REGIONAL MAP-BASED ANALYTICAL PLATFORM FOR STATEWIDE HIGHWAY SAFETY PERFORMANCE ASSESSMENT

FINAL PROJECT REPORT

by

Ali Hajbabaie
Washington State University

Yinhai Wang
University of Washington

SMA Bin Al Islam
Washington State University

Ziqiang Zeng
University of Washington

Sponsorship
Washington State Department of Transportation

for
Pacific Northwest Transportation Consortium (PacTrans)
USDOT University Transportation Center for Federal Region 10
University of Washington
More Hall 112, Box 352700
Seattle, WA 98195-2700

In cooperation with US Department of Transportation-Research
and Innovative Technology Administration (RITA)

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. The Pacific Northwest Transportation Consortium, the U.S. Government and matching sponsor assume no liability for the contents or use thereof.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.	

4. Title and Subtitle	5. Report Date
Regional Map-Based Analytical Platform for Statewide Highway Safety Performance Assessment	06/30/2016
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
Ali Hajbabaie, Yinhai Wang, SMA Bin Al Islam, and Ziqiang Zeng	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
PacTrans Pacific Northwest Transportation Consortium University Transportation Center for Region 10 University of Washington More Hall 112 Seattle, WA 98195-2700	**11. Contract or Grant No.** DTRT13-G-UTC40

12. Sponsoring Organization Name and Address	13. Type of Report and Period Covered
United States of America Department of Transportation Research and Innovative Technology Administration	Research
	14. Sponsoring Agency Code

15. Supplementary Notes

Report uploaded at www.pacTrans.org

16. Abstract

This research extended models for predicting current crash counts by severity (CCS) by developing a two-stage regression model and a generalized nonlinear regression model for formulating a new method for identifying CCS-based hotspots. The hotspot identification (HSID) method is designed to improve the safety performance module of the Digital Roadway Interactive Visualization and Evaluation Network (DRIVE Net) and to enable a regional, map-based, real-time analytical platform for statewide highway safety performance assessment. The most important contributing factors (static and dynamic) to traffic crashes of different severity types, including traffic characteristics, road conditions, and weather conditions, were identified by using the structured framework developed in this research. A total of 802 road segments on I-5, I-90, I-82, I-182, I-205, I-405 and I-705 in Washington state were selected as the candidate sites for data collection. A two-stage regression and logistics model was developed to predict crash counts on freeway segments by severity. The regression analysis found that annual average daily traffic per lane, number of lanes, curvature of segments, width of the outer shoulder, width of the inner shoulder, width of medians, average speed limit, lane surface type, outer shoulder type, inner shoulder type, and road surface conditions show strong relationships with the crash frequencies of different severity levels. A CCS-based HSID method was developed by employing a two-stage regression approach. A new safety performance index (SPI) and a new potential safety improvement index (PSII) were developed by introducing the risk weight factor and were compared with three indices by employing HSID evaluating methods. The results of four consistency tests revealed that the SPI method is the most consistent and reliable method for identifying hotspots. Finally, a generalized, nonlinear, model-based multinomial logistic regression approach was also developed to estimate the probability and frequency of crashes for different severity levels. It also showed that the significance and nonlinearity for each crash severity level are different among the contributing factors. This evaluation suggested that the SPI method (among the methods compared) has the potential to become the industry standard. Finally, a regional, map-based analytical platform was developed within the DRIVE Net system by expanding the safety performance module with the new SPI and PSII functions.

17. Key Words	18. Distribution Statement
Safety Performance Assessment, Generalized Nonlinear Model, Crash Frequency by Severity, Multinomial Logistic Regression, Regional Map Based Analytical Platform	No restrictions.

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified.	Unclassified.	78	NA

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AADT: Annual Average Daily Traffic
ARP: Accident Reduction Potential
SPL: Average Speed Limit
CCS: Crash Counts by Severity
CMF: Crash Modification Factors
COS: Curvature of Segment
IF: Injury-fatal
IST: Dominant Inner Shoulder Type
ISW: Inner Shoulder Width
LST: Dominant Lane Surface Type
MST: Dominant Median Surface Type
OST: Dominant Outer Shoulder Type
OSW: Outer Shoulder Width
DRIVE Net: Digital Roadway Interactive Visualization and Evaluation Network
EB: Empirical Bayes
GLM: Generalized Linear Model
GNM: Generalized Nonlinear Model
HCT: Horizontal Curve Type
HSID: Hotspot Identification
HSIS: Highway Safety Information System
MCT: Method Consistency Test
MNL: Multinomial Logit
MWD: Median Width
NB: Negative Binomial
NOL: Number of Lanes
PacTrans: Pacific Northwest Transportation Consortium
PDO: Property Damage Only
PSII: Potential Safety Improvement Index
RSC: Road Surface Condition
SCT: Site Consistency Test
SPF: Safety Performance Function
SPI: Safety Performance Index
TRDT: Total Rank Differences Test
TST: Total Score Test
USDOT: United States Department of Transportation
WITS: Washington Incident Tracking System
WSDOT: Washington State Department of Transportation

Acknowledgments

The PI and Co-PI would like to give their great appreciation to WSDOT for the effort of data collection and the WSU and UW team researchers and Ph.D. students, including S.M.A. Bin Al Islam, Mehrdad Tajalli, Rasool Mohebifar, Ziqiang Zeng, Wenbo Zhu, Ruimin Ke, Zhiyong Cui, John Ash, Jinjun Tang, Wenhui Zhang, Haifeng Guo, and Xinqiang Chenfor for their effort and contributions to this research.

Executive Summary

Traffic crashes cost billions of dollars annually in life and property damage worldwide. Hence, traffic safety improvement has been a top strategic goal of the U.S. Department of Transportation (USDOT) and an important goal for the Washington State Department of Transportation (WSDOT) over the past several years. To prioritize spending of limited safety improvement funds, a statewide analytical tool for highway safety performance assessment is needed. Although existing GIS-based tools are good for visualization, they do not have the desired functionality for statewide traffic safety performance assessment. This research extended the current state-of-the-art for predicting crash counts by severity (CCS) to develop a two-stage regression model (integrating logit and generalized linear models) and a generalized nonlinear regression model for formulating a new method for identifying CCS-based hotspots. The hotspot identification (HSID) method can be utilized to improve the safety performance module of the Digital Roadway Interactive Visualization and Evaluation Network (DRIVE Net) and enable a regional, map-based, real-time analytical platform for statewide highway safety performance assessment. The most important contributing factors (static and dynamic) to traffic crashes of different severity types, including traffic characteristics, road conditions, and weather conditions, were identified by using a structured framework developed in this research. A total of 802 road segments on I-5, I-90, I-82, I-182, I-205, I-405 and I-705 in Washington state were selected as candidate sites for data collection. A two-stage model was developed to predict crash counts on freeway segments by severity. The regression analysis showed that the contributing factors—including annual average daily traffic per lane (AADT/Lane), number of lanes (NOL), curvature of segment (COS), width of outer shoulder (OSW), width of inner shoulder (ISW), width of median (MWD), average speed limit (SPL), lane surface type (LST), outer shoulder type (OST),

inner shoulder type (IST), and road surface conditions (RSC)—have strong relationships with the crash frequencies of different severity levels. A CCS-based HSID method was developed by employing the two-stage and generalized nonlinear regression approaches. A new safety performance index (SPI) and a new potential safety improvement index (PSII) were developed by introducing a risk weight factor and were compared with three indices by employing HSID evaluating methods. The results of four consistency tests revealed that the SPI method is the most consistent and reliable method for identifying hotspots. Although it can only be applied to roadway segments where the crash data for different levels of severity are available, with the rapid development of intelligent transportation systems and data collection technologies, this method could become quite useful in identifying high-risk road sites. On several criteria, the SPI outperformed other methods by a wide margin. Next, a generalized, nonlinear, model-based multinomial logistic regression approach was developed to estimate the probability and frequency of crashes for different severity levels and was compared with traditional indices. It also showed that the significance and nonlinearity for each crash severity level were different among the contributing factors. This evaluation also suggested that the SPI method (among the methods compared) has the potential to become the industry standard. Finally, a regional, map-based analytical platform was developed within the DRIVE Net system by expanding the safety performance module with the new SPI and PSII functions.

Future work will focus on the following three directions: (1) developing a framework for a real-time safety performance analysis platform; (2) considering an analysis of crash frequency by collision type and severity; (3) developing new criteria for evaluating methods of identifying hotspots based on new safety performance indexes.

Chapter 1 Introduction

1.1 Background

Traffic collisions contribute to the loss of billions of dollars annually around the world. Researchers have sought ways to better understand the factors that affect the probability of crashes and their severities. The objective has been to predict the likelihood of crashes more accurately and to provide guidance for developing more effective countermeasures aimed at reducing the number of crashes (Lord and Mannering, 2010). As such, traffic safety improvement has been among the top strategic goals of the United States Department of Transportation (USDOT) and an important priority for the Washington State Department of Transportation (WSDOT) over the past several years. To prioritize spending of the limited safety improvement funds, a statewide analytical tool for highway safety performance assessment is needed. Although existing GIS-based tools are good for visualization, they do not have the desired functionality for statewide traffic safety performance assessment.

1.1.1 Challenges in Identifying Major Factors for High Risk Locations

Traffic crashes are rare events and thus multiple years of observations are typically required in order to conduct statistical analyses. Over the observation period, there are likely changes in traffic demand, roadway geometry, and other relevant factors that may influence both crash frequency and severity. Furthermore, dynamic variables are usually represented with their expected values (or another point estimate) during the data collection period. This makes it challenging to identify spots with relatively high crash risk, as well as to determine the potential contributing factors to crashes that occur at such locations. Without properly recognizing the major factors that lead to the high risk locations being identified, countermeasures selected to reduce traffic crashes may not be as effective as expected, and therefore, the expenditure of valuable resources may not be optimized.

1.1.2 Importance of Taking into Account Crash Severities for Hotspot Identification

Hotspot identification (HSID) is of great importance to transportation authorities in their efforts to improve highway safety. Qu and Meng (2014) concluded that the severity of crashes should not be neglected in the HSID process. Figure 1-1 shows the distribution and locations of traffic fatalities by county in Washington state for 2014 (NHTSA, 2014). However, hotspots corresponding to locations with high crash risk can be quite different when crash frequency by different levels of crash severity is considered. It is particularly important to take into account crash severities in site ranking because the costs of crashes are significantly different at various severity levels. This means that, for instance, a road segment with a few fatal crashes may be considered more hazardous than a road segment with a lot of property damage only crashes and zero fatalities. Therefore, it is necessary to estimate accident frequency for each severity type separately.

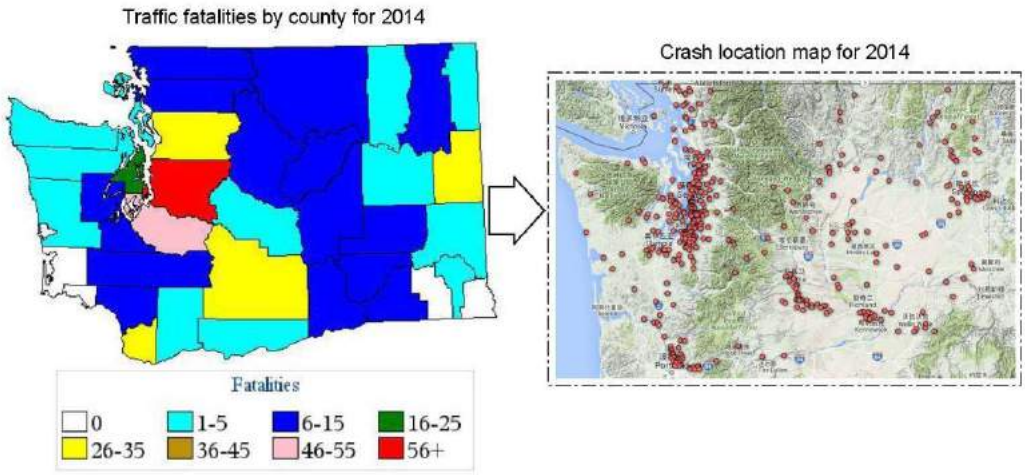


Figure 1-1 Distribution and location maps of traffic fatalities by county in Washington state for 2014 (NHTSA, 2014)

1.1.3 The Need for Developing a Regional, Map-Based Platform

According to the 2015 WSDOT Gray Notebook, there were 462 traffic fatalities and 2,010 serious injuries on Washington's public roads in 2014. These numbers represent an

increase from 2013 (a 6 percent increase for traffic fatalities and a 5 percent increase for serious injuries, as illustrated in figure 1-2). In 2010, WSDOT began a new Strategic Highway Safety Plan that aims to end traffic-related deaths and serious injuries by 2030.

Recently, WSDOT revised its safety program by instituting the Sustainable Highway Safety Program (Sustainable Safety), a more integrated and analytic multimodal approach. Sustainable Safety continues to evolve from a reactive approach, in which safety enhancements are applied to areas with a history of crashes, to a more proactive, risk-based approach in which WSDOT predicts and analyzes crash locations by evaluating the factors that contribute to crashes. To accomplish such a goal, an online platform that could enable quick and easy quantification of safety performance measures over the state highway network would be highly desirable. Such a platform would significantly cut down the time and labor hours needed to conduct safety analyses. The proposed regional, map-based analytical platform for highway safety performance assessment directly addresses this need.

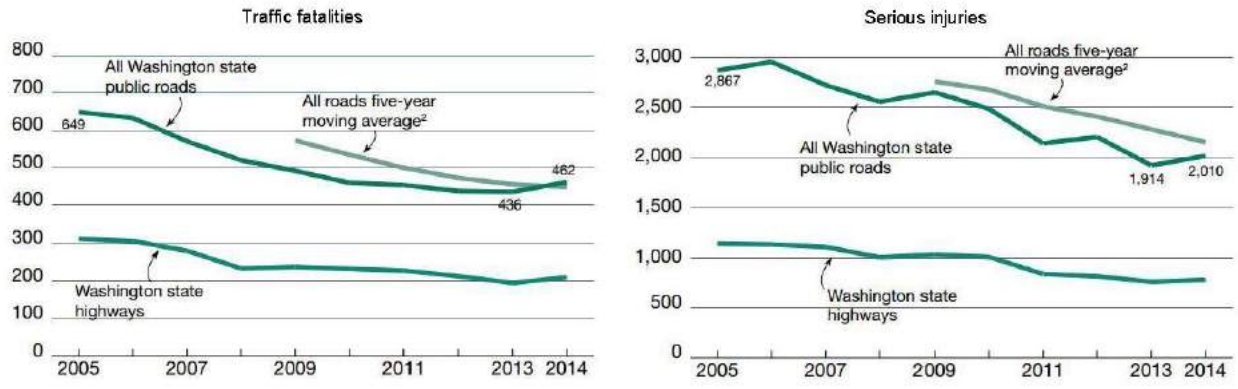


Figure 1-2 Trends of traffic fatalities and serious injuries on all Washington state public roads from 2005 to 2014

1.2 Problem Statement

Over the past several decades, a number of studies have focused on finding the relationships between traffic crash frequencies and potential contributing factors such as

roadway geometry and environmental, traffic, and human factors. A number of statistical modeling techniques have been developed on the basis of the diverse characteristics of collisions in different circumstances. The majority of these models are generalized linear models (GLMs), such as Poisson (Jovanis and Chang, 1986; Miaou and Lum, 1993; Miaou, 1994), gamma (Winkelmann and Zimmermann, 1995; Oh, Washington and Nam, 2006), negative binomial (Miaou, 1994; Maher and Summersgill, 1996; Milton and Mannering, 1998; Chin and Quddus, 2003; Wang, Ieda and Mannering, 2003; Wang and Nihan, 2004; Donnell and Mason, 2006; Kim, Wang and Ulfarsson, 2007; Daniels *et al.*, 2010; Malyshkina and Mannering, 2010a), random-parameters (ElBasyouny and Sayed, 2006; Anastasopoulos and Mannering, 2009) and bivariate/multivariate regression models (Park and Lord, 2007; Lao *et al.*, 2011).

Conventional GLMs provide a straightforward way to examine the relationship between crash frequency and crash contributing factors. However, the GLM-based approach is constrained by its linear model specifications. For example, the crash rate (number of crashes per million vehicles per year) is believed to increase with traffic volume until a certain point, after which it should start to decrease with increasing traffic volume (Wang *et al.*, 2010). The conventional GLMs are not capable of modeling such relationships. This inappropriate relationship enforced by conventional GLMs may lead not only to inaccurate modeling results but also inaccurate crash frequency elasticity analyses. The elasticity analyses of specific factors and locations are important since transportation agencies rely on estimated elasticity to understand the potential improvements in traffic safety that may be realized at a location.

Another issue for the traditional frequency-based modeling methods is their estimation assumption that fatality rates are identical across locations with different volumes (Milton, Shankar and Mannering, 2008). Significant error may be introduced when this assumption is violated. Thus, a new method to simultaneously consider both incident frequency and severity is

highly desired. As shown by our literature review, very few studies monitor and assess safety performance on a regional level (Ivan *et al.*, 2007)

1.3 Research Objectives

Specifically, the proposed research had the following objectives:

(1) Improve current crash modeling methods (by introducing a two-stage regression model and a non-linear prediction function in the form of generalized nonlinear models (GNMs)) to describe the relationship between injury-severity and its contributing factors.

(2) Develop a safety performance index (SPI) by considering crash frequency at different levels of crash severity, based on the accident frequency expected from the improved modeling approach and associated factors, to reflect safety condition changes with roadway, vehicle, and environmental factors.

(3) Monitor the safety performance of the state highway network on a regional map by using the SPI.

(4) Develop a potential safety improvement index (PSII) for identification of the key factors that may lead to safety improvements at each location on the basis of the elasticity estimates in the new model.

(5) Develop safety improvement analysis methods for accident hotspots based on the overlap of the SPI and PSII.

1.4 Methodology

The research team conducted a comprehensive review of literature to identify static and dynamic contributing factors to traffic crashes and appropriate segmentation and statistical analysis approaches. The research team developed data collection and analysis plans. Using the collected data, the team developed various prediction models for different crash severity types.

To accommodate large-scale safety analysis and performance monitoring, both SPI and PSII were developed by considering crash frequencies with different levels of crash severity, and they were further implemented on an interactive regional map. The proposed tool was built on the Digital Roadway Interactive Visualization and Evaluation Network (DRIVE Net) system (Ma, Wu and Wang, 2011), in which various layers of needed data have already been internally connected, to save cost for this study.

The highway SPI can be used as a basis for color-coding the map to show the safety performance of each road segment. The PSII, computed from elasticity estimates at each location, can be employed to highlight potential safety improvements on the state highway network. The highway SPI changes with traffic and roadway conditions and may be used to monitor safety performance of highway segments in real time. The PSII can be used to identify and prioritize safety improvement measures for a specific segment. By combining these two indices on the regional map, one can easily identify accident hotspots and the potential contributing factors to be considered in a safety improvement package to help WSDOT accomplish its goal of zero fatal collisions.

Chapter 2 Literature Review

Traffic safety performance assessment largely relies on incident prediction and hotspot identification (HSID). Recent advances in incident prediction models have focused on improving estimations of incident duration, crash severity, and crash frequency. Some researchers have also investigated the modeling of crash counts by severity (CCS) and crash counts by collision type. The key methods for incident duration prediction include parametric models, nonparametric models, and hazard-based duration modeling methods. For injury

severity, logit models and various extensions are the most commonly used. The research for crash frequency prediction has mainly focused on the development of generalized linear and nonlinear models. There have also been several identification and evaluation methods developed for HSID. Figure 2-1 depicts a research path in the field of traffic safety performance assessment.

