consensus on what model is the best, and each model has its own strengths and weaknesses (Savolainen et al. 2011). Therefore, this study uses four statistical models including ordered logit and probit models, which can control for the ordinal nature inherent in the level of injury, and multinomial logit and probit models, which do not account for ordering but allow the independent variables' effects to vary among the injury levels. The study also uses three machine learning approaches (i.e., random forest, naive Bayes, and artificial neural network) to quantify the relationship between contributing factors and pedestrian injury severity. This approach provides an opportunity to compare the performance of the ordered with unordered models, and statistical models with machine learning approaches for analyzing the injury severity.

Ordered Logit Regression (OL) Model. Let P_{ij} denote the probability of crash i experiencing severity level of j. The standard form of the ordered logit model is as follows (Haghighi et al. 2018):

$$P_{ij} = P(Y_i = j) = \frac{\exp(\alpha_j - X_i\beta)}{1 + [\exp(\alpha_j - X_i\beta)]} , \qquad j = 1, 2, \dots, J - 1 \tag{10}$$

where Y_i denotes the severity level for pedestrian crash i, X_i represents the vector of independent variables, β is the vector of coefficients, α_j indicates the cutoff term for the threshold in the model for j$^{\text{th}}$ severity level, and J is the number of severity levels. This model has a major assumption called the parallel regression or proportional odds assumption. This assumes the β values are constant across each severity level but the α values are different, which result in J-1 parallel regression lines with different intercepts on logit scale.

Ordered Probit Regression (OP) Model. The OP is different from the OL model in random error distribution and mathematical specification. The OL model assumes the random error follows a logistic distribution with a mean of zero and standard deviation of $\pi/\sqrt{3}$, while the OP model assumes the random error has a standard normal distribution with a mean of zero and standard deviation of 1.0. The OP model has the following structure (Abdel-Aty 2003),

$$P_{i1} = \Phi(X_i\beta - \alpha_1) \tag{11}$$

$$P_{ij} = \Phi(X_i\beta - \alpha_j) - \Phi(X_i\beta - \alpha_{j-1}), \quad j = 2, \dots, J-1 \tag{12}$$

$$P_{iJ} = 1 - \sum_{j=1}^{J-1} P_{ij} \tag{13}$$

where Φ is the cumulative standard normal distribution function.

Multinomial Logit (MNL) Model. The MNL model does not account for the ordinal nature of the independent variable but allows independent variables' effects to vary among outcome levels. Specifically, it estimates a series of binary models where one level of dependent variables is known as reference. The MLP model does not assume independence of irrelevant alternatives (IIA). An important effect of this assumption is that the odds ratios are fixed when other choices are added or dropped. When IIA is violated, MNL is an incorrectly specified model, and MNL coefficient estimates are biased and inconsistent (Dow and Endersby 2004). The MNL model is structured as follows (Sasidharan and Menéndez 2014),

$$P_{ij} = P(Y_i = j) = \frac{\exp(X_i \beta_j)}{\sum_{j=1}^{J} \exp(X_i \beta_j)}\ , \qquad j = 1, 2, \dots, J-1 \tag{14}$$

Multinomial Probit (MNP) Model. The MNP model specification is similar to the MNL model, but it uses the standard normal cumulative distribution function as given below,

$$P_{ij} = P(Y_i = j) = \Phi\ (X_i \beta_j) \tag{15}$$

Random Forest (RF) Technique. The Random Forest (RF) method is a well-known technique to either predict crash severity or rank the importance of independent variables affecting crash safety (Abdel-Aty et al. 2008; Haleem et al., 2015). Overall, it grows a number of trees on various sub samples of the data set. For each tree, it works by first allocating all the observations at the top of the tree (parent node) and then dividing the parent node into several child nodes based on which independent variables result in the best homogeneity (Ghasemzadeh et al. 2018). It means the algorithm selects the independent variables, based on which observations with the same outcome are assigned into the same category. There are different methods for dividing the nodes. The most well-known method is the Gini index, which measures inequality among values of the injury severity levels as given below (Kashani and Mohaymany 2011),

$$P(i|k) = \frac{p(i,k)}{p(k)} \tag{16}$$

$$p(i,k) = \frac{\pi(i) N_i(k)}{N_i} \tag{17}$$

$$p(k) = \sum_{i=1}^{I} p(i,k) \tag{18}$$

$$Gini\ (k) = 1 - \sum_{i=1}^{I} p^2(i|k) \tag{19}$$

where $\pi(i)$ is the prior probability of class i (=1, 2, …, *I*), $P(i|k)$ is the conditional probability of an object being in class *i* provided that it is in node k, $N_i(k)$ refers to the number of objects of class *i* in the node k, and $Gini\ (k)$ represents the Gini index of node k. When all observations in the child node belong to an injury severity level, the Gini index is equal to zero, which indicates the best possible homogeneity in that node. This procedure will be repeated for each child node until each node has the greatest possible homogeneity. The nodes that cannot be divided are called terminal nodes.

The RF method applies the Gini index criterion to determine the relative importance of covariates. The reductions in the Gini index for each individual variable are added up over all the trees in the forest. The Gini index reduction for each variable is equal to the importance score. The variables with higher Gini index reductions are ranked first.

Naive Bayes Technique. The Naive Bayes is a probabilistic technique that works on the basis of the Bayes theorem with strong (naive) independence assumptions between the independent variables. Using Bayes' theorem, the conditional probability of class (Y= j) given independent variables is formulated as follows (Chen et al. 2016),

$$P(Y = j|X = (x_1,\ x_2, \dots ,\ x_n)) = \frac{P(X = (x_1,\ x_2, \dots ,\ x_n)|Y = j)\ P(Y = j)}{P(X = (x_1,\ x_2, \dots ,\ x_n))} \tag{20}$$

where Y is the severity level variable, $X = (x_1,\ x_2, \dots,\ x_n)$ is the set of covariates, J is total severity levels. This technique calculates the probabilities of different severity levels for each unknown crash event and then assigns the most probable severity level to that specific crash. Although this technique strictly relies on independence assumptions, it performed effectively in classification of many real world data sets such as crash data sets (Chen et al., 2016; Taamneh, et al., 2017).

Artificial Neural Network (ANN) Technique. The Artificial Neural Network technique is capable of modeling complex nonlinear data sets. The ANN model transfers a series of input variables to a set of one or more output variables. It consists of three layers: input, hidden, and output as illustrated in Figure 1. Each neuron in any layer is connected to all neurons in the next layer through lines called "weight coefficients" (Moghaddam, et al., 2011). Basically, the output is formulated as given below,

$$Output_i = \frac{1}{1 + \exp(-\sum_j w_{ij} \times x_i)} \tag{21}$$

where w_{ij} represents the weight coefficient from *i*th neuron of the first layer to *j*th neuron of the next layer. For more information, see the literature such as Delen et al., (2006), Moghaddam et al., (2011), or Deka and Quddus, (2014).

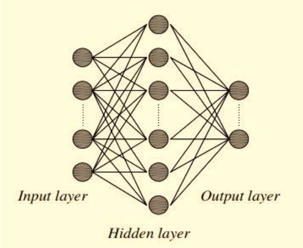


Figure 1. Three-layer neural network (Moghaddam et al., 2011).

Crash Frequency Analysis

The literature has commonly used count data models such as Poisson regression (Roshandeh et al. 2016), negative binomial (Abdulhafedh 2008; Malyshkina and Mannering 2010; Daniels et al. 2010), zero-inflated (Chen et al. 2016), and Hurdle models (Son et al. 2011; Boucher and Santolino 2010; Hosseinpour et al. 2014) for crash frequency analysis. The negative binomial, zero-inflated, and hurdle models can control for over-dispersion and excess zero problems that are common in crash data sets. This study examines the Hurdle negative binomial model, with a comparative look at negative binomial and zero-inflated models.

As with many other crash data sets, the three-year data that this study used is imbalanced, where 90 percent of the intersections in the study area reported no crash, 9 percent of the intersections reported one crash, and only one percent of the intersections reported two or more crashes. Imbalanced data may cause estimation issues. To address this issue, the gradient boosting approach can be used to improve the predictive power of machine learning models (Friedman 2001). Therefore, this study also explores the Boosted random forest (BRF) technique to estimate the pedestrian crash frequency.

Hurdle Negative Binomial (HNB) Model

. The HNB model is a count data model that can handle excess zeros and an over-dispersion problem (Son et al. 2011; Boucher and Santolino 2010; Hosseinpour et al. 2014). It is a two-state model, like the zero-inflated models. The first state is a binary logit model that models whether an intersection experiences a crash or not. The second state is a negative binomial that models the intersections with at least one crash. The overall HNB density is given as (Hosseinpour et al. 2014),

$$P(Y = y_i) = \begin{cases} P_i & y_i = 0 \\ (1 - P_i)\left(1 - \frac{1}{(1 + \alpha\mu_i)^{\frac{1}{\alpha}}}\right)\left(\frac{\Gamma\left(y_i + \frac{1}{\alpha}\right)}{\Gamma(y_i + 1)\Gamma\left(\frac{1}{\alpha}\right)}\right)\left(\frac{(\alpha\mu_i)^{y_i}}{(1 + \alpha\mu_i)^{y_i + \frac{1}{\alpha}}}\right) & y_i > 0 \end{cases}, \quad (22)$$

$$P_i = \frac{\exp(\lambda X_i)}{1 + \exp(\lambda X_i)}, \text{ and} \quad (23)$$

$$\mu_i = \exp(\beta X_i), \quad (24)$$

where α denotes a dispersion parameter, Γ is a gamma function, P_i represents the probability of an intersection *i* having zero crashes, μ_i is the estimated crash frequency, λ and β are the estimated coefficients, and X represents the vector of the explanatory variables.

The HNB model provides two sets of coefficients for explanatory variables: a zero model, and a count model. The zero model examines the effect of explanatory variables on whether a crash occurs or not, while the count model explores the impacts of variables on positive numbers of crash frequency. However, the zero-inflated model assumes zeros originate from two types of zeros: a binary distribution that generates zeros and a negative binomial distribution that generates counts, some of which could be zero.

<u>Boosted Random Forest (BRF).</u> As with many other crash data sets, the three-year data that this study used is imbalanced, where 90 percent of the intersections in the study

area reported no crash, 9 percent of the intersections reported one crash, and only one percent of the intersections reported two or more crashes. Imbalanced data may cause estimation issues. To address this issue, the gradient boosting approach can be used to improve the predictive power of machine learning models (Friedman 2001). Boosting refers to combining the rough and inaccurate rules to create more effective rules in a iterative procedure. In any further steps, the residual of the model is calculated, accounted as a target variable for the subsequent iteration (Friedman 2001). Basically, Gradient Boosting first trains the data to create a model (*F*) and calculates the loss (difference between predicted and observed value). The algorithm then improves the model *F* by adding an estimator *k* to provide a better model:

$$F_{m+1} = F_m + k \tag{25}$$

where *m* is the number of iteration. The solution starts with a perfect *k* as given below,

$$k = y - F_m \tag{26}$$

where *y* is the average response variable.

Moreover, Synthetic Minority Over-sampling Technique (SMOTE) is a well-known approach for over-sampling the minor classes to increase the number of instances in minor classes (Chawla et al. 2002). Basically, the SMOTE randomly selects a subset of data from the minority class and then creates new synthetic similar instances through K-nearest neighbors of minor class. The synthetic instances are introduced to the original dataset and the classification algorithm model the new dataset. Therefore, this study employs the gradient boosted approach with SMOTE technique to improve the Random forest algorithm for crash frequency estimation.

Hotspot Locations Identification

As noted in the literature, a good hotspot identification method should control for a range of crash contributory factors (e.g., road geometry and traffic exposure), while considering different severity levels and spatial heterogeneity that are unique to crash counts. To achieve these goals, this study proposes a new method that fills the aforementioned gaps with a case study provided to identify hotspots of pedestrian crashes in rural and small urban areas.

Two-Step Floating Catchment Area (2SFCA) Method

This study combines spatial variation, pedestrian exposure, and severity levels into a new measure to identify crash hotspots. In particular, the new method utilizes the two-step floating catchment area (2SFCA) method (a special case of the gravity model). The generic form of a gravity model is shown as (Luo and Qi 2009):

$$A_i^G = \sum_{j=1}^{n} \frac{S_j d_{ij}^{-\beta}}{\sum_{k=1}^{m} P_k d_{kj}^{-\beta}} \tag{27}$$

where A_i^G is the gravity-based index of accessibility at demand location i (e.g., block group), S_j represents the variable of interest (e.g., number of physicians) at supply location *j* (e.g., health center), P_k indicates the population at demand location k, d_{ij} is the distance between demand and supply locations, β is the frictional coefficient, and *n* and *m* are the total number of supply and demand locations, respectively. A large index value indicates that a (demand) location has more access to supply of medical resources. Since the frictional coefficient (β) is not as straightforward to interpret and requires extensive data to calibrate, the 2SFCA assumes β equals one in the catchment area and 0 otherwise

(Luo and Qi 2009). Luo and Qi (2009) enhanced the 2SFCA method by assigning distance-based weights to neighboring zones, as shown in the following steps:

The first step calculates the supply-to-demand ratio at each supply location, considering only those demand locations that fall within a threshold distance (d_0) from the supply location,

$$R_j = \frac{S_j}{\sum_{k\in(d_{ij}\leq d_0)} P_k W_{kj}} \tag{28}$$

where, W_{kj} represents a weight parameter between locations k and j and is generated from the Gaussian function as given below (Wang 2007),

$$W_{kj} = \exp(-{d_{kj}}^2/\beta) \tag{29}$$

where d_{ij} denotes the distance between locations k and j; and β is an empirical parameter that represents distance friction.

The threshold distance is set on a case by case basis: Kocatepe et al. (2017) used a 5-mile threshold in their crash hotspot analysis and Luo and Qi (2009) considered a threshold distance of 30-minute travel time in their medical service analysis.

The second step calculates the accessibility index for each demand location by summing up the supply-to-demand ratio of all supply locations that are within the threshold distance from the demand location (Luo and Qi 2009),

$$R_j = \frac{S_j}{\sum_{k\in(d_{ij}\leq d_0)} P_k W_{kj}} \tag{30}$$

$$A_i^F = \sum_{j\in(d_{ij}\leq d_0)} (R_j) = \sum_{j\in(d_{ij}\leq d_0)} \frac{S_j}{\sum_{k\in(d_{ij}\leq d_0)} P_k\, W_{kj}} \tag{31}$$

In the context of hotspot identification, pedestrian crash locations supply crash hazards (underlying risk factors like poor visibility, high traffic speed, etc.) and pedestrian areas generate walking trips (i.e., pedestrian demand). Intuitively, risk factors from a crash location may spill over to increase crash risk at nearby locations (e.g., a missing traffic control device at an intersection can encourage speeding at nearby downstream intersections). The crash risk index (CRI) is defined as follows,

$$CRI_j = SI_j \; / \; (\textstyle\sum_{i \in neighbors_j} P_i \, W_{ij}) \tag{32}$$

$$\overline{CRI_\iota} = \sum_{j \in neighbors_i} CRI_j \, W_{ij} \tag{33}$$

where, CRI_j indicates the crash risk index at crash location *j*, $\overline{CRI_\iota}$ is the weighted summation of crash risk index at pedestrian area *i*, W_{ij} represents the spatial weight between locations *i* and *j*, *SI* denotes the severity index, and *P* is the pedestrian exposure, with details explained in the following sections.

Weight Matrix

Previous 2SFCA studies usually considered the census block groups as the demand locations (i.e., population centers) and larger zones (e.g., zip codes or census tracts) as the supply locations (where medical resources exist). However, such large areas are not suitable for crash hotspot analysis because crash risk factors tend to exert more localized influence (e.g., among nearby intersections or road segments), compared with the broader influence area of a medical center that serves residents from nearby block groups or census tracts. This spatial influence can be further diluted in rural and small urban areas where the traffic network is sparsely connected.

Therefore, to more effectively reveal spatial influence, this study uses smaller grid cells for both crash locations and pedestrian areas, rather than any predefined administrative boundary, throughout the two steps of the 2SFCA method. The cell size is 660 by 660 feet, which reflects the average block size in the study areas (Census 2010).

The weight matrices describe the strength of influence between cells i and j. This study used contiguity neighboring to reduce a dense matrix (in which many cells have values close to zero) to a parsimonious, first-order or higher-order matrix, which keeps the weights between grids that share a border and sets other cell values to zero. Furthermore, the study experimented with an array of weight definitions, including inverse distance, exponential distance, and double-power distance. An inverse distance definition (Cho 1983) produced the most significant spatial variation and hence was selected for the analysis as shown given below,

$$W_{ij} = \left\{ \begin{matrix} \frac{1}{d_{ij}} & if \;\; i \; and \; j \; are \; neighbors \\ 0 & otherwise \end{matrix} \right\} \tag{34}$$

where, W_{ij} indicates the spatial weight between grid cells of i and j, d_{ij} is the distance between grid cell centroids.

Severity Index

Conventionally, a severity index is calculated by weighting observed crash frequency with insurance/hospital-reported crash costs, which biases the results toward locations that experienced crashes. This study adopts the Empirical Bayesian approach to control for both crash history data and expected crash data to smooth biased results. Empirical Bayesian estimation combines expected crash frequency with observed crash counts to obtain a long term mean value of crash frequency. Therefore, this study utilized an Empirical Bayesian estimation of crash frequency for each severity level i as follows (Cheng and Washington 2005),

$$\hat{y}_i = \alpha \times E(y_i) + (1 - \alpha) \times x_i \text{ }, \forall i = 1,2, \dots, I \tag{35}$$

$$\alpha_i = \frac{1}{1 + (\frac{E(y_i)}{k})} \tag{36}$$

where $\hat{y}$ denotes the estimated mean crash frequency, α indicates the weight factor, x refers to observed crash frequency, $E(y)$ is the expectation of crash frequency, and k is the over-dispersion parameter of the developed crash frequency model.

The severity index can be calculated by weighting the mean crash frequency of different severity levels. The conversion factors based on insurance/hospital-reported crash costs (Table 10, FHWA 2010) are used to convert the estimated mean crash frequency into the equivalent property damage only (PDO) crash severity:

$$\text{SI} = \sum_i \rho_i \times \hat{y}_i \tag{37}$$

where, SI represents the severity index and ρ_i denotes the conversion factor for severity level i. Accounting for crash severity in crash hotspot analysis can compensate for two kinds of errors. First, it addresses the errors generated by the random distribution of crashes because severe crashes are rarely distributed by chance (Da Costa et al. 2015; Soltani and Askari 2017), especially in rural and small urban areas where pedestrian crashes are more sparsely distributed. Second, it compensates for the errors generated by under-reporting less severe crashes (Oh et al. 2010).

Table 10. Average comprehensive costs by injury severity.

Severity Level	**Cost (USD)**	**Weight (ρ)**
Fatality (K)	$4,008,900	541.7
Disabling Injury (A)	$216,000	29.2
Evident Injury (B)	$79,000	10.7
Fatal/Injury (K/A/B)	$158,200	21.3
Possible Injury (C)	$44,900	6.1
PDO (O)	$7,400	1.0

Countermeasure Selection

This step aims to select the candidate countermeasures to target the risk factors associated with pedestrian collisions. The candidate countermeasures can be identified through some well-known sources such as the NCHRP report 500 series (Zegeer et al. 2004), the FHWA crash modification factor (CMF) Clearinghouse, and the Highway Safety Manual (HSM). These sources develop detailed information about the potential countermeasures, like target crash type (e.g., vehicle, bicycle, pedestrian crashes) and target crash severity levels (fatal, injury crashes). In addition, to identify candidate countermeasures, the feasibility of implementation should be considered, such as the impact of environmental issues. For example, it may not be feasible to construct a sidewalk along some roadways due to insufficient structural room. Hence, a variety of factors should be considered to identify and select potential countermeasures.

Project Prioritization

This study adopted a mixed linear programing (LP) problem that accounts for safety and economic concerns simultaneously. The objective function is to minimize the crash risk index at hotspot locations. The LP problem is formulated through Equations (38) to (43):

$$Min\ Z = \sum_{j} CRI_j \tag{38}$$

s.t. $$CRI_j = \{\textstyle\sum_k \hat{y}_{jk} \times \rho_k \times (\ 1 - \sum_l CRF_{lk} \times X_{lj})\} / \left(\sum_{i \in neighbors_j} P_i\, W_{ij}\right), \tag{39}$$

$$\textstyle\sum_j \sum_l C_{lj} X_{lj} \leq Budget, \tag{40}$$

$$\textstyle\sum_l X_{lj} \leq 1, \forall j, \tag{41}$$

$$X_{lj} + a_{lj} \leq 1, \forall j\ and\ \forall l, \tag{42}$$

$$X_{lj} = 0\ or\ 1, \tag{43}$$

where CRI_j is the crash risk index at location j; $\hat{y}_{jk}$ is the expected crash frequency of severity level *k* at location *j*; ρ_k and C_{lj} represent the cost associated with severity level *k* and countermeasure *l* at location *j*, respectively; a_{lj} is a binary variable that takes 1 if countermeasure *l* at location *j* is already installed, and 0 otherwise; X_{lj} is the decision variable that takes 1 if countermeasure *l* at location *j* is selected, and 0 otherwise; and CRF represents the crash reduction factor.

Equation (38) is an objective function that minimizes the crash risk index at hotspot locations. Equation (39) calculates the safety improvement, which is the difference between current crash risk index and crash risk index after implementing countermeasures. Equation (40) ensures the costs associated with selective countermeasures do not exceed the available budget. Equation (41) ensures only one countermeasure is selected at any location. Equation (42) ensures the existing countermeasures are not selected again for each location. Equation (43) defines the binary decision variable such that 1 means that countermeasure is selected and 0 means not selected.

This methodology has two main advantages over previous methodologies (Saha and Ksaibati 2016; Sadeghi and Mohammadzadeh Moghaddam 2016; Mishra et al. 2015) in the literature. First, it gives priority to the hotspot locations with a higher crash risk index. According to the definition of a crash risk index, a higher crash risk index indicates not only that a location is prone to pedestrian crashes, but also that its neighboring locations are at high risk for pedestrian crash occurrence. Second, the proposed methodology distinguishes between different severity levels. This prioritizes safety projects at locations with a higher chance of fatality.

CHAPTER FIVE

RESULTS AND ANALYSIS

This chapter discusses the results of applying the methodologies developed in Chapter Four of this dissertation.

Pedestrian Exposure Estimation

Statistical Models

The negative binomial (NB), zero-inflated NB (ZINB), and finite mixture NB (FMNB) models were used to explain the number of walk trips at the household level while controlling for household characteristics, socio-economic factors, and transportation network attributes for the block group where each household resides. The NB, ZINB, and FMNB models were implemented using “MASS” (Venables and Ripley 2002), “pscl “ (Zeileis et al. 2008), “flexmix” (Leisch 2004) packages in R, an open source software for statistical computing, with results summarized in Table 11.