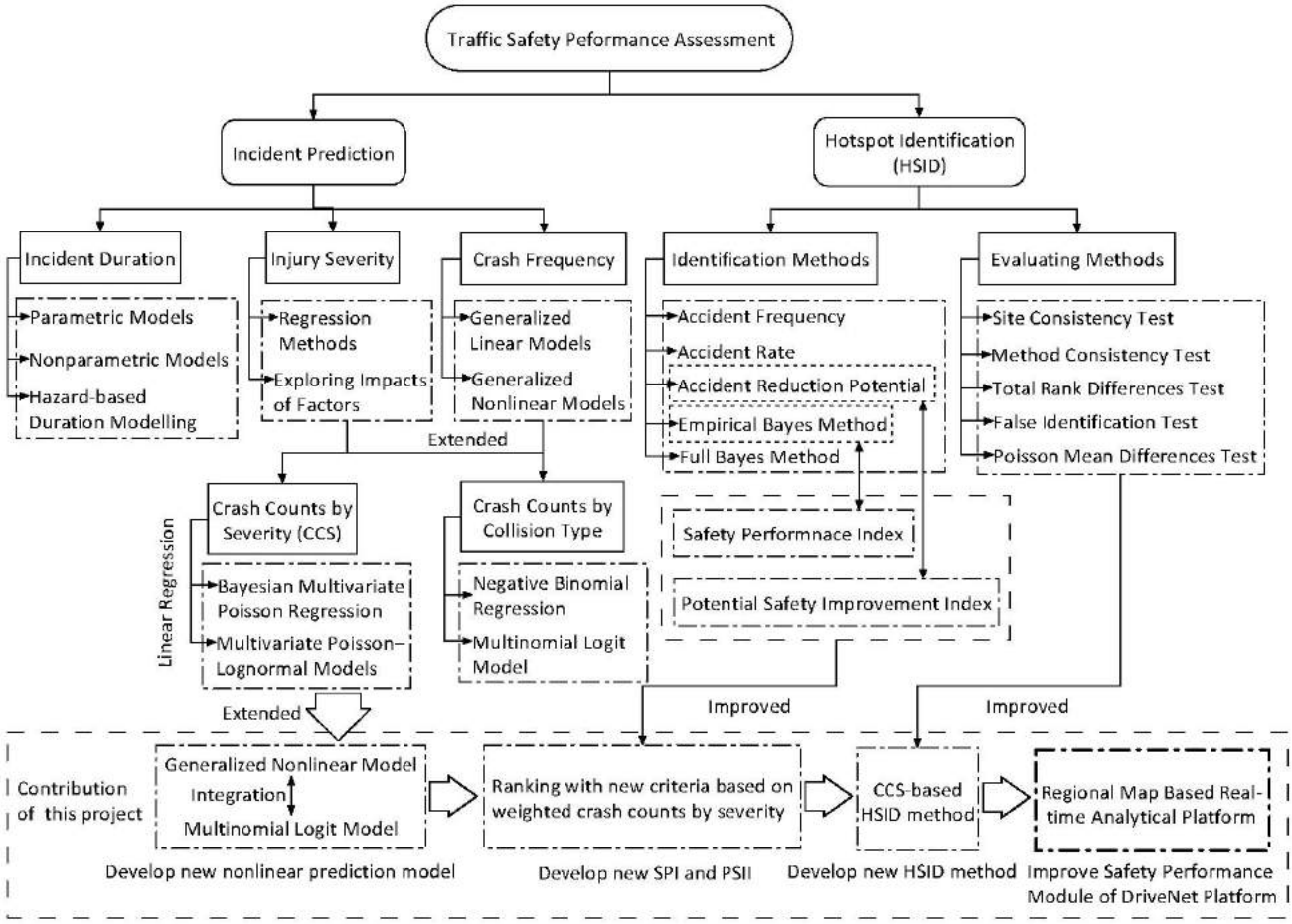


Figure 2-1 The research path in the field of traffic safety performance assessment

2.1. Incident Duration

A variety of methods have been developed to model traffic incident duration in the last several decades. These methods include analysis of variance, parametric regression, nonparametric regression, hazard-based methods, decision trees, fuzzy logic, artificial neural networks, and Bayesian network models. A review of recent research on the prediction of incident duration showed that the majority of previous research activities have found that the distribution of incident durations to be skewed, with a long tail to the right, a shape that is similar to the lognormal or Weibull distribution. Incident duration typically has large variance in

comparison to the average duration. Most results have shown that the standard deviation is around 70 percent of the mean.

Poor data quality is a common problem that many researchers have acknowledged and dealt with in different ways. Some data sets are not complete (i.e., they are missing select observations entirely or have incomplete observations for which not all measures were recorded), and others may contain inaccurate data, which can drastically affect prediction performance. Other data sets are insufficient in size, which leads to issues with statistical power and making reliable predictions. The aforementioned data problems may be caused by integrating different incident data sources from parties that include law enforcement, transportation authorities and insurance companies, as was noted in previous research.

2.1.1 Parametric Models

To date, various parametric models have been proposed for the analysis of traffic incident durations. Lord and Mannering (2010) and Savolainen, Mannering et al. (2011) documented a review of a number of methodological alternatives used in traffic safety analysis. Among these parametric models, the accelerated failure time (AFT) models have been the most widely used in previous studies (Weng *et al.*, 2014). Yang *et al.*, (2015) indicated that the Weibull AFT model with shared frailty is appropriate for modeling pedestrian waiting durations. Junhua, Haozhe and Shi (2013) estimated freeway incident duration by using AFT modeling, which also helped overcome difficulties associated with missing data. Statistical models have been the widest used parametric techniques in traffic safety analysis for many years. Some examples follows:

- Poison regression (Jovanis and Chang, 1986; Joshua and Garber, 1990; Miaou and Lum, 1993; Miaou, 1994)

- Negative binomial regression (Lord, Washington and Ivan, 2005; El-Basyouny and Sayed, 2006; Kim and Washington, 2006; Malyshkina and Mannering, 2010a)
- Zero-inflated Poison and negative binomial regression (Shankar *et al.*, 2003; Lord, Washington and Ivan, 2005; Malyshkina and Mannering, 2010b)
- Random effects model (Li *et al.*, 2008; Wang *et al.*, 2010) etc.

However, parametric models rely on lots of assumptions that may not represent reality.

2.1.2 Nonparametric Models

Nonparametric models gained popularity to eliminate some drawbacks of parametric models. Stewart (1996) pointed out the advantages of using the classification and regression tree (CART) model in traffic safety analysis for determining complex interactions among the variables. The application of nonparametric regression trees in the other safety literature includes prediction of crash frequency (Karlaftis and Golias, 2002; Chang and Chen, 2005), rear-end crash analysis (Yan and Radwan, 2006), and exploration of the effects of drivers', vehicles', and environments' characteristics associated with crash avoidance maneuvers (Harb *et al.*, 2009). Another non-parametric approach, Bayesian networks (BN), can better interpret the complex relationships among variables. Gregoriades (2007) used BN to identify accident prone spots on roadway networks. Other applications of BN include severity analysis (de Oña, Mujalli and Calvo, 2011; de Oña *et al.*, 2013) and traffic crash causality mechanisms (Hongguo, Huiyong and Fang, 2010). Other nonparametric models, such as artificial neural networks (Wei and Lee, 2007; Lao *et al.*, 2011), fuzzy logic models (Wu and Chen, 2008), text analysis approach (Pereira, Rodrigues and BenAkiva, 2013), and support vector machine (Li *et al.*, 2008, 2012) have been proposed for crash analysis. While nonparametric models do not assume a functional form and thus are more flexible, it is harder to interpret the marginal effects of

independent variables with a nonparametric model (Weng *et al.*, 2015). They are also harder for generalization to other data sets (Lord and Mannering, 2010).

2.1.3 Hazard-Based Duration Modeling Method

Hazard-based duration modeling is used to study the conditional probability of a time duration ending at some time *t*, given that the duration has continued until time *t* (Hensher and Mannering, 1994), and it has been extensively applied to problems within the transportation discipline including modeling the time between individuals' traffic accidents, the time between incident occurrence and clearance, and the time between trips (Alkaabi, Dissanayake and Bird, 2011; Psarros, Kepaptsoglou and Karlaftis, 2011; Hojati *et al.*, 2013; Li and Shang, 2014; Lin, Wang and Sadek, 2016). Because of the fact that the time variable is connected with a conditional probability, hazard-based duration modeling has an advantage in that it allows the explicit study of the relationship between incident duration and the explanatory variables. Nam and Mannering (2000) removed impractical variables from a hazard-based model and developed a sub-model for each stage of incident duration, including incident detection, duration, the time of response, and clearance.

2.2. Injury Severity

Many studies have focused on estimating crash severity (Anarkooli and Hosseinlou, 2016). A large number of these studies have been conducted to determine significant factors influencing the increased levels of injury severity of crashes. Also, various techniques have been employed in order to explore the effects of these factors on injury severity. These techniques can be classified into four major groups: discrete outcome models, data mining methods, soft computing, and regression methods (Mujalli and de Oña, 2013). In addition, we can categorize the contributing factors into five groups: driver conditions, vehicle characteristics, roadway

geometry and traffic conditions, environmental conditions, and types of collisions (Lee and Li, 2014).

Discrete choice models are popular for predicting injury severity because of the discrete outcome. Krull, Khattak and Council (2000) used a logistic model to study the impact of driver, vehicle, roadway geometrics, and traffic-related impacts on the probability of fatality and injury. Similarly, Bedard *et al.* (2002) applied multivariate logistic regression to determine the fatality risk of drivers involved in crashes.

Some researches adapted the multinomial logit model (MNL) to predict crash severity. The outcomes of MNL models are not ordinal. So, MNL ignores some restrictions assumed by standard ordered models (Zhang, 2010). Bham, Javvadi and Manepalli (2011) studied the differences in crash-contributing factors in collision types using MNL. MNL has also been used to analyze the severity of crashes in rural and urban areas for crashes involving large trucks (Khorashadi *et al.*, 2005), the effects of increasing the speed limit (Malyshkina and Mannering, 2008), and factors related to work zones (Robin, 2014). Kockelman and Kweon (2002) investigated the effects of vehicle types, crash type, weather, speed, and occupant characteristics on crash severity (i.e., no injury, not severe injury, severe injury, and death). In the MNL model, the odds between any two outcomes are independent of the number and nature of other outcomes being simultaneously considered. On the other hand, nested logit (NL) allows correlations between choices by nesting them. Several studies, such as those by Nassar, Saccomanno and Shortreed, 1994; Abdel-Aty and Abdelwahab, 2004; Savolainen and Mannering, 2007; and Patil, Geedipally and Lord, 2012, have used NL structure to analyze crash severity.

O'Donnell and Connor (1996) used ordered logit and ordered probit models while comparing the injury severity as a function of drivers' characteristics. They identified the

victim's age and vehicle speed as the main contributing factors to crash severity. An ordered probit model was also used by Duncan, Khattak and Council (1998) to examine the effects of occupant characteristics and roadway and environmental conditions on crash severity in rear-end crashes between trucks and passengers. Khattak (1999) presented another ordered probit model to examine the effects of information accuracy on rear-end crash propagation. Abdel-Aty (2003) tested multinomial logit and nested logit models with different nesting structures and compared their results to the results of a probit model. This comparison showed that the ordered probit model was simpler and at the same time produced better results than the multinomial logit model.

The results of Ma *et al.* (2015) showed that several explanatory variables, including at-fault driver's age, at-fault driver having a license or not, alcohol usage, speeding, involvement of pedestrians, type of area, weather condition, pavement type, and collision type, significantly affect crash severity. Mergia *et al.* (2013) found that semi-truck-related crashes, higher number of lanes on freeways, higher number of lanes on ramps, speeding-related crashes, and alcohol-related crashes tend to increase the likelihood of sustaining severe injuries at freeway merging locations. Wang, Chen and Lu (2009) found that the factors that significantly influence injury severity at freeway diverge areas include length of deceleration and ramp lanes, curve and grade at diverge areas, light and weather conditions, alcohol or drug involvement, heavy-vehicle involvement, number of lanes on the mainline freeway section, average daily traffic on the mainline, pavement surface condition, land type, and crash type.

2.3. Crash Frequency

Existing research on crash frequency prediction is mostly regression analyses based. All models divide the roadway into several segments and aggregate all the roadway, environmental, and crash data on the basis of road segments. Different types of regression models are used to

predict crash frequency in the form of safety performance functions (SPF). An SPF is an equation that predicts crash density (typically number of crashes each year) at a location as a function of prevailing conditions and roadway characteristics. Multiple linear regression analysis techniques have been extensively used to develop crash prediction models (Okamoto and Koshi, 1989; Miaou and Lum, 1993; Golob and Recker, 2003). Jovanis and Chang (1986) identified a number of disadvantages of using linear regression in the accident prediction context, such as violation of homoscedasticity and prediction of a negative number of accidents. To deal with this problem, researchers adapted count data models such as the Poisson, negative binomial, zero-inflated Poisson, zero-inflated negative binomial, and hurdle regression models, assuming log-linear relationships between crash frequency and explanatory variables. Poisson and negative binomial regression models have been used to predict numbers of traffic crashes (Jones, Janssen and Mannering, 1991; Hadi *et al.*, 1995; Shankar, Mannering and Barfield, 1995; Poch and Mannering, 1996; Milton and Mannering, 1998; Abdel-Aty and Radwan, 2000; Savolainen and Tarko, 2005). Similarly, Savolainen and Tarko (2005) found a statistical relationship between crash occurrence and intersection geometric characteristics, including curvature of the main road. Lord, Guikema and Geedipally (2008) proposed a model that describes random, discrete, and non-negative accidents, assuming an exponential relationship between the number of accidents and the contributing factors. This approach solved the problem associated with homoscedasticity in linear regression model; however, these models were associated with over-dispersion and heterogeneity.

The negative Binomial regression model was developed to address the over-dispersion issue. Miaou and Lum (1993) showed that the negative binomial model is limited in dealing with the randomness of the shape parameters. On the other hand, El-Basyouny and Sayed (2006) found that in terms of model application (identification and ranking of accident-prone locations),

randomness in the shape parameter of the negative binomial regression can give satisfactory results. Anastasopoulos and Mannering (2009) concluded that random shape parameters need to be used when the standard deviation of density is statistically significant. Kumara and Chin (2003) applied zero-inflated negative binomial models to capture the apparent "excess" zeros that commonly arise in crash data. The zero-inflated Poisson and negative binomial models were also applied to estimate crash counts in Shankar *et al.* (2003); Qin, Ivan and Ravishanker (2004); and Lord, Washington and Ivan (2005, 2007). Malyshkina and Mannering (2010b) proposed a two-state Markov switching count-data model as an alternative to zero-inflated models to account for the excess zeros in crash frequency analysis. The Markov switching approach allows direct statistical estimation to switch between zero and a crash count, whereas traditional zero-inflated models do not.

As discussed previously, there may be reason to expect spatial and temporal correlation among observations. To account for such correlation, using random and fixed effects in crash frequency analysis on panel data is quite common. Random-effects models assume a spatial and temporal distribution of the unobserved effects with explanatory variables. On the other hand, fixed-effects models account for unobserved effects by indicator variables and assume correlation of unobserved effects with individual variables. Random effects in the context of crash frequencies have been studied by a number of researchers, including Johansson (1996); Shankar *et al.* (1998); Miaou, Song and Mallick (2003); and Kweon and Kockelman (2004). Yaacob, Lazim and Wah (2012) analyzed crash frequency by using the fixed effect model.

Anastasopoulos and Mannering (2009) explored the use of random-parameters count models as another methodological alternative in analyzing accident frequency. They concluded from their empirical results that random-parameters count models can provide a full understanding about the factors influencing crash frequencies.

Generalized linear models have also been employed to address collision prediction and crash frequency at intersections (Geedipally, Lord and Dhavala, 2012). Heydari, Miranda-Moreno and Liping (2014) used two Bayesian generalized mixed linear models to predict crash speed and frequency. These models have the advantage of addressing the heterogeneity problem in observations and efficiently capturing potential intra-site correlations. Chen and Tjandra (2014) developed generalized linear models on the basis of temporal and weather variables to predict daily total collisions. Their models were applied to support scheduling of traffic operations, maintenance and enforcement, and dispatch of material and personnel resources. Zhang *et al.* (2014) use generalized linear regression models to predict the frequency of opposing left-turn conflicts at signalized intersections. They found that the use of conflict predictive models has potential to expand the uses of surrogate safety measures in safety estimation and evaluation.

For GNMs, only limited studies in the field of incident prediction have been conducted. Lee *et al.* (2015) used generalized nonlinear models to develop crash modification factors (CMF) for changing lane width on roadway segments. The study demonstrated that the CMFs estimated with GNMs clearly reflect variations in crashes with lane width, which cannot be captured by the CMFs estimated with GLMs. Lao *et al.* (2014) formulated a generalized nonlinear model-based approach for modeling highway rear-end crash risk using Washington state traffic safety data. Their results showed that truck percentage and grade have a parabolic impact: they increase crash risk initially, but decrease it after certain thresholds. Such non-monotonic relationships cannot be captured by regular GLMs, which further demonstrate the flexibility of GNM-based approaches in analyzing the nonlinear relationships among data and providing more reasonable explanations.

2.4. Crash Counts by Severity

Numerous efforts have been devoted to investigating crash occurrences as they relate to roadway design features, environmental factors and traffic conditions (Ma and Kockelman, 2006). However, most research has modeled crash counts at different severity levels separately, which can lead to biased results in terms of parameter estimates and other model aspects (Ma, Kockelman and Damien, 2008). Ma, Kockelman and Damien (2008) further offered a multivariate Poisson-lognormal (MVPLN) specification that simultaneously modeled crash counts by injury severity. The MVPLN specification allowed for a more general correlation structure as well as accounting for over-dispersion.

2.5. Crash Counts by Collision Type

Over the past 20 years, a few researchers have developed crash prediction models by collision type. Hauer, Ng and Lovell (1988) were the first to develop such models. They developed models for 15 crash patterns at urban and suburban signalized intersections in Toronto, Ontario, Canada. Shankar, Mannering and Barfield (1995) developed models for six crash types. They concluded that models that predict crashes for different crash types have a greater explanatory power than a single model that predicts for all crash types combined together. Kockelman and Kweon (2002) developed crash type models (e.g., total, single-vehicle, and multi-vehicle crashes) by using ordered-probit models to examine the risks associated with various driver injury severity levels. Their study estimated the safety effects on drivers of different types of vehicles. Geedipally, Patil and Lord (2010) investigated the applicability of multinomial logit (MNL) models to predict the proportion of crashes by collision type and to estimate crash counts by collision type. Their method based on the MNL model was found useful to estimate crash counts by collision type, and it performed better than the method based on the use of fixed proportions.

2.6 Contributing Factors to Traffic Crashes

The risk of crashes during rain and snow is greater than that in dry weather, even though motorists overtake less, drive more slowly, and maintain more reasonable following distances under inclement weather conditions (Hogema and Van der Horst, 1994; Agarwal, Maze and Souleyrette, 2005). The changes in driving behavior are, apparently, insufficient to compensate for the greater risk during bad weather (Bijleveld and Churchill, 2009). First, visibility decreases during rainfall. This is even more intense at night, since the light reflection on a wet road makes the detection of the road and objects nearby more difficult (Brodsky and Hakkert, 1988). Reduced visibility during precipitation, splashing water from other vehicles, and clouded windows as a result of high humidity during rain are the causes that lead to crashes. A layer of water on the road surface can also cause vehicles to lose contact with the road surface and skid (Bijleveld and Churchill, 2009).

Many studies have focused on contributing factors to traffic crash frequency and severity (Tay and Rifaat, 2007; Haleem and Gan, 2015; Ma *et al.*, 2015). Abdel-Aty and Radwan (2000) showed that heavy traffic volumes, speeding, narrow lane width, more lanes, urban roadway sections, narrow shoulder width, and reduced median width increase the likelihood of accident involvement. Rainfall constitutes a driving hazard for a number of reasons. Jung *et al.* (2014) combined vehicle to vehicle crash frequency and severity estimations to examine factor impacts on Wisconsin highway safety in rainy weather. They found that higher levels of average daily rainfall per month and wider left shoulder widths are factors that decrease the likelihood of vehicle to vehicle crashes.

2.7. Hotspot Identification

The identification of crash hotspots, also referred to as hazardous road locations, high-risk locations, accident-prone locations, black spots, sites with promise, or priority investigation

locations, is the first step in the highway safety management process (Montella, 2010). There is a fairly extensive body of literature focused on methods for hotspot identification (HSID) (Cheng and Washington, 2005, 2008; Park, Lord and Lee, 2014). Studies have discussed methods based on accident count or frequency (Deacon, Zegeer and Deen, 1974), employed both accident rate (AR) and rate quality control (Stokes and Mutabazi, 1996), and adopted the joint use of accident frequency and rate to flag sites. To correct for the regressionto-the-mean bias associated with typical HSID methods (Hauer, 1980), some researchers have suggested using the empirical Bayes (EB) techniques (Hauer *et al.*, 1991). This method combines clues from both the accident history of a specific site and expected safety of similar sites, and has the advantage of revealing underlying safety problems that otherwise would not be detected.

2.8 Safety Performance Indices and Potential Safety Improvement Indices

Safety performance indices are increasingly used to identify and combat the rising problems of road safety. Put simply, a road safety performance index is defined as a quantitative or qualitative metric based upon/developed from a series of observed characteristics from a specific collision (Wegman *et al.*, 2008; Coll, Moutari and Marshall, 2013). Examples of road safety performance indices include number, frequency, and rate of crashes, number and severity of injuries, number of vehicles involved in collisions, type of collision, etc. Safety performance indices are useful in the sense that they can simplify the presentation of larger amounts of data. That said, because many factors affect traffic crashes, it can be difficult to evaluate such indicators individually. Therefore, decision-makers may prefer a unified, composite index. Such an index is commonly called the Composite Safety Performance Index (CSPI) (Wegman *et al.*, 2008; Coll, Moutari and Marshall, 2013). Some researchers have suggested using the EB method and accident reduction potential (ARP) to develop the CSPI (Cheng and Washington, 2008; Coll, Moutari and Marshall, 2013). The current safety performance module in DRIVE

Net, however, uses the estimated crash frequency as the SPI and the ARP as the PSII. New safety performance and potential safety improvement indexes based on crash counts by severity were developed in this research.

Chapter 3 Identify Static and Dynamic Contributing Factors

The factors contributing to traffic crashes of different severity types were identified by using the framework developed in the following sections. The structural framework for identifying factors contributing to traffic crashes and the methods for identifying static and dynamic contributing factors are explained in detail below.

3.1 Structural Framework for Identifying Contributing Factors

Many studies have investigated the factors that influence traffic crashes (Ma *et al.*, 2015). Karlaftis and Golias (2002) found that roadway geometry and pavement condition significantly affect accident rate. Farah, Bekhor and Polus (2009) analyzed drivers' passing decisions on rural two-lane highways based on data collected from an interactive driving simulator. They found that traffic conditions, roadway geometry, and drivers' characteristics—such as speed of the subject vehicle, gender, and age—have a significant effect on the risk associated with the passing behavior. De Oña, Mujalli and Calvo (2011) found that other factors, such as driver age, crash type, and lighting condition, also affect injury severity. On the basis of the state-of-the-art research on these factors influencing traffic crashes, we classified contributing factors into three types, i.e., traffic characteristics, road conditions, and weather conditions, as shown in figure 3-1. Road conditions were regarded as static contributing factors, since road types and geometry rarely change over short time periods. On the other hand, traffic characteristics change over the year, but they never change drastically unless a major change happens in geometric characteristics. So we included annual average daily traffic (AADT) in the analysis as an average AADT for the study period. Weather conditions usually vary with time and were regarded as a dynamic contributing factor. Crash frequencies, by severity, were considered to be the dependent variables in the incident prediction model. In fact, crash injury severity is usually categorized into five levels, including fatal, incapacitating injury, non-

incapacitating injury, possible injury, and no injury or property damage only. Because of limited numbers of crash records, they were classified into two categories for this study: injury-fatal (IF) and property damage only (PDO).

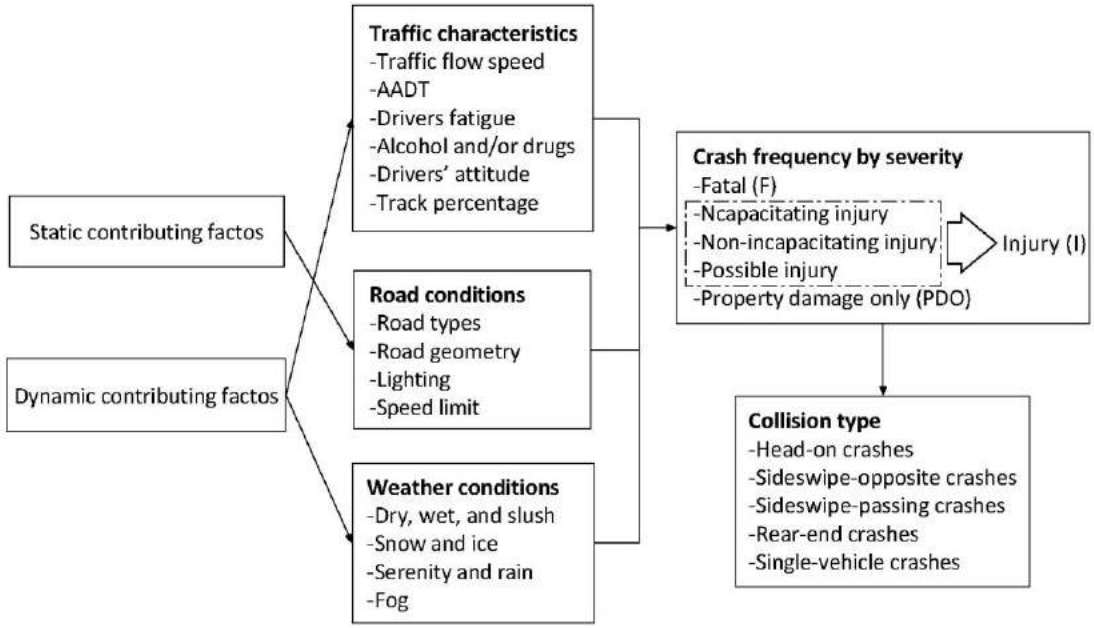


Figure 3-1 Structural framework for contributing factors to traffic crashes

Based on the structural framework in figure 3-1, the approach to identifying contributing factors to traffic crashes was developed as illustrated in figure 3-2. There were six main steps for identifying the factors that contribute to different types of crash severities: (1) review and selection of factors; (2) classification of factors and development of hypotheses; (3) data collection; (4) detection of multicollinearity among factors; (5) regression modeling for each level of crash severity; and (6) significance testing for the factors. The selected contributing factors, which were classified into the three categories of traffic characteristics, road geometry, and weather conditions, were assumed to have a linear or nonlinear relationship with crash frequency at different severity levels. A data collection plan was made for the selected factors (see Chapter 4) and multicollinearity tests were performed. Various statistical models, including a generalized nonlinear model-based multinomial logit regression approach, were developed to

predict the crash rate and crash frequency of each type of severity (see Chapter 5). Finally, significance testing was implemented on the regression model to remove the insignificant predictors under a certain confidence level.

3.2 Identify Static Contributing Factors

The contributing factors were classified into road conditions road geometry (e.g., horizontal curve type, length of the segment, curvature of the segment, average number of lanes, lane width, and shoulder width), and speed limit. Because these factors are usually constant over longer periods of time, they were identified as static contributing factors. AADT remains relatively constant over time as long as the roadway type and land use characteristics stay constant. Over the whole study period, we assumed all the roadway factors to be constant. Consequently, the average value of AADT over the study period was considered to be a static contributory factor.

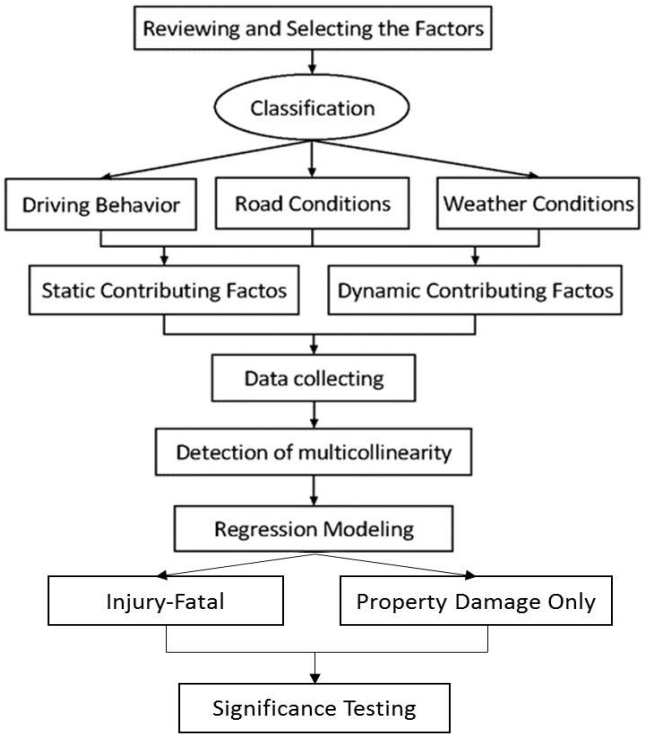


Figure 3-2 Flowchart for identifying contributing factors to traffic crashes

3.3 Identify Dynamic Contributing Factors

Weather conditions are usually time-dependent factors and thus were identified as dynamic contributing factors. The weather conditions were classified into two types, good weather condition (e.g., clear and dry) and adverse weather condition (e.g., rain, snow, ice, slush, wet). Dealing with dynamic contributory factors in a crash frequency model is relatively difficult. It is common that the largest number of crashes occurs in clear weather, but clear weather is also very common on road segments. So giving equal weight to crashes in both clear and adverse weather condition may result in a model that over-predicts crashes in normal weather conditions. However, according to the literature discussed in Chapter 2, crashes are more likely to happen in adverse weather conditions. To deal with this problem, crashes were weighted on the basis of the duration of normal and inclement weather conditions. The data were gathered from airports in Washington State.

Chapter 4 Data Collection and Analysis

This chapter explains the data collection plan, data analysis plan, and the summary statistics of the collected data.

4.1 Data Collection Plan

In this research, a data collection plan is developed to ensure that (1) appropriate data on the most important contributing factors to incidents were collected; (2) incidents with different severity types were included in the data collection plan; (3) highway facilities with different geometric conditions were included (e.g., different pavement types, different grades, different horizontal curvatures, etc.); and (4) acceptable ranges of dynamic contributing factors were included (e.g., different traffic volume levels, different weather conditions, etc.).

In this task, the research team worked together with Washington State Department of Transportation (WSDOT) to identify various data sources and to collect the required data. Crash data were collected in terms of crash frequencies and severity levels. Usually, statistical models are produced for all crash severity levels (often referred to as KABCO, i.e., fatal (K), incapacitating injury (A), non-incapacitating injury (B), minor injury (C), and property damage only (PDO or O)) or for different crash severity levels, such as fatal and nonfatal injury crashes (e.g., KABC) or for PDO crashes. Although the data on crash frequency by severity are multivariate in nature, they have often been analyzed by modeling each severity level separately, without taking into account correlations that exist among different severity levels. In this research, the collected crash frequency data were classified into two categories: injury-fatal (IF) and property damage only (PDO).

Crash data, roadway geometric characteristics, and AADT values were obtained from WSDOT for all Washington state owned roadways from 2011 to 2014. The candidate sites for data collection were selected as a total of 802 road segments on various Interstate highways in

Washington, including I-5, I-90, I-82, I-182, I-205, I-405 and I-705. The data collection period was from 2011-2014 for a total of four years of data. The total number of crashes recorded during the data collection period was 45,270, including 129 fatal crashes, 13,189 injury crashes, and 31,952 PDO crashes. Figure 4.1 illustrates the study area. The orange, red, and green lines denote the selected road segments on I-5, I-90, and I-82, respectively. It is well known that I-5 has one of the highest rates of fatal crashes across all Interstates in the U.S. A total of 29,020 crashes occurred on I-5 over 276.54 miles. All of the crash data were provided by WSDOT.

Figure 4-1 Study area for I-5, I-90, and I-82 in Washington

Weather conditions during the crashes can be found from the WSDOT crash report. But WSDOT does not maintain any weather information for the freeways throughout the year. So the team collected weather data from the ASOS website (*Automated Surface Observing System : ASOS user's guide*, 1998). This automated data collection system provides 1-minute, 5- minute,

hourly, and special observation data 24 hours a day. ASOS collects and archives data from various weather stations located at airports. In this analysis, the team assumed that the weather data of a station is representative of its neighboring roads. We collected the archived precipitation data for every 15 minutes. Then the team calculated the probability of the precipitation of a site from total hour of precipitation. In this analysis, the team assumed that the roadway surface remained wet throughout the precipitation. Figure 4-2 shows the locations of all the stations in Washington state (*Automated Surface Observing System : ASOS user's guide*, 1998).

Figure 4-2 Locations of all the stations in Washington state *(*Automated Surface Observing System : ASOS user's guide, *1998)*

4.2 Data Preparation Plan

After collecting the data, a sequence of steps, starting with data cleaning, data integration, and data selection, were required to process the data for modeling purposes. These steps are addressed in detail.

4.2.1 Data Cleaning

It is quite impossible to collect a large amount of data without any noise and inconsistency. The data set might be missing important information, such as the milepost and travel direction of the involved vehicles, road surface conditions during a crash, or precipitation duration in the weather data, as all the information is recorded manually. Hence, the records with missing information were discarded during the data preprocessing, as they could have been misleading in the estimataion models. Some of the information that was excluded from the data set are included the following:

1. crash records with missing inputs were not included for analysis
2. missing precipitation data were excluded from analysis.

4.2.2 Data Quality Control Based on Segmentation

Success in developing crash prediction model depends on data quality (Cafiso, D'Agostino and Persaud, 2013). The most common technique for developing a crash prediction model is to divide the road sections into a number of segments. Then, all the crashes on any road segment are analyzed on the basis of roadway geometrics, traffic, and weather characteristics. Generally, segmentation of the roadway can be done in different ways (Cafiso, D'Agostino and Persaud, 2013):

- Completely homogeneous, i.e., a new segment will start if any attribute changes;

- Based on curvature, i.e., a new segment starts when a tangent section transitions into a curve;
- Homogeneous based on AADT, i.e. a new segment starts when AADT changes;
- Fixed length segment, i.e., all the segments are of fixed length (typically 1 or 2 miles);
- Two curves, two tangents, i.e., each segment consists a curve and a tangent.

Crash locations are identified on the basis of police reports. Police reports estimate crash location on the basis of the spot where the vehicle was located upon police arrival and therefore do not always show the exact location. Furthermore, the number of accidents is often proportional to segment length. Therefore, if a segment is too short (less than 100 feet), the probability of an accident will be very small. As a result, in order to improve the quality of the data sets, the road segments with short lengths should be removed or aggregated to create longer segments.

Segmentation, when based on multiple variables, may lead to very short homogeneous segments (Resende and Benekohal, 1997). For example, when the segmentation approach proposed by the *Highway Safety Manual* (HSM) is used, the presence of very short segments does not allow proper statistical inference for several reasons (AASHTO, 2010). The most important are the imperfect identification of crash locations, which are often taken from police reports (Qin and Wellner, 2012), and the fact that crashes are rare events, resulting in a great number of segments with zero crashes. Lengthening segments to avoid these issues will sacrifice homogeneity.

In the literature there are a number of different segmentation approaches. Miaou and Lum (1993) suggested that short sections, less than or equal to 80 meters, could create bias in the estimation of linear models, but not when using Poisson models. Similarly, Ogle, Alluri and

Sarasua, (2011) demonstrated that short segment lengths, less than 160 meters, can lead to uncertain results in crash analyses. Cafiso and Di Silvestro (2011) showed that to increase performance in identifying correct positives as black spots, segment length should be related to AADT, with lower AADT values requiring longer segment lengths. Qin and Wellner (2012) studied the relationship between segmentation and safety screening analysis by using different lengths of sliding windows to identify hazardous sites, and they concluded that short segments, as well as those that are too long, create a bias in the identification of sites with safety problems.

Some studies have focused on the relationship between crashes and road geometry in addressing segmentation. For example, Cenek *et al.* (1997) investigated this relationship for rural roads by using a fixed segment length of 200 meters. A similar study was done by Cafiso *et al.* (2008) using homogeneous sections with different lengths on a sample of Italian two-lane rural roads and aggregating variables related to curvature and roadside hazards. They concluded that models that contain geometry and design consistency variables are more reliable than those that do not. Other studies suggested different ways to aggregate segment data to avoid lengths that are too short. For example, Koorey (2009) proposed the aggregation of curves and tangents when the radius of curves exceeds a predetermined threshold value.

The HSM (Manual, 2010) recommends the use of homogeneous segments with respect to AADT, number of lanes, curvature, presence of ramps at the interchanges, lane width, outside and inside shoulder widths, median width, and clear zone width. There is no prescribed minimum segment length for application of the predictive models, but there is the suggestion of a segment length of no less than 0.10 miles.

Based on the literature, there is no exact value for the defining the length of a short section. The best segment length depends on the particular data set. Cafiso, D'Agostino and Persaud (2013) compared five different segmentation techniques with three different model

forms. The best results were obtained for the segmentation based on two curves and two tangents. Referring to Cafiso, D'Agostino and Persaud (2013), in this research, the two curve, two tangent segmentation technique was used. The minimum length of the segment was 0.15 mile and the average length of each segment was approximately 1.78 mile. The roadway segment lengths were reasonable according to the HSM suggestion.

4.2.3 Integrating Dynamic Variables

It should be noted that weather conditions are regarded as dynamic factors. The team considered two types of road surface conditions, dry and wet, and two types of visibility conditions, good visibility and bad visibility.

4.3 Data Description

Fourteen variables were included in the initial analysis in an attempt to identify the important ones that affect crash frequency more significantly. All variables are described briefly below.

4.3.1 Roadway Geometry

Horizontal Alignment Variables: The data set included mileposts on the beginnings and ends of all horizontal curves. Each horizontal curve segment had information on the direction of curvature (left or right), curvature, length, and design speed. As a result, this information could be used to divide the roadway into two curves and two-tangent segments. Variables related to roadway geometry included the number of left directed curves, portion of the curve length, and average curvature of a segment.

Vertical Alignment Variables: Within a segment, vertical alignment (e.g., grade) may change. As a result, to represent the grade percentage of horizontal segments, a length-weighted average grade value was used.

Roadway Features: Numerical variables related to roadway features included the number of lanes, median width,and inner and outer shoulder widths (in the analysis direction). Similar to vertical alignment, roadway features may change within a horizontal segment. Therefore, a length-weighted average was used. For categorical variables, each of the roadway segments was represented by dominant lane surface type, dominant outer shoulder type, dominant inner shoulder type, and dominant median type. The categorical variable that had the highest length on a segment was considered to be the dominant type.

4.3.2 Traffic Characteristics (AADT)

The average AADT value on each roadway segment for years 2011 to 2014 was used. Again, a length-weighted average was used when a roadway segment was associated with more than one AADT value.

4.3.3 Weather Related Variables

Each crash record that was collected from WSDOT included road surface conditions during the crash. The team calculated the proportion of time a particular segment experienced wet or dry pavement and incorporated it in the data set.

4.3.4 Crash Frequency

Each crash is associated with a milepost. Therefore, the total number of crashes that occurred on each segment during the analysis period of 2011 to 2014 was determined and divided by four to calculate yearly crash frequency on each segment. For the years from 2011 to 2014, crash data were extracted for road segments by severity (fatal, injury, and property damage only).

Summary statistics of the numerical and categorical variables for the road segments are shown in table 4-1 and table 4-2, respectively. The data for crash frequency by severity for the road segments are shown in table 4-3.