The number of components for the finite mixture model should be selected such that the goodness-of-fit measures are optimized (Park et al. 2014). Therefore, a series of models with different numbers of components were applied, and the model with two components was selected due to its better performance on AIC metric. The results revealed that as number of components increases, the model’s goodness-of-fit decreases. For example, the AIC for two-component and three-component models were equal to 62,049.2 and 62,261.8.

Table 11 also reports the elasticity averaged over the households. Elasticity is defined as the percentage change in the average walking trip counts in response to a one percent change in the explanatory variable, mathematically:

$$Elasticity(y) = \frac{\partial \mu}{\partial x}\frac{x}{\mu} = \beta * x \quad (44)$$

Elasticity for the FM model is obtained by a weighted average of the components' elasticities.

The overall goodness-of-fit (pseudo R-squared) suggests that models can explain about 14 percent of the variation in the daily walk trips of a household, which is comparable with the R-squared values reported in prior NHTS studies. Kim and Susilo (2013) developed a negative binomial model to estimate pedestrian trip frequency using NHTS 2001 Baltimore add-on data, which included 3,519 sample households. The pseudo R-squared for their model was equal to 0.13. (Mwakalonge 2012) also developed a linear regression to estimate walk trip frequency using a NHTS 2009 data set including a sample of 109,321 households. The obtained R-squared was approximately 0.09. The low value of R-squared reported might be attributed to the inherent bias of NHTS data toward denser areas (Sullivan and Dowds 2012).

Table 11 also shows a comparison between NB, ZINB, and FMNB model results according to different goodness-of-fit measures. The pseudo R-squared and RMSE showed that the FMNB model outperforms the NB model. A difference greater than 2 or 4 between information-based criteria (AIC) is strong evidence to show that the model with lower value performs much better than the other model. Accordingly, AIC also confirmed the FMNB improved the walk trip frequency estimation over the NB model.

Table 11. NB, ZINB, and FMNB models of HB walking trips.

Variable		Negative Binomial				Zero-inflated Negative Binomial				Finite Mixture Negative Binomial			
		Coefficient	Pr(>|z|)	Significant	Elasticity	Coefficient	Pr(>|z|)	Significant	Elasticity	Coefficient	Pr(>|z|)	Significant	Elasticity
		-*				Count Model				Component 1, weight=0.84			
Constant		-4.6e-01	8.7e-06	***		7.3e-01	< 2e-16	***	-	-2.310	0.0003	***	-
Household Characteristics													
Log of household size		4.2e-01	1.9e-10	***	0.30	3.1e-01	6.0e-13	***	1.18	1.7508	2.0e-06	***	1.31
Vehicle count		-1.6e-01	< 2e-16	***	-0.40	-2.8e-02	0.0048	**	-0.47	-5.4e-01	9.7e-09	***	-1.22
Average age		-1.2e-02	< 2e-16	***	-0.67	-3.8e-04	0.7007		-0.76	-9.9e-02	<2e-16	***	-5.50
Household income (dollar/year)	Low	-5.7e-01	< 2e-16	***	-0.14	-1.9e-02	0.5437		-0.20	-2.9e-01	0.2210		-0.07
	Median	-2.5e-01	5.0e-12	***	-0.13	-3.6e-02	0.1154		-0.13	-4.5e-01	0.0136	*	-0.22
	High	-	-	-	-	-	-	-	-	-	-	-	-
# of adults		1.6e-01	4.4e-05	***	0.32	-2.2e-02	0.3186		0.33	6.8e-01	2.1e-07	***	1.32
# of workers		-1.4e-02	0.5349		-0.01	-4.4e-02	0.0019	**	-0.01	3.3e-01	0.0007	***	0.31
Block group Characteristics													
Population Density		4.0e-05	8.0e-06	***	0.03	9.1e-06	0.0745		0.04	6.3e-05	0.0351	*	0.05
Land use	Commercial	4.9e-01	0.0454	*	0.01	4.9e-01	0.0003	***	0.02	2.1945	0.0135	*	0.05
	Industrial	1.7e-01	0.7710		0.02	-1.3e-01	0.7010		0.01	1.9608	0.3477		0.02
	Mixed	1.1e-01	0.7765		0.01	-2.8e-01	0.2400		0.01	7.56e-01	0.6125		0.01
	Residential	2.2e-01	0.0433	*	0.03	1.0e-01	0.0293	*	0.10	2.51e-01	0.4317		0.03
Setting	Small Urban	1.9e-01	5.6e-06	***	0.10	3.2e-02	0.2215		0.10	5.88e-01	0.0025	**	0.15
	Rural	-	-	-	-	-	-	-	-	-	-	-	-
Regional effect	West	3.3e-01	5.1e-09	***	0.03	1.1e-02	0.7564		0.02	6.15e-01	0.0164	*	0.05
	Southwest	-2.3e-02	0.5886		-0.01	-2.9e-02	0.3180		-0.01	1.47e-01	0.4901		0.03
	Midwest	9.2e-02	0.0568	.	0.01	-2.2e-02	0.4973		0.02	5.07e-01	0.0497	*	0.06
	Northeast	3.2e-01	2.0e-13	***	0.10	7.1e-02	0.0102	*	0.05	9.69e-01	4.9e-06	***	0.15
	Southeast	-	-	-	-	-	-	-	-	-	-	-	-

Table 11 Continued.

		-				Zero Model				Component 2, weight=0.16			
Constant		-	-	-	-	1.047	< 2e-16	***	-	8.86e-01	< 2e-16	***	-
Household Characteristics													
Log of household size		-	-	-	-	-2.2e-01	0.0002	***	-	2.23e-01	2.2e-06	***	0.16
Vehicle count		-	-	-	-	1.7e-01	< 2e-16	***	-	-4.5e-02	2.1e-05	***	-0.10
Average age		-	-	-	-	1.2e-02	< 2e-16	***	-	-2.6e-03	0.0072	**	-0.17
Household income (dollar/year)	Low	-	-	-	-	5.7e-01	< 2e-16	***	-	-2.0e-01	6.1e-09	***	-0.05
	Median	-	-	-	-	2.3e-01	2.0e-12	***	-	-8.0e-02	0.0005	***	-0.04
	High	-	-	-	-	-	-	-	-	-	-	-	-
# of adults		-	-	-	-	-2.2e-01	7.6e-10	***	-	4.38e-02	0.1072		0.08
# of workers		-	-	-	-	-5.8e-02	0.0058	**	-	-4.4e-02	0.0031	**	-0.04
Block group Characteristics													
Population Density		-	-	-	-	-4.3e-05	0.3461		-	9.84e-06	0.1122		0.01
Land use	Commercial	-	-	-	-	-1.3e-01	0.5389		-	1.88e-01	0.2426		0.01
	Industrial	-	-	-	-	-6.2e-01	0.2255		-	2.66e-02	0.9413		0.001
	Mixed	-	-	-	-	-2.8e-01	0.4177		-	-8.6e-02	0.7327		-0.01
	Residential	-	-	-	-	7.4e-02	0.0078	**	-	1.35e-01	0.0180	*	0.02
Setting	Small Urban	-	-	-	-	-2.2e-01	1.2e-08	***	-	6.12e-02	0.0244	*	0.02
	Rural	-	-	-	-	-	-	-	-	-	-	-	-
Regional effect	West	-	-	-	-	-3.6e-01	8.5e-13	***	-	7.36e-02	0.0389	*	0.005
	Southwest	-	-	-	-	-2.6e-03	0.9487		-	-1.7e-02	0.5677		-0.01
	Midwest	-	-	-	-	-1.4e-01	0.0016	**	-	9.21e-03	0.7860		0.001
	Northeast	-	-	-	-	-2.9e-01	9.1e-14	***	-	9.88e-02	0.0004	***	0.02
	Southeast	-	-	-	-	-	-	-	-	-	-	-	-
Goodness-of-fit measures													
Pseudo R square		0.12				0.14				0.15			
RMSE		0.93				0.92				0.87			
AIC		66,733.9				63,162.5				62,049.2			
BIC		66,899.9				63,484.5				62,388.6			
-2*Log Likelihood		66,695.9				65,080.0				63,971.2			
Significant. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 No. of Observations = 52,064 households (within 7,721 block groups)													

The results confirmed the size effect of the exposure term (household size): a larger household size is naturally associated with more walk trips, holding everything else constant. After controlling for exposure, trip rate (i.e., number of trips per person) increases with population density of the block group where the household resides. This finding is consistent with that of Ewing and Cervero (2010), who found that individuals are more likely to take non-motorized modes for work trips when either the residential or job location is in a densely populated location. As expected, households that own more vehicles tend to make fewer walk trips (Mwakalonge 2012). Higher-income households were estimated to make more walk trips, possibly because higher-income neighborhoods enjoy greater provisions of pedestrian facilities (Estabrooks et al. 2003; Zhu and Lee 2008). This finding echoes work by Cerin and Leslie (2008) on recreational walk trips, but contradicts with others (Agrawal and Schimek 2007; Etminani-Ghasrodashti and Ardeshiri 2016). Mature households are prone to make fewer walk trips, which is consistent with previous studies on the impact of age on walking (Hatamzadeh et al. 2014; Mwakalonge 2012). Furthermore, daily walk trips decrease with the number of workers in the household, consistent with New et al.'s (2016) work based on the 2012 Utah Household Travel Survey. In contrast, the number of adults is positively associated with more walk trips.

Machine Learning Technique

This study explored the K-NN classification algorithm to estimate pedestrian exposure using the NHTS data set. The response variable, number of walk trips per household, was grouped into five classes, as shown in Table 12.

Table 12. Walk trip frequency in rural and small urban areas.

Class	Reported Walk Trips	# of Households
1	0	42,489
2	1	1,001
3	2	4,287
4	3	243
5	4 or more	1,893

The study performed three K-NN analyses (under K=1, 2, and 3), using Weka, an open source software for machine learning algorithms (Witten et al. 1999). Only three K values were considered because the algorithm cannot predict walk trip frequency well with larger K values. For validation, 25 percent of the total households (No. of Obs. = 12,478) were randomly drawn to serve as the testing data, with the remaining 75 percent of households (No. of Obs. =37,435) used to calibrate the models, with results summarized in Table 13.

In general, K-NN can correctly predict walk frequency class 75 percent of the time, regardless of the K value. However, the prediction accuracy can be too high because the algorithm favors one class, as in cases where the data set has a preponderance of one particular class (Mujalli and De Oña 2011). Therefore, sensitivity and precision indicators are also used to assess the performance. The results suggest that 2-NN and 3-NN algorithms perform slightly better than the 1-NN algorithm for the households with zero walk trip frequency, but not as well for the prediction of households making one or more daily walk trips. Since there is a trade-off between measurements' results, the area under a Receiver Operating Characteristic (ROC) curve can be provided to measure overall performance. The ROC curve is created by plotting sensitivity versus incorrectly classified instances rate (Mujalli and De Oña 2011). It ranges from 0.5, which describes weak performance, to 1.0, which describes the perfect performance. As ROC suggests, the 1-NN algorithm has a better performance than 2-NN and 3-NN to estimate walk trip frequency.

The relative mean square error (RMSE) measure was used to compare the K-NN with statistical models. The K-NN classification method had the lowest values and the NB model had the highest values of RMSE. Therefore, this study selects the K-NN algorithm to estimate pedestrian exposure in rural and small urban areas for further analysis.

Table 13. Result summary of K-NN classification models.

Training Data							Testing Data						
K=1													
Prediction Accuracy[*]= 99.95%, RMSE=0.09							Prediction Accuracy= 74.16%, RMSE=0.32						
Observed walk trip class	**Predicted walk trip class**					**Sensitivity**	**Observed walk trip class**	**Predicted walk trip class**					**Sensitivity**
	1	**2**	**3**	**4**	**5**			**1**	**2**	**3**	**4**	**5**	
1	31833	0	1	0	0	99.98	**1**	9131	188	905	50	381	85.70
2	1	734	0	0	0	99.86	**2**	218	9	17	1	21	3.38
3	11	0	3260	0	0	99.66	**3**	847	17	95	12	45	9.35
4	0	0	0	178	0	100.00	**4**	53	4	5	0	3	0.00
5	3	0	1	0	1413	99.72	**5**	383	15	56	3	19	3.99
Precision	99.95	100.00	99.94	100.00	99.72	ROC=1.00	**Precision**	85.88	3.81	8.85	0.00	4.05	ROC=0.62
K=2													
Prediction Accuracy= 86.87%, RMSE=0.16							Prediction Accuracy= 83.49%, RMSE=0.27						
Observed walk trip class	**Predicted walk trip class**					**Sensitivity**	**Observed walk trip class**	**Predicted walk trip class**					**Sensitivity**
	1	**2**	**3**	**4**	**5**			**1**	**2**	**3**	**4**	**5**	
1	31831	0	1	0	2	99.99	**1**	10399	16	203	16	21	97.60
2	604	48	76	7	0	6.53	**2**	256	0	9	0	1	0.00
3	2734	0	532	5	0	16.26	**3**	989	2	17	5	3	1.67
4	151	0	0	27	0	15.17	**4**	63	0	2	0	0	0.00
5	1149	42	134	8	84	5.93	**5**	456	3	14	1	2	0.42
Precision	87.28	53.33	71.60	57.45	97.67	ROC=0.95	**Precision**	85.50	0.00	6.94	0.00	7.41	ROC=0.61
K=3													
Prediction Accuracy= 86.26%, RMSE=0.19							Prediction Accuracy= 83.30%, RMSE=0.26						
Observed walk trip class	**Predicted walk trip class**					**Sensitivity**	**Observed walk trip class**	**Predicted walk trip class**					**Sensitivity**
	1	**2**	**3**	**4**	**5**			**1**	**2**	**3**	**4**	**5**	
1	31518	12	250	2	52	99.01	**1**	10365	11	227	1	51	97.28

Table 13 Continued.

2	690	27	12	1	5	3.67
3	2667	3	590	2	9	18.04
4	168	0	1	9	0	5.06
5	1247	0	19	2	149	10.52
Precision	86.85	64.29	67.66	56.25	69.30	ROC=0.90

2	258	0	6	0	2	0.00
3	984	1	25	2	4	2.46
4	63	0	2	0	0	0.00
5	454	1	16	1	4	0.84
Precision	85.49	0.00	9.06	0.00	6.56	ROC=0.61

* Three metrics are used to evaluate K-NN performance:.

$$\text{Prediction Accuracy (PA)} = \frac{\sum_i t_i}{\sum_i (t_i + \sum_{j \neq i} f_{ij})} \, 100\%$$

$$\text{Sensitivity}_i = \frac{t_i}{t_i + \sum_{j \neq i} f_{ij}} \, 100\% \qquad i = 1, 2, \ldots, N$$

$$\text{Precision}_i = \frac{t_i}{t_i + \sum_{j \neq i} f_{ji}} \, 100\% \qquad i = 1, 2, \ldots, N$$

where t_i is the true classified instances of class i; f_{ij} represents the instances of class i that classified incorrectly to class j; and N is the total number of considered classes.

Crash Risk Factors Identification

Crash Severity Analysis

The order logit (OR) and order probit (OP) model were implemented using “MASS” (Venables and Ripley 2002), multinomial logit (MNL) using “mlogit” (Croissant 2018), multinomial probit (MNP) using “MNP” (Imai and Dyk 2005) packages in R, an open source software for statistical computing. The random forest (RF), Naïve Bayes, and artificial neural network (ANN) were implemented in Weka (Witten et al. 1999), an open source software for machine learning algorithms. For evaluation, about 20 percent of the total crashes (No. of Obs. = 408) were randomly drawn to serve as the testing data, with the remaining 80 percent of crashes (No. of Obs. =1,633) used to estimate the models.

Multicollinearity. Multicollinearity occurs when two highly correlated predictors are used simultaneously in a regression model that can lead to inaccurate estimates of the regression coefficients, standard errors, and insignificant p-values (Greene 2012). As a result, it causes misleading conclusions about the effects of independent variables. Since this study aims to investigate the role of factors in occurrence of a severity level, it is necessary to check for existence of multicollinearity among independent variables.

The variance inflation factor (VIF) is the most widely used measurement to determine how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related (Abdulhafedh 2017). The VIF is calculated for each factor by running a linear regression of that predictor on all the other predictors, as given below,

$$VIF_i = \frac{1}{1 - R_i^2} \tag{45}$$

where R_i^2 denotes the R^2-value obtained by regressing the *ith* factor on the remaining factors. The obtained VIF from linear regression can be used in a logistic regression model because VIF measures the relationship among the independent variables regardless of the functional forms that are used in the estimation model (Greene 2012). The VIF ranges from 1.0, which refers to no collinearity, to unbounded maximum value. As a rule of thumb, a VIF value of more than 10.0 indicates a severe multicollinearity issue, and necessitates countermeasures to reduce multicollinearity (e.g., removing variables with high VIF value) (Greene 2012; Abdulhafedh 2017; Menard 2002).

Regarding some variables that have more than one degree of freedom, such as categorical variables, Fox and Monette (1992) suggested generalized VIF (GVIF), which accounts for variables' degree of freedom (*df*) as given below,

$$GVIF_i = VIF_i^{\,1/2\times df} \tag{46}$$

The GVIF values of crash severity factors are illustrated in Table 14. Note that the threshold value for variables is equal to $10^{1/2\times df}$ to determine highly correlated variables. The variables with GVIF value of higher than the threshold are shown in red.

Table 14. GVIF measurement of crash severity factors.

Variable	GVIF	Variable	GVIF	Variable	GVIF
Driver characteristics				Built-environment characteristics	
Driver age	1.36	Driver gender	1.94	Area setting	9.52
Roadway characteristics				Centerline	6.72
Surface	1.46	Operation	3.48	# of intersections	1.15
Control	1.67	Driveway	1.61	Hospital distance	1.41

Table 14 Continued.

Shoulder	1.27	Skew	1.05	Land use share	
Sidewalk	1.53	Extension	2.17	Commercial	1.07
Bicycle	1.15	Type	1.99	Industrial	1.08
Sign	2.03	Lighting	1.31	Mixed	1.52
Opening	1.17	Median	1.31	Residential	1.28
Parking	1.16	Pavement	1.09		
Natural-environment characteristics				Natural-environment characteristics	
Pop. density	1.44	Male	1.24		
Income	1.19	Employment density	2.35	Season	1.36
Average age	1.22	Household density	1.34	Time	1.34
Traffic characteristics				Weather	1.34
AADT	1.28	Exposure	1.18	Day	1.37

As Table 14 shows, “traffic control type,” “operation,” “intersection type,” “sign presence,” “sidewalk extension,” “area setting,” “employment density,” and “centerline” factors have high GVIF values. A GVIF stepwise approach was used to remove factors with high GVIF until all GVIF values were below the desired threshold. Accordingly, after removing “operation,” “intersection type,” “area setting,” “employment density,” and “centerline,” remaining variables are not highly correlated as illustrated in Table 15.

Table 15. GVIF measurement of crash severity factors after removing highly correlated factors.

Variable	GVIF	Variable	GVIF	Variable	GVIF
Driver characteristics				Built-environment characteristics	
Driver age	1.36	Driver gender	1.94	Area setting	-
Roadway characteristics				Centerline	-
Surface	1.44	Operation	-	# of intersections	1.14
Control	1.46	Driveway	1.60	Hospital distance	1.34
Shoulder	1.26	Skew	1.05	Land use share	
Sidewalk	1.51	Extension	1.38	Commercial	1.06
Bicycle	1.14	Type	-	Industrial	1.06
Sign	1.13	Lighting	1.30	Mixed	1.42
Opening	1.15	Median	1.28	Residential	1.14

Table 15 Continued.

Parking	1.13	Pavement	1.67	Natural-environment characteristics	
Natural-environment characteristics					
Pop. density	1.39	Male	1.20		
Income	1.14	Employment density	-	Season	1.36
Average age	1.24	Household density	1.39	Time	1.34
Traffic characteristics				Weather	1.33
AADT	1.27	Exposure	1.17	Day	1.37

Statistical Models. Table 16 and Table 17 show the estimation results of the OL, MNL, OP, and MNP models, respectively. The non-injury case was selected as the reference case for MNL and MNP models. Hence, The MNL and MNP results are shown in two columns: one for the probability of injury and one for the probability of fatality. In addition to the estimated coefficients and significance levels, the tables also summarize the Odds Ratio (OR), which is defined as the ratio between two probabilities. The OR represents the effect of a variable on the odds of being involved in higher severity level crashes for a pedestrian. It is the exponential of a variable's coefficient as follows (Sasidharan and Menéndez 2014; Yu 2015),

$$OR_i = e^{\beta_i} \tag{47}$$

where β_i is the coefficient of variable i. Values of OR greater than 1 indicate that the risk of fatality will increase and values less than 1 indicate that the risk of fatality will decrease. For example, based on the OL model, Table 16 shows that the odds of being involved in higher severity level crashes are 1.083 times higher for elderly (older than 65 years) pedestrians, while holding other variables constant. The OR for the MNL model indicates that the effect of a variable on the odds of a crash can experience different

Table 16. The OL and MNL models of pedestrian crash severity.