Table 4-1: Summary statistics of numerical variables for Interstate freeway segments in Washington state for years 2011 – 2014

Roadway geometry						
	Notation	mean	Std. dev.	Median	min	max
Segment length (mile)	LEN	1.78	1.56	1.35	0.15	1.56
Inner Shoulder Width (ft)	ISW	3.80	2.64	4.00	0.00	15.26
Outer Shoulder width (ft)	OSW	6.60	3.53	7.50	0.00	12.97
Median Width (ft)	MWD	93.54	148.36	70.00	6.17	999.00
Speed Limit (mph)	SPL	65.67	4.74	70.00	60.00	70.00
Grade (%)	GRD	0.67	2.22	0.11	0.00	36.24
No. of left curve/Segment	HCD	1.32	0.61	1.00	0.00	2.00
Length of the curve (%)	CUR	0.36	0.18	0.34	0.00	1.00
Average Curvature	ACU	0.56	0.48	0.43	0.00	3.02
Traffic Characteristics						
AADT/Lane	APL	12239.28	8846.59	9600.00	2500.00	55000.00
Crash frequency						
Log(Normalized Crash Density)	lnCR	2.17	1.65	2.32	-1.88	6.44

Table 4-2: Summary statistics of categorical variables for Interstate freeway segments in Washington state for years 2011 – 2014

Roadway geometry			
Variable	Description	Notation	No. of segments
Number of Lane (NOL)	⩽ 2		
	2 < NOL < 4	NOL2	653
	NOL ⩾ 4	NOL3	
Lane Surface Type (LST)	Portland Cement Concrete	LSTPD	158
			886
	Others (Asphalt)	LSTOD	718
Outer Shoulder Type (OST)	Asphalt	OSTAD	1434
	Others	OSTOD	170
Median Surface Type (MST)	Soil	MSTSD	1135
	Others	MSTOD	469
			634
Speed Limit (SPL)	⩽ 60 mph	SPL1	
	> 60 mph	SPL2	970
Weather Characteristics			
Road Surface Condition (RSC)	Dry	RSCD	802
	Non-Dry	RSCND	802

Table 4-3 Data of crash frequency for Interstate freeway segments in Washington state for years 2011 – 2014

Total crashes	Fatal crashes	Injury crashes	PDO crashes	Average crashes per year	Number of road segments
45270	129	13189	31952	11317.5	1604

Chapter 5 Crash Frequency Model

This chapter describes the use of regression models to fit best models to predict total crash frequency based on static and dynamic contributing factors. The proposed models were proven to have better goodness-of-fit than those found in the existing literature and to provide a better fit according to the assumptions of the regression model.

5.1 Models for Predicting the Total Number of Crashes on Freeways

Regression equations model the relationship between a dependent variable and a collection of independent variables. This relationship among the variables can be modelled as linear, nonlinear, and/or a combination of both functions. According to classical linear regression models, the expectation of crash frequency (or rate) is formulated as an ordinary least squares (OLS), which tries to minimize the error in model estimation. This model specification can be expressed as follows:

$$\mathrm{E}(y_i) = \mu_i = L_i * \sum_{j=1}^{J} x_{ij}\beta_j + \beta_0 \qquad (5.1)$$

where y_i denotes the crash frequency (or rate) along roadway segment i; (y_i) or μ_i is the expected crash frequency (or rate: number of crashes per year in this study) along segment i during a certain time period; L_i is the segment length in miles; x_{ij} is the j^{th} explanatory variable for segment i; β_j is the corresponding coefficient for the j^{th} explanatory variable; and J is the total number of explanatory variables considered in the model.

In this research, crashes were divided into two groups on the basis of the road surface conditions dry and wet. Dry conditions were more common in the area and so are the number of crashes in such conditions. Therefore, the number of crashes needed to be normalized on the basis of the durations of dry and wet surface conditions. For this purpose, the proportion of time

a certain segment was wet or dry was found as follows in equations (5.2) and (5.3). Weather data from Washington state airports stations were used.

$$Proportion\ of\ tine\ segment\ i\ is\ wet,\ \Pr(RSCW) = \frac{Duration\ of\ precipitation\ in\ road\ segment\ i}{Total\ recorded\ duration\ in\ road\ segment\ i} \tag{5.2}$$

$$Probability\ of\ road\ segment\ i\ being\ wet,\ \Pr(RSCD) = 1 - \Pr(RSCW) \tag{5.3}$$

Finally, to make road segments comparable, the logarithmic function of normalized crash density (e.g., normalized crash per mile-year) was considered as the dependent variable, as shown in equation (5.4). A very common method to handle the nonlinear relationship between the independent and dependent variables is logarithmic transformation of the variables in a regression model. Using the logarithm of one or more variables can capture the effective nonlinear relationship while preserving the linear model's assumptions. To approximate highly skewed variables to normal distribution, logarithmic transformation is considered a convenient way (Benoit, 2011). The normalized expected crash density (E_{CD}) is expressed as follows:

$$E_{CD} = \ln \frac{\mu_i}{L_i * \Pr(RSC_i)} = \sum_{j=1}^{J} x_{ij}\beta_j + \beta_0 \tag{5.4}$$

where Pr (RSC_i) is the proportion of time a segment is dry/wet, μ_i is the crash density per year, and all other variables are explained previously. There are four main assumptions involved in using an OLS regression approach:

1. Linearity and additivity of the relationship between dependent and independent variables:

    a. The expected value of the dependent variable is linearly related with the function of independent variable, holding the others fixed.

b. The slope of that line does not depend on the values of the other variables.

c. The effects of different independent variables on the expected value of the dependent variable are additive. They are not linearly dependent on each other.

2. Statistical independence of the errors
3. Homoscedasticity (constant variance) of the errors
4. Normality of the error distribution.

After fitting each model, the team will go through testing each assumption to ensure it is not violated.

5.2 Estimation Results

To develop the best-fit statistical models, initially, we fitted simple statistical models including a single explanatory variable. Next, multiple explanatory variables were systematically added to the regression models. Explanatory variables in the regression model were used as both continuous and dummy variables. Furthermore, non-linear forms such as logarithmic, quadratic, etc. were tested. The models with the highest adjusted R-square, as well as the fewest number of significant variables, were considered to be the best models. Finally, the effect of one independent variable on a dependent variable as a function of a second independent variable was tested by introducing interaction variables. In addition, a multicollinearity test was conducted while including any independent variable to ensure the independence among variables.

Table 5-1 shows the regression result for predicting total number of crashes. All the variables were statistically significant at a 0.1 significance level except for inner shoulder type and the interaction term of curve length and inner shoulder width. The adjusted R^2 value of 0.75 indicates that the model fit the data adequately.

Table 5-1: Summary of regression models for predicting the total number of crashes

Variables	Estimate	Std. Error	t-value	Pr(>|t|)

(Intercept)	2.96600	0.10130	29.29100	0.00000	***
Numerical variable					
AADT/Lane	0.00005	0.00000	12.41300	0.00000	***
Inner shoulder width (ft)	-0.03034	0.01318	-2.30200	0.02147	*
Median Width (ft)	0.00031	0.00015	2.08400	0.03730	*
Absolute value of Grade	-0.08305	0.03341	-2.48600	0.01303	*
Categorical Variable					
No. of lane (1 if more than 4; 0 otherwise)	0.62940	0.08706	7.23000	0.00000	***
No. of lane (1 if more than 2 and less than 4; 0 otherwise)	0.17080	0.05540	3.08300	0.00208	**
Outer shoulder type (1 if Asphalt; 0 otherwise)	-0.15220	0.08736	-1.74200	0.08168	.
Inner shoulder type (1 if Asphalt; 0 otherwise)	-0.13680	0.08732	-1.56600	0.11750	
Lane Surface Type (1 if Portland cement; 0 otherwise)	-0.17220	0.04492	-3.83400	0.00013	***
Road Surface Condition (1 if Dry, 0 Non-Dry) Intera	-2.98900 cting	0.07004	-42.68200	0.00000	***
/ariables					
Length of the curve in a segment*outer shoulder width (1 if less than 4 ft)	0.18640	0.13020	1.43200	0.15242	
Road Surface Condition (1 if Dry, 0 Non-Dry)* AADT/Lane	0.00007	0.00000	16.03800	0.00000	***
Number of observations			1604		
Adjusted R^2			0.75		

5.3 Interpretation of Causal Effects

As shown in table 5-1, a variety of variables were found to be significant determinants of crash frequency, and these variables were of plausible signs in terms of their tendency to increase or decrease crash density. Two numerical variable AADT/lane and median width showed positive relationships with the dependent variable (log transformed normalized crash density). Thus, the number of crashes increased with an increase in AADT/lane and median width, which supports the results of previous research in this field (e.g., Qin, Ivan and Ravishanker, 2004; Kononov, Bailey and Allery, 2008).

The negative sign for the inner shoulder variable demonstrates a reduction of crashes associated with an increase in shoulder width. Hadi *et al.* (1995) and Stamatiadis (2009) also

concluded that wider shoulder width is relatively safer. This is consistent with the fact that drivers have more room to take corrective actions after making an errant maneuver. It is also apparent that drivers are more apt to encounter roadside obstacles with reduced widths (Milton and Mannering, 1998). Similarly, asphalt paved inner and outer shoulders reduce the total number of crashes. This finding is consistent with previous research. The shoulder material, and thus the surface condition, has significant impact on the recovery of an errant driver leaving the travel lane.

The negative sign for rigid pavement type indicates that rigid pavements are relatively safer. Rigid pavements are more skid resistant, which offers more friction against uncontrolled maneuvers.

The two indicator variables on the number of lanes in a section were highly significant, with a positive coefficient. This indicates that crash frequency increases as the number of lanes increases from two lanes. As such, roadways with two lanes per direction are safer. According to Milton and Mannering (1998) this variable is likely to act as a proxy for limited access control and ramp merging, which are the indicators of a higher number of lane changing, passing, passing/turning, and merging activities. Those could lead to the difficulty of visually determining following vehicles in an adjacent lane (Sen, Smith and Najm, 2003).

The product of a curved length in a road segment with outer shoulder width has a positive coefficient. So, a curve segment with narrower outer shoulders is more likely to cause crashes. In curves, vehicles have a tendency to move away from the road. Curved road segments with wider outer shoulders offer drivers more room to be safe as the tendency is to leave the roadway. This finding is also consistent with findings of Easa and You (2009) and Bauer and Harwood (2013).

Finally, wet road surfaces are more prone to crashes than dry surfaces. Wet surfaces are associated with precipitation. According to Brodsky and Hakkert (1988) reduced visibility during precipitation, glare from reflecting light from wet surfaces, splashing water from other vehicles, and clouded windows as a result of high humidity are some causes that lead to crashes.

5. 4. Regression Diagnostics: Testing the Assumptions

This section describes analysis of all assumptions of an OLS linear regression to find out whether they were met or violated. Residual analysis is used to test the assumptions of constant variance and the independence of variables and normality of the distribution. In addition, error terms need to be independent and identically distributed. A random distribution of residuals around zero indicates that these assumption are not violated. Figure 5-1 shows that residuals were normally distributed with a mean of zero, thereby satisfying the assumptions of OLS estimation.

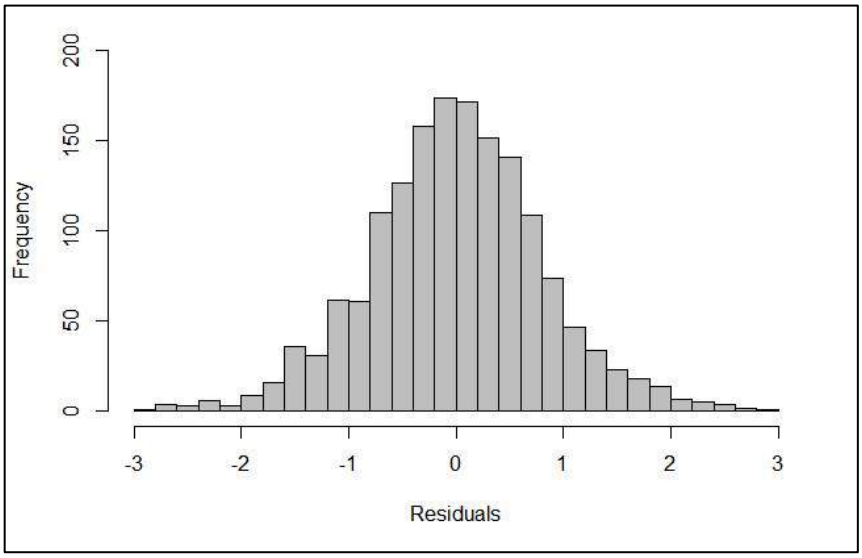


Figure 5-1 Histogram of residuals

To examine the homoscedasticity of the regression model, the residuals were plotted against fitted values. Plotting residuals versus the value of a fitted response should produce a

distribution of points scattered randomly around zero, regardless of the size of the fitted value. The residual plot in figure 5-2 shows a fairly random pattern around zero. This random pattern indicates that the proposed model provided an appropriate fit to the data.

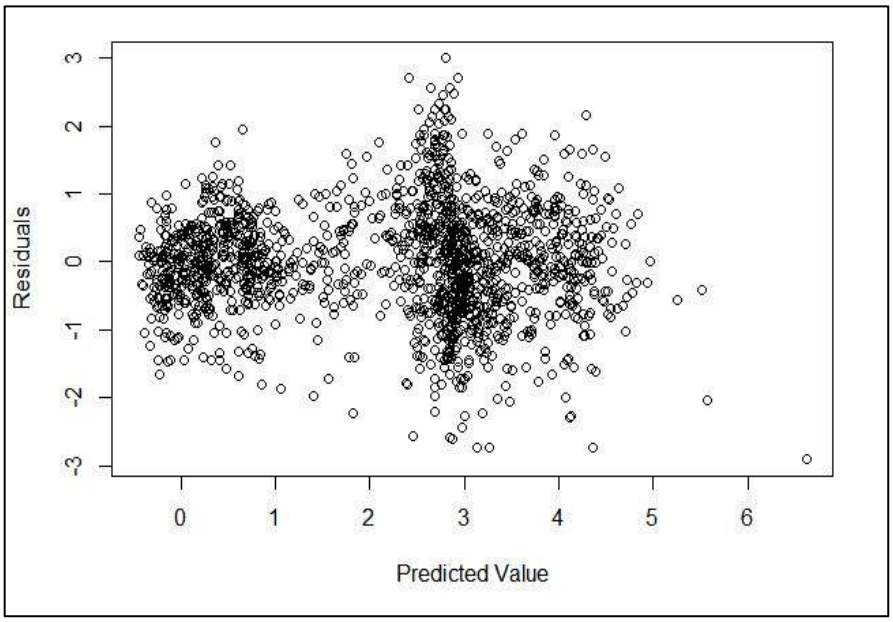


Figure 5-2 Residuals vs fitted plot

A randomly varied scatterplot of the standardized residuals against each of the independent variables also confirms the homoscedasticity of the regression model. The team plotted the residuals against all independent variables and found a non-systematic pattern. For instance, figure 5-3 shows residuals vs AADT/lane where the residuals appear on the y-axis and the predictor AADT/lane appear on the x-axis. The residuals are randomly scattered along the zero value.

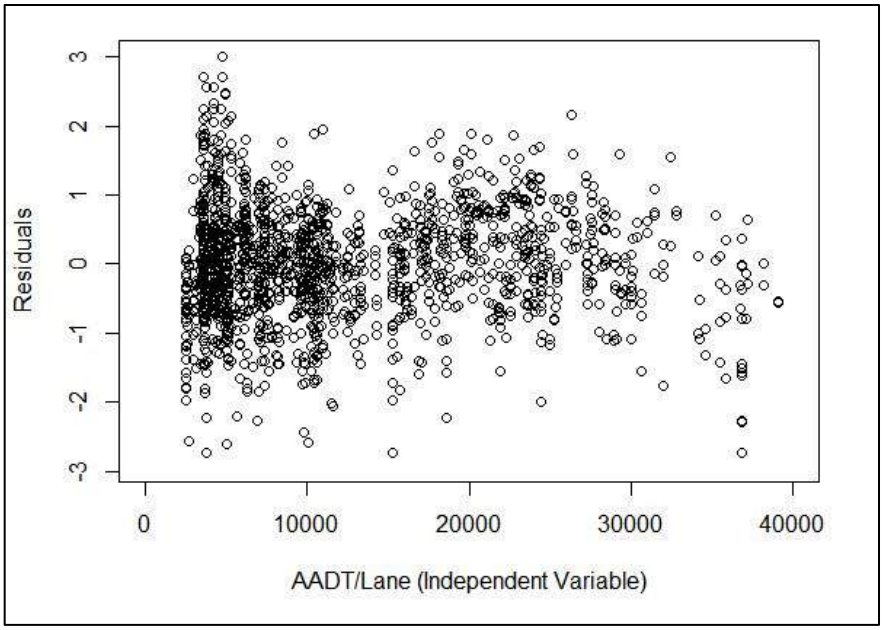


Figure 5-3 Residuals vs independent variables AADT/lane

The comparison of the expected number of total crashes and the observed number of crashes is shown in figure 5-4. A well fitted model should produce a slope close to 1.0 with 0.0 intercept.

The proposed model yielded an intercept of 0.91.

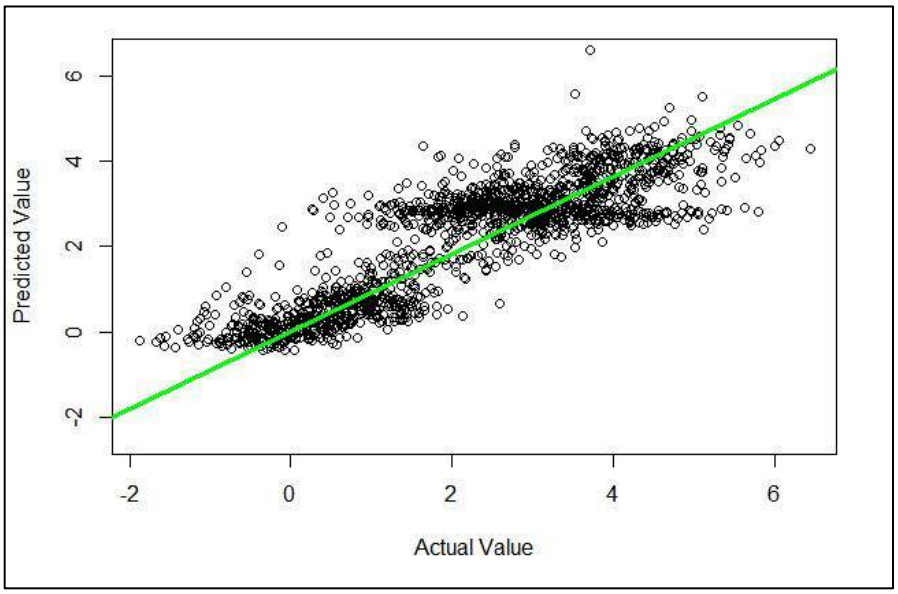


Figure 5-4 Fitted vs actual value

In the figure, all points do not lie on the optimum line because the number of observed accidents cannot be predicted with 100 percent accuracy using the selected indicator variables. It is likely that this variability can be reduced by using additional indicator variables. It should be noted that variables describing driver characteristics were not considered in the model. However, the results of the analyses indicated that the assumptions of OLS regression were not violated, and the proposed model adequately predicted the total number of crashes on each segment.

Chapter 6 Two-Stage Model for Predicting Crash Frequency by Severity

Using a combination of generalized linear regression and logit estimation techniques, this work modeled correlated traffic crash counts at different levels of severity. First, two logistic models were fitted, one for property damage only and one for injury-fatal crashes. The role of these logistic models is to predict the probability of having one or more crashes of a certain severity. Then generalized linear regression models were used to predict the number of PDO and IF crashes within each group. Details of the approach follow.

6.1 Models for Predicting the Number of Crashes by Severities on Freeways

Log-transformed linear regression models cannot predict the frequency of zero. Therefore, the research team incorporated a logistic regression model to tackle this problem. In order to get the total crash counts by severity level, the estimated PDO and IF crash densities from the regression models were used with corresponding logistic regressions, as shown in equations (6.1) and (6.2):

Estimated PDO crash density

$$E(PDO_i) = \Pr(Y_{PDOi} = 0) * 0 + \Pr(Y_{PDOi} = 1) * E(PDO_i \,|Y_{PDOi} = 1) \qquad (6.1)$$

Estimated Injury-fatal(IF) crash density

$$E\,(IF_i) = \Pr(Y_{IFi} = 0) * 0 + \Pr(Y_{IFi} = 1) * E(IF_i \,|Y_{IFi} = 1) \qquad (6.2)$$

where $E(PDO_i \,|Y_{PDOi} = 1)$ is the expected PDO crash density (number per mile-year) on segment i given there was at least one PDO crash on the segment. Note that $(IF_i \,|Y_{IFi} = 1)$ is

defined in the same way. Furthermore, $\Pr(Y_{PDOi} = 1)$ and $\Pr(Y_{IFi} = 1)$ respectively denote the probability of having at least one PDO and IF crash on the segment.

6.1.1 Logistic Regression Models for Predicting the Probability of Severity Events

In this research, logistic regression models were used to predict the probability of an event occurrence (i.e., PDO and injury-fatal). The basic idea is similar to a Bernoulli probability that governs the binary outcome of whether a variate has a zero or positive realization. If the realization is positive, we multiply this realization by the outcome from the regression model to get expected value. Two logistic models were fitted to predict the probability of PDO and injury-fatal occurrence.

Let Y_S represent the response variable of severity type S (S can take on PDO or IF types), whereas contributing factors for road conditions, traffic characteristics, and weather conditions are denoted by x_{ij} (j=1, 2, …, J), where i represents the segment and j denotes the number of independent variables. The expression of Y is defined as below:

$$Y_{S_i} = \begin{cases} 1, & \text{if severity type } S \text{ occur during the study period in road segment } i \\ 0, & \text{otherwise} \end{cases}$$

When the response categories 1 or 0 are unordered, Y_{Si} is related to independent variables through a set of baseline category logits as shown below:

$$\pi_{S_i} = \Pr(Y_{S_i} = 1|X) = \frac{\exp(\beta_0 + \sum_{j=1}^{J} x_{ij}\beta_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{J} x_{ij}\beta_j)} \quad (6.3)$$

Or,

$$logit(\pi_{S_i}) = log\left(\frac{\pi_{S_i}}{1-\pi_{S_i}}\right) = \beta_0 + \sum_{j=1}^{J} x_{ij}\beta_j \quad (6.4)$$

6.1.2 Regression Models for Predicting Crash Frequency by Severity

The basics of these crash count by severities models are similar to those in Section 5.1. We fitted two regression models (crash count for PDO and IF). The expected logarithmic normalized PDO and IF crash densities in a segment i are expressed as follows:

$$E_{PDO} = \ln\frac{\mu_{PDO_i}}{L_i * \Pr(RSC_i)} = \sum_{j=1}^{J} x_{ij}\beta_j + \beta_0 \quad (6.5)$$

$$E_{\mathrm{IF}} = \ln\frac{\mu_{IF_i}}{L_i * Pr(RSC_i)} = \sum_{j=1}^{J} x_{ij}\beta_j + \beta_0 \quad (6.6)$$

where, μ_{PDOi} and μ_{IFi} are the expected PDO and IF crash density (number of crashes per mile-year) of a segment i.

6.2 Model Estimation and Diagnosis

6.2.1. Logistic Model for PDO

This section presents a fitted logistic regression model for predicting the probability of having at least one PDO crash on a segment. Table 6-1 shows the summary of the logistic model for PDO crashes. All the variables were statistically significant. The ρ^2 value of 0.41 indicates

that the model fit the data adequately. This logistic model included linear, nonlinear, categorical and interacting variables.

Table 6-1 Summary of the logit model for predicting the probability of PDO crash occurrence

Variables	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	-1.77400	0.31960	-5.551	0.000000	***
Numerical variable					
AADT/Lane	0.00027	0.00002	13.92	0.000000	***
Inner Shoulder Width	0.14890	0.04641	3.209	0.001334	**
Curved portion in a segment	-2.73000	0.28600	-9.545	0.000000	***
No of left directed curve in a segment	-0.61500	0.09104	-6.755	0.000000	***
Absolute value of Grade	0.47370	0.07644	6.197	0.000000	***
Categorical Variable					
No. of lane (1 if more than 4; 0 otherwise)	1.22800	0.33790	3.635	0.000278	***
No. of lane (1 if more than 2 and less than 4; 0 otherwise)	1.30800	0.17370	7.532	0.000000	***
Inner Shoulder Type (1 if Asphalt, 0 otherwise)	-1.87700	0.25940	-7.234	0.000000	***
Outer Shoulder Type (1 if Asphalt, 0 otherwise)	1.60700	0.26820	5.992	0.000000	***
Road Surface Condition (1 if Dry, 0 Non-Dry)	0.49920	0.23080	2.163	0.030552	*
Interacting variables					
Road Surface Condition (1 if Dry, 0 Non-Dry)* AADT/Lane	0.00009	0.00004	2.373	0.017650	*
Number of observations			1604		
ρ^2			0.41		

6.2.2. Regression Model for PDO

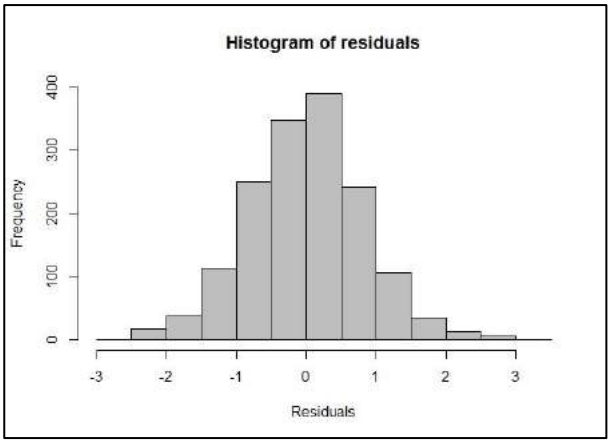
This section presents the best regression model for estimating logarithmic normalized PDO crash density. Table 6-2 shows the regression results for predicting log-transformed,

normalized PDO density. All the variables were statistically significant with a 95 percent confidence level. The adjusted R^2 value of 0.76 indicates that the model fit the data adequately. This regression model included linear, nonlinear, categorical, and interacting variables.

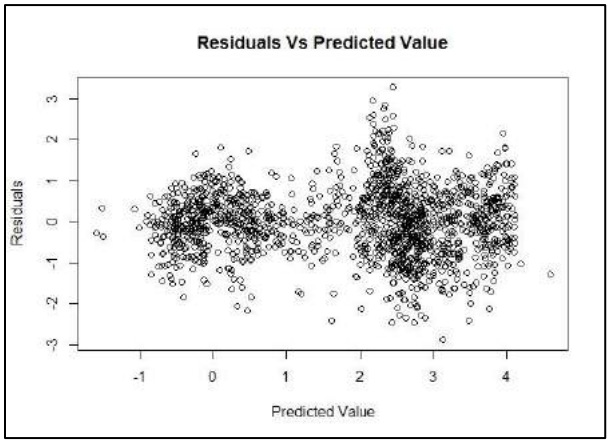
Table 6-2: Summary of regression model for predicting log-transformed, normalized PDO density

Variables	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	2.48700	0.12280	20.246	2.00E-16	***
Numerical variable					
AADT/Lane	15.82000	1.34700	11.742	2.00E-16	***
$AADT/Lane^2$	-4.24900	0.92630	-4.587	4.86E-06	***
Number of Lane	0.23710	0.04012	5.911	4.17E-09	***
Inner shoulder width (ft)	-0.03387	0.01225	-2.764	0.00579	**
Median Width (ft)	0.00042	0.00016	2.633	0.00856	**
Curved portion in a segment	0.25250	0.12080	2.09	0.0368	*
Absolute value of Grade	-0.10260	0.03408	-3.011	0.00265	**
Categorical Variable					
Lane Surface Type (1 if Asphalt cement; 0 otherwise)	-0.75510	0.31690	-2.383	0.01729	*
Inner Shoulder Type (1 if Asphalt, 0 otherwise)	-0.17740	0.07417	-2.392	0.01688	*
Road Surface Condition (1 if Dry, 0 Non-Dry)	-3.04100	0.07222	-42.107	2.00E-16	***
Interacting variables					
Road Surface Condition (1 if Dry, 0 Non-Dry)* AADT/Lane	0.00007	0.00000	15.842	2.00E-16	***
Number of observations			1562		
R_2			0.76		

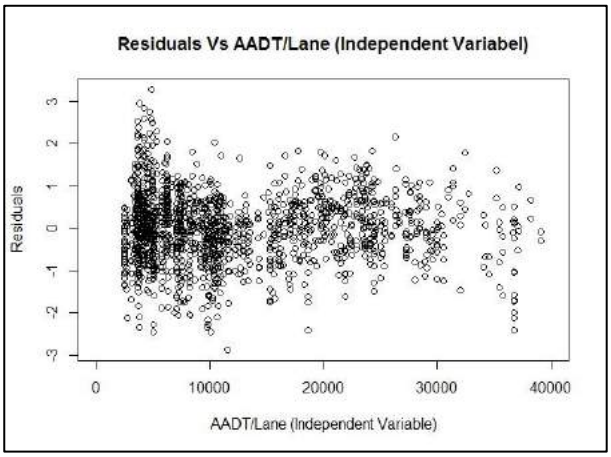
The team tested all the assumptions of OLS regression models. Residuals were found to be normally distributed with a mean of zero. Both residuals against fitted value and residuals against independent variables plots showed a random pattern around zero. The residuals fell in a symmetric pattern and had a constant spread throughout the range. Finally, the adjusted R^2 of the actual vs fitted value plot was 0.79 with a slope 0.8 and 0.0 intercept. Figure 6-1 shows the diagnosis of the regression model.



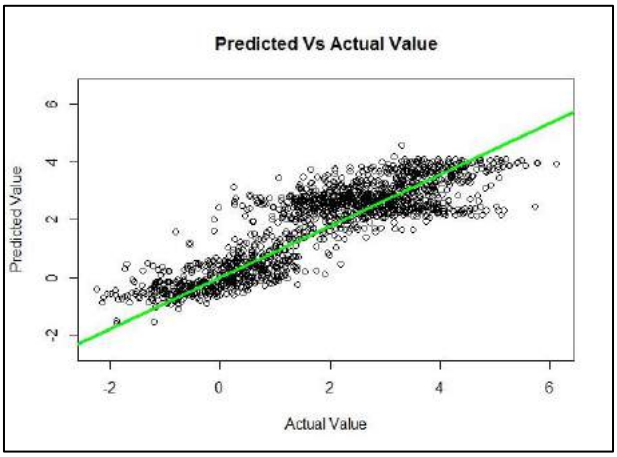
(a) Histogram of the residuals



(b) Residual vs predicted value



(c) Residuals vs AADT/Lane



(d) Predicted vs actual value

Figure 6-1 Model diagnosis of regression model for log-transformed, normalized PDO density

6.2.3. Logistic Model for IF Crashes

Table 6-3 summarizes the estimation of the logistic model for the estimation of the probability of IF crashes. Like the previous models, all the variables listed here were statistically significant. The model yielded a ρ^2 value of 0.2 and included linear, nonlinear, categorical, and interacting variables.

Table 6-3 Summary of the logit model for predicting the probability of IF crash occurrence

Variables	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	-0.69680	0.26510	-2.628	0.00859	**
Numerical variable					
AADT/Lane	35.07000	3.27000	1.07E+01	0.00000	***
$AADT/Lane^2$	-8.45500	2.42500	-3.486	0.00049	***
Inner shoulder width (ft)	-0.12810	0.02534	-5.057	0.00000	***
Logarithm of Median Width (ft)	0.30570	0.07035	4.346	0.00001	***
Curved portion in a segment	-1.63400	0.26060	-6.269	0.00000	***
Absolute value of Grade	0.34090	0.07063	4.827	0.00000	***
Categorical Variable					
No. of lane (1 if more than 4; 0 otherwise)	1.27600	0.22220	5.74	0.00000	***
No. of lane (1 if more than 2 and less than 4; 0 otherwise)	0.89420	0.11500	7.774	0.00000	***
Outer Shoulder Width (1 if greater than or equal 10 ft, 0 otherwise)	-0.21100	0.10030	-2.104	0.03534	*
Inner Shoulder Type (1 if Portland cement; 0 otherwise)	0.63340	0.30290	2.091	0.03651	*
Road Surface Condition (1 if Dry, 0 Non-Dry)	0.00003	0.00002	1.876	0.06067	.
Interacting variables					
Road Surface Condition (1 if Dry, 0 Non-Dry)* AADT/Lane	0.00007	0.00001	14.113	0.00000	***
Number of observations			1604		
ρ_2			0.2		

6.2.4. Regression Model for IF Crashes

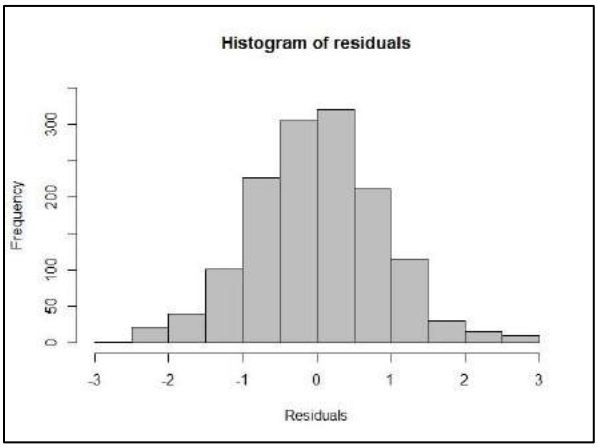
Table 6-4 summarizes the regression model for predicting the number of IF crashes. All the variables were statistically significant at a 90 percent confidence level. The value of adjusted R^2 was 0.71. This model included linear, categorical, and interaction variables.

Table 6-4 Summary of regression model for predicting log-transformed, normalized IF density

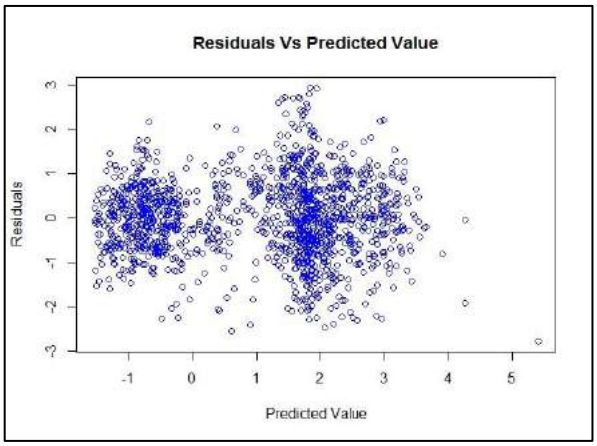
Variables	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	1.76400	0.12540	14.072	0.00000	***
Numerical variable					
AADT/Lane	0.00004	0.00000	9.931	0.00000	***
Inner shoulder width (ft)	-0.04083	0.01021	-3.999	0.00007	***
Outer Shoulder width (ft)	-0.00911	0.00735	-1.239	0.09537	
Median Width (ft)	0.00037	0.00017	2.17	0.03017	*
Curved portion in a segment	0.40440	0.13670	2.959	0.00314	**
Absolute value of Grade	-0.07148	0.03818	-1.872	0.06137	.
Categorical Variable					
No. of lane (1 if greater than 2 and less than 4; 0 otherwise)	0.09238	0.05781	1.598	0.10027	
No. of lane (1 if greater than 4; 0 otherwise)	0.63150	0.09033	6.991	0.00000	***
Lane Surface Type (1 if Portland cement; 0 otherwise)	-0.21180	0.05270	-4.019	0.00006	***
Outer Shoulder Type (1 if Asphalt, 0 otherwise)	-0.13910	0.08443	-1.647	0.09974	.
Road Surface Condition (1 if Dry, 0 Non-Dry)	-2.98900	0.08336	-35.859	0.00000	***
Interacting variables					
Road Surface Condition (1 if Dry, 0 Non-Dry)* AADT/Lane	0.00008	0.00001	14.319	0.00000	***
Number of observations			1404		
R^2			0.71		

The team diagnosed the regression model in the same way as previous models. Residual diagnosis and actual vs fitted plots showed that assumptions were not violated. Figure 6-2

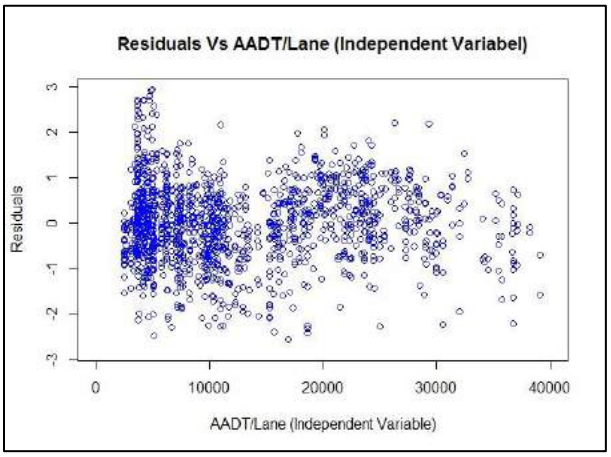
summarizes the results of all the diagnosis tests. The adjusted R^2 of the actual vs fitted value plot was 0.79 with a slope of 0.8 and 0.0 intercept.



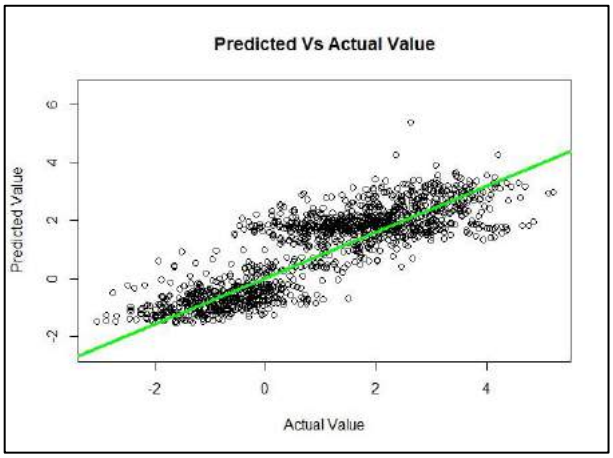
(a) Histogram of the residuals



(b) Residual vs predicted value



(c) Residuals vs AADT/Lane



(d) Predicted vs actual value

Figure 6-2 Model diagnosis of the regression model for log-transformed, normalized injury-fatality (IF) density

6.3 Interpretation of the Causal Effects of the Two-Stage Model

6.3.1. Logistic Model for PDO and Injury-Fatal (IF) Crashes

The probability of both PDO and IF occurrence increases with increases in traffic volume. However, above a certain traffic volume level, the probability of IF crashes occurring drops as the term $(AADT/Lane)^2$ dominates $AADT/Lane$. This outcome is logical, as with high

volume levels, vehicles are forced to follow each other closely, which contributes to an increase in the probability of minor and mostly PDO rear-end crashes. Increasing the volume level also reduces lane changing and overtaking maneuvers, actions that can lead to severe crashes. Therefore, the probability of IF crashes occurring decreases as traffic volumes increase.

The probabilities of PDO and IF crashes occuring increase as the number of lanes increases. This is as expected. A larger number of lanes is an indicator of merging and weaving activities with more lane changes.

Crashes are likely to be more severe on road segments with narrow shoulders. Narrow shoulders provide drivers with less room and time to correct a mistake, which can lead to severe crashes. Curvature and grades also increase the probability of PDO and IF crashes occurring as curves and grades restrict visibility and the sight distance of drivers.

6.3.2. Regression Model for PDO and Injury-Fatal (IF) Crashes

The model showed that PDO and IF crash frequency increases as AADT/lane, number of lanes, and median width increase. Road segments with a relatively higher proportion of curved lengths are more prone to both PDO and IF crashes. Factors contributing to crashes on curved segments often include loss of control or misjudging a curve. This type of crash results in a vehicle striking a fixed object or rollovers.

Wider inner and outer shoulder width reduce the risk of both PDO and IF crashes. Wide shoulders offer recovery measures to an errant driver going out of the travel lane. Both PDO and IF frequencies depend on the absolute value of the gradient of the road segments.

The model indicated that asphalt paved shoulders are safer. Finally, both PDO and IF crashes are more likely to happen on wet road surface conditions.