Variable	Ordered Logit Model				Multinomial Logit Model							
					Injury Crash				*Fatal Crash*			
	Coefficient	Pr(>|z|)	Significance	OR	Coefficient	Pr(>|z|)	Significance	OR	Coefficient	Pr(>|z|)	Significance	OR
Intercept												
Non-injury | Injury	-1.9683	2.1e-14	***	-	1.8288	1.4e-07	***		0.8401	5.9e-02		-
Injury | Fatal	2.0455	2.5e-15	***	-								
Driver Characteristics												
Driver age												
<= 65	-	-	-	-	-	-	-	-	-	-	-	-
> 65	0.0801	4.2e-02	*	1.083	-	-	-	-	-	-	-	-
Driver gender												
male	-	-	-	-	-	-	-	-	-	-	-	-
female	-0.3128	1.0e-02	*	0.731	-0.136	0.37074		0.873	-0.829	0.0006		0.436
Intersection Characteristics												
Traffic Control Type												
signal	-0.9462	4.1e-04	***	0.388	-0.693	0.04723	*	0.500	-1.609	0.0011	**	0.200
two-way stop	-0.7847	2.5e-02	*	0.456	-0.838	0.04992	*	0.432	-1.405	0.0531		0.245
four-way stop	-0.5594	2.1e-02	*	0.571	-0.520	0.10261		0.594	-0.706	0.0876		0.493
none	-	-	-	-	-	-	-	-	-	-	-	-
Shoulder presence												
on 1 approach	0.9388	5.5e-09	***	2.556	0.6101	3.1e-03	**	1.841	1.5386	2.0e-08	***	4.658
on 2 approaches	0.5381	7.8e-02		1.712	0.9181	0.03583	*	2.504	1.0013	0.0812		2.721
none	-	-	-	-	-	-	-	-	-	-	-	-
Sidewalk presence												
on 1 approach	-0.0188	9.3e-02		0.981	0.1772	0.41702		0.838	-0.101	0.0229	*	1.106
on 2 approaches	-0.1281	4.0e-02	*	0.879	0.4671	0.76171		0.627	-0.201	0.0476	*	1.222
none	-	-	-	-	-	-	-	-	-	-	-	-
Bicycle lane presence												

Table 16 Continued.

on 1 approach	**0.6583**	**2.1e-03**	**	**0.932**	**1.7858**	**3.9e-05**	***	**5.964**	**2.1400**	**1.3e-04**	***	**8.499**
on 2 approaches	0.5776	1.9e-01		1.782	1.1035	0.17040		3.014	0.8738	0.5045		2.395
none	-	-	-	-	-	-	-	-	-	-	-	-
Sign presence												
on 1 approach	-1.7793	8.1e-26	***	0.168	-1.9616	< 2e-16	***	0.141	-2.912	2.2e-11	***	0.054
on 2 approaches	-2.1455	1.3e-11	***	0.117	-2.335	4.4e-09	***	0.097	-3.225	2.4e-03	***	0.039
none	-	-	-	-	-	-	-	-	-	-	-	-
Surface Condition												
wet	0.4403	1.1e-02	*	1.553	0.7008	0.01198	*	2.015	1.0394	0.0037	**	2.827
snow	-0.4034	6.0e-01		0.668	-0.122	0.90795		0.885	-1.9055	< 2e-16	***	0.149
ice	-1.5484	4.0e-02	*	0.212	-0.929	0.24402		0.394	-1.4967	< 2e-16	***	0.225
clear	-	-	-	-	-	-	-	-	-	-	-	-
Traffic Characteristics												
AADT	-0.1771	6.8e-03	**	0.837	-0.254	0.00069	***	0.775	-0.192	0.0588		0.825
Ped. Exposure	0.0764	2.3e-02	*	1.079	0.3677	0.00445	**	1.444	0.2845	0.1118		1.329
Built Environment Characteristics												
Hospital	0.3578	3.8e-08	***	1.430	-0.404	3.3e-08		0.667	-0.586	3.6e-04		0.556
Land use												
Undeveloped	-	-	-	-	-	-	-	-	-	-	-	-
Commercial	0.0500	4.4e-02	*	1.051	0.0301	0.72178		1.031	0.0943	0.4747		1.098
Industrial	0.0695	2.3e-02	*	1.072	0.4039	0.00463	**	1.497	0.2903	0.0943		1.336
Residential	-0.0115	4.2e-02	*	0.988	0.2251	0.00867	**	1.252	-0.120	0.3746		0.886
HH Density	-0.2156	2.6e-03	**	0.806	-0.423	0.00002	***	0.655	-0.097	0.3316		0.907
Natural Environment Characteristics												
Time												
down	0.2753	3.2e-01		1.316	0.3520	0.33984		1.421	0.5425	0.5307		1.720
dusk	0.1327	5.5e-01		1.141	0.3034	0.16218		1.354	0.4729	0.8449		1.604
dark	0.4841	7.3e-04	***	1.622	0.1758	0.15172		1.192	0.2525	0.0004	***	1.287
daylight	-	-	-	-	-	-	-	-	-	-	-	-

Note: Number of Observations=1,814.

OL: Log Likelihood= -1,001.69, AIC= 2,059.37, and Pseudo R^2=0.139.

MNL: Log Likelihood= -929.62, AIC= 1,963.23, Pseudo R^2=0.200, and Non-injury crash used as reference case.

Significant. Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’.

Table 17. The ordered Probit and multinomial Probit models of pedestrian crash severity.

Variable	Ordered Probit Model			Multinomial Probit Model					
				Injury Crash			*Fatal Crash*		
	Coefficient	Pr(>|z|)	Significance	Coefficient	Pr(>|z|)	Significance	Coefficient	Pr(>|z|)	Significance
Intercept									
Non-injury | Injury	-1.1357	6.6e-16	***	0.8805	1.5e-05	***	0.3948	0.36055	
Injury | Fatal	1.1859	4.0e-17	***						
Driver Characteristics									
Driver age									
<= 65	-	-	-	-	-	-	-	-	-
> 65	0.0200	4.2e-02	*	-	-	-	-	-	-
Driver gender									
male	-	-	-	-	-	-	-	-	-
female	-0.1845	6.6e-03	**	-	-	-	-	-	-
Intersection Characteristics									
Traffic Control Type									
signal	-0.5138	4.7e-04	***	-0.3812	0.0599		-0.5463	0.00864	**
two-way stop	-0.4011	3.8e-02	*	-0.3889	0.1054		-0.4735	0.05896	
four-way stop	-0.2860	3.0e-02	*	-0.2946	0.0829		-0.2975	0.07220	
none	-	-	-	-	-	-	-	-	-
Shoulder presence									
on 1 approach	0.5192	1.7e-09	***	0.3222	0.0449	*	0.51795	0.00121	**
on 2 approaches	0.3295	5.4e-02		0.4828	0.0408	*	0.46137	0.05396	
none	-	-	-	-	-	-	-	-	-
Sidewalk presence									
on 1 approach	-0.0098	9.2e-01		-0.1219	0.2922		-0.0160	0.89132	
on 2 approaches	-0.0595	3.2e-02	*	-0.2855	0.0116	*	-0.0739	0.64165	
none	-	-	-	-	-	-	-	-	-
Bicycle lane presence									

Table 17 Continued.

on 1 approach	**0.4082**	**8.5e-04**	***	**0.9617**	**5.2e-05**	***		**0.91431**	**0.00301**	**
on 2 approaches	0.3432	1.9e-01		0.6876	0.1492			0.63467	0.17975	
none	-	-	-	-	-	-		-	-	-
Sign presence										
on 1 approach	-1.0088	6.9e-25	***	-1.0592	1.0e-14	***		-1.0154	0.00024	***
on 2 approaches	-1.2521	1.9e-11	***	-1.3122	2.6e-09	***		-1.2054	0.00639	**
none	-	-	-	-	-	-	-	-	-	-
Surface Condition										
wet	0.2501	1.3e-02	*	0.6103	0.2256			2.1571	0.0123	*
snow	-0.2747	5.3e-01		-0.1121	0.3145			-1.5467	1.2e-03	***
ice	-0.9140	4.8e-02	*	-0.8751	0.5721			-2.4328	34e-04	***
clear	-	-	-	-	-	-		-	-	-
Traffic Characteristics										
AADT	-0.0905	1.0e-02	*	-0.1371	0.0013	**		-0.0905	0.07281	
Ped. Exposure	0.0416	2.4e-02	*	0.1540	0.0080	**		0.11558	0.08176	
Built Environment Characteristics										
Hospital	0.1945	9.4e-08	***	-0.2175	1.8e-08	***		-0.2051	0.00170	**
Land use										
Undeveloped	-	-	-	-	-	-	-	-	-	-
Commercial	0.0288	4.3e-02	*	0.0170	0.7414			0.02958	0.56502	
Industrial	0.0395	2.4e-02	*	0.1874	0.0030	**		0.14733	0.04419	*
Residential	-0.0086	8.2e-01		0.1357	0.0043	**		0.02167	0.75247	
HH Density	-0.1081	3.4e-03	**	-0.1936	4.2e-10	***		-0.1039	0.10617	
Natural Environment Characteristics										
Time										
down	0.1204	4.4e-01		0.2133	0.2724			0.17885	0.35156	
dusk	0.0784	5.3e-01		0.2596	0.1120			0.14000	0.44289	
dark	0.2709	5.0e-04	***	0.1226	0.2987			0.28110	0.01848	*
daylight	-	-	-	-	-	-		-	-	-

Note: Number of Observations=1,814.

OP: Log Likelihood= -1,002.43, AIC= 2,060.85, and Pseudo R^2=0.137

MNP: Log Likelihood= -943.28, AIC= 1980.58, Pseudo R^2=0.189, and Non-injury crash used as reference case.

Significant. Codes: 0 '***' 0.001 '**' 0.01 '*'.

severity levels relative to reference level (or non-injury crash). For example, the odds of an injury crash compared to a non-injury crash are 2.01 times higher on wet surfaces, while holding other variables constant. Similarly, the odds of a fatal crash are 2.827 times higher on wet surfaces relative to non-injury crashes.

The overall goodness-of-fit (pseudo R-squared) suggests that the OL and MNL models can explain about 16 and 23 percent, respectively, of the variations in the injury severity of pedestrian crashes. According to the Akaike Information Criterion (AIC), the MNL model outperforms the OL models. Note that a difference greater than 2 or 4 between AIC is strong evidence to show that the model with the lower value outperforms the other model. This might be due to increased flexibility of the MNL specification model as compared to the OL model, because MNL allows the independent variables to have different effects on different levels of response variables. Moreover, as Table 17 shows, the probit versions of the models have a weaker performance and do not show any improvement over logit models to estimate the severity level of pedestrian crashes. Therefore, this study used logit models for further analysis.

In terms of driver demographics, elderly and male drivers positively correlated with fatal and serious injury levels of a pedestrian in the event of a collision. This is as expected, because elderly drivers require longer reaction time (Rifaat et al., 2011) and male drivers are more susceptible to speeding (Kim et al., 2017; Dai 2012), which would increase severity levels.

Among roadway characteristics, traffic control type, sidewalk presence, shoulder presence, bicycle lane presence, sign presence, and surface condition were found to influence crash severity in statistically and practically significant ways. As expected, traffic signalization can decrease the propensity of a fatal crash since drivers reduce vehicle speed and drive more cautiously when approaching a signalized intersection (Sarkar et al., 2011; Rifaat et al., 2011) than a stop-controlled or a non-controlled intersection. As expected, the risk for pedestrian injuries and fatality (when a crash happens) is decreased when sidewalks and warning signs are present, which is consistent with previous findings (Sarkar et al., 2011; Yu 2015; Rifaat et al., 2011).

However, shoulder and bicycle lane presence are estimated to increase severity levels of pedestrian crashes. Shoulders increase usable roadway width for motorists and therefore it is plausible that they encourage higher traveling speed (Zajac and Ivan 2003). Although pedestrians can walk on shoulders in the absence of sidewalks, they have little protection against motor vehicles (traveling at high speed) in the event of a collision. Moreover, the effect of a bike lane is consistent with previous studies (Bennet and Yiannakoulias 2015). The increased severity level could be due to the fact that a bicycle lane makes the interaction of roadway users more complex, especially at intersections where users are confused about their right of way (Jensen 2007).

Surface condition also showed a significant effect in injury severity for pedestrians. A wet surface is associated with an increase in the probability of fatality or severe injury, possibly as a result of skidding and vehicle loss of control (Rifaat et al., 2011). By contrast, snowy and icy surfaces are associated with a decrease in severity levels. This is as expected as drivers often reduce speed and pay more attention to roadway conditions (Rifaat et al., 2011), although such weather conditions tend to increase overall crash frequency (Li and Fernie 2010).

In terms of traffic characteristics, higher AADT is associated with lower risk of serious injury or fatality, which might be due to lower vehicular speed and cautious driving on congested roadways (Moudon et al. 2011). Note that the results are based on severity models, as opposed to frequency models. In a frequency model, one would expect that higher AADT values correlate positively with higher crash frequency. The severity models revealed the effects of AADT on the severity of a crash, provided that a crash happens.

However, higher pedestrian exposure is associated with a propensity for higher injury severity. Higher pedestrian exposure (i.e., more walk trips) could lead to crowded walk paths and longer waiting time to cross streets. As waiting time increases, more pedestrians will accept the risk of crossing streets during unsafe conditions (Tiwari et al., 2007; Brosseau et al., 2013), which may increase the risk of fatality.

Pedestrians in commercial and industrial areas are more likely to experience severe injuries and deaths. This might be attributed to the fact that the interaction between pedestrians and vehicles is more complex, and that both drivers and pedestrians might be more distracted in commercial and industrial areas (Bennet and Yiannakoulias 2015). In addition, pedestrians are less likely to experience severe injuries in residential areas and areas with greater household density, which might be due to the lower traffic speed in such areas. Moreover, the nearest (network) distance to hospitals, which is a rarely considered variable in previous studies, showed a positively related to more severe injuries.

The result of natural environment characteristics showed that injury severity was relatively higher during dark hours, which might reflect speeding, low visibility, and driving/walking under the influence of alcohol or other substances that cause impairment (Jang et al. 2013; Rifaat et al., 2011).

Machine Learning Techniques. Figure 2 shows the most important independent variables according to the Gini impurity reduction through random forest technique. The first ranked variables are "Sign," which indicates the presence of a warning sign at intersections and "Hospital," which refers to the distance of the crash location to the nearest hospital. The traffic characteristics (i.e., AADT and pedestrian volume) and network connectivity variables (i.e., total centerline miles and the number of intersections within half-mile buffer) are also known as important factors to the severity of pedestrian crashes. Land use variables, such as "Residential" and "Undeveloped," also ranked within the most important variables. Furthermore, demographic variables, such as block-group level population density and household density where the crash occurred, emerged as key variables for estimating pedestrian crash severity. Therefore, these factors should be considered in the pedestrian crash severity machine learning models.

The statistical models found driver, roadway, and natural environmental characteristics as significant factors. However, these were found to be less important via machine learning methods. These differences between the two methods may be due to the fact that they use distinctly different approaches. Statistical models take the specification of main effects and interaction terms into account; however, machine learning techniques aim to discover patterns in data without requiring explicit model specification.

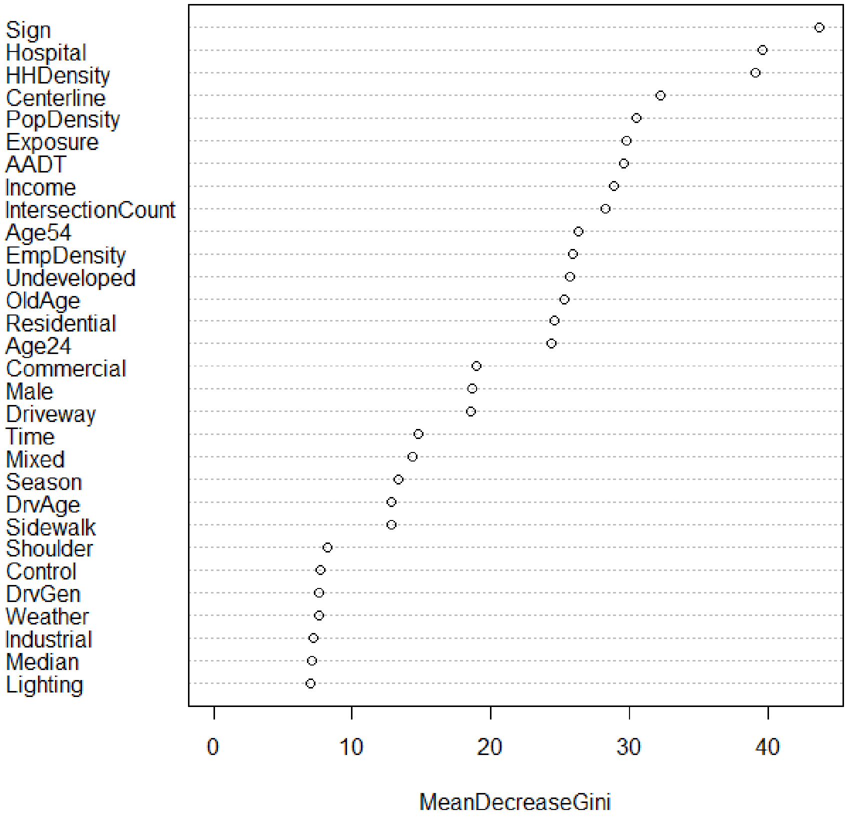


Figure 2. Variable importance score plot from random forest technique.

Table 18 provides the prediction results of the machine learning models with different evaluation indicators, including confusion matrix, prediction accuracy, sensitivity and precision. Since the ability of models in prediction is being examined, the evaluation indicators for testing data are considered. The prediction accuracy (PA) measures the percentage of instances that are correctly classified by each model (shown as the underscored values). Based on the PA values, the RF model performed moderately better than other machine learning and statistical models. However, the prediction accuracy can be too high because the algorithm favors one severity level, as in cases where the data set is overly represented a specific severity level (Mujalli and De Oña 2011). In the current data set, about 72 percent of cases are injury crashes, which shows a highly skewed data set. Therefore, sensitivity and precision metrics should also be considered to assess prediction accuracy of the models for different severity levels. The sensitivity measures the proportion of correctly classified instances to the total observed instances of a given class, whereas the precision measures the proportion of correctly classified instances to the total predicted instances for each class. The results imply that the MNL, RF, and Naïve Bayes models are the best ones to classify non-injury, injury, and fatality levels, respectively.

Since crash severity levels do not have equal monetary cost (Iranitalab and Khattak 2017), this study innovatively provides a new weighted prediction accuracy (WPA) to account for the tradeoff between overall prediction accuracy and sensitivity as given below,

$$WPA = \frac{\sum_i W_i \times ({}^{C_i}/_{N})}{\sum_i W_i} \tag{48}$$

where N_i is the total number of instances of severity level i, C_i represents the number of correctly classified instances of crash severity level *i*, and W denotes the weight factor that is estimated through average monetary costs of each crash severity level. The weight factors based on insurance/hospital-reported crash costs (

Table *10*, FHWA 2010) are used to convert estimated severity level into the equivalent property damage only (PDO) crash severity.

Table 18. Result summary of OL, MNL, RF, Naïve Bayes, and ANN models (A=non-injury, B=injury, C=fatal).

Training Data					Test Data				
OL Model, Prediction Accuracy = 71%					OL Model, Prediction Accuracy = 70%				
Observed crash severity	**Predicted crash severity**				**Observed crash severity**	**Predicted crash severity**			
	A	**B**	**C**	**Sensitivity**		**A**	**B**	**C**	**Sensitivity**
A	129	213	0	37.7	**A**	35	56	0	38.5
B	65	895	1	93.1	B	13	217	0	94.3
C	3	146	0	0.0	C	2	39	0	0.0
Precision	65.5	71.2	0.0	ROC=62%	**Precision**	70.0	69.6	0.0	ROC=63%
MNL Model, Prediction Accuracy = 73%					MNL Model, Prediction Accuracy = 70%				
Observed crash severity	**Predicted crash severity**				**Observed crash severity**	**Predicted crash severity**			
	A	**B**	**C**	**Sensitivity**		**A**	**B**	**C**	**Sensitivity**
A	148	193	1	43.3	**A**	43	47	1	47.3
B	59	894	8	93.0	**B**	13	216	1	93.9
C	2	132	15	10.1	**C**	2	132	15	10.1
Precision	70.8	73.3	62.5	ROC=66%	**Precision**	74.1	54.7	88.2	ROC=66%
RF Model, Prediction Accuracy = 99%					RF Model, Prediction Accuracy = 72%				
Observed crash severity	**Predicted crash severity**				**Observed crash severity**	**Predicted crash severity**			
	A	**B**	**C**	**Sensitivity**		**A**	**B**	**C**	**Sensitivity**
A	342	0	0	100	**A**	36	54	1	39.6
B	1	960	0	99.9	**B**	6	222	2	96.5
C	0	0	149	100	**C**	3	34	4	9.8
Precision	99.7	100	100	ROC=100%	**Precision**	80	71.6	57.1	ROC=76%
Naïve Bayes Model, Prediction Accuracy = 52%					Naïve Bayes Model, Prediction Accuracy = 54%				
Observed crash severity	**Predicted crash severity**				**Observed crash severity**	**Predicted crash severity**			
	A	**B**	**C**	**Sensitivity**		**A**	**B**	**C**	**Sensitivity**
A	150	84	108	43.9	**A**	42	19	30	46.2
B	109	488	364	50.8	**B**	27	125	78	54.3
C	5	34	110	73.8	**C**	2	11	28	68.2
Precision	56.8	80.5	18.9	ROC=70%	**Precision**	59.2	80.6	20.6	ROC=70%
ANN Model, Prediction Accuracy = 92%					ANN Model, Prediction Accuracy = 61%				

Table 18 Continued.

Observed crash severity	Predicted crash severity				Observed crash severity	Predicted crash severity			
	A	B	C	Sensitivity		A	B	C	Sensitivity
A	310	26	6	90.6	**A**	41	45	5	45.1
B	13	920	28	95.7	**B**	44	167	19	72.6
C	**12**	**30**	**107**	**71.8**	**C**	**7**	**20**	**14**	**34.1**
Precision	92.5	94.2	75.9	ROC=94%	**Precision**	44.6	71.9	36.8	ROC=69%

Figure 3 compares the overall PA and weighted PA associated with different models. Although OL, MNL, and RF have high PA values, their associated low WPA values imply that their predictions are biased in favor of low injury severity levels, because 72 percent of the observed cashes in the data set were non-injury. In contrast, the Naïve Bayes model has very good performance in predicting the fatal severity level, which is the most costly severity level. In conclusion, Naïve Bayes outperforms the statistical models and other machine learning models to predict pedestrian crash severity levels.

In sum, the RF model outperformed other models according to the ROC measurement, which is a tradeoff measurement among sensitivity values of different levels. Specifically, the area under the ROC curve measures the ability of the models to classify correctly the random objects. However, if different injury levels are not equally weighted, WSP measurement suggested that the Naïve Bayes model is the best one to predict pedestrian crash severity levels.

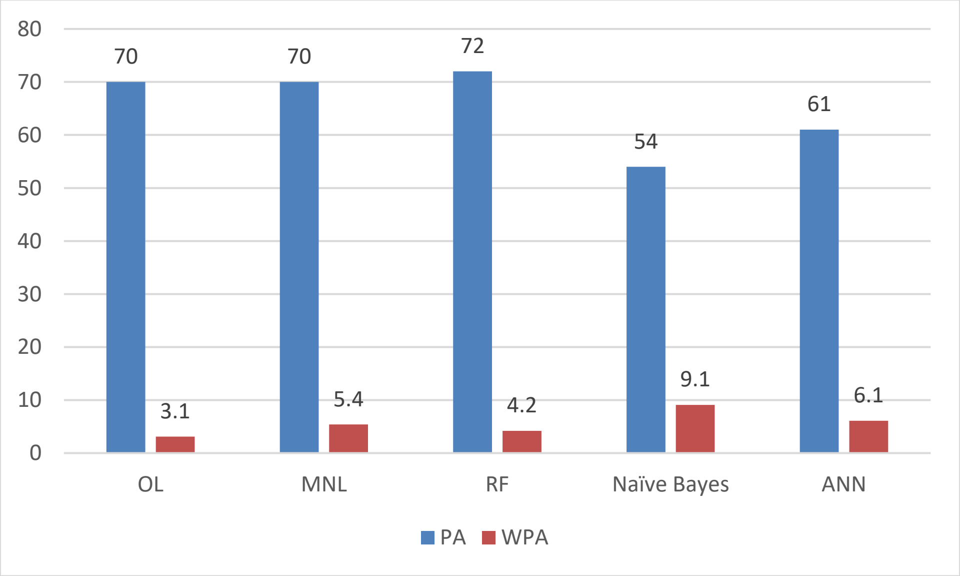


Figure 3. Overall prediction accuracy percentage and weighted prediction accuracy percentage measures.