Chapter 7 Crash Counts by Severity-Based HSID Method

This chapter describes the development of a crash count by severity-based hotspot identification (CCS-based HSID) method by employing the frequency estimation and binary logistic regression approaches discussed in previous chapters. First, a binary logistic model was fitted to predict the probability of a PDO or injury/fatal crash occurring on the basis of the explanatory variables. Then, PDO and injury-fatal crashes were modeled separately to predict their frequency. A new safety performance index and a new potential safety improvement index were developed and compared with traditional ones by employing HSID evaluation methods.

7.1 Safety Performance Index

In order to develop a new safety performance index that can reflect crash counts by severity level, the estimated PDO and IF crash densities from the generalized nonlinear models were employed with corresponding logistic regressions to estimate expected crash densities according to severity type on a roadway segment *i*, as shown below:

1. *Estimated PDO crash density*:

$$N_{PDOi} = (PDO_i \mid Y_{PDOi} = 1) * \Pr(Y_{PDOi} = 1)$$

$$= \Pr(RSC_i) * \exp\left(\beta_0 + \sum_{j=1}^{J} x_{ij}\beta_j\right) * \frac{\exp(\beta_0 + \sum_{j=1}^{J} t_{ij}\beta_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{J} t_{ij}\beta_j)} \quad (7.1)$$

2. *Estimated IF crash density*:

$$N_{IFi} = (IF_i \mid Y_{IFi} = 1) * \Pr(Y_{IFi} = 1)$$

$$= \Pr(RSC_i) * \exp\left(\beta_0 + \sum_{j=1}^{J} x_{ij}\beta_j\right) * \frac{\exp(\beta_0 + \sum_{j=1}^{J} t_{ij}\beta_j)}{1 + \exp(\beta_0 + \sum_{j=1}^{J} t_{ij}\beta_j)} \quad (7.2)$$

where, N_{PDOi} and N_{IFi} are the estimated number of PDO and IF crashes per mile per year, respectively. If equations (7.1) and (7.2) are multiplied by the length of the segment i, then the total estimated number of crashes in a segment is found.

Based on equations (7.1) and (7.2), the equivalent property damage only crash frequency measure was modified and employed to assign weight to crashes according to their severity (fatal, injury, and PDO) to develop a combined crash density and severity score (CCDSS) for each site (Washington *et al.*, 2014). The weight factors were based on PDO crash costs. An EPDO value summarized the crash costs and severity. In the calculations, weight factors were assessed from the crash cost estimates developed by WSDOT in the Annual Collision Data Summary Reports (2011-2014). Using average crash costs for motorways, fatal crashes ($2,227,851) had a weight factor equal to 981, injury crashes ($20,439) had a weight factor equal to 9, and PDO crashes ($2,271) had a weight factor equal to 1. However, if only the average crash costs were considered to be the weight factor, then inconsistencies could occur when the HSID methods were evaluated in different time periods, since the traditional EPDO method over-emphasizes sites with a low frequency of fatal or severe crashes (Montella, 2010). As a result, a risk weight factor was developed in this research by combining the average crash cost with the corresponding probability for each type of crash severity. Let F_w denote the fatality risk weight factor, I_w denote the injury risk weight factor, and P_w, the PDO risk weight factor as follows:

$$F_w = \frac{c_F \cdot \eta_F}{c_P \cdot \eta_P}, \quad I_w = \frac{c_I \cdot \eta_I}{c_P \cdot \eta_P}, \quad P_w = 1, \tag{7.3}$$

where c_F =\$2,227,851 , c_I =\$20,439 , and c_P = \$2,271 are the average costs for fatal, injury, and PDO crashes, respectively; η_F, η_I, and η_P are the probabilities of occurrence for fatal, injury, and PDO crashes, respectively. In this research, we considered using the proportion of the crash frequency for each type of severity based on the collected crash data of 21,396 road segments along I-5, I-90, I-82, I-182, I-205, I-405 and I-705 in Washington to represent the probability of crash occurrence for each severity level in this area. Based on the crash counts in table 4-3, we can calculate the following:

- $\eta_I = 0.291$,
- $\eta_F = 0.003$, and
- $\eta_P = 0.706$

The values of the risk weight factors were obtained by employing equation (7.2) as $F_w = 4.169$, $I_w = 3.709$ and $P_w = 1$. As we analyzed injury and fatal crashes together, we calculated the injury-fatality risk weight factor based on the equation given below:

$$IF_w = I_w \cdot R_I + F_w \cdot R_F \tag{7.4}$$

Here, R_I and R_F denote ratio of the injury and fatal crashes with respect to total injury-fatal crash, respectively. The calculated value of $IF_W = 3.72$.

Based on the preceding analysis, the expected CCDSS (ECCDSS) for roadway segment *i* can be defined as:

$$ECCDSS_i = EPDO_i . P_w + EPDO_i . IF_w, \qquad i = 1, 2, \dots \dots, n \tag{7.5}$$

Equation (7.5) is regarded as the safety performance function (SPF). In fact, the ECCDSS is an extension of the expected crash density based on the two-stage regression and logistic models.

The EB method is a statistical method that combines the observed crash frequency with the predicted crash frequency using the SPF to calculate the expected crash frequency for a site of interest. The EB method pulls the crash count towards the mean, accounting for the regression to the mean (RTM) bias. A lot of studies have proved that the EB approach is the most consistent and reliable method for identifying sites with promise (Cheng and Washington, 2008; Montella, 2010). In this research, the EB method was employed to develop the new SPI as shown in the following:

$$SPI_i = \lambda_i ECCDSS_i + (1 - \lambda_i) OCCDSS_i, \quad i = 1,2,\cdots, n \tag{7.6}$$

where $OCCDSS_i$ is the observed combined crash density and severity score (OCCDSS) for roadway segment i and is defined as below:

$$OCCDSS_i = \sigma_{PDO_i}.P_w + \sigma_{IF_i}.IF_w \tag{7.7}$$

where σ_{PDO_i} and σ_{IF_i} are the observed PDO and injury-fatal crash density along segment i during a certain time period respectively; λ_i is a weighting factor that is calculated through the following equation:

$$\lambda_i = \frac{1}{1 + \alpha_i ECCDSS_i} \tag{7.8}$$

where α_i is the over-dispersion parameter, which is a constant for a given model and is derived during the regression calibration process.

7.2 Potential Safety Improvement Index

The PSII was developed as the difference between the SPI and the ECCDSS, as follows:

$$PSII_i = \lambda_i ECCDSS_i + (1 - \lambda_i) OCCDSS_i - ECCDSS_i = SPI_i - ECCDSS_i, i = 1,2,\cdots, n \quad (7.9)$$

When the potential safety improvement index value is greater than zero, a site experiences a higher combined frequency and severity score than expected. When the potential safety improvement index value is less than zero, a site experiences a lower combined frequency and severity score than expected.

7.3 Evaluation Tests of Performance of HSID Methods

In order to demonstrate the effectiveness of the SPI and PSII as developed in this research, they were compared with three other models that were developed in this research. All these five regression models are given below:

Table 7-1 Description of different HSID methods

Model	Description
I	$\mu_{Total_i} = \mu_{PDO_i} + \mu_{IF_i}$
II	$N_{Total_i} = EPDO_i + EIF_i$
SPF	$SPF_i = EPDO_i . P_w + EIF_i . IF_w$
SPI	$SPI_i = \lambda_i ECCDSS_i = (1- \lambda_i) OCCDSS_i, \ i = 1,2,\cdots, n$ where, $ECCDSS_i = EPDO_i P_w + EIF_i IF_w, \quad i = 1, 2, \ldots \ldots, n$ $OCCDSS_i = \sigma_{PDO} P_w + \sigma_{IF} IF_w$ $\lambda_i = \frac{1}{1+\alpha_i ECCDSS_i}$
PSII	$PSII_i = \lambda_i ECCDSS_i + (1 - \lambda_i) OCCDSS_i - ECCDSS_i = SPI_i - ECCDSS_i, \ i = 1,2,\cdots, n$

Cheng and Washington (2008) developed four new evaluation tests for HSID. In this research, the site consistency test, method consistency test, total rank differences test, and the total score test were employed to evaluate the effectiveness of the developed safety performance indexes and reference performance indexes.

The evaluation experiment used the following procedure, which closely mimicked how reactive safety management programs are conducted in practice:

1. For the purpose of comparing alternative HSID approaches, the four-year accident data were separated into two periods, Period 1 (years 2011-2012) and Period 2 (years 2013-2014).
2. For each HSID method, road sections were sorted in descending order of estimated safety (note that the four HSID methods rank sites according to different criteria).
3. Sections with the highest rankings were flagged as hotspots (in practice these sites would be further scrutinized). Typically, a threshold is assigned according to safety funds available for improvement, such as the top 10 percent of sites. In this evaluation, both the top 10 percent and 20 percent of the locations are used as experimental values.

7.3.1 Site Consistency Test

The site consistency test (SCT) measures the ability of an HSID method to consistently identify a high-risk site over repeated observation periods. The test rests on the premise that a site identified as high risk during time period t should also reveal an inferior safety performance in a subsequent time period $t+1$, given that the site is in fact high risk and no significant changes have occurred at the site. The method that identifies sites in a future period with the highest crash frequency is the most consistent. In this research, the SPI developed above was employed as the safety performance criterion in the subsequent time period. The test statistic was given as:

$$SCT_{h,t+1} = \sum_{q=n-n\gamma+1}^{n} SPI_{q,h,t+1}, \quad h = 1, 2, \cdots, H, \tag{7.10}$$

where h is the HSID method index being compared; n is the total number of roadway segments, γ is the threshold of identified hotspots (e.g., $\gamma = 0.01$ corresponds with top 1 percent of n roadway segments identified as hotspots, and $n\gamma$ is the number of identified hotspots).

Table 7-2 Results of site consistency test of various HSID methods

HSID Method Index (h)	HSID Method Name	γ = 0.1			γ = 0.2		
		$SCT_{h\,t,}$ Period 1 (2011-2012)	$SCT_{h\,t,\,t1}$ Period 2 (2013-2014)	% of change	$SCT_{h\,t,}$ Period 1 (2011-2012)	$SCT_{h\,t,\,t1}$ Period 2 (2013-2014)	% of change
1	Model-I	14813.9	14569.5	1.65	21811.27	20871	4.31
2	Model-II	14607.4	16872.83	15.51	21809.46	24362.4	11.71
3	SPF	14942.3	16880	12.97	21987.9	24296	10.50
4	SPI	17643.4	20366.2	15.43	22995.72	27053.79	17.65
5	PSII	15640.02	18364.65	17.42	20439.81	23377.68	14.37

The site consistency test results, as shown in table 7-2, indicated that Model-I outperformed other HSID methods in identifying both the top 10 percent and 20 percent of segments. Model-I was more consistent, as the SCT value changed 1.65 percent from Period 1 to Period 2 for $\gamma = 0.1$ and changed 4.31 percent from Period 1 to 2 for $\gamma = 0.2$.

7.3.2 Method Consistency Test

The method consistency test (MCT) evaluates a method's performance by measuring the number of the same hotspots identified in both time periods. It is assumed that road sections are in the same or similar underlying operational state and their expected safety performance remains virtually unaltered over the two analysis periods. With this assumption of homogeneity, the greater the number of hotspots identified in both periods, the more consistent the performance of the HSID method. The test statistic is given as:

$$MCT_h = \{s_{n-n\gamma+1}, s_{n-n\gamma}, \cdots, s_n\}_{h,t} \cap \{s_{n-n\gamma+1}, s_{n-n\gamma}, \cdots, s_n\}_{h,t+1} \quad h = 1, 2, \cdots, H, \tag{7.11}$$

where only segments $\{s_{n-n\gamma+1}, s_{n-n\gamma}, \ldots, s_n\}$ identified in the top threshold γ are compared.

Table 7-3 shows the number of similarly identified hotspots identified by alternative HSID methods over the two periods. The SPF method was superior in this test by identifying the largest number of the same hotspots in both cases of $\gamma = 0.1$ and $\gamma = 0.2$, with 87 percent and 88 percent matched hotspots, respectively. The SPF method identified 144 segments in 2011-2012 that were also identified as hotspots in 2013-2014 for $\gamma = 0.1$. Model-II and the SPI method were the second and third best models in terms of site consistency. Table 7-3 shows the result of the consistency test of all five HSID methods.

Table 7-3 Results of method consistency test of various HSID methods

HSID Method Index (h)	HSID Method Name	$\gamma = 0.1$ (166 sites)	$\gamma = 0.2$ (332 sites)
1	Model-I	106 (64%)	195 (59%)
2	Model-II	140 (84%)	289 (87%)
3	SPF	144 (87%)	291 (88%)
4	SPI	133 (80%)	286 (86%)
5	PSII	119 (72%)	232 (70%)

7.3.3 Total Rank Differences Test

The total rank differences test (TRDT) takes into account the safety performance rankings of the road sections in the two periods. The test is conducted by calculating the sum of the total rank differences of the hotspots identified across the two periods. The smaller the total rank difference, the more consistent the HSID method. The test statistic is given as:

$$TRDT_h = \sum_{q=n-n\gamma+1}^{n} \left| R(q_{h,t}) - R(q_{h,t+1}) \right|, \quad h = 1, 2, \cdots, H, \tag{7.12}$$

where $R\,q(_{h\,t,})$ is the rank of segment q in period t for method h. The difference in ranks is summed over all identified segments for threshold level γ for period t. Table 7-4 illustrates that the PSII method was superior in the total rank differences test. In both the $\gamma = 0.1$ and $\gamma = 0.2$ cases, the PSII method had significantly smaller summed ranked differences, by 11.4 percent and 14.2 percent, in comparison to the SPI method. This result suggests that the PSII method is the best HSID method for ranking roadway segments consistently from period to period.

Table 7-4 Results of total rank differences test of various HSID methods

HSID Method Index (h)	HSID Method Name	γ = 0.1 (166 sites)	γ = 0.2 (332 sites)
1	Model-I	98913	207720
2	Model-II	96196	196200
3	SPF	103648	211727
4	SPI	107826	213520
5	PSII	95510	183202

7.3.4 Total Score Test

The total score test (TST) combines the site consistency test, the method consistency test, and the total rank difference test in order to provide a synthetic index. The test statistic is given as:

$$TST_h = \frac{100}{3}\left[\left(\frac{\overline{SCT_{h,t+1}}}{\max_h\{SCT_{h,t+1}\}}\right) + \left(\frac{\overline{MCT_h}}{\max_h\{MCT_h\}}\right) + \left(1 - \frac{TRDT_h - \min_h\{TRDT_h\}}{\max_h\{TRDT_h\}}\right)\right], \quad h = 1, 2, \cdots, H, \tag{7.13}$$

where the test assumes that the SCT, MCT, and TRDT have the same weight. The former three tests provide absolute measures of effectiveness, whereas the total score test gives an effectiveness measure relative to the methods being compared. If method *h* performed best in all of the previous tests, the TST value would be equal to 100. If method *h* performed worst in all of the tests, the TST value would be positive since all three components of the test had a positive value.

Indeed, SCT and MCT, which should be maximized by the HSID methods, are weighted in relation to the maximum values in the tests, whereas TRDT, which should be minimized by the HSID methods, is weighted in relation to its difference from the minimum value in the test. Table 7-5 illustrates the results of the total score test of the five HSID methods, in which SPI performed best in both the $\gamma = 0.1$ and $\gamma = 0.2$ cases and was followed closely by the Model-II method. The SPI method has a score of 93.64 and 98.37 for $\gamma = 0.1$ and $\gamma = 0.2$, respectively.

Table 7-5 Results of total score test of various HSID methods

HSID Method Index (h)	HSID Method Name	γ = 0.1 (166 sites)	γ = 0.2 (332 sites)
1	Model-I	80.66	80.39
2	Model-II	93.14	97.73
3	SPF	91.77	95.45
4	SPI	93.64	98.37
5	PSII	90.93	91.89

Overall, the total score tests revealed that the SPI method is the most consistent and reliable method for identifying hotspots. Although it can only be applied to roadway segments where the crash data for different levels of severity are available, with the rapid development of intelligent

transportation systems and data collection technologies, this method could become quite useful in identifying high-risk road sites. The performance of Model-II was also relatively better than PSII, SPF, and Model-I. This evaluation suggests that the SPI method and Model-II (of the methods compared) have the potential to become the industry standard.

Chapter 8 Generalized Nonlinear Model for Incident Prediction

This chapter describes the use of regression models to fit generalized nonlinear models to predict crash frequency by severity based on static and dynamic contributing factors. The proposed incident prediction functions were used to distinguish different crash severity types. The proposed models were proven to have better goodness-of-fit than those found in the existing literature and to provide a better fit for different incident severity types. In this effort, we used a different roadway segmentation technique. So this chapter introduces the data set first, then describes the generalized nonlinear models. Note that unlike the two-stage regression-logistic models, no weight was given for different road surface conditions.

8.1 Data Description

In this effort, segmentation of the roadways based on curvature was employed, i.e., a new segment started when a tangent section transitioned into a curve. Thus, the segment lengths were not fixed. Next, a sensitivity analysis was conducted to determine the threshold value for short road segments. In order to avoid losing too much information in the data sets, we selected 0.05 mile as the threshold value and removed all of the road segments having a length of less than 0.05 mile.

Tables 8.1 and 8.2 show summaries of all numerical and categorical variables, respectively.

Table 8-1 Summary statistics of numerical variables for Interstate freeway segments in Washington state for years 2011 – 2014 after removing outliers and short segments

Factor Type	Classification Type	Explanatory variables	Min.	Max.	Mean	Median	St.Dev.
Static	Road conditions	Segment length (ft)	264	71227.20	2342.11	1372.8	3615.54
		NOL	2	5	2.6	2	0.715
		COS	0	6.030	0.726	0	1.009
		OSW	0	18	6.596	8	3.741
		ISW	0	18	3.770	4	2.843
		MWD	5.59	999	94.89	68	156.79
		SPL	46.67	70	65.58	70	4.87
Dynamic	Traffic characteristics	AADT	6700	229500	68305.42	44000	58910.24
		AADT/Lane	1675	57375	12118.15	9500	8999.12
		Truck percentage	0%	29%	11.86%	8%	10.35%

Note: NOL=number of lanes; COS=curvature of the segment; OSW=weighted average width of outer shoulder; ISW= weighted average width of inner shoulder; MWD=weighted average width of median; SPL=average speed limit; AADT=annual average daily traffic.

Table 8-2 Summary statistics of categorical variables for Interstate freeway segments in Washington state for years 2011 – 2014 after removing outliers and short segments

Factor Type	Classification Type	Explanatory variables	Number of categories	Category Types
Static	Road conditions	HCT	3	S=Straight, L=Left, R=Right
		LST	3	A= Asphalt, B= Bituminous, P= Portland Cement Concrete
		OST	6	A= Asphalt, B= Bituminous, C= Curb, P=Portland Cement Concrete, W=Wall, O=Other
		IST	6	A= Asphalt, B= Bituminous, C= Curb, P=Portland Cement Concrete, W=Wall, O=Other
		MST	4	A= Asphalt, P=Portland Cement Concrete, S=Soil, O=Other
Dynamic	Weather conditions	RSC	3	Dry, Wet, Snow/Ice/Slush
		Visibility	2	Good, Bad

Note: HCT=horizontal curve type; LST=dominant lane surface type; OST=dominant outer shoulder type; IST=dominant inner shoulder type; MST=dominant median type; RSC=road surface conditions

8.2 Generalized Nonlinear Models for Incident Prediction

In classical linear regression models, the expectation of crash frequency (or rate) is formulated as an ordinary linear model. This model specification can be expressed as follows (McCullagh, 1984):

$$E(y_i) = \mu_i = L_i \sum_{j=1}^{J} x_{ij}\beta_j \tag{8.1}$$

where y_i denotes the crash frequency (or rate) along roadway segment *i*, $E(y)_i$ or μ_i is the expected crash frequency (or rate) along segment i during a certain time period; L_i is the

segment length in miles; x_{ij} is the jth explanatory variable for segment i; β_j is the corresponding coefficient for the jth explanatory variable; and J is the total number of explanatory variables considered in the model. In comparison to the simplest linear regression, more complicated models, such as Poisson and negative binomial models for crash frequency and logit and probit models for crash severity, have been used to interpret crash data. These models can be generalized by using a smooth and invertible linearizing link function to transform the expectation of the response variable, μ_i, to its linear predictor:

$$g(\mu_i) = L_i \sum_{j=1}^{J} x_{ij}\beta_j \tag{8.2}$$

where $g(.)$ is the link function, which is monotonic, differentiable, and used to connect the linear predictor of the explanatory variables with the expected crash frequency (or rate) in various formats, such as identity, log, logit, etc. In this research, the log function was used for crash analysis.

As we have discussed earlier, in many scenarios the relationship between the expected crash frequency (or rate) by severity level and its associated factors cannot be simply expressed by GLMs. GNMs are proposed as an extension of GLMs in order to satisfy such a specific requirement by changing the linear predictor to be nonlinear in the parameters, β_j, in Equation (8.3).

The GNM-based method uses a user-defined, customized function to extract the relationship between crash risks and contributing factors with more general assumptions. For the other explanatory variables, the diverse nonlinear predictor, $U(x)$, such as the polynomial

function, exponential function, parabolic function, logarithmic function, etc., may be utilized to extract proper data features. In general, the model format of $U(x)$ can be determined on the basis of a statistical analysis of the crash rate and a specific explanatory variable. Notice that the defined nonlinear function $U(x)$ is an assumed relationship. This defined function can be revised on the basis of further statistical analysis. Aggregating the nonlinear predictors for all the independent variables, equation (8.3) can be rearranged as:

$$E(y_i) = \mu_i = L_i \sum_{j=1}^{J} U_j(x_{ij})\omega_j, \quad i = 1, 2, \cdots, n, \tag{8.3}$$

where $U_j(x_{ij})$ is a nonlinear predictor for the jth explanatory variable; and ω_j is the corresponding weight for $U_j(x_{ij})$. Consequently, the GNM link functions becomes:

$$g(\mu_i) = \sum_{j=1}^{J} U_j(x_{ij})\omega_j, \quad i = 1, 2, \cdots, n, \tag{8.4}$$

If all of the $U_j(x_{ij})$ in the model are linear regressions of x_{ij}, a GNM will degrade to a GLM. Therefore, in this research, GLMs were special cases of GNMs. To make road sections comparable, in this research, $g(\mu_i)$ was considered as a logarithmic function and applied on the basis of accident density (i.e., accidents per kilometer-year) as show below:

$$g(\mu_i) = \ln d_i = \ln \frac{\mu_i}{L_i y_i} = \sum_{j-1}^{J} U_j(x_{ij})\omega_j = U_i\omega, \quad i = 1, 2, \cdots, n, \tag{8.5}$$

where $d_i = \mu_i / L_i y_i$, L_i, and y_i are the crash density, segment length, and time period length (years) of crash frequency of roadway segment *i* respectively; and $\omega = [\omega_1, \omega_2, \ldots, \omega_J]^T$ is the coefficient vector for $U_i = [U x_1(_{i1}), U x_2(_{i2}), \ldots, U x_J(_{iJ})]$ when the expected crash density is estimated.

8.3 GNM-Based Multinomial Logistic Regression Approach

Logistic regression is generally used to handle categorical data (Bham, Javvadi and Manepalli, 2011). It can handle bivariate response variables, i.e., variables with two possible values, and can be extended to handle a polytomous response variable Y that takes a discrete set of values reflecting *K* categories (*K* can be greater than two). Since the response variable is nominal (unordered), a generalized logit model is suitable. This approach frames *K*-1 logits for the response variable to compare each categorical level with a reference category.

In this research, three categories were considered for the crash severity (i.e., fatal (*k*=1), injury (*k*=2), and PDO (*k*=3)), in which PDO crashes were used as the base category for comparison with the other categories. The crash severity type, denoted by Y, was the response variable, whereas contributing factors for road conditions, traffic characteristics, and weather conditions were the independent variables denoted by x_{ij} (*j*=1, 2, …, *J*), where *i* denotes the observation and *J* denotes the number of independent variables. The expression of Y is defined as below:

$$Y = \begin{cases} 1, & if \ \ crash \ \ is \ \ fatal, \\ 2, & if \ \ crash \ \ is \ \ injury, \\ 3, & if \ \ crash \ \ is \ \ PDO, \end{cases} \tag{8.6}$$

Based on the GNM link functions in equation (8.5), when the response categories 1, …, K (K=3) are unordered, Y is related to independent variables through a set of K-1 baseline category logits as shown below:

$$\ln \frac{\Pr(Y_i = k)}{\Pr(Y_i = K)} = \sum_{j=1}^{J} U_{kj}(x_{ij})\omega_{kj} = U_{ki}\omega_k, \quad i = 1,2,\cdots,n; \quad k = 1,\cdots K-1 \tag{8.7}$$

where $\Pr(Y = k_i)$ is the probability of crash of severity type k; $U_{ki} = [U_{k1}(x_{i1}), U_{k2}(x_{i2}), \ldots, U_{kJ}(x_{iJ})]$ is the nonlinear predictor vector of observation i; $\omega_k = [\omega_{k1}, \omega_{k2}, \ldots, \omega_{kJ}]^T$ is the coefficient vector for kth level of the response variable.

By exponentiating both sides of equation (8.7) and solving for the probabilities, we get:

$$\Pr(Y_i = k) = \Pr(Y_i = K\, e)^{U_{ki}\omega_k}, \quad i = 1,2,\ldots,n; \; k = 1,\ldots K-1 \tag{8.8}$$

Using the fact that all k of the probabilities must sum to one, we find:

$$\Pr(Y_i = K) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{U_{ki}\omega_k}}, \quad i = 1, 2, \cdots, n \tag{8.9}$$

According to equations (8.8) and (8.9), the other probabilities can be expressed as below:

$$\Pr(Y_i = k) = \frac{e^{U_{ki}\omega_k}}{1 + \sum_{k=1}^{K-1} e^{U_{ki}\omega_k}}, \quad i = 1, 2, \cdots, n; \quad k = 1, \cdots K-1 \tag{8.10}$$

Equations (8.9) and (8.10) are called the prediction functions of the GNM-based multinomial logistic regression approach. In equations (8.9) and (8.10), it can be seen that again, if all the $U_j(x_{ij})$ in the nonlinear predictor vector are linear regressions of x_{ij}, the GNM-based multinomial logistic regression approach will degrade to a normal multinomial logistic regression approach.

8.4 Estimation of Nonlinear Predictors

It is necessary to determine the appropriate predictors $U_j(x_{ij})$ and $U_{kj}(x_{ij})$ in equations (8.9) and (8.10) before the corresponding coefficients ω_j and ω_{kj} can be calibrated. To better illustrate the nonlinear contribution function estimation process, an example is detailed to formulate the contribution function for the variable,AADT per lane as follows.

Assume the number of crashes for each severity level follows the Poisson model, and all other dependent variables are approximately consistent across different AADT levels when the sample data are large enough. Let *j*=1 denote the index of contributing factor AADT per lane. To develop an appropriate format of the predictor $U x_1(_{i1})$, the visualized comparisons between

the logarithm of the expectation of crash density (number of crashes per mile per year) and AADT per lane are illustrated in figure 8-1. The scatter points show the logarithm of crash density from the Interstate freeway segments in Washington for the years 2011-2014, classified by AADT per lane. As we can see, the logarithm of the average crash density tends to increase when the AADT per lane increases at a variable rate. The increase rate becomes smaller with the higher AADT per lane, which indicates the inappropriateness of using a linear contribution function. To address this issue, we adopted a logarithmic calculation as the nonlinear predictor to approximate the impacts of AADT on crash density:

$$U\,x_1(\ _1) = -0.371x_1^2 + 0.29x_1 - 2.07 \qquad (8.11)$$

In comparison to the linear predictor $L\,(x_{1)} = 3.82x_1 - 1.58$, the value of R^2 increases from 0.8413 to 0.9051 when the nonlinear predictor, $U\,x_1(\ _1)$, is utilized, as shown in figure 8-1. Therefore, the nonlinear predictor is more suitable for describing the relationship between crash density and AADT, and it was employed in this study. The logarithm of AADT and its impacts on crash frequencies have been found significant, and thiese results are consistent with many previous studies (Wong, Sze and Li, 2007; Abdel-Aty and Haleem, 2011; Lao *et al.*, 2014).

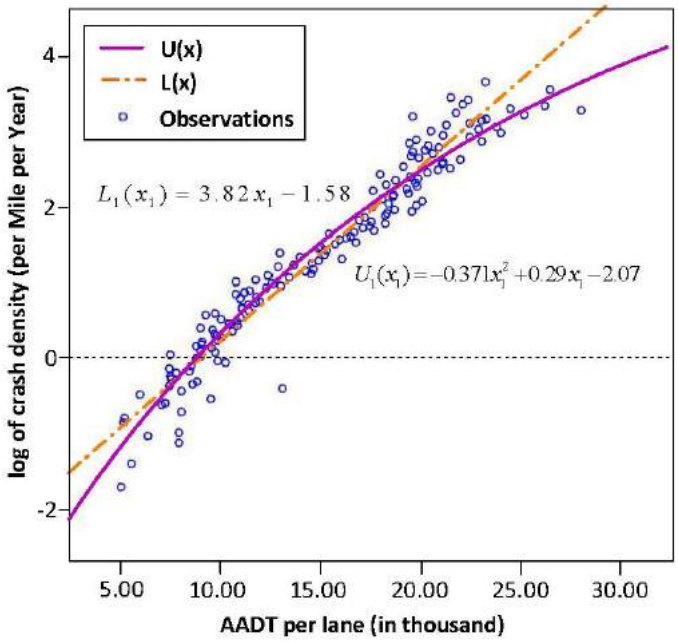


Figure 8-1 Logarithm of the expectation of crash density (number of crashes per mile per year) from Interstate freeway segments in Washington state for years 2011 – 2014, by AADT per lane

Similar procedures can be applied to develop the formats of the predictors $U\,x_{11}(\,)_{i1}$ and $U\,x_{21}(\,_{i1})$ as shown below:

$$U_{11}(x_1) = 0.159x_1^2 - 0.1x_1 + 0.0796 \tag{8.12}$$

$$U_{21}(x_1) = 0.132x_1^2 - 0.126x_1 + 0.108 \tag{8.13}$$

For the other contributing factors, similar studies can be conducted to develop the nonlinear predictor functions. The other contribution functions for the variables truck percentage (j=2), NOL (j=3), COS (j=4), WOS (j=5), WIS (j=6), WM (j=7), and ASL (j=8) are shown in equations (8.14) through (8.34), respectively:

Truck percentage:

$$U_2(x_2)=\begin{cases}1.32+\dfrac{1.78(x_2-1)}{3}, & x_2\le 4\\ -1.563\ln(\dfrac{x_2}{2}-1)+3.1, & x_2>4\end{cases} \quad (8.14)$$

$$U_{12}(x_2) = 0.245x_2^2 - 0.114x_2 + 0.0987 \quad (8.15)$$

$$U_{22}(x_2) = 0.163x_2^2 - 0.096x_2 + 0.117 \quad (8.16)$$

NOL:

$$U_3(x_3) = 0.35x_3^2 - 1.99x_3 + 3.18 \quad (8.17)$$

$$U_{13}(x_3) = -0.217x_3^2 - 0.043x_3 + 0.137 \quad (8.18)$$

$$U_{23}(x_3) = -0.188x_3^2 + 0.163x_3 - 1.02 \quad (8.19)$$

COS:

$$U_4(x_4) = 0.342x_4^2\, e^{-0.278x_{43}} \quad (8.20)$$

$$U_{14}(x_4) = -0.043x_4^2 + 0.028x_4 - 0.359 \quad (8.21)$$

$$U_{24}(x_4) = -0.036x_4^2 + 0.0673x_4 - 0.882 \quad (8.22)$$

OSW:

$$U_5(x_5) = -0.274x_5^2 + 0.785x_5 - 1.06 \quad (8.23)$$

$$U_{15}(x_5) = 0.104x_5^2 - 0.263x_5 + 0.985 \quad (8.24)$$

$$U_{25}(x_5) = 0.307x_5^2 - 0.088x_5 + 0.774 \quad (8.25)$$

ISW:

$$U_6(x_6) = -0.013x_6^2 + 0.025x_6 - 0.054 \quad (8.26)$$

$$U_{16}(x_6) = 0.047x_6^2 - 0.061x_6 + 0.093 \quad (8.27)$$

$$U_{26}(x_6) = 0.012x_6^2 - 0.069x_6 + 0.084 \quad (8.28)$$

MWD:

$$U_7(x_7) = -0.476x_7^2 + 0.584x_7 - 0.137 \quad (8.29)$$

$$U_{17}(x_7) = 0.353x_7^2 - 0.183x_7 + 0.564 \quad (8.30)$$

$$U_{27}(x_7) = 0.145x_7^2 - 0.718x_7 + 0.244 \quad (8.31)$$

SPL:

$$U_8(x_8) = -0.065x_8^2 + 0.079x_8 - 0.085 \quad (8.32)$$

$$U_{18}(x_8) = 0.084x_8^2 - 0.098x_8 + 0.112 \quad (8.33)$$

$$U_{28}(x_8) = 0.101x_8^2 - 0.099x_8 + 0.113 \quad (8.34)$$

For the categorical contributing factors, including HCT (j=9), LST (j=10), OST (j=11), IST (j=12), MST (j=13), RSC (j=14), and visibility (j=15), an N-category class function was developed as below:

$$U_j(x_{ij}) = \begin{cases} \pi_{1j}, & category\ 1 \\ \cdots & \\ \pi_{Nj}, & category\ N \end{cases} \quad (8.35)$$

Table 8-3 summarizes the estimated parameters for each categorical predictor function, including HCT (j=9), LST (j=10), OST (j=11), IST (j=12), MST (j=13), RSC (j=14), and visibility (j=15).

Table 8-3 Summary of the estimated parameters for each categorical predictor function

Contributing Factors	Fatal	Injury	Crash density
Road conditions			
HCT			
Straight ($\pi_{1,9}$)			
Left ($\pi_{2,9}$)	-0.741	-0.279	0.412
Right ($\pi_{3,9}$)	-0.828	-0.219	0.875
LST	-0.678	-0.221	0.851
Asphalt ($\pi_{1,10}$)			
Bituminous ($\pi_{2,10}$)	-0.758	-0.270	0.421
Portland Cement ($\pi_{3,10}$)	-0.001	-0.050	-1.106
OST	-0.727	-0.247	0.772
Asphalt ($\pi_{1,11}$)			
Bituminous ($\pi_{2,11}$)	-0.734	-0.259	0.478
Curb ($\pi_{3,11}$)	-0.001	-0.098	0.362
Portland Cement ($\pi_{4,11}$)	-1.393	-0.323	1.471
Wall ($\pi_{5,11}$)	-0.762	-0.243	1.198
Other ($\pi_{6,11}$)	-1.223	-0.227	1.700
IST	-0.001	-0.202	0.673
Asphalt ($\pi_{1,12}$)			
Bituminous ($\pi_{2,12}$)	-0.702	-0.231	0.232
Curb ($\pi_{3,12}$)	-0.001	-0.129	0.335
Portland Cement ($\pi_{4,12}$)	-1.393	-0.337	1.658
Wall ($\pi_{5,12}$)	-1.223	-0.230	1.746
Other ($\pi_{6,12}$)	-0.743	-0.221	1.333
MST	-0.823	-0.296	1.036
Asphalt ($\pi_{1,13}$)			
Portland Cement ($\pi_{2,13}$)	-0.799	-0.303	1.028
Soil ($\pi_{3,13}$)	-0.798	-0.263	1.097
Other ($\pi_{4,13}$)	-0.703	-0.230	0.282
	-1.308	-0.253	1.552

Weather conditions			
RSC			
Dry ($\pi_{1,14}$)	-0.749	-0.259	0.795
Wet ($\pi_{2,14}$)	-0.759	-0.269	0.739
Snow/Ice/Slush ($\pi_{3,14}$)	-0.668	-0.205	-0.080
Visibility			
Good ($\pi_{1,15}$)	-0.907	-0.267	0.825
Bad ($\pi_{2,15}$)	-0.524	-0.246	0.430

8.5 Estimation of Coefficients in GNMs

On the basis of the preceding results and estimated nonlinear predictors for each contributing factor, the coefficient vector $\omega = [\omega_1, \omega_2, \ldots, \omega_J]^T$ in equation (8.5) and coefficient vector $\omega_k = [\omega_{k1}, \omega_{k2}, \ldots, \omega_{kJ}]^T$ in equation (8.7) could be estimated by employing a multivariate regression method based on the collected data from Interstate freeways in Washington from 2011 to 2014. The results for significant factors (α=0.05) are shown in table 8-4.

Table 8-4 Summation of the estimated parameters for each categorical predictor function

Contributing Factors	Fatal (ω_{1j})			Injury (ω_{2j})			sh density (ω_j)		
	Coeff.	Std. error	t-value	Coeff.	Std. error	t-value	Coeff.	Std. error	t-value
Traffic characteristics									
AADT/Lane	0.114	0.093	15.651	0.402	0.068	19.154	0.714	0.047	21.865
Truck percentage	0.065	0.087	4.318	0.157	0.056	6.315	0.265	0.033	8.231
Road conditions									
NOL	-0.012	0.132	-8.213	-0.278	0.098	-9.334	-0.312	0.076	-5.892
COS	0.109	0.078	4.332	0.312	0.092	6.124	0.437	0.083	5.982
OSW	0.145	0.108	6.784	0.398	0.132	8.767	0.576	0.102	7.818
ISW	-0.089	0.093	-3.424	-0.021	0.045	-4.233	-0.091	0.106	-4.983
MWD	0.005	0.037	2.987	0.038	0.009	4.732	0.105	0.051	5.875
SPL	-0.003	0.012	-2.874	-0.08	0.023	-4.675	-0.012	0.007	-3.987
HCT	0.010	0.013	3.418	0.032	0.031	4.762	0.068	0.018	5.383
LST	-0.024	0.028	-2.873	-0.047	0.036	-8.165	-0.172	0.212	9.289
OST	0.078	0.030	3.589	0.121	0.046	5.327	0.239	0.075	10.231
IST	0.102	0.045	6.732	0.245	0.068	8.404	0.415	0.066	13.538
MST	0.091	0.076	4.383	0.184	0.093	9.673	0.205	0.142	12.665
Weather conditions									
RSC	-0.302	0.078	-5.884	-0.428	0.066	-6.893	-0.398	0.058	-7.545
Visibility	0.213	0.155	4.672	0.353	0.098	5.387	1.325	0.106	8.342
R_2		0.856			0.902			0.913	

8.6 Safety Performance Index

In order to develop a new safety performance index that can reflect crash counts by severity level, the generalized nonlinear model was employed to estimate the expected crash density in a roadway segment *i* as shown below:

$$\ln d_i = \ln \frac{\mu_i}{L_i y_i} = \sum_{j=1}^{J} U_j(x_{ij})\omega_j, \quad i = 1, 2, \cdots, n, \tag{8.36}$$

i.e.,

$$d_i = e^{\sum_{j=1}^{J} U_j(x_{ij})\omega_j} = e^{U_i \omega}, \quad i = 1, 2, \cdots, n, \tag{8.37}$$

where $d_i = \mu_i / L_i y_i$ is the crash density of roadway segment *i* during a certain time period.

Based on equations (8.5), (8.7), and (8.37), we can get the expected crash density for different severity types.

(1) *Expected Fatal Crash Density*:

$$d_{i1} = d_i \cdot \Pr(Y_i = 1) = \frac{e^{U_i \omega + U_{1i} \omega_1}}{1 + \sum_{k=1}^{2} e^{U_{ki} \omega_k}}, \quad i = 1, 2, \cdots, n, \tag{8.38}$$

where d_{i1} is the expected fatal crash density along segment *i* during a certain time period; $\omega = [\omega_1, \omega_2, \ldots, \omega_J]^T$ is the coefficient vector for $U_i = [U_1(x_{i1}), U_2(x_{i2}), \ldots, U_J(x_{iJ})]$ when estimating the expected crash density.