Crash Frequency Analysis

The Negative Binomial (NB), Zero-Inflated Negative Binomial (ZINB), and Hurdle Negative Binomial (HNB) models were used to explain number of pedestrian crashes while controlling for intersection characteristics, and socioeconomic and transportation network attributes for the block group where each intersection is located. The NB, ZINB, and HNB models were implemented in R using “MASS” (Venables and Ripley 2002) and “pscl “ (Zeileis et al. 2008) packages. For evaluation, about 20 percent of the total instances were randomly drawn to serve as hold-out sample, with the remaining 80 percent of instances used to calibrate the models. Additionally, multicollinearity among explanatory variables was checked due to a high number of existing factors.

Multicollinearity. Table 19 shows the GVIF values of independent variables in the crash frequency data set. Only three factors including “centerline miles,” “number of intersections” and “road type” have higher GVIF than the desired value.

Table 19. GVIF measurement of crash frequency factors.

Variable	GVIF	Variable	GVIF	Variable	GVIF
Traffic Characteristics				Built-Environment Characteristics	
AADT	1.99	Exposure	1.45	Area setting	1.32
Roadway characteristics				Centerline	3.61
Median Type	2.74	Traffic Control	1.03	# of intersections	3.46
Shoulder Type	1.09	Road Alignment	1.02	Hospital distance	1.09
Surface	1.05	Road Type	1.67	Land use share	
Socio-economic Characteristics				Commercial	1.27
Pop. density	1.67	Income	1.13	Industrial	1.01
HH density	2.07	Average age	1.13	Mixed	1.06
				Residential	1.67

A GVIF stepwise approach was used to remove factors with high GVIF until all GVIF values were below the desired threshold. The only variable with high GVIF was “road type” when the “centerline miles” factor was removed. After removing “road type,” remaining variables are not highly correlated as illustrated in Table 20.

Table 20. GVIF measurement of crash frequency factors after removing highly correlated factors.

Variable	GVIF	Variable	GVIF	Variable	GVIF
Traffic Characteristics				Built-Environment Characteristics	
AADT	1.29	Exposure	1.53	Area setting	1.27
Roadway characteristics				Centerline	-
Median Type	1.33	Traffic Control	1.02	# of intersections	1.80
Shoulder Type	1.04	Road Alignment	1.02	Hospital distance	1.09
Surface	1.05	Road Type	-	Land use share	
Socio-Economic Characteristics				Commercial	1.25

Table 20 Continued.

Pop. density	1.74	Income	1.14	Industrial	1.01
HH density	2.24	Average age	1.12	Mixed	1.05
				Residential	1.63

Statistical Models. The Hurdle NB model was implemented in R, with results presented in Table 21. Table 21 also represents the NB and ZINB model of crash frequency to evaluate the HNB model’s performance. In addition to the estimated coefficients and significance level, the Table 21 also summarized the arc elasticity, which uses the midpoint between two states to normalize the amount of change. The arc elasticity is formulated as follows,

$$Arc\ Elasticity = \frac{\Delta y}{\Delta x} \times \frac{\bar{x}}{\bar{y}} \tag{49}$$

where x and y are independent and dependent variables, respectively.

The goodness-of-fit measurements suggested that the Hurdle NB model outperformed the other two models. The pseudo R-squared of Hurdle NB was reported as 0.455, which was higher than that of NB and ZINB models, with 0.329 and 0.405, respectively. The AIC values also suggested Hurdle NB performed better than NB and ZINB to estimate pedestrian crash frequency at intersections in rural and small urban areas. Moreover, the HNB model provided the best value of RMSE, which is based on a 20% hold-out-sample.

Moreover, Vuong's test (Vuong 1989) was used to determine whether zero-inflation is present in the data. The Vuong test suggested that the zero-inflated negative binomial model is a significant improvement over a standard negative binomial model, with a very small p-value (=<0.0001). Therefore, the zero-inflated and HNB models are needed options to model the pedestrian crash frequency.

With respect to HNB models, the count model represents a negative binomial version that models the intersections with at least one crash. As expected, since only a few intersections have experienced more than one crash, there is not much of a significant relationship between dependent and independent variables. However, the zero model, which is a binary logit model for determining whether an intersection experiences a crash or not, has found more significant variables.

Table 21. The NB, ZINB, and Hurdle NB models of pedestrian crash frequency (# of observations = 15,288 intersections)

Variable	Negative Binomial				Zero-Inflated NB				Hurdle NB			
	Coefficient	Pr(>|z|)	Significance	Elasticity	Coefficient	Pr(>|z|)	Significance	Elasticity	Coefficient	Pr(>|z|)	Significance	Elasticity
	-				Count Model				Count Model			
Constant	2.272	4.7e-14	***	-	-0.468	0.19042		-	-10.883	0.3147		-
Exposure Term												
log (AADT × Ped. Exposure)	0.310	< 2e-16	***	0.310	0.131	< 2e-16	***	0.131	0.043	0.6067		0.043
Roadway Characteristics												
Traffic Control												
Signal	-0.719	< 2e-16	***	0.086	-0.291	0.00287	**	0.035	0.8642	0.0045	**	0.104
Stop Sign	-1.483	< 2e-16	***	0.549	-0.867	1.11e-15	***	0.321	-1.515	0.0433	*	0.561
Yield Sign	-1.692	8.0e-14	***	0.102	-0.985	0.00025	***	0.059	0.226	0.8351		0.014
None	-	-	-	-	-	-	-	-	-	-	-	-
Shoulder Presence												
available	1.733	< 2e-16	***	1.022	1.619	< 2e-16	***	0.955	-0.4725	0.1614		0.279
None	-	-	-	-	-	-	-	-	-	-	-	-
Median Presence												
available	-0.126	0.0195	*	0.019	-0.133	0.02463	*	0.020	0.019	0.9646		0.003
None	-	-	-	-	-	-	-	-	-	-	-	-
Speed Limit	0.036	< 2e-16	***	1.411	0.009	0.00274	**	0.352	0.01645	0.1836		0.645
# of lanes	0.453	< 2e-16	***	1.268	0.191	3.55e-07	***	0.535	-0.1070	0.4853		0.299
Socioeconomic Characteristics												
Pop. Density	0.0001	< 2e-16	***	0.053	9.4e-05	2.31e-08	***	0.049	-8.9e-05	0.4928		0.046
Avg. Age	-0.046	< 2e-16	***	1.716	-0.002	0.00054	***	0.075	-9.6e-03	0.6389		0.358
Income	1.0e-05	< 2e-16	***	0.555	5.2e-06	4.21e-09	***	0.288	5.4e-06	5.2e-06		0.299
Built-Environment Characteristics												
# intersections	0.002	0.0159	*	0.083	4.9e-04	0.56339		0.020	5.5e-03	0.1374		0.229
Commercial area	1.528	9.5e-15	***	0.061	1.084	3.82e-07	***	0.043	1.926	0.0131	*	0.077
Residential area	0.557	0.0001	***	0.078	0.244	0.07912		0.034	0.295	0.6376		0.041

Table 21 Continued.

Variable	-				Zero Model				Zero Model			
Constant	-	-	-	-	-10.901	< 2e-16	***	-	7.108	<2e-16	***	-
Exposure Term												
log (AADT × Ped. Exposure)					1.306	< 2e-16	***	1.306	0.7155	<2e-16	***	0.716
Roadway Characteristics												
Traffic Control												
Signal	-	-	-	-	2.741	2.5e-16	***	0.329	-1.582	<2e-16	***	0.190
Stop Sign	-	-	-	-	2.689	1.3e-15	***	0.994	-2.407	<2e-16	***	0.891
Yield Sign	-	-	-	-	2.724	3.9e-05	***	0.163	-2.557	<2e-16	***	0.153
None	-	-	-	-	-	-	-	-	-	-	-	-
Shoulder Presence												
available	-	-	-	-	-1.733	1.3e-06	***	1.022	3.285	<2e-16	***	1.938
None	-	-	-	-	-	-	-	-	-	-	-	-
Median Presence												
available	-	-	-	-	0.107	0.7686		0.016	-0.174	0.0197	*	0.026
None	-	-	-	-	-	-	-	-	-	-	-	-
Speed Limit	-	-	-	-	0.088	< 2e-16	***	3.449	0.063	<2e-16	***	2.469
# of lanes	-	-	-	-	-1.339	< 2e-16	***	3.749	0.918	<2e-16	***	2.570
Socioeconomic Characteristics												
Pop. Density	-	-	-	-	-4.8e-03	< 2e-16	***	2.526	8.44e-04	<2e-16	***	0.444
Avg. Age	-	-	-	-	0.111	9.1e-12	***	4.140	-0.071	<2e-16	***	2.648
Income	-	-	-	-	-1.8e-05	0.0441	*	0.998	1.74e-05	<2e-16	***	0.965
Built-Environment Characteristics												
# intersections	-	-	-	-	-2.0e-03	0.5501		0.083	1.08e-04	0.9371		0.004
Commercial area	-	-	-	-	-2.538	0.0144	*	0.101	3.064	<2e-16	***	0.123
Residential area	-	-	-	-	1.094	0.0675		0.153	0.541	0.0187	*	0.076
Goodness-of-fit Measurements												
R-Squared	0.329				0.405				0.455			
AIC	6811.887				6076.851				5463.519			
RMSE	0.189				0.185				0.171			

In terms of roadway characteristics, traffic control type, shoulder presence, median presence, speed limit, and number of lanes were associated significantly with pedestrian crash occurrence. The results suggested that traffic signals in rural areas are conducive to pedestrian safety because drivers may operate more carefully at signalized intersections, compared with un-signalized intersections (Lee and Abdel-Aty 2005). A median decreased the chance of pedestrian crash occurrence, which can serve as a safe resting point for pedestrians while crossing roadways (Chimba and Ajieh 2017). Shoulder presence increases the probability of crash occurrence. One explanation is that shoulders increase usable roadway width for motorists, and therefore may encourage higher traveling speed (Zajac and Ivan 2003).

As expected, the speed limit was significantly positively correlated with pedestrian crashes due to shorter perception–reaction time on high speed roadways (Huang et al. 2017). Number of lanes also showed a positive effect on pedestrian crash occurrence. A higher number of lanes increases the crossing distance for pedestrians, hence prolonging the time during which pedestrians are exposed to vehicles (Aziz et al. 2013).

Moreover, traffic characteristics (i.e., AADT and pedestrian exposure) were significantly related to pedestrian crash occurrence. As traffic volume increases, the frequency of conflicts between pedestrians and motor vehicles increases when all the other factors stay constant (Lee and Abdel-Aty 2005). Therefore, it is expected that pedestrian crashes are more prevalent in areas with denser development (e.g., shopping, businesses, and hospitals).

Among socio-economic characteristics, population density and median income were positively associated, and average age was negatively associated with the occurrence of injury and fatal pedestrian crashes. Areas with higher population density increase the chance of pedestrians involved in a traffic crash (Loukaitou-Sideris et al. 2007; Siddiqui et al. 2012). Regarding income impact, Rhee et al. (2016) explained that high income areas are more likely to be served by high speed roads, leading to more crash occurrence. With respect to the impact of age, older individuals are likely to drive and walk more cautiously (e.g., lower driving speed, lower risky crossing behavior) (Wier et al. 2009).

In terms of built-environment attributes, only number of intersections and commercial land use share were identified as significant variables. As expected, as the number of intersections increases, the likelihood of crash occurrence increases, because intersections increase the pedestrian exposure to vehicles (Chen and Zhou 2016). Commercial land use was also positively correlated with pedestrian crash occurrence since such areas have a higher level of pedestrian activity, which in turn leads to more pedestrian crashes (Ukkusuri et al. 2011; X. Wang et al. 2016).

Machine Learning Models. Due to the severely unbalanced crash frequency data and relatively bad performance of regression machine learning models, the count response was converted to a categorical response that has three classes: intersections with zero crash (which represents 90 percent of the total observations); intersections with one crash (9 percent of the total observations); and intersections with two or more crashes (1 percent). As explained in methodology section, The BRF model with SMOTE technique is able to moderate the frequency of instances in different classes to produce a more balanced dataset. Table 22 summarizes the HNB prediction result versus Boosted RF prediction performance.

Table 22 Result Summary of NB, ZINB, HNB, RF, and BRF Models (A=0 crash, B=1 crash, C=2+ crashes).

Training Data					**Test Data**				
NB Model, Prediction Accuracy = 90.85%					NB Model, Prediction Accuracy = 90.72%				
Observed crash frequency	**Predicted crash severity**				**Observed crash frequency**	**Predicted crash severity**			
	A	**B**	**C**	**Sensitivity**		**A**	**B**	**C**	**Sensitivity**
A	10553	7	0	99.9	**A**	2647	4	0	99.8
B	1007	93	0	9.1	**B**	252	20	0	6.7
C	51	5	0	0.0	**C**	14	2	0	0.0
Precision	90.9	89.4	0.0	ROC=0.64	**Precision**	90.9	75.0	0.0	ROC=0.63
ZI Model, Prediction Accuracy = 90.85%					ZI Model, Prediction Accuracy = 90.72%				
Observed crash frequency	**Predicted crash severity**				**Observed crash frequency**	**Predicted crash severity**			
	A	**B**	**C**	**Sensitivity**		**A**	**B**	**C**	**Sensitivity**
A	10553	7	0	99.9	**A**	2647	4	0	99.8
B	1007	93	0	9.1	**B**	252	20	0	6.7
C	51	5	0	0.0	**C**	14	2	0	0.0
Precision	90.9	89.4	0.0	ROC=0.64	**Precision**	90.9	75.0	0.0	ROC=0.63
HNB Model, Prediction Accuracy = 90.94%					HNB Model, Prediction Accuracy = 90.78%				
Observed crash frequency	**Predicted crash severity**				**Observed crash frequency**	**Predicted crash severity**			
	A	**B**	**C**	**Sensitivity**		**A**	**B**	**C**	**Sensitivity**
A	10551	9	0	99.9	**A**	2650	1	0	99.9
B	993	103	4	9.4	**B**	253	18	1	6.6
C	51	5	0	0.0	**C**	13	3	0	0.0

Table 22 Continued.

Precision	**90.9**	**88.0**	**0.0**	**ROC=0.66**	Precision	**90.8**	**81.8**	**0.0**	**ROC=0.64**
RF Model, Prediction Accuracy = 100.0%					RF Model, Prediction Accuracy = 99.42%				
Observed crash frequency	**Predicted crash severity**				**Observed crash frequency**	**Predicted crash severity**			
	A	**B**	**C**	**Sensitivity**		**A**	**B**	**C**	**Sensitivity**
A	10569	0	0	100	**A**	2642	0	0	100
B	0	1098	0	100	**B**	2	272	0	99.3
C	0	0	57	100	**C**	0	15	0	0.0
Precision	100	100	100	ROC=1.00	**Precision**	99.9	94.7	0	ROC=0.90
BRF Model, Prediction Accuracy = 89.7%					BRF Model, Prediction Accuracy = 89.6%				
Observed crash frequency	Predicted crash severity				Observed crash frequency	Predicted crash severity			
	A	**B**	**C**	**Sensitivity**		**A**	**B**	**C**	**Sensitivity**
A	10378	0	191	98.2	**A**	2600	0	42	98.4
B	98	78	922	7.1	**B**	26	15	233	5.5
C	0	0	57	100	**C**	1	0	14	93.3
Precision	99.1	100	4.9	ROC=0.83	**Precision**	98.9	100	4.8	ROC=0.82

The results indicated that the RF model outperformed the other models to estimate the crash frequency. However, the RF model prediction results, like statistical models, are biased toward the zero-crash and one-crash classes, which are the most prevalent classes in this particular data. While the BRF model predicted the two-crash class very well, neither BRF nor HNB models produced reasonable accuracy for the two-or-more-crash category.

Hotspot Locations Identification

The data from the City of Bozeman, Montana, a small urban area, are used as a case study. The Montana Department of Transportation (MDT) reported 15 intersection-related pedestrian crashes from 2011 to 2013. The crash locations were coded with latitude and longitude coordinates and visualized in a GIS environment. Roadway network shapefiles are available from the U.S. Census Tiger/Line archive, and they cover the entire state (including local streets). The shapefiles were used to divide the study area into 285 grid cells (cell size 660 ft. by 660 ft.). To implement the hotspot identification method, the grid cell layer was joined spatially with the crash location layer to calculate the severity index based on Equation (37).

This study used the Hurdle negative binomial to estimate the total crash frequency and the multinomial logit model to estimate the probability of different severity levels due to their better performance than other models. The crash frequency ($\hat{y}$) of each severity level *i* at intersection *j* is formulated as below

$$\hat{y}_{ij} = \hat{Y}_j \times \hat{P}_i \tag{50}$$

where $\hat{Y}_j$ is total crash frequency at intersection *j* and $\hat{P}_i$ is the probability of severity level i.

The proposed methodology assumed that spatial dependence exists among neighboring cells, that is, severity indices are more similar among neighboring cells than they are from distant cells. To check this assumption, this study used Moran’s I to estimate the strength of spatial dependence (Xie et al. 2014; Truong and Somenahalli 2011; Soltani and Askari 2017; Flahaut et al. 2003; Yu et al. 2014). The Moran’s I test produces a z-score (to indicate the type of spatial correlation, cluster or dispersion) and the corresponding p-value (to measure whether the spatial correlation is statistically significant). However, this statistic test has a skewed distribution that can lead to erratic p-values (Xie et al. 2014). To address this issue, Anselin (2003) developed a premutation p-value, which is a more robust measure, using a random permutation test. This study used the Geoda software (Anselin 2003) and tested different orders of contiguity neighboring to select the most significant neighboring structure. It was found that the Moran's I statistics across all the grid cells were close to 0.1616, 0.0909, and 0.03198 for first-order, second-order, and third order neighboring with associated p-values of 0.001, 0.01, and 0.037, respectively. This indicates first-order contiguity neighboring has the strongest statistical significance at the 5% level.

As noted earlier, pedestrian exposure was estimated at the block-group level using the NHTS 2009 sample. To overlay those estimates onto the grid cell level, a grid cell has the same pedestrian exposure density (pedestrian exposure per land area) as the block group which it falls completely within. If a grid cell lies between two or more block groups, it takes on a weighted average of the pedestrian exposure densities of those block groups with,

$$Pdensity_i = \sum_{\substack{j \in intersected\ block \\ groups\ with\ grid\ cell\ i}} \frac{A_{ij}}{A_i} \times density_j \tag{51}$$

$$P_i = Pdensity_i * A_i \tag{52}$$

where $Pdensity_i$ and P_i note the pedestrian exposure density and pedestrian exposure, respectively; and A_{ij} indicates the area of grid cell *i* that lies within block group *j*.

Finally, the hotspot locations were identified using Equations (32) and (33) and applying the pedestrian exposure obtained from the 1-NN results, the severity index calculated by Equation (37), and the weight matrix obtained based on Equation (34). Figure 4 illustrates the hotspot map across Bozeman, which reveals the proposed methodology was able to detect unsafe locations well. Zones that have experienced pedestrian crashes (e.g., zone B) and zones that are susceptible to fatal injuries (zones A) are both identified as unsafe locations (red and orange zones). As Figure 5 shows, zone A suffers from a lack of sidewalks, warning signs, appropriate traffic control, and medians. Many locations also have shoulders present. All of these characteristics are known as pedestrian crash risk and severity factors in rural and small urban areas. Zone B suffers from the presence of some risk and severity factors (e.g., commercial and residential area, lack of warning sign); additionally, it is close to crash locations as illustrated in Figure 6. However, green zones (e.g., zone C) that experience low pedestrian exposure, are designed with safety considerations, and are far away from crash locations are identified as safe locations (Figure 7).

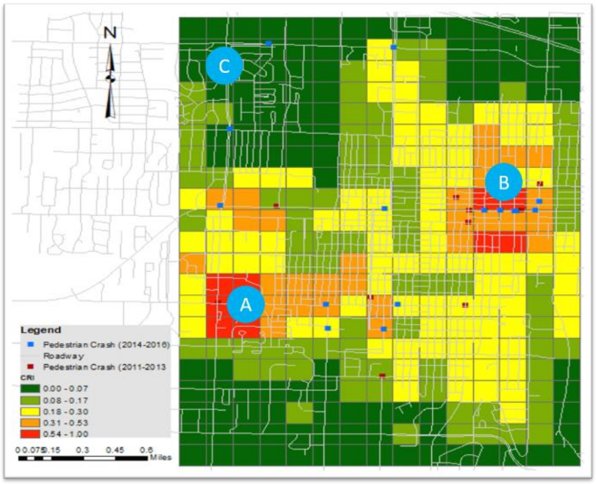


Figure 4. Hotspot locations map across Bozeman, MT.

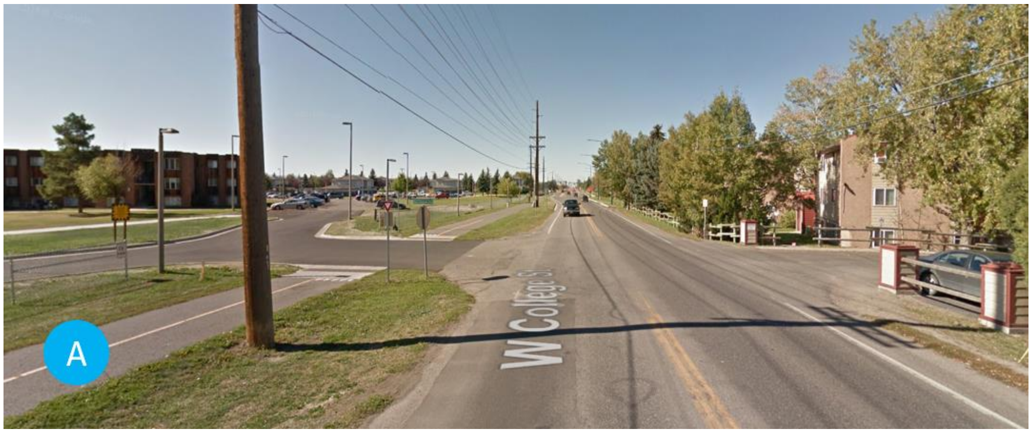


Figure 5. Hotspot location - zone A (W College St. – Grant Chamberlain Dr.)

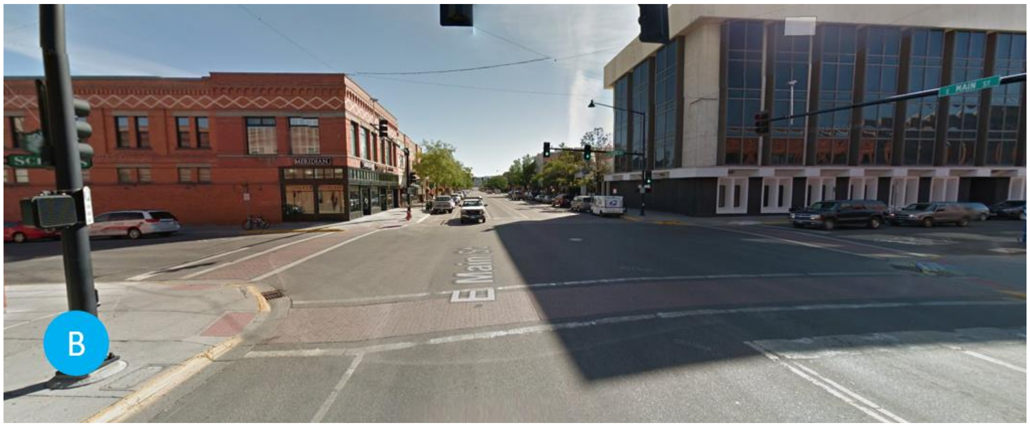


Figure 6. Hotspot location - zone B (E Main St. – Black Ave.)

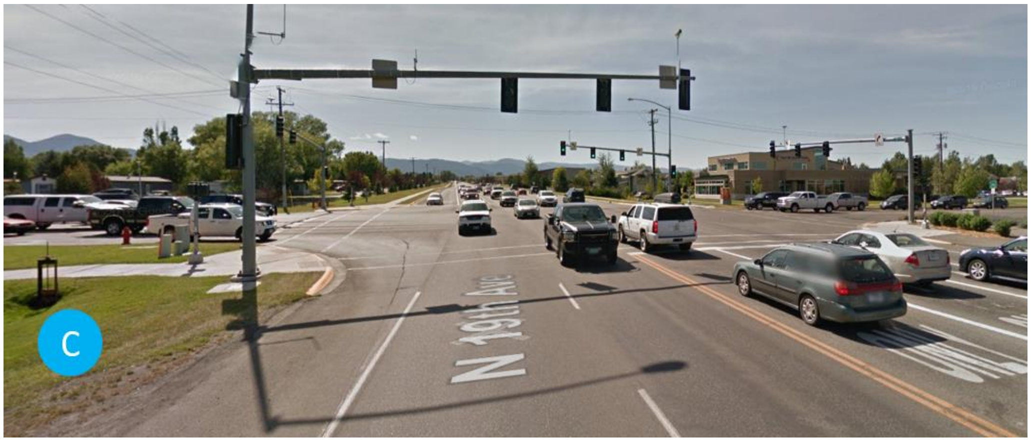


Figure 7. Hotspot location - zone C (N 19th Ave. – W Oak St.)

This study also used the pedestrian crash data set from 2014 to 2016 to validate the proposed methodology results. The crash locations were overlaid with X and Y coordinates on the hotspot map in Figure 4. It was found that about 60 percent of crash locations fall within the red and orange areas that show locations with high crash risk index values. Twenty percent of crashes are in yellow areas with medium crash risk index values and the remaining are in green areas. Hence, the proposed methodology can identify the risky sites for pedestrians with a strong level of accuracy.

Countermeasure Selection

The candidate countermeasures for each severity and risk factor, which were identified in previous sections, are shown in Table 23. For example, installing a pedestrian hybrid beacon, high-visibility crosswalk, raised pedestrian crosswalks, warning signs, and sequential flashing beacons (FB) are candidate safety improvements at locations where lack of a warning sign is a factor. Table 23 also presents the target severity and crash reduction factor (CRF) (reported by FHWA [2008]), and cost (reported by Bushell et al. [2013]) associated with each countermeasure. Note that all costs have been changed to 2012 US Dollar equivalents using the United States Consumer Price Index (Bushell et al. 2013). It is also assumed that cost values contain both construction and maintenance costs.

Table 23. Proposed countermeasures for severity and risk factors.

Risk Factor	Countermeasure	Severity	CRF (%)	Cost (USD)	Site-specific condition
Sign Presence	Install high-visibility crosswalk	All	40	2,540/ea	-
	Install raised pedestrian crosswalks	Injury	46	5000 to 7000/ea	only on local roads
	Install a warning sign, and/or sequential flashing beacons (FB)	All	39.4	Sign=300/ea, FB= 10,010/ea	-
Traffic Control Type	Convert two-way to all-way STOP control	All	39	300/stop sign	limited on arterials and collectors
	Convert un-signalized intersection to roundabout	Fatal/Injury	27	85,370/ea	-
	Install a pedestrian hybrid beacon	All	69	57,680/ea	-
Sidewalk Presence	Install sidewalk	All	74	32/ft (concrete)	-
Pedestrian exposure	Left turn prohibitions	All	10	Sign= 220/ea, diverter= 15,060/ea	needs traffic study
	Turn on red prohibition	All	88	220/ea	needs traffic study
	Pedestrian scramble	All	51	5.85/ft	
	Install pedestrian crossing (signed and marked with curb ramps and extensions)	All	37	Ramp=42/ft^2 stripped crosswalk= 5.85/ft	-
	Underpass/overpass	Fatal/injury	90	124,670 to 206,290/ea	-

Table 7 Continued.

	Modify signal phasing (Install a leading pedestrian volume)	All	5	1,480/ea	-
Speed Limit	Raised intersection	Fatal/Injury	8	50,540/ea	only on local roads
	Speed hump	All	50	2,640/ea	only on local roads
	Dynamic Speed Feedback Signs	All	5	2,500/ea	-
Median presence / Number of lanes	Refuge island	All	56	13,520/ea	-
	Raised median	All	27	7.26 / ft^2	-
Shoulder presence	Install sidewalk	All	74	32/ft (concrete)	-
Bicycle lane	Install separated bicycle lane (Bollard)	All	25.9	730/ea (10 - 40 ft Spacing)	-
AADT	Raised intersection	Fatal/Injury	8	50,540/ea	only on local roads
	Raised median	All	25	7.26/ft^2	-
	Refuge island	All	56	13,520/ea	-

Project Prioritization

The proposed mixed linear programming was applied to a set of intersections in the City of Bozeman as a case study. Since safety budgets are typically limited, only twenty intersections (out of 567 intersections) with highest crash risk index were selected for safety improvement projects. Selecting a limited number of locations is consistent with the literature. For example, Saha and Ksaibati (2016) selected 41 sites (out of 3762 sites) with higher crash frequency and Mishra et al. (2015) selected 20 intersections for each county for safety countermeasures.

The proposed methodology was implemented using the Excel Solver tool. Table 24 provides the objective function reduction for different budgets. The first column lists a range of values representing the budget available for safety countermeasures. The second column shows exactly how much the objective function is reduced if the right-hand side of the cost constraint changes. For example, if a city spends $200,000 on countermeasures instead of $100,000, the objective function reduction will increase by 10 percent. The third column indicates the percentage of the budget spent, which shows that the model uses almost the entire budget for each budget constraint level.

Decision makers can use these parameters to help select a budget allocation that provides the maximum benefit. Figure 8 illustrates the crash risk index reduction trend for different budget levels. The slope of the trend line clearly changes after it reaches a budget level of $100,000, such that the slope is greater for budgets under $100,000. Therefore, a budget of $100,000 is an optimal option for the City of Bozeman according to the model’s assumptions.

Table 24. Sensitivity analysis on cost constraint.

Cost Constraint	Objective Function Reduction Percent	Budget Usage Percent
25,000	44	99.9
50,000	52	99.8
75,000	56	96.9
100,000	61	99.8
125,000	63	99.9
150,000	64	92.2
175,000	66	99.9
200,000	70	97.7

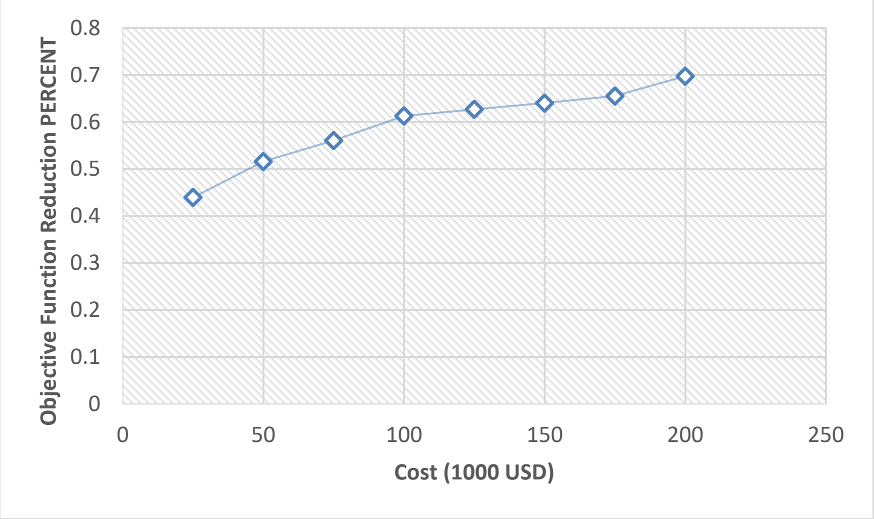


Figure 8. Optimization model’s performance for different budget levels.

Table 25 shows the recommended improvements for each intersection that can be implemented within a budget of $100,000. The optimal solution provides a 63 percent reduction in total crash risk index and uses 99 pecent of resources available. In this scenario, one countermeasure was selected for 18 intersections and no countermeasures were selected for two intersections. As expected, the proposed methodology assigned a lower priority to these two intersections (e.g., E Olive St. - S Black Ave), which are located farther away from downtown areas and experience lower pedestrian exposure.

With respect to countermeasures, a pedestrian scramble countermeasure was selected most often for signalized intersections due to its relatively higher CRF and lower cost. Pedestrian crossing with extensons was chosen at the intersections with two-way stop control. This change can reduce the vehicular speed and increase the driver's attention especially on Mendenhall Street, which is a one-way approach and located in the downtown area. The sidewalk installation was suggested for two intersections (such as W College St. - Grant Chamberlain) that are located in residential areas and experience high pedestrian exposure. Based on the results, a hybrid beacon should be installed at the intersection of W College St. and S 8th Ave., which is near campus and several commercial stores.

Table 25. Selected countermeasures for hotspot locations.

Intersection	Selected Countermeasure	Picture
W Main St. - 19th Ave.	Pedestrian scramble	
W Peach St - N Tracy Ave.	Install a warning sign	
College St. - Grant Chamberlain	Install sidewalk	
W College St. - S 7th Ave.	Install a warning sign	

Table 25 Continued.

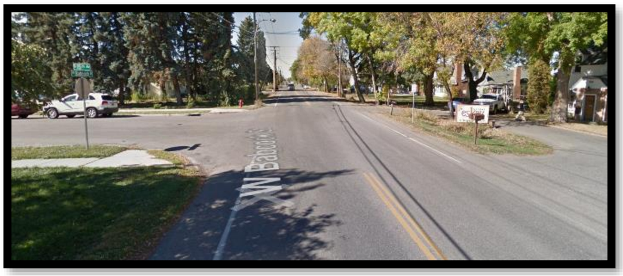
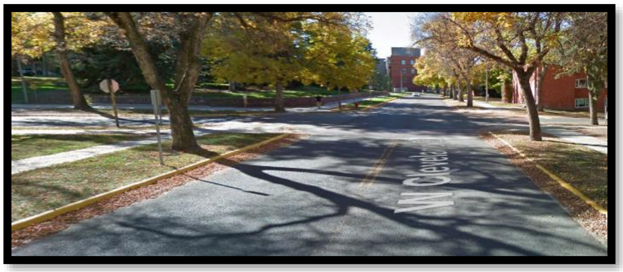
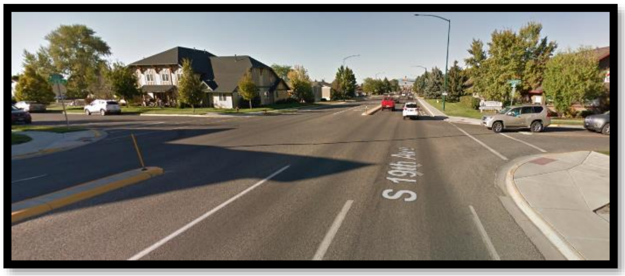
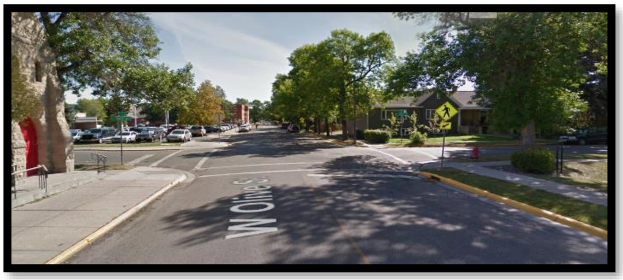
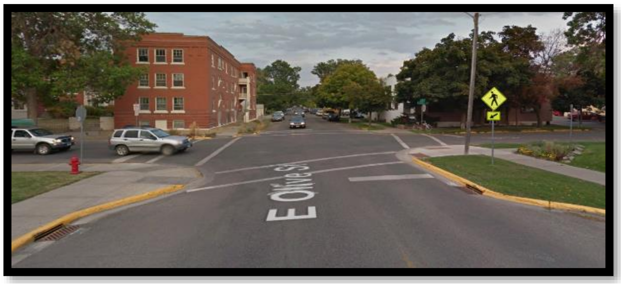
W Babcock St. - S 15 Ave.	Install sidewalk	
W Cleveland St. - S 7th Ave.	Install pedestrian crossing (signed and marked with curb ramps and extensions)	
S 19th Ave. - W Dickerson St.	Install a warning sign	
W Olive St. - S Tracy Ave	-	
E Olive St. - S Black Ave	-	

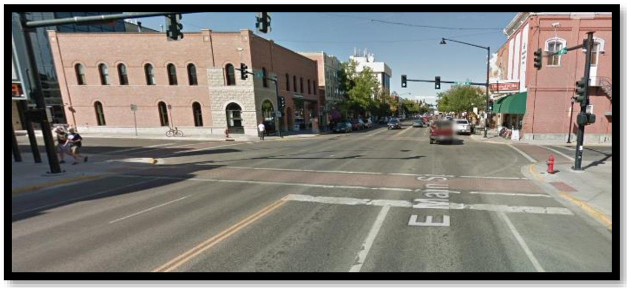
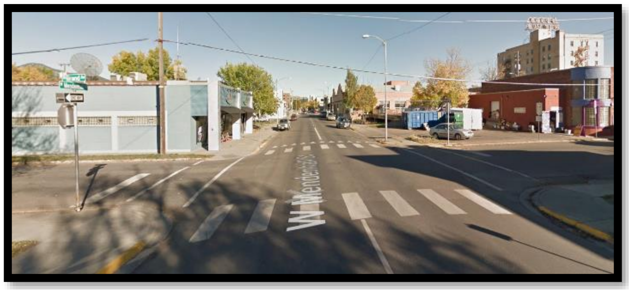
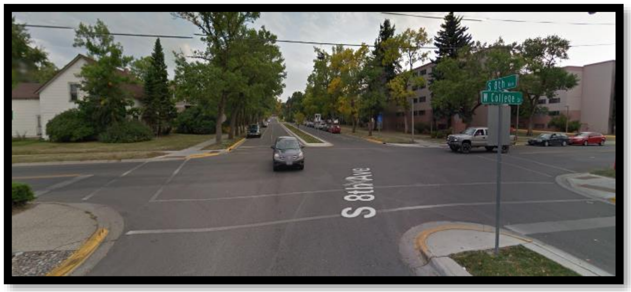
Table 25 Continued.

E Main St. - N Rouse Ave.	Pedestrian scramble	
E Mendenhall - N Black Ave.	Install pedestrian crossing with extensions	
E Mendenhall - N Bozeman Ave.	Install pedestrian crossing with extensions	
E Mendenhall - N Rouse Ave.	Pedestrian scramble	

Table 25 Continued.

W College St. - S 19th Ave.	Pedestrian scramble	
S Wilson Ave. - W Babcock St.	Pedestrian scramble	
E Main St. - N Tracy Ave.	Pedestrian scramble	
E Main St. - N Black Ave.	Pedestrian scramble	

Table 25 Continued.

E Main St. - N Bozeman Ave.	Pedestrian scramble	
W Mendenhall - N Grand Ave.	Install pedestrian crossing with extensions	
W College St. - S 8th Ave.	Install a pedestrian hybrid beacon	

In conclusion, the proposed methodology worked very well for allocating safety funding to hotspot locations in rural and small urban areas. However, the use of this methodology is not limited to these crash types or area settings. Therefore, it can be used for other crash types (e.g., vehicle crashes) or in urban areas, especially when the budget for implementing countermeasures is limited.

CHAPTER SIX

CONCLUSION

This study adopted a six-step systemic safety planning tool to customize the process for pedestrian safety in small urban and rural settings. The proposed systemic approach is a step by step process that begins by identifying target facilities and associated risk factors, and then evaluates the entire road system with a set of criteria (i.e., primary risk factors) to identify high risk locations. It then recommends cost-effective measures for those candidate locations and concludes by prioritizing locations for implementation.

The proposed methodology was implemented in the City of Bozeman, Montana, a small urban area. The results validated that the methodology is able to identify hotspot locations well and suggest reasonable countermeasures to improve pedestrian safety across the entire network. In addition, this safety tool is easy to implement and can be used to develop a spreadsheet tool to facilitate applications in local transportation agencies. The products will assist local safety improvement programs in rural and small urban areas to effectively improve pedestrian safety, with modest requirements for input parameters and computing resources.

Generally, this study contributes to the literature in five ways:

- This study provided an area-level pedestrian exposure metric for rural and small urban areas. Several studies have emphasized developing estimation models for pedestrian exposure. These studies chiefly targeted urban areas, with one exception found in Ivan et al. (2001), which focused on rural areas. Ivan et al. (2001) estimated the weekly crossing pedestrian volume at rural intersections by controlling for site characteristics (e.g., sidewalk provision and traffic control type), median household income, area type (e.g., downtown area and residential area), and road attributes (e.g., number of lanes and lane width) in 32 intersections from rural areas in Connecticut. However, this study has used the data from only 32 intersections from rural areas in Connecticut, which limits the application of their model. Moreover, Ivan et al. (2001) study has been done in 2001, which shows new study like this dissertation is needed to analyze most recent pedestrian trip patterns in traffic networks. These limitations make us develop a more generalizable exposure estimation tool. Since, the data (e.g., pedestrian count data) is usually limited in rural areas, an area-based metrics such as number of trips are more appropriate to estimate pedestrian exposure in such areas.

The proposed exposure estimation methodology can be generalized to other rural areas thanks to the national travel behavior data used to calibrate the models and the standard covariates that can be easily accessed from the U.S. Census data.

- This study employed statistical models and machine learning techniques, as opposed to trends and descriptive analysis (which failed to account for correlations among the many factors that influence pedestrian crash severity and rates, as used in previous systemic safety studies [Preston et al. 2013; AASHTO 2010]). This dissertation work also collected road and land use factors across 2,200 intersections through Google

Street Views, a much larger and more detailed sample than previous studies [Preston et al. 2013; AASHTO 2010; Walden et al. 2015].

- This study provided an innovative hotspot identification framework based on a two-step floating catchment area (2SFCA) method. As noted in the literature, a good hotspot identification method should control for a range of crash contributory factors (e.g., road geometry and traffic exposure), while considering different severity levels and spatial heterogeneity that are unique to crash counts. However, previous methods only consider one of these factors simultaneously. For example, Getis-Ord Gi only consider spatial dependence but ignores the exposure and roadway characteristics.

  Kocatepe et al. (2017) first used the Gaussian-based two-step floating catchment area (2SFCA) method to estimate crash propensity while accounting for spatial heterogeneity and population (as a proxy for crash exposure). However, Kocatepe et al. (2017) have overlooked three important issues. Firstly, population does not necessarily represent pedestrian exposure, because not all residents in a block group walk or drive. Secondly, Census block groups, employed in Kocatepe et al. (2017) as the analysis unit, may be too large for crash analysis because traffic volume, network connectivity, and other common contributory factors can change measurably within a block group. Thirdly, the proposed methodology accounted for severity index, which distinguishes the effects of different severity levels. A case study conducted in Bozeman showed that the 2SFCA technique can more effectively pinpoint high-risk locations.

- This study innovatively estimated the severity index based on an Empirical Bayesian approach. Previous studies (Truong and Somenahalli 2011; Manepalli et al. 2011) calculated the severity index only based on observed crash frequency, which ignores locations that harbor crash risk factors but have not experienced any crashes (due to short observation periods or sheer statistical randomness). However, the EB approach is able to reveal locations that have experienced high crash frequency and/or have high severity potential.
- Lastly, another contribution of this study is to inject an optimization program into the systemic safety planning process to achieve optimal allocation of limited budget. Coded in an Excel spreadsheet, the optimization program developed in this dissertation work can handle 20 sites (intersections) across large geography. According to the definition of a crash risk index, a higher crash risk index indicates not only that a location is prone to pedestrian crashes but also its neighboring locations are at high risk for pedestrian crash occurrence. This also gives priority to the hotspot locations with higher pedestrian exposure. Moreover, the proposed methodology distinguishes between different crash severity levels. This prioritizes the safety projects at locations with higher chance of fatality. Previous studies usually ranked the safety projects based on crash frequency or total cost (Saha and Ksaibati 2016; Cook and Green 2000).

The following sections separately summarize the major findings, contributions, and limitations associated with different steps of the proposed systemic tool.

Pedestrian Exposure Estimation

This step addresses estimation of pedestrian exposure, with a focus on practices suitable for rural and small urban areas. Pedestrian exposure serves a critical role in traffic safety analysis (Loukaitou-Sideris et al. 2007; Kerr et al. 2013) and offers insight for pedestrian planning through anticipating areas or routes with higher pedestrian demand (Raford and Ragland 2006). There is a large volume of work on pedestrian exposure but limited information on what measures should be adopted under different conditions (data quality, spatial resolution, and estimation accuracy). Therefore, this study synthesized relevant literature with the goal of offering best practices and tools that are broadly useful. These tools generally fall into five groups (area-based, segment-based, point-based, distance-based, and trip-based), with strengths and limitations summarized in Table 2.

In short, area-based approaches (e.g., population density and number of walk trips) broadly apply in situations that involve limited data and coarser spatial resolutions. Exposure metrics under this category can be inferred using household travel surveys and census data. Point-based, segment-based, and distance-based approaches capture pedestrian exposure with great details (e.g., number of pedestrian crossings at midblock or intersections or walking along streets) but require detailed road information (e.g., number of lanes), pedestrian attributes (e.g., walking speed), and fine-grained traffic volume to calibrate local models or apply existing coefficients. Trip-based approaches exploit pedestrians' decisions at the microlevel. The trip-based approach can produce improved estimation accuracy but requires data that may often be unavailable (e.g., fine-grained information about origins and destinations).

In addition, this study developed an area-based model to estimate household-level walk trip frequency for rural and small urban areas using the NHTS 2009 data. In this respect, three statistical models including NB, ZINB, and FMNB and one machine learning technique (i.e., K-NN) were implemented to investigate their application. Since walk trip data suffer from over-dispersion, it was hypothesized that using more flexible models such as the FMNB, which assume over-dispersion arises from two or more components, might lead to better fits. The results also confirmed that the FMNB model outperformed the two other models, which have been widely used in previous studies to model count data. However, the statistical models' performance was limited by weak overall goodness-of-fit. The results revealed that K-NN showed improvement over the statistical models thanks to its ability to not assume any predefined relationship between the response and explanatory variables.

This model can be used in two ways for estimating pedestrian exposure. First, it provides coefficient estimates that can be used broadly to estimate walk trips across zones as small as census block groups, similar to Kerr et al. (2013) and Beck et al. (2007), but with coefficients specific to less dense areas. Second, this model provides information on trip generation at the block group level and can be inserted into the four-step travel demand procedure to estimate link-level pedestrian exposure where a pedestrian network is available (or can be reasonably approximated). The four-step travel demand model should be relatively easy to implement because pedestrian networks involve only one transportation mode (walking) and do not have volume constraints (Clifton et al. 2008).

This area-based model utilizes standard socioeconomic variables available from published data (e.g., census and NHTS), facilitating its applications in rural and small urban areas where local travel survey and network details (e.g., sidewalks) are unavailable. The model was calibrated with travel behavior data that span five different regions of the country to reduce biases associated with localized data. Admittedly, the NHTS data employed here have a few challenges, such as overrepresentation of older people, which might bias the results.

While the proposed model offers reasonable estimates with minimal data requirements, it suffers from a few limitations. First, it did not consider transit-related walk trips in part because of the inability to obtain transit data from the study area that spans a large part of the United States. Second, a more fine-grained study would be valuable to determine how walk trip decisions are modified under sidewalk provisions and natural environment characteristics (e.g., temperature) for rural and small urban areas. Third, future studies need to address the deficiencies inherent in area-based estimates and investigate ways to infer node- or link-level exposure using these estimates. One possible solution is to utilize techniques (e.g., the four-step travel demand model) to disaggregate the area-based estimates into smaller geographic units while controlling for pedestrians’ route choices [e.g., nonmotorized travelers consider level of stress when making route decisions (Kuzmyak et al. 2014)]. It would be valuable to explore these options while factoring in the data and resource constraints faced by rural and small urban areas.

Crash Risk Factors Identification

The first objective of this step was to examine the contributing factors associated with both pedestrian injury severity levels and pedestrian crash occurrence. This study identified some contributing factors associated with pedestrian injury severity levels through probabilistic and classification models. The crash severity can play a key role in health outcomes, treatment costs, and long-term repercussions for pedestrians (Dai 2012). In addition, the factors associated with pedestrian crash occurrence were identified by statistical count models. The crash frequency estimation can be used to identify hotspot locations and the crash reduction benefits associated with countermeasures.

This study explored a battery of methods to analyze all the models described above. Prior studies and practice have also used diverse methods, and the performance results have varied substantially across studies. It was very challenging to determine the best method or set of methods. Therefore, the second objective of this step was to compare the effectiveness of the different models in the estimation of pedestrian crash occurrence and severity levels using hold-out-sample predictions and statistical testing.

With respect to the severity analysis, the results showed that in rural and small urban areas pedestrian fatality risk is associated positively with male drivers, elderly drivers, shoulder presence, bike lane presence, wet surfaces, pedestrian exposure, hospital distance, intersection density, industrial and commercial areas, and dark hours and weekends. In contrast, signal control, sidewalk and warning sign presence, icy and snowy surfaces, higher AADT, share of residential land use, and high-density household areas decrease the probability of pedestrian fatality.

Moreover, the results showed that machine learning models hold promise for enhancing prediction accuracy of crash data, possibly because they do not assume any predefined underlying relationship between the dependent and independent variables, reducing biases that arise from parameterization. The ORL and ORP models describe the ordinal nature inherent in the data but require that the independent variables have the same effect on different severity levels. In contrast, the MNL and MNP models overlook the ordinal nature of crash severity but allow the independent variables' effects to vary among the injury levels. The results showed that the MNL outperformed the MNP, ORL and ORP models, which was perhaps due to the higher flexibility of the MNL specification model.

With respect to crash frequency analysis, it was found that in rural and small urban areas pedestrian crash occurrance is associated positively with shoulder presence, speed limit, number of lanes, higher AADT and pedestrian exposure, population density, median income, intersection density, and residential and commercial areas. In contrast, signal control, median presence, and higher average age decrease the likelihood of pedestrian crashes.

Additionally, the Hurdle negative binomial model outperformed the NB and ZINB models for estimating crash frequency. It was expected due to highly over-dispersed crash data in rural and small urban areas. The results also indicated that the RF model outperformed the other models to estimate the crash frequency. However, the RF model prediction results, like statistical models, are biased toward the zero-crash and one-crash classes, which are the most prevalent classes in this particular data. While the boosted RF model predicted the two-crash class very well, neither BRF nor HNB models produced reasonable accuracy for the two-or-more-crash category.

The findings of this step can help traffic planners and engineers to identify appropriate countermeasures (e.g., traffic warning signs) to improve pedestrian safety in rural and small urban areas. In addition, the results illuminated the benefits of combining different methods when making inferences from crash data (e.g., employing machine learning tools to improve prediction accuracy while utilizing statistical models to infer effects of an individual variable on the response variable. However, this study contains a few limitations (e.g., the inability of machine learning models to estimate an individual variable's effect on crash severity). Although this study identified several significant factors of pedestrian injury severity, some other predictors were missing (e.g., pedestrian individual level characteristics). Future studies can consider larger data sets with more observations or multi-year data to control for temporal variations.

Hotspot Locations Identification

This study utilized the two-step floating catchment area (2SFCA) method to identify high risk locations for pedestrian crashes in rural and small urban areas. The 2SFCA method has been widely used to measure accessibility to medical resources (Luo and Qi 2009; Yang et al. 2006; Radke and Mu 2000), and it is starting to enjoy applications in the transportation field (Kocatepe et al. 2017). This study utilized the 2SFCA method to reveal crash hotspots while simultaneously controlling for spatial heterogeneity, crash severity level, crash risk factors and pedestrian exposure. The method is especially useful for sparse areas with low crash density because it does not rely only on crash history; it also accounts for crash risk factors such as sidewalk and warning sign presence to measure risk at a given location.

Accordingly, a hotspot was defined as any location that has a higher severity index per exposure unit than other similar locations, with awareness of spatial interaction among neigboring crash locations. The pedestrian exposure was used rather than population, which is usually used in the 2SFCA method, because population does not represent the number of people who walk. The empirical bayesian approach, which is known as the best approach for hotspot identification in the literature, was also employed to calculate the sevrity index. These innovative changes can increase the 2SFCA method's accuracy to identify hotspots. The results confirmed that this methodology performed very well to identify crash prone locations and reduce the errors associated with simple hotspot identification methods.

Safety agencies can use this framework to improve their safety programs and their methods for selecting and assigning countermesures. Although this study used pedestrian crash data, this methodology can be applied to other crash types. Therefore, it can enhance roadway network safety for both pedestrians or motorized vehicle users.

Although the proposed methodology has several advantages over previous work, some limitations need to be noted and addressed in future work. For example, the pedestrian exposure model provided area-based estimates, and future work might explore enhancements such as node- or link-level exposure. Additionally, it can be compared with other well-known hotspot identification approaches to evaluate the accuracy of its performance.

Project Prioritization

This study proposed a mixed linear programming to rank safety improvement projects in rural and small urban areas. The objective function minimizes the crash risk index at hotspot locations, while accounting for budget constraints. According to the definition of a crash risk index, a higher crash risk index indicates not only that a location is prone to pedestrian crashes but also its neighboring locations are at high risk for pedestrian crash occurrence. This also gives priority to the hotspot locations with higher pedestrian exposure. Moreover, the proposed methodology distinguishes between different crash severity levels. This prioritizes the safety projects at locations with higher chance of fatality.

The proposed methodology was applied in the City of Bozeman as a case study. Reasonable countermeasures were suggested for twenty intersections with the highest crash risk index. The methodology identified $100,000 as the optimal budget, which would reduce the crash risk index by 63 percent. Using this budget cap, a final list of 17 countermeasures were recommended.

In conclusion, the proposed methodology worked very well for allocating safety funding to hotspot locations in rural and small urban areas. However, it is not limited to these crash types or area settings. Therefore, it can be used for other crash types (e.g., vehicle crashes) or in urban areas, especially when the budget for implementing countermeasures is limited. Further, additional research is required to expand the proposed methodology by testing it with a larger case study, identifying more alternatives per location, and considering other criteria such as traffic mobility.

REFERENCES CITED

AASHTO. 2010. “What Is the Difference between CMFs in the HSM and CMFs in the CMF Clearinghouse?” http://highwaysafetymanual.org/support_answers.aspx.

Abay, Kibrom A. 2013. “Examining Pedestrian-Injury Severity Using Alternative Disaggregate Models.” *Research in Transportation Economics* 43 (1). Elsevier Ltd: 123–36. doi:10.1016/j.retrec.2012.12.002.

Abdel-Aty, Mohamed. 2003. “Analysis of Driver Injury Severity Levels at Multiple Locations Using Ordered Probit Models.” *Journal of Safety Research* 34 (5): 597–603. doi:10.1016/j.jsr.2003.05.009.

Abdel-aty, Mohamed, Anurag Pande, Abhishek Das, and Willem Jan Knibbe. 2008. “Assessing Safety on Dutch Freeways with Data from Infrastructure-Based Intelligent Transportation Systems.” *Transportation Research Record: Journal of the Transportation Research Board* 2083: 153–61. doi:10.3141/2083-18.

Abdulhafedh, Azad. 2008. “Crash Frequency Analysis.” *Journal of Transportation Technologies* 6 (6): 169–80. doi:10.4236/jtts.2016.64017.

———. 2017. “Incorporating the Multinomial Logistic Regression in Vehicle Crash Severity Modeling: A Detailed Overview.” *Journal of Transportation Technologies* 7 (3): 279–303. doi:10.4236/jtts.2017.73019.

Agrawal, Asha Weinstein, and Paul Schimek. 2007. “Extent and Correlates of Walking in the USA.” *Transportation Research Part D: Transport and Environment* 12 (8): 548–63. doi:10.1016/j.trd.2007.07.005.

Al-Kaisy, A., L. Ewan, and F. Hossain. 2017. “Economic Feasibility of Safety Improvements on Low-Volume Roads.” *Journal of Transportation Safety and Security* 9 (3). Taylor & Francis: 369–82. doi:10.1080/19439962.2016.1212446.

Anselin, Luc. 2003. “GeoDa^{TM} 0.9 User’s Guide.” *Urbana* 51 (61801): 126. http://www.unc.edu/~emch/gisph/geoda093.pdf.

Arvin, Ramin, Mostafa Khademi, and Hesamoddin Razi-Ardakani. 2017. “Study on Mobile Phone Use While Driving in a Sample of Iranian Drivers.” *International Journal of Injury Control and Safety Promotion* 24 (2). Taylor & Francis: 256–62. doi:10.1080/17457300.2016.1175480.

Awasthi, Anjali, and Satyaveer S. Chauhan. 2011. “Using AHP and Dempster-Shafer Theory for Evaluating Sustainable Transport Solutions.” *Environmental Modelling and Software* 26 (6). Elsevier Ltd: 787–96. doi:10.1016/j.envsoft.2010.11.010.

Aziz, H. M Abdul, Satish V. Ukkusuri, and Samiul Hasan. 2013. “Exploring the Determinants of Pedestrian-Vehicle Crash Severity in New York City.” *Accident Analysis and Prevention* 50. Elsevier Ltd: 1298–1309. doi:10.1016/j.aap.2012.09.034.

Ballesteros, Michael F., Patricia C. Dischinger, and Patricia Langenberg. 2004. “Pedestrian Injuries and Vehicle Type in Maryland, 1995-1999.” *Accident Analysis and Prevention* 36 (1): 73–81. doi:10.1016/S0001-4575(02)00129-X.

Bauman, Adrian E., Rodrigo S. Reis, James F. Sallis, Jonathan C. Wells, Ruth J.F. Loos, and Brian W. Martin. 2012. “Correlates of Physical Activity: Why Are Some People Physically Active and Others Not?” *The Lancet* 380 (9838). Elsevier Ltd: 258–71. doi:10.1016/S0140-6736(12)60735-1.

Beck, Laurie F., Ann M. Dellinger, and Mary E. O’Neil. 2007. “Motor Vehicle Crash Injury Rates by Mode of Travel, United States: Using Exposure-Based Methods to Quantify Differences.” *American Journal of Epidemiology* 166 (2): 212–18. doi:10.1093/aje/kwm064.

Bennet, Scott A., and Nikolaos Yiannakoulias. 2015. “Motor-Vehicle Collisions Involving Child Pedestrians at Intersection and Mid-Block Locations.” *Accident Analysis and Prevention* 78. Elsevier Ltd: 94–103. doi:10.1016/j.aap.2015.03.001.

Boucher, Jean Philippe, and Miguel Santolino. 2010. “Discrete Distributions When Modeling the Disability Severity Score of Motor Victims.” *Accident Analysis and Prevention* 42 (6). Elsevier Ltd: 2041–49. doi:10.1016/j.aap.2010.06.015.

Bushell, Authors Max a, Bryan W Poole, Charles V Zegeer, and Daniel a Rodriguez. 2013. “Costs for Pedestrian and Bicyclist Infrastructure Improvements.” *UNC Highway Safety Research Center*.

Cao, Xinyu (Jason), Patricia L. Mokhtarian, and Susan L. Handy. 2009. “The Relationship between the Built Environment and Nonwork Travel: A Case Study of Northern California.” *Transportation Research Part A: Policy and Practice* 43 (5). Elsevier Ltd: 548–59. doi:10.1016/j.tra.2009.02.001.

Carlson, Jordan A., Brian E. Saelens, Jacqueline Kerr, Jasper Schipperijn, Terry L. Conway, Lawrence D. Frank, Jim E. Chapman, Karen Glanz, Kelli L. Cain, and James F. Sallis. 2015. “Association between Neighborhood Walkability and GPS-Measured Walking, Bicycling and Vehicle Time in Adolescents.” *Health and Place* 32. Elsevier: 1–7. doi:10.1016/j.healthplace.2014.12.008.

Census, U.S. 2000. “Census 2000 Urban and Rural Classification.”

Cerin, Ester, and Eva Leslie. 2008. “How Socio-Economic Status Contributes to Participation in Leisure-Time Physical Activity.” *Social Science and Medicine* 66 (12): 2596–2609. doi:10.1016/j.socscimed.2008.02.012.

Chang, Li Yen. 2005. “Analysis of Freeway Accident Frequencies: Negative Binomial Regression versus Artificial Neural Network.” *Safety Science* 43 (8): 541–57. doi:10.1016/j.ssci.2005.04.004.

Chang, Li Yen, and Wen Chieh Chen. 2005. “Data Mining of Tree-Based Models to Analyze Freeway Accident Frequency.” *Journal of Safety Research* 36 (4): 365–75. doi:10.1016/j.jsr.2005.06.013.

Chawla, Nitesh V., Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. “SMOTE: Synthetic Minority over-Sampling Technique.” *Journal of Artificial Intelligence Research* 16: 321–57. doi:10.1613/jair.953.

Chen, Cong, Guohui Zhang, Jinfu Yang, John C. Milton, and Adélamar Dely Alcántara. 2016. “An Explanatory Analysis of Driver Injury Severity in Rear-End Crashes Using a Decision table/Naïve Bayes (DTNB) Hybrid Classifier.” *Accident Analysis and Prevention* 90. Elsevier Ltd: 95–107. doi:10.1016/j.aap.2016.02.002.

Chen, Feng, Suren Chen, and Xiaoxiang Ma. 2016. “Crash Frequency Modeling Using Real-Time Environmental and Traffic Data and Unbalanced Panel Data Models.” *International Journal of Environmental Research and Public Health* 13 (6): 1–16. doi:10.3390/ijerph13060609.

Chen, Li, Cynthia Chen, Reid Ewing, Claire E. McKnight, Raghavan Srinivasan, and Matthew Roe. 2013. “Safety Countermeasures and Crash Reduction in New York City - Experience and Lessons Learned.” *Accident Analysis and Prevention* 50. Elsevier Ltd: 312–22. doi:10.1016/j.aap.2012.05.009.

Chen, Peng, and Jiangping Zhou. 2016. “Effects of the Built Environment on Automobile-Involved Pedestrian Crash Frequency and Risk.” *Journal of Transport & Health* 3 (4). Elsevier: 448–56. doi:10.1016/j.jth.2016.06.008.

Cheng, Wen, and Simon Washington. 2008. “New Criteria for Evaluating Methods of Identifying Hot Spots.” *Transportation Research Record: Journal of the Transportation Research Board* 2083 (1): 76–85. doi:10.3141/2083-09.

Cheng, Wen, and Simon P. Washington. 2005. “Experimental Evaluation of Hotspot

Identification Methods." *Accident Analysis and Prevention* 37 (5): 870–81. doi:10.1016/j.aap.2005.04.015.

Chimba, Deo, and Henry Ajieh. 2017. "Impact of Access Management Practices to Pedestrian Safety." *Transportation Research Center for Livable Communities (TRCLC).*

Chu, X. 2003. "The Fatality Risk of Walking in America: A Time-Based Comparative Approach." In *Walk21 Conference: Health, Equity and the Environment*.

Clifton, K.J., Burnier, C.V., Schneider, R., Huang, S., and Kang, M.W. 2008. "Pedestrian Demand Model for Evaluating Pedestrian Risk Exposure." *The National Center for Smart Growth Research and Education, University of Maryland, College Park*.

Clifton, Kelly J., Carolina V. Burnier, and Gulsah Akar. 2009. "Severity of Injury Resulting from Pedestrian-Vehicle Crashes: What Can We Learn from Examining the Built Environment?" *Transportation Research Part D: Transport and Environment* 14 (6). Elsevier Ltd: 425–36. doi:10.1016/j.trd.2009.01.001.

Clifton, Kelly J., and Kandice Kreamer-Fults. 2007. "An Examination of the Environmental Attributes Associated with Pedestrian-Vehicular Crashes near Public Schools." *Accident Analysis and Prevention* 39 (4): 708–15. doi:10.1016/j.aap.2006.11.003.

Cook, Wade D., and Rodney H. Green. 2000. "Project Prioritization: A Resource-Constrained Data Envelopment Analysis Approach." *Socio-Economic Planning Sciences* 34 (2): 85–99. doi:10.1016/S0038-0121(99)00020-8.

Costa, Stephen Da, Xiaobo Qu, and Partha Mani Parajuli. 2015. "A Crash Severity-Based Black Spot Identification Model." *Journal of Transportation Safety & Security* 7 (3): 268–77. doi:10.1080/19439962.2014.911230.

Cottrill, Caitlin D., and Piyushimita Thakuriah. 2010. "Evaluating Pedestrian Crashes in Areas with High Low-Income or Minority Populations." *Accident Analysis and Prevention* 42 (6). Elsevier Ltd: 1718–28. doi:10.1016/j.aap.2010.04.012.

Croissant, Yves. 2018. "Estimation of Multinomial Logit Models in R: The Mlogit Packages." *Universit´e de La R´eunion*.

Dai, Dajun. 2012. "Identifying Clusters and Risk Factors of Injuries in Pedestrian-Vehicle Crashes in a GIS Environment." *Journal of Transport Geography* 24. Elsevier Ltd: 206–14. doi:10.1016/j.jtrangeo.2012.02.005.

Daniels, Stijn, Tom Brijs, Erik Nuyts, and Geert Wets. 2010. “Explaining Variation in Safety Performance of Roundabouts.” *Accident Analysis and Prevention* 42 (2): 393–402. doi:10.1016/j.aap.2009.08.019.

Deka, Lipika, and Mohammed Quddus. 2014. “Network-Level Accident-Mapping: Distance Based Pattern Matching Using Artificial Neural Network.” *Accident Analysis and Prevention* 65. Elsevier Ltd: 105–13. doi:10.1016/j.aap.2013.12.001.

Delen, Dursun, Ramesh Sharda, and Max Bessonov. 2006. “Identifying Significant Predictors of Injury Severity in Traffic Accidents Using a Series of Artificial Neural Networks.” *Accident Analysis and Prevention* 38 (3): 434–44. doi:10.1016/j.aap.2005.06.024.

Dow, Jay K., and James W. Endersby. 2004. “Multinomial Probit and Multinomial Logit: A Comparison of Choice Models for Voting Research.” *Electoral Studies* 23 (1): 107–22. doi:10.1016/S0261-3794(03)00040-4.

Eluru, Naveen, Chandra R. Bhat, and David A. Hensher. 2008. “A Mixed Generalized Ordered Response Model for Examining Pedestrian and Bicyclist Injury Severity Level in Traffic Crashes.” *Accident Analysis and Prevention* 40 (3): 1033–54. doi:10.1016/j.aap.2007.11.010.

Elvik, Rune. 2008. “Comparative Analysis of Techniques for Identifying Locations of Hazardous Roads.” *Transportation Research Record: Journal of the Transportation Research Board* 2083 (1): 72–75. doi:10.3141/2083-08.

Estabrooks, Paul a, Rebecca E Lee, and Nancy C Gyurcsik. 2003. “Resources for Physical Activity Participation: Does Availability and Accessibility Differ by Neighborhood Socioeconomic Status?” *Annals of Behavioral Medicine : A Publication of the Society of Behavioral Medicine* 25 (August): 100–104. doi:10.1207/S15324796ABM2502_05.

Etminani-Ghasrodashti, Roya, and Mahyar Ardeshiri. 2016. “The Impacts of Built Environment on Home-Based Work and Non-Work Trips: An Empirical Study from Iran.” *Transportation Research Part A: Policy and Practice* 85. Elsevier Ltd: 196–207. doi:10.1016/j.tra.2016.01.013.

Ewing, Reid, and Robert Cervero. 2010. “Travel and the Built Environment.” *Journal of the American Planning Association* 76 (3): 265–94. doi:10.1080/01944361003766766.

Federal Highway Administration. 2010. “Highway Safety Improvement Program (HSIP)

Manual.” http://safety.fhwa.dot.gov/hsip/resources/fhwasa09029/sec2.cfm.

FHWA. 2008. “Desktop Reference for Crash Reduction Factors.”

Fitzpatrick, Kay, Susan T Chrysler, Ron Van Houten, William W Hunter, and Shawn Turner. 2011. “Evaluation of Pedestrian and Bicycle Engineering Countermeasures: Rectangular Rapid-Flashing Beacons, HAWKs, Sharrows, Crosswalk Markings, and the Development of an Evaluation Methods Report.” http://www.fhwa.dot.gov/publications/research/safety/pedbike/11039/11039.pdf.

Flahaut, Benoît, Michel Mouchart, Ernesto San Martin, and Isabelle Thomas. 2003. “The Local Spatial Autocorrelation and the Kernel Method for Identifying Black Zones.” *Accident Analysis & Prevention* 35 (6): 991–1004. doi:10.1016/S0001-4575(02)00107-0.

Fox, John, and Georges Monette. 1992. “Generalized Collinearity Diagnostics.” *Journal of the American Statistical Association* 87 (417): 178–83.

Frank, Lawrence D., James F. Sallis, Terry L. Conway, James E. Chapman, Brian E. Saelens, and William Bachman. 2006. “Many Pathways from Land Use to Health: Associations between Neighborhood Walkability and Active Transportation, Body Mass Index, and Air Quality.” *Journal of the American Planning Association* 72 (1): 75–87. doi:10.1080/01944360608976725.

Friedman, Jerome H. 2001. “Greedy Boosting Approximation: A Gradient Boosting Machine.” *Ann. Stat* 29: 1189–1232. doi:10.1214/aos/1013203451.

Gårder, Per E. 2004. “The Impact of Speed and Other Variables on Pedestrian Safety in Maine.” *Accident Analysis and Prevention* 36 (4): 533–42. doi:10.1016/S0001-4575(03)00059-9.

Ghasemzadeh, Ali, Britton E. Hammit, Mohamed M. Ahmed, and Rhonda Kae Young. 2018. “Parametric Ordinal Logistic Regression and Non-Parametric Decision Tree Approaches for Assessing the Impact of Weather Conditions on Driver Speed Selection Using Naturalistic Driving Data.” *Transportation Research Record*. doi:10.1177/0361198118758035.

Greene-Roesel, Ryan, Mara Chagas Diogenes, and David R Ragland. 2007. “Estimating Pedestrian Accident Exposure: Protocol Report.” *University of California Traffic Safety Center*.

Greene-Roesel, Ryan, Simon Washington, Megan Wier, Rajiv Bhatia, Md. Mazharul

Haque, and Beth Wemple. 2013. “Benefit Cost Analysis Applied to Behavioral and Engineering Safety Countermeasures in San Francisco , California.” In *92th Annual Meeting of Transportation Research Board (TRB), Washington DC, USA*.

Greene, William H. 2012. *Econometric Analysis*. 7th ed. Prentice Hall, Upper Saddle River.

Haddak, Mohamed Mouloud. 2016. “Exposure-Based Road Traffic Fatality Rates by Mode of Travel in France.” *Transportation Research Procedia* 14. Elsevier B.V.: 2025–34. doi:10.1016/j.trpro.2016.05.170.

Haghighi, Nima, Xiaoyue Cathy Liu, Guohui Zhang, and Richard J. Porter. 2018. “Impact of Roadway Geometric Features on Crash Severity on Rural Two-Lane Highways.” *Accident Analysis and Prevention* 111. Elsevier: 34–42. doi:10.1016/j.aap.2017.11.014.

Haleem, Kirolos, Priyanka Alluri, and Albert Gan. 2015. “Analyzing Pedestrian Crash Injury Severity at Signalized and Non-Signalized Locations.” *Accident Analysis and Prevention* 81. Elsevier Ltd: 14–23. doi:10.1016/j.aap.2015.04.025.

Harwood, Dw, Fm Council, E Hauer, W E Hughes, and a Vogt. 2000. “Prediction of the Expected Safety Performance of Rural Two-Lane Highways,” no. December: 197. doi:FHWA-RD-99-207.

Hatamzadeh, Yaser, Meeghat Habibian, and Ali Khodaii. 2014. “Walking Behaviors in Different Trip Purposes.” *Transportation Research Record: Journal of the Transportation Research Board* 2464 (2464): 118–25. doi:10.3141/2464-15.

Hosseinpour, Mehdi, Ahmad Shukri Yahaya, and Ahmad Farhan Sadullah. 2014. “Exploring the Effects of Roadway Characteristics on the Frequency and Severity of Head-on Crashes: Case Studies from Malaysian Federal Roads.” *Accident Analysis and Prevention* 62. Elsevier Ltd: 209–22. doi:10.1016/j.aap.2013.10.001.

Hough, Jill a., Xinyu Cao, and Susan L. Handy. 2008. “Exploring Travel Behavior of Elderly Women in Rural and Small Urban North Dakota: An Ecological Modeling Approach.” *Transportation Research Record: Journal of the Transportation Research Board* 2082: 125–31. doi:10.3141/2082-15.

Huang, Helai, Hanchu Zhou, Jie Wang, Fangrong Chang, and Ming Ma. 2017. “A Multivariate Spatial Model of Crash Frequency by Transportation Modes for Urban Intersections.” *Analytic Methods in Accident Research* 14. Elsevier Ltd: 10–21. doi:10.1016/j.amar.2017.01.001.

Imai, Kosuke, and David A. van Dyk. 2005. “MNP: R Package for Fitting the Multinomial Probit Model.” *Journal of Statistical Software* 14 (3): 1–32.

Iranitalab, Amirfarrokh, and Aemal Khattak. 2017. “Comparison of Four Statistical and Machine Learning Methods for Crash Severity Prediction.” *Accident Analysis & Prevention* 108 (February). Elsevier: 27–36. doi:10.1016/j.aap.2017.08.008.

Islam, Samantha, and Steven L. Jones. 2014. “Pedestrian at-Fault Crashes on Rural and Urban Roadways in Alabama.” *Accident Analysis and Prevention* 72. Elsevier Ltd: 267–76. doi:10.1016/j.aap.2014.07.003.

Jamali, Amir, and Yiyi Wang. 2017. “Estimating Pedestrian Exposure for Small Urban and Rural Areas.” *Transportation Research Record: Journal of the Transportation Research Board* 2661–10: 84–94.

Jang, Kitae, Shin Hyoung Park, Sanghyeok Kang, Ki Han Song, Seungmo Kang, and Sungbong Chung. 2013. “Evaluation of Pedestrian Safety.” *Transportation Research Record: Journal of the Transportation Research Board* 2393 (1): 104–16. doi:10.3141/2393-12.

Jensen, Soren Underlien. 2007. “Bicycle Tracks and Lanes : A Before-After Study.” *TRB Annual Meeting Compendium of Papers CD-ROM*, no. November.

Jonah, Brian A., and G. Ray Engel. 1983. “Measuring the Relative Risk of Pedestrian Accidents.” *Accident Analysis and Prevention* 15 (3): 193–206. doi:10.1016/0001-4575(83)90019-2.

Kashani, Ali Tavakoli, and Afshin Shariat Mohaymany. 2011. “Analysis of the Traffic Injury Severity on Two-Lane, Two-Way Rural Roads Based on Classification Tree Models.” *Safety Science* 49 (10). Elsevier Ltd: 1314–20. doi:10.1016/j.ssci.2011.04.019.

Kerr, Zachary Y., Daniel A. Rodriguez, Kelly R. Evenson, and Semra A. Aytur. 2013. “Pedestrian and Bicycle Plans and the Incidence of Crash-Related Injuries.” *Accident Analysis and Prevention* 50. Elsevier Ltd: 1252–58. doi:10.1016/j.aap.2012.09.028.

Khattak, Aemal. 2012. “Severity of Pedestrian Crashes At Highway-Rail Grade Crossings” 54 (2): 91–100.

Kim, Joon Ki, Gudmundur F. Ulfarssom, Venkataraman N. Shankar, and Fred L. Mannering. 2010. “A Note on Modeling Pedestrian-Injury Severity in Motor-

Vehicle Crashes with the Mixed Logit Model." *Accident Analysis and Prevention* 42 (6). Elsevier Ltd: 1751–58. doi:10.1016/j.aap.2010.04.016.

Kim, Joon Ki, Gudmundur F. Ulfarsson, Venkataraman N. Shankar, and Sungyop Kim. 2008. "Age and Pedestrian Injury Severity in Motor-Vehicle Crashes: A Heteroskedastic Logit Analysis." *Accident Analysis and Prevention* 40 (5): 1695–1702. doi:10.1016/j.aap.2008.06.005.

Kim, Karl, I. Made Brunner, and Eric Yamashita. 2008. "Modeling Fault among Accident-Involved Pedestrians and Motorists in Hawaii." *Accident Analysis and Prevention* 40 (6): 2043–49. doi:10.1016/j.aap.2008.08.021.

Kim, Myeonghyeon, Seung-Young Kho, and Dong-Kyu Kim. 2017. "Hierarchical Ordered Model for Injury Severity of Pedestrian Crashes in South Korea." *Journal of Safety Research* 61. The Authors: 33–40. doi:10.1016/j.jsr.2017.02.011.

Kim, Nam Seok, and Yusak O. Susilo. 2013. "Comparison of Pedestrian Trip Generation Models." *JOURNAL OF ADVANCED TRANSPORTATION* 47: 399–412.

Kingham, Simon, Clive E. Sabel, and Phil Bartie. 2011. "The Impact of the 'School Run' on Road Traffic Accidents: A Spatio-Temporal Analysis." *Journal of Transport Geography* 19 (4). Elsevier Ltd: 705–11. doi:10.1016/j.jtrangeo.2010.08.011.

Kocatepe, Ayberk, Mehmet Baran Ulak, Eren Erman Ozguven, and Mark Horner. 2017. "SOCIOECONOMIC CHARACTERISTICS AND CRASH PRONENESS : A CASE STUDY IN FLORIDA USING TWO-STEP FLOATING." *Transportation Research Board's 96th Annual Meeting, Washington, D.C.*

Kröyer, Höskuldur R G. 2014. "Is 30 Km/h a 'Safe' Speed? Injury Severity of Pedestrians Struck by a Vehicle and the Relation to Travel Speed and Age." *IATSS Research* 39 (1). International Association of Traffic and Safety Sciences: 42–50. doi:10.1016/j.iatssr.2014.08.001.

Kuo, Pei Fen, Dominique Lord, and Troy Duane Walden. 2013. "Using Geographical Information Systems to Organize Police Patrol Routes Effectively by Grouping Hotspots of Crash and Crime Data." *Journal of Transport Geography* 30. Elsevier Ltd: 138–48. doi:10.1016/j.jtrangeo.2013.04.006.

Kuzmyak, J. R., J. Walters, M. Bradley, and K. M. Kockelman. 2014. "Estimating Bicycling and Walking for Planning and Project Development: A Guidebook." *Transportation Research Board of the National Academies, Washington, D.C.*

Kwigizile, V., T. Sando, and D. Chimba. 2011. “Inconsistencies of Ordered and Unordered Probability Models for Pedestrian Injury Severity.” *Transportation Research Record*, no. 2264: 110–18. doi:10.3141/2264-13.

Lam, Winnie W.Y., Becky P.Y. Loo, and Shenjun Yao. 2013. “Towards Exposure-Based Time-Space Pedestrian Crash Analysis in Facing the Challenges of Ageing Societies in Asia.” *Asian Geographer* 30 (2): 105–25. doi:10.1080/10225706.2012.735436.

Lam, Winnie W Y, Shenjun Yao, and Becky P Y Loo. 2014. “Pedestrian Exposure Measures: A Time-Space Framework.” *Travel Behaviour and Society* 1 (1). Hong Kong Society for Transportation Studies: 22–30. doi:10.1016/j.tbs.2013.10.004.

Lassarre, Sylvain, Eleonora Papadimitriou, George Yannis, and John Golias. 2007. “Measuring Accident Risk Exposure for Pedestrians in Different Micro-Environments.” *Accident Analysis and Prevention* 39 (6): 1226–38. doi:10.1016/j.aap.2007.03.009.

Lee, Chris, and Mohamed Abdel-Aty. 2005. “Comprehensive Analysis of Vehicle-Pedestrian Crashes at Intersections in Florida.” *Accident Analysis and Prevention* 37 (4): 775–86. doi:10.1016/j.aap.2005.03.019.

Lefler, Devon E., and Hampton C. Gabler. 2004. “The Fatality and Injury Risk of Light Truck Impacts with Pedestrians in the United States.” *Accident Analysis and Prevention* 36 (2): 295–304. doi:10.1016/S0001-4575(03)00007-1.

Leisch, Friedrich. 2004. “A General Framework for Finite Mixture Models and Latent Class Regression in {R}.” *Journal of Statistical Software* 11 (8): 1–18. doi:10.18637/jss.v011.i08.

Li, Yue, and Geoff Fernie. 2010. “Pedestrian Behavior and Safety on a Two-Stage Crossing with a Center Refuge Island and the Effect of Winter Weather on Pedestrian Compliance Rate.” *Accident Analysis and Prevention* 42 (4). Elsevier Ltd: 1156–63. doi:10.1016/j.aap.2010.01.004.

Loukaitou-Sideris, Anastasia, Robin Liggett, and Hyun-Gun Sung. 2007. “Death on the Crosswalk.” *Journal of Planning Education and Research* 26 (3): 338–51. doi:10.1177/0739456X06297008.

Luo, Wei, and Yi Qi. 2009. “An Enhanced Two-Step Floating Catchment Area (E2SFCA) Method for Measuring Spatial Accessibility to Primary Care Physicians.” *Health and Place* 15 (4). Elsevier: 1100–1107. doi:10.1016/j.healthplace.2009.06.002.

Ma, Wen Jun, Shao Ping Nie, Hao Feng Xu, Yan Jun Xu, and Yu Run Zhang. 2010. “Socioeconomic Status and the Occurrence of Non-Fatal Child Pedestrian Injury: Results from a Cross-Sectional Survey.” *Safety Science* 48 (6). Elsevier Ltd: 823–28. doi:10.1016/j.ssci.2010.02.021.

Malyshkina, Nataliya V., and Fred L. Mannering. 2010. “Zero-State Markov Switching Count-Data Models: An Empirical Assessment.” *Accident Analysis and Prevention* 42 (1): 122–30. doi:10.1016/j.aap.2009.07.012.

Manepalli, Uday R. R., Ghulam H. Bham, and Srinadh Kandada. 2011. “Evaluation of Hotspots Identification Using Kernel Density Estimation (K) and Getis-Ord (Gi*) on I-630.” *The 3rd International Conference on Road Safety and Simulation* 1750: 17.

McAndrews, Carolyn, Kirsten Beyer, Clare E. Guse, and Peter Layde. 2013. “Revisiting Exposure: Fatal and Non-Fatal Traffic Injury Risk across Different Populations of Travelers in Wisconsin, 2001-2009.” *Accident Analysis and Prevention* 60. Elsevier Ltd: 103–12. doi:10.1016/j.aap.2013.08.005.

McGrail, Matthew R., and John S. Humphreys. 2009. “Measuring Spatial Accessibility to Primary Care in Rural Areas: Improving the Effectiveness of the Two-Step Floating Catchment Area Method.” *Applied Geography* 29 (4). Elsevier Ltd: 533–41. doi:10.1016/j.apgeog.2008.12.003.

Mcnally, Michael G. 2007. “The Four Step Model.” *Report UCI-ITS-WP-07-2. Institute of Transportation Studies, University of California, Irvine*. doi:https://escholarship.org/uc/item/0r75311t.

Menard, Scott. 2002. *Applied Logistic Regression Analysis*. Sage Publications, Thousand Oaks.

Millward, Hugh, and Jamie Spinney. 2011. “Time Use, Travel Behavior, and the Rural-Urban Continuum: Results from the Halifax STAR Project.” *Journal of Transport Geography* 19 (1). Elsevier Ltd: 51–58. doi:10.1016/j.jtrangeo.2009.12.005.

Miranda-Moreno, L.F. 2006. “Statistical Models and Methods for Identifying Hazardous Locations for Safety Improvements.” *PhD Dissertation. University of Waterloo, Ontario, Canada.*

Miranda-Moreno, Luis F., Patrick Morency, and Ahmed M. El-Geneidy. 2011. “The Link between Built Environment, Pedestrian Activity and Pedestrian-Vehicle Collision Occurrence at Signalized Intersections.” *Accident Analysis and Prevention* 43 (5). Elsevier Ltd: 1624–34. doi:10.1016/j.aap.2011.02.005.

Mishra, Sabyasachee, Mihalis M. Golias, Sushant Sharma, and Stephen D. Boyles. 2015. “Optimal Funding Allocation Strategies for Safety Improvements on Urban Intersections.” *Transportation Research Part A: Policy and Practice* 75. Elsevier Ltd: 113–33. doi:10.1016/j.tra.2015.03.001.

Moghaddam, F. Rezaie, S. Afandizadeh, and M. Ziyadi. 2011. “Prediction of Accident Severity Using Artificial Neural Networks.” *International Journal of Civil Engineering* 9 (1): 41–49.

Mohamed, Mohamed Gomaa, Nicolas Saunier, Luis F. Miranda-Moreno, and Satish V. Ukkusuri. 2013. “A Clustering Regression Approach: A Comprehensive Injury Severity Analysis of Pedestrian-Vehicle Crashes in New York, US and Montreal, Canada.” *Safety Science* 54. Elsevier Ltd: 27–37. doi:10.1016/j.ssci.2012.11.001.

Molino, John A., Jason F. Kennedy, Patches J. Inge, Mary Anne Bertola, Pascal A. Beuse, Nicole L. Fowler, Amanda K. Emo, and Ann Do. 2012. “A Distance-Based Method to Estimate Annual Pedestrian and Bicyclist Exposure in an Urban Environment.” *FHWA-HRT-11-043. FHWA, U.S. Department of Transportation*.

Montella, Alfonso. 2010. “A Comparative Analysis of Hotspot Identification Methods.” *Accident Analysis and Prevention* 42 (2): 571–81. doi:10.1016/j.aap.2009.09.025.

Montigny, Luc de, Richard Ling, and John Zacharias. 2012. “The Effects of Weather on Walking Rates in Nine Cities.” *Environment and Behavior* 44 (6): 821–40. doi:10.1177/0013916511409033.

Moudon, A.V., C. Lee, A.D. Cheadle, C.W. Collier, D. Johnson, T.L. Schmid, and R.D Weather. 2005. “Cycling and the Built Environment.” *A US Perspective. Transportation Research Part D: Transport and Environment* 10 (3): 245–61.

Moudon, Anne Vernez, Lin Lin, Junfeng Jiao, Philip Hurvitz, and Paula Reeves. 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 and Prevention* 43 (1). Elsevier Ltd: 11–24. doi:10.1016/j.aap.2009.12.008.

Mujalli, Randa Oqab, and Juan De Oña. 2011. “A Method for Simplifying the Analysis of Traffic Accidents Injury Severity on Two-Lane Highways Using Bayesian Networks.” *Journal of Safety Research* 42 (5). Elsevier Ltd: 317–26. doi:10.1016/j.jsr.2011.06.010.

Murat, Yetis Sazi. 2011. “An Entropy (Shannon) Based Traffic Safety Level

Determination Approach for Black Spots." *Procedia - Social and Behavioral Sciences* 20: 786–95. doi:10.1016/j.sbspro.2011.08.087.

Mwakalonge, Judith L. 2012. "Temporal Stability and Transferability of Non-Motorized and Total Trip Generation Models." *Journal of Transportation Technologies* 2 (4): 285–96. doi:10.4236/jtts.2012.24031.

New, Callie, Andy Li, Jonathan Larsen, and Yong Li. 2016. "Examining Walk Travel Behavior and Land Use in Utah." In *95th Annual Meeting of the Transportation Research Board, Washington, D.C.*

NHTSA. 2015. "Traffic Safety Facts." *National Highway Traffic Safety Administration, NHTSA's National Center for Statistics and Analysis*.

Noland, Robert, and Mohammed Quddus. 2004. "Analysis of Pedestrian and Bicycle Casualties with Regional Panel Data." *Transportation Research Record: Journal of the Transportation Research Board* 1897: 28–33. doi:10.3141/1897-04.

Obeng, Kofi, and Md Rokonuzzaman. 2013. "Pedestrian Injury Severity in Automobile Crashes." *Safety Science and Technology* 3: 9–17. doi:https://escholarship.org/uc/item/0r75311t.

Oh, Jutaek, Simon Washington, and Dongmin Lee. 2010. "Property Damage Crash Equivalency Factors to Solve Crash Frequency-Severity Dilemma." *Transportation Research Record: Journal of the Transportation Research Board* 2148 (1): 83–92. doi:10.3141/2148-10.

Oikawa, S, and Y Matsui. 2017. "Features of Fatal Pedestrian Injuries in Vehicle-to-Pedestrian Accidents in Japan." *International Journal of Crashworthiness* 1 (2). Taylor & Francis: 297–308. doi:10.4271/2013-01-0777.

Park, Byung Jung, Dominique Lord, and Jeffrey D. Hart. 2010. "Bias Properties of Bayesian Statistics in Finite Mixture of Negative Binomial Regression Models in Crash Data Analysis." *Accident Analysis and Prevention* 42 (2): 741–49. doi:10.1016/j.aap.2009.11.002.

Park, Byung Jung, Dominique Lord, and Chungwon Lee. 2014. "Finite Mixture Modeling for Vehicle Crash Data with Application to Hotspot Identification." *Accident Analysis and Prevention* 71. Elsevier Ltd: 319–26. doi:10.1016/j.aap.2014.05.030.

Poudel-Tandukar, Kalpana, Shinji Nakahara, Masao Ichikawa, Krishna C Poudel, and

Masamine Jimba. 2007. “Risk Perception, Road Behavior, and Pedestrian Injury among Adolescent Students in Kathmandu, Nepal.” *Injury Prevention : Journal of the International Society for Child and Adolescent Injury Prevention* 13 (4): 258–63. doi:10.1136/ip.2006.014662.

Pour-Rouholamin, Mahdi, and Huaguo Zhou. 2016. “Investigating the Risk Factors Associated with Pedestrian Injury Severity in Illinois.” *Journal of Safety Research* 57. Elsevier Ltd and National Safety Council: 9–17. doi:10.1016/j.jsr.2016.03.004.

Prasannakumar, V., H. Vijith, R. Charutha, and N. Geetha. 2011. “Spatio-Temporal Clustering of Road Accidents: GIS Based Analysis and Assessment.” *Procedia - Social and Behavioral Sciences* 21: 317–25. doi:10.1016/j.sbspro.2011.07.020.

Preston, Howard, Richard Storm, Jacqueline Dowds Bennett, and Elizabeth Wemple. 2013. “Systemic Safety Project Selection Tool.”

Pulugurtha, Srinivas S., Vanjeeswaran K. Krishnakumar, and Shashi S. Nambisan. 2007. “New Methods to Identify and Rank High Pedestrian Crash Zones: An Illustration.” *Accident Analysis and Prevention* 39 (4): 800–811. doi:10.1016/j.aap.2006.12.001.

Qin, Xiao, and John Ivan. 2001. “Estimating Pedestrian Exposure Prediction Model in Rural Areas.” *Transportation Research Record* 1773 (1): 89–96. doi:10.3141/1773-11.

Qu, Xiaobo, and Qiang Meng. 2014. “A Note on Hotspot Identification for Urban Expressways.” *Safety Science* 66. Elsevier Ltd: 87–91. doi:10.1016/j.ssci.2014.02.006.

Radke, John, and Lan Mu. 2000. “Spatial Decompositions, Modeling and Mapping Service Regions to Predict Access to Social Programs.” *Geographic Information Sciences*. doi:10.1080/10824000009480538.

Raford, Noah, and David R Ragland. 2004. “Space Syntax : An Innovative Pedestrian Volume Modeling Tool for Pedestrian Safety.” *Transportation Research Record: Journal of the Transportation Research Board* 1878: 66–74.

Raford, Noah, and David R. Ragland. 2006. “Pedestrian Volume Modeling for Traffic Safety and Exposure Analysis : The Case of Boston , Massachusetts.” In *Transportation Research Board 85th Annual Meeting*.

Raj, Jithin, Hareesh Bahuleyan, and Lelitha Devi Vanajakshi. 2016. “Application of Data Mining Techniques for Traffic Density Estimation and Prediction.” *Transportation*

Research Procedia 17: 321 – 330. http://ac.els-cdn.com/S2352146516307177/1-s2.0-S2352146516307177-main.pdf?_tid=45918b24-02d4-11e7-be6b-00000aab0f01&acdnat=1488849818_d530fd06303811f0963ca7d4cb478cf9.

Rhee, Kyoung Ah, Joon Ki Kim, Young Ihn Lee, and Gudmundur F. Ulfarsson. 2016. “Spatial Regression Analysis of Traffic Crashes in Seoul.” *Accident Analysis and Prevention* 91. Elsevier Ltd: 190–99. doi:10.1016/j.aap.2016.02.023.

Rifaat, Shakil Mohammad, Richard Tay, and Alexandre De Barros. 2011. “Effect of Street Pattern on the Severity of Crashes Involving Vulnerable Road Users.” *Accident Analysis and Prevention* 43 (1). Elsevier Ltd: 276–83. doi:10.1016/j.aap.2010.08.024.

Rosén, Erik, and Ulrich Sander. 2009. “Pedestrian Fatality Risk as a Function of Car Impact Speed.” *Accident Analysis and Prevention* 41 (1): 536–42. doi:10.1016/j.aap.2010.04.003.

Roshandeh, Arash M., Bismark R D K Agbelie, and Yongdoo Lee. 2016. “Statistical Modeling of Total Crash Frequency at Highway Intersections.” *Journal of Traffic and Transportation Engineering (English Edition)* 3 (2). Elsevier Ltd: 166–71. doi:10.1016/j.jtte.2016.03.003.

Roudsari, Bahman, Robert Kaufman, and Thomas Koepsell. 2006. “Turning at Intersections and Pedestrian Injuries.” *Traffic Injury Prevention* 7 (3): 283–89. doi:10.1080/15389580600660153.

Roudsari, B S, C N Mock, R Kaufman, D Grossman, B Y Henary, and J Crandall. 2004. “Pedestrian Crashes: Higher Injury Severity and Mortality Rate for Light Truck Vehicles Compared with Passenger Vehicles.” *Injury Prevention : Journal of the International Society for Child and Adolescent Injury Prevention* 10 (3): 154–58. doi:10.1136/ip.2003.003814.

Saaty, R. W. 1987. “The Analytic Hierarchy Process-What It Is and How It Is Used.” *Mathematical Modelling* 9 (3–5): 161–76. doi:10.1016/0270-0255(87)90473-8.

Sabir, Muhammad, Mark J Koetse, and Piet Rietveld. 2011. “The Impact of Weather Conditions on Mode Choice : Empirical Evidence for the Netherlands.” *VU University, Amsterdam, Netherlands*.

Sadeghi, Aliasghar, and Abolfazl Mohammadzadeh Moghaddam. 2016. “Uncertainty-Based Prioritization of Road Safety Projects: An Application of Data Envelopment Analysis.” *Transport Policy* 52. Elsevier: 28–36. doi:10.1016/j.tranpol.2016.07.003.

Saha, Promothes, and Khaled Ksaibati. 2016. “An Optimization Model for Improving Highway Safety.” *Journal of Traffic and Transportation Engineering (English Edition)* 3 (6). Elsevier Ltd: 549–58. doi:10.1016/j.jtte.2016.01.004.

Sallis, James F., Brian E. Saelens, Lawrence D. Frank, Terry L. Conway, Donald J. Slymen, Kelli L. Cain, James E. Chapman, and Jacqueline Kerr. 2009. “Neighborhood Built Environment and Income: Examining Multiple Health Outcomes.” *Social Science and Medicine* 68 (7). Elsevier Ltd: 1285–93. doi:10.1016/j.socscimed.2009.01.017.

Sarkar, Sudipta, Richard Tay, and John Hunt. 2011. “Logistic Regression Model of Risk of Fatality in Vehicle-Pedestrian Crashes on National Highways in Bangladesh.” *Transportation Research Record: Journal of the Transportation Research Board* 2264 (2264): 128–37. doi:10.3141/2264-15.

Sasidharan, Lekshmi, Wu K.-F., and Menendez M. 2015. “Exploring the Application of Latent Class Cluster Analysis for Investigating Pedestrian Crash Injury Severities in Switzerland.” *Accident Analysis and Prevention* 85. Elsevier Ltd: 219–28. doi:10.1016/j.aap.2015.09.020.

Sasidharan, Lekshmi, and Mónica Menéndez. 2014. “Partial Proportional Odds Model - An Alternate Choice for Analyzing Pedestrian Crash Injury Severities.” *Accident Analysis and Prevention* 72. Elsevier Ltd: 330–40. doi:10.1016/j.aap.2014.07.025.

Savolainen, Peter T., Fred L. Mannering, Dominique Lord, and Mohammed A. Quddus. 2011. “The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives.” *Accident Analysis and Prevention* 43 (5). Elsevier Ltd: 1666–76. doi:10.1016/j.aap.2011.03.025.

Siddiqui, Chowdhury, Mohamed Abdel-Aty, and Helai Huang. 2012. “Aggregate Nonparametric Safety Analysis of Traffic Zones.” *Accident Analysis and Prevention* 45. Elsevier Ltd: 317–25. doi:10.1016/j.aap.2011.07.019.

Siddiqui, Naved, Xuehao Chu, and Martin Guttenplan. 2006. “Crossing Locations, Light Conditions, and Pedestrian Injury Severity.” *Transportation Research Record* 1982 (1): 141–49. doi:10.3141/1982-19.

Silcock, D. T., R. Walker, and T. Selby. 1996. “Pedestrians at Risk.” In *PTRC 24th European Transport Conference*, 1–12.

Singh, Richa. 2016. “Factors Affecting Walkability of Neighborhoods.” *Procedia - Social and Behavioral Sciences* 216 (October 2015). Elsevier B.V.: 643–54.

doi:10.1016/j.sbspro.2015.12.048.

Smith, Ken R., Barbara B. Brown, Ikuho Yamada, Lori Kowaleski-Jones, Cathleen D. Zick, and Jessie X. Fan. 2008. “Walkability and Body Mass Index. Density, Design, and New Diversity Measures.” *American Journal of Preventive Medicine* 35 (3): 237–44. doi:10.1016/j.amepre.2008.05.028.

Soltani, Ali, and Sajad Askari. 2017. “Exploring Spatial Autocorrelation of Traffic Crashes Based on Severity.” *Injury*, no. 2016. Elsevier Ltd: 1–11. doi:10.1016/j.injury.2017.01.032.

Son, Hojun Daniel, Young Jun Kweon, and Byungkyu Brian Park. 2011. “Development of Crash Prediction Models with Individual Vehicular Data.” *Transportation Research Part C: Emerging Technologies* 19 (6): 1353–63. doi:10.1016/j.trc.2011.03.002.

Sullivan, James, and Jonathan Dowds. 2012. “Applying a Vehicle Miles of Travel Calculation Methodology to a Countywide Calculation of Bicycle and Pedestrian Miles of Travel.” In *91st Annual Meeting of the Transportation Research Board, Washington, D.C.*

Sze, N. N., and S. C. Wong. 2007. “Diagnostic Analysis of the Logistic Model for Pedestrian Injury Severity in Traffic Crashes.” *Accident Analysis and Prevention* 39 (6): 1267–78. doi:10.1016/j.aap.2007.03.017.

Taamneh, Madhar, Sharaf Alkheder, and Salah Taamneh. 2017. “Data-Mining Techniques for Traffic Accident Modeling and Prediction in the United Arab Emirates.” *Journal of Transportation Safety and Security* 9 (2). Taylor & Francis: 146–66. doi:10.1080/19439962.2016.1152338.

Tarko, Andrew, and Md Shafiul Azam. 2011. “Pedestrian Injury Analysis with Consideration of the Selectivity Bias in Linked Police-Hospital Data.” *Accident Analysis and Prevention* 43 (5). Elsevier Ltd: 1689–95. doi:10.1016/j.aap.2011.03.027.

Tay, R, J Choi, L Kattan, and A Khan. 2011. “A Multinomial Logit Model of Pedestrian-Vehicle Crash Severity.” *International Journal of Sustainable Transportation* 5 (4): 233–49. doi:10.1080/15568318.2010.497547.

Tefft, Brian C. 2013. “Impact Speed and a Pedestrian’s Risk of Severe Injury or Death.” *Accident Analysis and Prevention* 50. Elsevier Ltd: 871–78. doi:10.1016/j.aap.2012.07.022.

Toran Pour, Alireza, Sara Moridpour, Richard Tay, and Abbas Rajabifard. 2016. "Modelling Pedestrian Crash Severity at Mid-Blocks." *Transportmetrica A: Transport Science* 9935 (April): 1–25. doi:10.1080/23249935.2016.1256355.

Truong, Long T, and Sekhar V Somenahalli. 2011. "Using GIS to Identify Pedestrian-Vehicle Crash Hot Spots and Unsafe Bus Stops." *Journal of Public Transportation* 14 (1): 99–114. doi:http://dx.doi.org/10.5038/2375-0901.14.1.6.

Ukkusuri, Satish, Samiul Hasan, and H. Aziz. 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: 98–106. doi:10.3141/2237-11.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S. Springer, New York*. Fourth. New York. http://www.stats.ox.ac.uk/pub/MASS4.

Verzosa, Nina, and Rebecca Miles. 2016. "Severity of Road Crashes Involving Pedestrians in Metro Manila, Philippines." *Accident Analysis and Prevention* 94. Elsevier Ltd: 216–26. doi:10.1016/j.aap.2016.06.006.

Vuong, Quang H. 1989. "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses." *Econometrica* 57 (2): 307–33.

Walden, Troy D., Dominique Lord, Myunghoon Ko, Srinivas Geedipally, and Lingtao Wu. 2015. "Developing Methodology for Identifying, Evaluating, and Prioritizing Systemic Improvements." *TRAFFIC OPERATION DIVISION, TEXAS DEPARTMENT OF TRANSPORTATION*.

Wang, Lu. 2007. "Immigration, Ethnicity, and Accessibility to Culturally Diverse Family Physicians." *Health and Place* 13 (3): 656–71. doi:10.1016/j.healthplace.2006.10.001.

Wang, Xuesong, Junguang Yang, Chris Lee, Zhuoran Ji, and Shikai You. 2016. "Macro-Level Safety Analysis of Pedestrian Crashes in Shanghai, China." *Accident Analysis and Prevention* 96. Elsevier Ltd: 12–21. doi:10.1016/j.aap.2016.07.028.

Wang, Yiyi, and Kara M. Kockelman. 2013. "A Poisson-Lognormal Conditional-Autoregressive Model for Multivariate Spatial Analysis of Pedestrian Crash Counts across Neighborhoods." *Accident Analysis and Prevention* 60. Elsevier Ltd: 71–84. doi:10.1016/j.aap.2013.07.030.

Warburton, Darren E R, Crystal Whitney Nicol, and Shannon S D Bredin. 2006. "Health

Benefits of Physical Activity: The Evidence." *CMAJ : Canadian Medical Association Journal = Journal de l'Association Medicale Canadienne* 174 (6): 801–9. doi:10.1503/cmaj.051351.

Wei, Yehua, Weiye Xiao, Ming Wen, and Ran Wei. 2016. "Walkability, Land Use and Physical Activity." *Sustainability* 8 (1): 65–80. doi:10.3390/su8010065.

Wier, Megan, June Weintraub, Elizabeth H. Humphreys, Edmund Seto, and Rajiv Bhatia. 2009. "An Area-Level Model of Vehicle-Pedestrian Injury Collisions with Implications for Land Use and Transportation Planning." *Accident Analysis and Prevention* 41 (1): 137–45. doi:10.1016/j.aap.2008.10.001.

Witten, Ian H, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham. 1999. "Weka : Practical Machine Learning Tools and Techniques with Java Implementations." *Department of Computer Science, University of Waikato, New Zealand*, 192–96. doi:10.1.1.16.949.

Xie, Kun, Xuesong Wang, Kaan Ozbay, and Hong Yang. 2014. "Crash Frequency Modeling for Signalized Intersections in a High-Density Urban Road Network." *Analytic Methods in Accident Research* 2. Elsevier: 39–51. doi:10.1016/j.amar.2014.06.001.

Xie, Zhixiao, and Jun Yan. 2008. "Kernel Density Estimation of Traffic Accidents in a Network Space." *Computers, Environment and Urban Systems* 32 (5). Elsevier Ltd: 396–406. doi:10.1016/j.compenvurbsys.2008.05.001.

Yan, Xuedong, Bin Wang, Meiwu An, and Cuiping Zhang. 2012. "Distinguishing between Rural and Urban Road Segment Traffic Safety Based on Zero-Inflated Negative Binomial Regression Models." *Discrete Dynamics in Nature and Society* 2012. doi:10.1155/2012/789140.

Yang, Duck Hye, Robert Goerge, and Ross Mullner. 2006. "Comparing GIS-Based Methods of Measuring Spatial Accessibility to Health Services." *Journal of Medical Systems* 30 (1): 23–32. doi:10.1007/s10916-006-7400-5.

Yang, Yong, Amy H. Auchincloss, Daniel A. Rodriguez, Daniel G. Brown, Rick Riolo, and Ana V. Diez-Roux. 2015. "Modeling Spatial Segregation and Travel Cost Influences on Utilitarian Walking: Towards Policy Intervention." *Computers, Environment and Urban Systems* 51. Elsevier Ltd: 59–69. doi:10.1016/j.compenvurbsys.2015.01.007.

Yang, Zhao, Yuanyuan Zhang, and Offer Grembek. 2016. "Combining Traffic Efficiency and Traffic Safety in Countermeasure Selection to Improve Pedestrian Safety at Two-Way Stop Controlled Intersections." *Transportation Research Part A: Policy and Practice* 91. Elsevier Ltd: 286–301. doi:10.1016/j.tra.2016.07.002.

Yao, Shenjun, Becky P.Y. Loo, and Winnie W.Y. Lam. 2015. "Measures of Activity-Based Pedestrian Exposure to the Risk of Vehicle-Pedestrian Collisions: Space-Time Path vs. Potential Path Tree Methods." *Accident Analysis and Prevention* 75. Elsevier Ltd: 320–32. doi:10.1016/j.aap.2014.12.005.

Yasmin, Shamsunnahar, Naveen Eluru, and Satish V Ukkusuri. 2014. "Alternative Ordered Response Frameworks for Examining Pedestrian Injury Severity in New York City." *Journal of Transportation Safety & Security* 6 (4): 275–300. doi:10.1080/19439962.2013.839590.

Yedla, Sudhakar, and Ram M. Shrestha. 2003. "Multi-Criteria Approach for the Selection of Alternative Options for Environmentally Sustainable Transport System in Delhi." *Transportation Research Part A: Policy and Practice* 37 (8): 717–29. doi:10.1016/S0965-8564(03)00027-2.

Yi, Sul Ki, Hwa Young Sin, and Eunnyeong Heo. 2011. "Selecting Sustainable Renewable Energy Source for Energy Assistance to North Korea." *Renewable and Sustainable Energy Reviews* 15 (1). Elsevier Ltd: 554–63. doi:10.1016/j.rser.2010.08.021.

Yu, Chia Yuan. 2015. "Built Environmental Designs in Promoting Pedestrian Safety." *Sustainability (Switzerland)* 7 (7): 9444–60. doi:10.3390/su7079444.

Yu, Hao, Pan Liu, Jun Chen, and Hao Wang. 2014. "Comparative Analysis of the Spatial Analysis Methods for Hotspot Identification." *Accident Analysis and Prevention* 66. Elsevier Ltd: 80–88. doi:10.1016/j.aap.2014.01.017.

Yu, Jie, and Yue Liu. 2012. "Prioritizing Highway Safety Improvement Projects: A Multi-Criteria Model and Case Study with SafetyAnalyst." *Safety Science* 50 (4). Elsevier Ltd: 1085–92. doi:10.1016/j.ssci.2011.11.018.

Zahabi, Seyed Amir H., Jillian Strauss, Kevin Manaugh, and Luis F. Miranda-Moreno. 2011. "Estimating Potential Effect of Speed Limits, Built Environment, and Other Factors on Severity of Pedestrian and Cyclist Injuries in Crashes:" *Transportation Research Record: Journal of the Transportation Research Board* 2247 (1): 81–90. doi:10.3141/2247-10.

Zajac, Sylvia S., and John N. Ivan. 2003. "Factors Influencing Injury Severity of Motor

Vehicle-Crossing Pedestrian Crashes in Rural Connecticut." *Accident Analysis and Prevention* 35 (3): 369–79. doi:10.1016/S0001-4575(02)00013-1.

Zegeer, Charles V., J. Richard Stewart, Herman H. Huang, and Peter A. Lagerwey. 2005. "Safety Effects of Marked vs Unmarked Crosswalks at Uncontrolled Locations: Executive Summary and Recommended Guidelines." *Report FHWA-HRT-04-100. FHWA, U.S. Department of Transportation*.

Zegeer, CHARLES V., JANE STUTTS, HERMAN HUANG, MICHAEL J. CYNECKI, RON VAN HOUTEN, BARBARA ALBERSON, RONALD PFEFER, TIMOTHY R. NEUMAN, KEVIN L. SLACK, and KELLY K. HARDY. 2004. "A Guide for Reducing Collisions Involving Pedestrians." doi:10.17226/13545.

Zeileis, Achim, Christian Kleiber, and Simon Jackman. 2008. "Regression Models for Count Data in {R}." *Journal of Statistical Software* 27 (8). http://www.jstatsoft.org/v27/i08/.

Zhang, Guangnan, Kelvin K W Yau, and Xun Zhang. 2014. "Analyzing Fault and Severity in Pedestrian-Motor Vehicle Accidents in China." *Accident Analysis and Prevention* 73. Elsevier Ltd: 141–50. doi:10.1016/j.aap.2014.08.018.

Zheng, Zuduo, and Dongcai Su. 2014. "Short-Term Traffic Volume Forecasting: A K-Nearest Neighbor Approach Enhanced by Constrained Linearly Sewing Principle Component Algorithm." *Transportation Research Part C: Emerging Technologies* 43. Elsevier Ltd: 143–57. doi:10.1016/j.trc.2014.02.009.

Zhu, Xuemei, and Chanam Lee. 2008. "Walkability and Safety Around Elementary Schools. Economic and Ethnic Disparities." *American Journal of Preventive Medicine* 34 (4): 282–90. doi:10.1016/j.amepre.2008.01.024.

Zou, Yajie, Yunlong Zhang, and Dominique Lord. 2013. "Application of Finite Mixture of Negative Binomial Regression Models with Varying Weight Parameters for Vehicle Crash Data Analysis." *Accident Analysis and Prevention* 50. Elsevier Ltd: 1042–51. doi:10.1016/j.aap.2012.08.004.

APPENDIX A: DATA DICTIONARY

Table 26. Site characteristics data obtained for intersections.

Name	Label	Code
FID*[1]	FID on intersection shapefile	Numerical (e.g., 388404)
GEOID*	Block group 12-digit ID number	Numerical (e.g., 301110004023)
Crash*	Number of crashes occurred at intersection and its influence area (i.e., within 250 ft.)	Numerical
Urban *	Urban indicator	1=Urban 2 =Urban Cluster 3 =Rural
Leg	Number of intersection legs	Numerical UN=Unknown
Control	Traffic control	1=Signalized 2=Four-way stop control 3= Two-way stop control 0=None UN=Unknown
Land use*	A set of land use variables (entropy, mix, and % of land area by type)	1=Undeveloped 2=Commercial 3=Industrial 4= Mixed 5=Residential 6 = Agriculture UN=Unknown
Skew angle	Angle between intersecting streets	1=90° 2=less than 90° UN=Unknown
Intersection Type*	Whether it is a driveway intersection or not.	1= Driveway 0=Otherwise UN=Unknown
Railroad	Railroad crossing in the vicinity of the intersection	0=none 1= if railroad exists
Road ID	FID on roadway shapefile (For consistency between Google street view map and Arc GIS map, first collect the roadway closer to vertical axis [longitudinal]. Note that North is directed to up.)	Numerical (e.g., 944408)
Name*	Roadway name	UN=Unknown

[1] * indicates items that are calculated using ArcGIS or already available from archival data.

Table 26 Continued.

Road Type*	Functional Class (Tiger/Line category)	1=Highway 2=local 3= Driveway 4=Other UN=Unknown
Operation	One-way or two-way operation	1= One-way 2=Two-way UN=Unknown
Pavement	Pavement condition	1=Paved 2=Gravel 3=Dirt UN=Unknown
Through Lanes	Number of through lanes	Numerical UN=Unknown
Right Lanes	Number of right lanes	Numerical UN=Unknown
Left Lanes	Number of left lanes	Numerical UN=Unknown
Median	Median type	1=Raised 2=Depressed 3=Flush 0=None UN=Unknown
Opening	Number of median openings	Numerical UN=Unknown
Shoulder	Shoulder type	1=Paved 2=Gravel 3=Composition (half paved) 0=None UN=Unknown
Sidewalk	Presence of sidewalk	1=if sidewalk exists on one side of the approach (i.e., within 250 ft. of the intersection) 2=if sidewalk exists on both sides of the approach 0=non-existent UN=Unknown
Bicycle	Presence of bicycle lane	1=if bicycle lane exists 0=otherwise UN=Unknown

Table 26 Continued.

Median Crossing within the intersection influence area (i.e., 250 ft)	Pedestrian crossing type	1=Un-signalized marked with refuge area 2=Un-signalized marked without refuge area 3=Signalized marked with refuge area 4=Signalized marked without refuge area 0=None UN=Unknown
Crossing with curb extensions	Presence of curb extension for pedestrian crossing at intersection	1=if curb extension exists 0=none UN=Unknown
Raised Crossing	Presence of raised crossing at intersection	1=if crossing is raised 0=none UN=Unknown e
Driveways	Number of driveways within the intersection and its influence area.	Numerical UN=Unknown
Speed limit	Speed limit for vehicular traffic	Numerical UN=Unknown
Parking	Presence of curb parking	1=Parallel parking 2=Angle parking 0=No parking UN=Unknown
Lighting	Presence of lighting at the intersection	1=if street lighting exists (including street lights or other ambient lights) 0=none UN=Unknown
Horizontal Curve	Presence of curve	1=Curved 0=Straight UN=Unknown
Vertical Curve	Presence of curve	1=Present (sloped) 0=None (level)) UN=Unknown
Signs	Presence of advance warning signs such as crossroad, STOP ahead, or signal ahead	1=if pedestrian signs exist 0=otherwise UN=Unknown
Transit stop	Presence of transit stops within the influence area	1=if any transit stop is present near the intersection 0=otherwise UN=Unknown

Table 26 Continued.

LT Treatment	Configuration of left-turn (LT) lane on the intersection approach	0= none 1=Painted 2 = Curbed 3= Prohibited UN=Unknown
LT Offset	Offset of the LT lane	0= none 1= Offset UN=Unknown
RT Treatment	Configuration of right-turn lane on the intersection approach	1= right-turn lane only 2= channelizing island only 3= right turn lane and channelizing island 0=None UN=Unknown
Right turn on red	Presence of sign to prohibit turning on red	1=if Right turn on red prohibited 0=otherwise UN=Unknown