(2) *Expected Injury Crash Density*:

$$d_{i2} = d_i \cdot \Pr(Y_i = 2) = \frac{e^{U_i \omega + U_{2i} \omega_2}}{1 + \sum_{k=1}^{2} e^{U_{ki} \omega_k}}, \quad i = 1, 2, \cdots, n, \tag{8.39}$$

where d_{i2} is the expected injury crash density along segment *i* during a certain time period.

(3) *Expected PDO Crash Density*

$$d_{i3} = d_i \cdot \Pr(Y_i = 3) = \frac{e^{U_i \omega}}{1 + \sum_{k=1}^{2} e^{U_{ik} \omega_k}}, \quad i = 1, 2, \cdots, n, \tag{8.40}$$

where d_{i3} is the expected PDO crash density along segment i during a certain time period. Based on equations (8.38) through (8.40), the equivalent property damage only (EPDO) crash frequency measure was modified and employed to weight crashes according to severity (fatal, injury, and PDO) to develop a combined crash density and severity score (CCDSS) for each site. On the basis of the crash counts and equation (7.3), we can calculate that $\eta_F = 0.0028$, $\eta_I = 0.29$, $\eta_P = 0.7072$; then, the values of the risk weight factors are obtained as $F_w = 3.884$, $I_w = 3.691$, $P_w = 1$.

Table 8-5 presents the formulation of Safety Performance Function (SPF), Safety Performance Index (SPI) and Potential Safety Improvement Index (PSII) based on GNM.

Table 8-5 Description of SPF, SPI and PSII based on GNM

Model	Description
SPF	$ECCDSS_i = d_{i1}F_w + d_{i2}I_w + d_{i3}P_w, \quad i = 1,2, \ldots, n,$
SPI_i	$SPI_i = \lambda_i ECCDSS_i + (1-\lambda_i)OCCDSS_i, \; i = 1,2, \ldots, n,$ where $OCCDSS_i = \sigma_{i1}F_w + \sigma_{i2}I_w + \sigma_{i3}P_w, i = 1,2, \ldots, n,$ $\lambda_i = \frac{1}{1+\alpha_i ECCDSS_i}$ σ_{i1}, σ_{i2}, σ_{i4} are the observed fatal, injury, and PDO crash density along segment *i* λ_i = a weighting factor α_i = over-dispersion parameter
$PSII_i$	$PSII_i = \lambda_i ECCDSS_i + (1-\lambda_i)OCCDSS_i - ECCDSS_i = SPI_i - ECCDSS_i, \quad i = 1,2, \ldots, n,$

Note: $OCCDSS_i$ is the observed combined crash density and severity score for roadway segment *i*, $ECCDSS_i$ = expected combined crash density and severity score for roadway segment *i*.

8.7 Evaluation Tests of the Performance of HSID Methods

In order to demonstrate the effectiveness of the SPI and PSII as developed in this research, they were compared with six reference performance indexes, which included the expected crash density based on the conventional safety performance function from the *HSM* (i.e., NB GLM), expected crash density based on the GNM, EB estimated crash density based on the NB GLM, EB estimated crash density based on the GNM, ARP based on the NB GLM, and ARP based on

the GNM. In this evaluation, both the top 1 percent and 5 percent of the locations were used as experimental values.

8.7.1 Site Consistency Test

Following equation (7.10), the site consistency test was done for all eight reference models. Table 8-6 shows that the SPI method outperformed other HSID methods in identifying both the top 1 percent and 5 percent of hotspots with the highest SCT values, 21521.79 and 44251.68, respectively, in Period 2, followed closely by the EB CD (GNM) method. The ARP (NB GLM) performed the worst in both cases, with the identified hotspots experiencing the lowest number of SCT values, say, 19034.25 and 42106.73, respectively.

Table 8-6 Results of site consistency test of various HSID methods

HSID Method Index (h)	HSID Method Name	$\gamma = 0.01$		$\gamma = 0.05$	
		$SCT_{h,t,}$ Period 1 (2011-2012)	$SCT_{h,t,\square 1}$ Period 2 (2013-2014)	$SCT_{h,t,}$ Period 1 (2011-2012)	$SCT_{h,t,\square 1}$ Period 2 (2013-2014)
1	SPI	22943.06	21521.79	46310.12	44251.68
2	PSII	22198.11	20879.23	45389.79	43921.56
3	CD (NB GLM)	21036.33	20105.38	43897.54	42868.83
4	CD (GNM)	21901.86	20993.89	44890.67	43571.45
5	EB CD (NB GLM)	21896.77	20981.31	45303.46	43784.09
6	EB CD (GNM)	22856.34	21349.04	45987.98	44012.49
7	ARP (NB GLM)	20131.67	19034.25	43015.82	42106.73
8	ARP (GNM)	20358.78	19734.55	43823.17	42871.29

8.7.2 Method Consistency Test

Table 8-7 shows the number of similarly identified hotspots identified by alternative HSID methods over the two periods, which was determined on the basis of the method described in Section 7.3.2.

Table 8-7 Results of method consistency test of various HSID methods

HSID Method Index (*h*)	HSID Method Name	□□ 0.01	□□ 0.05
1	SPI	124 (60.7%)	546 (53.5%)
2	PSII	103 (50.5%)	452 (44.3%)
3	CD (NB GLM)	92 (45.1%)	406 (39.8%)
4	CD (GNM)	101 (49.5%)	423 (41.5%)
5	EB CD (NB GLM)	109 (53.4%)	441 (43.2%)
6	EB CD (GNM)	118 (57.8%)	472 (46.3%)
7	ARP (NB GLM)	83 (40.7%)	382 (37.5%)
8	ARP (GNM)	88 (43.1%)	393 (38.5%)

The SPI method was superior in this test by identifying the largest number of the same hotspots in both cases of γ = 0.01 and γ= 0.05, with 124 and 546 roadway segments, respectively. In other words, the SPI method identified 124 segments in 2011-2012 that were also identified as hotspots in 2013-2014. The ED CD (GNM), which performed slightly better than the ED CD (GLM) method, placed second, identifying 118 consistent hotspots (in the case of γ = 0.01) and 472 consistent hotspots (in the case of γ = 0.05).

8.7.3 Total Rank Differences Test

Table 8-8 illustrates that the SPI method was superior in the total rank differences test. In both the $\gamma = 0.01$ and $\gamma = 0.05$ cases, the SPI method had significantly smaller-summed ranked differences. The total rank difference test was described in Section 7.3.3.

Table 8-8 Results of total rank differences test of various HSID methods

HSID Method Index (*h*)	HSID Method Name	□□ 0.01	□□ 0.05
1	SPI	2354	10237
2	PSII	3031	12798
3	CD (NB GLM)	3672	14781
4	CD (GNM)	3298	13587
5	EB CD (NB GLM)	3158	13016
6	EB CD (GNM)	2887	11973
7	ARP (NB GLM)	4123	18167
8	ARP (GNM)	3887	17105

8.7.4 Total Score Test

The total score test was performed according to the procedure described in Section 7.3.4 . Table 8-9 illustrates the results of the total score test of the eight HSID methods, in which SPI performed best in both the $\gamma = 0.01$ and $\gamma = 0.05$ cases, followed closely by the EB CD (GNM) method, with a 93.81 score (in the case of $\gamma = 0.01$) and 92.12 score (in the case of $\gamma = 0.05$). ARP (NB GLM) performed the worst in both cases, with a 70.82 score and 73.82 score, respectively.

Table 8-9 Results of the total score test of various HSID methods

HSID Method Index (*h*)	HSID Method Name	$\gamma = 0.01$	$\gamma = 0.05$
1	SPI	100	100
2	PSII	87.89	89.31
3	CD (NB GLM)	78.55	82.07
4	CD (GNM)	85.37	85.83
5	EB CD (NB GLM)	88.63	88.14
6	EB CD (GNM)	93.81	92.12
7	ARP (NB GLM)	70.82	73.82
8	ARP (GNM)	75.16	77.02

On several criteria, the SPI outperformed other methods by a wide margin. This evaluation suggests that the SPI method (of the methods compared) has the potential to become the industry standard.

Chapter 9 Regional, Map-Based Analytical Platform

A regional, map-based analytical platform was developed on the DRIVE Net system to highlight the methodology developed under this project. Ultimately, the existing safety performance analysis function under the system's "Safety Performance" module was expanded. The developed SPI color codes the regional map on the basis of safety performance. The PSII highlights potential safety improvements on the map. By combining the two indices on the regional map, one can easily identify accident hotspots and the key influencing factors to consider in an improvement package.

The interface of the safety performance module on the regional, map-based analytical platform is illustrated in figure 9-1. There are three sub-functions implemented on this panel: Incident Frequency (NB GLM), Estimated Crash Mean, and Potential Safety Improvement Index (ARP NB GLM). The new SPI and PSII were added as expanded safety performance analysis options. As stated previously, within a selected time range and corridor, the SPI shows a more comprehensive view of safety performance on a given corridor. The accident/incident data were from the Washington Incident Tracking System (WITS) and HSIS database. The SPI level ranges from Level A to Level F, where Level A (light green) corresponds to the highest safety performance and Level F (dark red) corresponds to the lowest safety performance expected, as shown in figure 9-2.

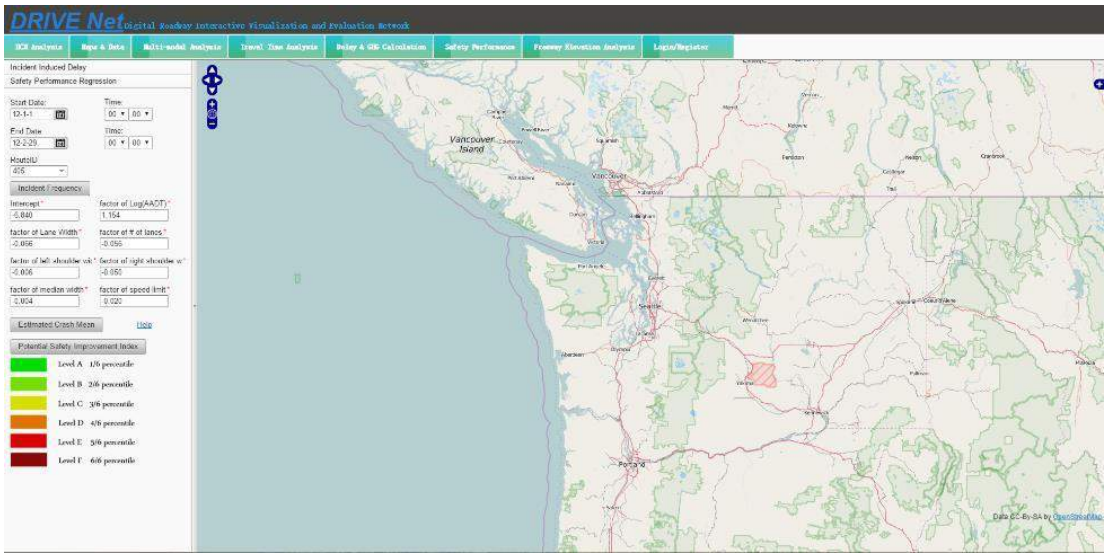


Figure 9-1 Interface of the safety performance module on the regional, map-based analytical platform

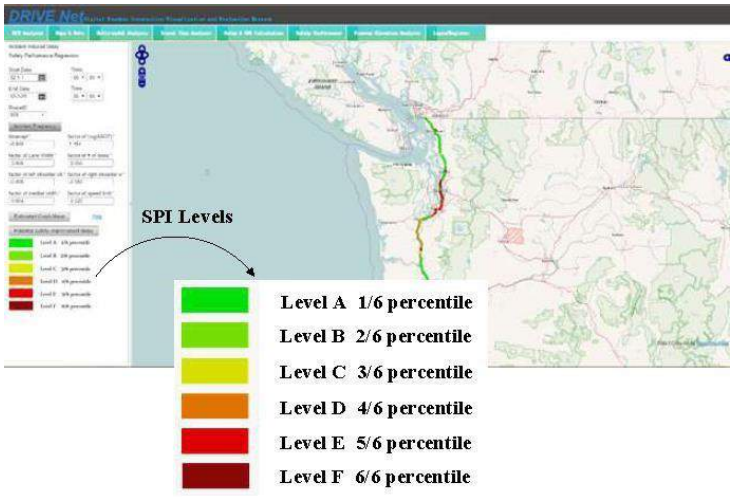


Figure 9-2 SPI levels range from Level A to Level F in the safety performance module**.**

The PSII implements the EB method in the modeling part. In this function, both the historical incident data and the characteristics of the selected corridor are used as model inputs. The output format still uses the six different colors representing Level A to Level F to show the potential safety improvement index on the map, where Level A shows the segments have the least potential to improve safety, and Level F shows the segments have the most potential to improve safety. Figure 9-3 shows an example of this function.

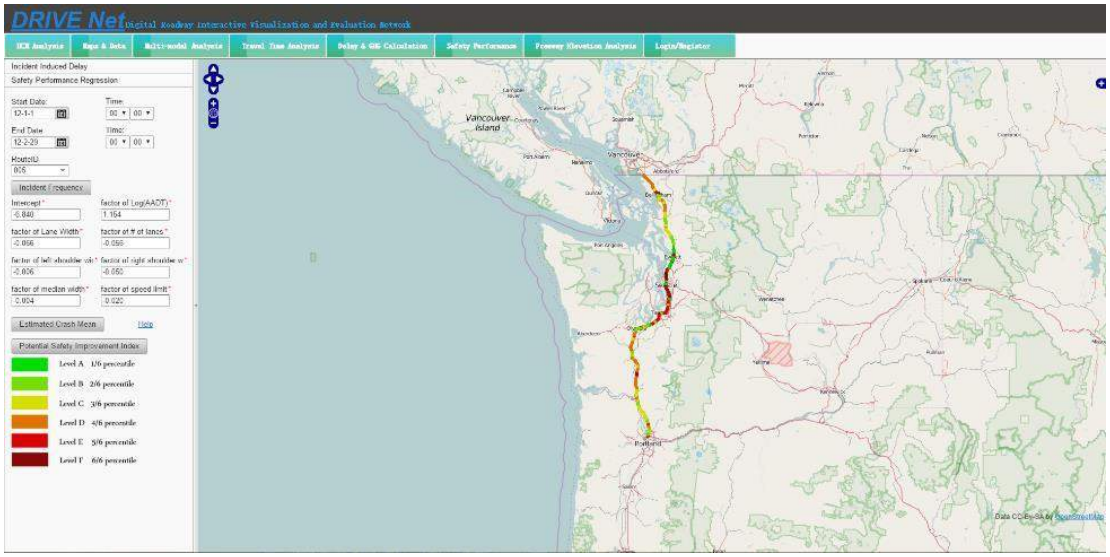


Figure 9-3 An example of the PSII function**.**

Chapter 10 Conclusions and Recommendations

This chapter discusses the conclusions of this research and explains some potential research directions for future studies. The outcome of this project has been and will be disseminated to different audiences through making presentations at different conferences, organizing workshops at state DOTs involved in the PacTrans consortium, publishing papers in reputable transportation safety journals, and development of a platform to implement the safety performance assessment.

10.1 Conclusions

This research extended the current state-of-the-art of crash count modeling by considering severity prediction to develop a two-stage (integrating logit and generalized linear regression models) and generalized nonlinear regression model for formulating a new CCS-based HSID method. This method can be utilized to improve the safety performance module of DRIVE Net, an analytical platform that allows for state-wide highway safety performance assessment. The most important contributing factors (static and dynamic) to traffic crashes with different severity types, including traffic characteristics, road conditions, and weather conditions, are identified by using a structured framework developed in this research. A total of 802 road segments on I-5, I-90, I-82, I-182, I-205, I-405 and I-705 in Washington state were selected as the candidate sites for data collection. A data quality control method was employed to remove short road segments, and a sensitivity analysis was conducted to determine the threshold values for short road segments. Two-stage regression and GNM-based multinomial logistic regression approaches were developed to estimate the probability and frequency of crashes for different severity levels. The regression analysis found that the contributing factors, including AADT per lane, NOL, COS, OSW, ISW, MWD, SPL, HCT, LST, OST, IST, MST, RSC, and visibility, showed strong relationships with the crash frequency of different severity

levels. It also showed that the significance and degrees of nonlinearity for each crash severity level were different among the contributing factors. A CCS-based HSID method as developed by employing the two-stage regression and the GNM-based multinomial logistic regression approaches. A new safety performance index and a new potential safety improvement index were developed by introducing a risk weight factor and were compared by employing HSID evaluating methods. The results of four consistency tests revealed that the SPI method is the most consistent and reliable method for identifying hotspots. Although it can only be applied to roadway segments where crash data for different levels of severity are available, with the rapid development of intelligent transportation systems and data collection technologies, this method could become useful in identifying high-risk road segments. This evaluation suggests that the SPI method (among the methods compared) has the potential to become the industry standard. Finally, a regional, map-based analytical platform esd developed in the DRIVE Net system by expanding the safety performance module with the new SPI and PSII functions.

10.2 Recommendations for Future Research

The team finds future work focusing on the following four areas to be promising: (1) developing a framework for a real-time safety performance analysis platform; (2) considering the analysis of crash frequency by collision type and severity; (3) including driver characteristics in crash prediction models; and (4) developing new criteria for evaluating methods of identifying hotspots based on new safety performance indexes.

References

AASHTO 2010

Abdel-Aty, M. (2003) 'Analysis of driver injury severity levels at multiple locations using ordered probit models', *Journal of safety research*, 34(5), pp. 597–603.

Abdel-Aty, M. A. and Radwan, A. E. (2000) 'Modeling traffic accident occurrence and involvement', *Accident Analysis & Prevention*, 32(5), pp. 633–642.

Abdel-Aty, M. and Abdelwahab, H. (2004) 'Modeling rear-end collisions including the role of driver's visibility and light truck vehicles using a nested logit structure', *Accident Analysis & Prevention*, 36(3), pp. 447–456.

Abdel-Aty, M. and Haleem, K. (2011) 'Analyzing angle crashes at unsignalized intersections using machine learning techniques', *Accident Analysis & Prevention*. Elsevier, 43(1), pp. 461– 470.

Agarwal, M., Maze, T. H. and Souleyrette, R. (2005) 'Impacts of weather on urban freeway traffic flow characteristics and facility capacity', in *Proceedings of the 2005 mid-continent transportation research symposium*, pp. 18–19.

Alkaabi, A., Dissanayake, D. and Bird, R. (2011) 'Analyzing clearance time of urban traffic accidents in Abu Dhabi, United Arab Emirates, with hazard-based duration modeling method', *Transportation Research Record: Journal of the Transportation Research Board*. Transportation Research Board of the National Academies, (2229), pp. 46–54.

Anarkooli, A. J. and Hosseinlou, M. H. (2016) 'Analysis of the injury severity of crashes by considering different lighting conditions on two-lane rural roads', *Journal of safety research*. Elsevier, 56, pp. 57–65.

Anastasopoulos, P. C. and Mannering, F. L. (2009) 'A note on modeling vehicle accident frequencies with random-parameters count models', *Accident Analysis & Prevention*. Elsevier, 41(1), pp. 153–159.

Automated Surface Observing System : ASOS user's guide (1998). [Washington, D.C.?] : U.S. Dept. of Commerce, National Oceanic and Atmospheric Administration : Federal Aviation Administration : U.S. Navy : U.S. Dept. of the Air Force, [1998].

Bauer, K. and Harwood, D. (2013) 'Safety effects of horizontal curve and grade combinations on rural two-lane highways', *Transportation Research Record: Journal of the Transportation Research Board*, (2398), pp. 37–49.

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

Benoit, K. (2011) 'Linear regression models with logarithmic transformations', *London School of Economics, London*.

Bham, G. H., Javvadi, B. S. and Manepalli, U. R. R. (2011) 'Multinomial logistic regression model for single-vehicle and multivehicle collisions on urban US highways in Arkansas', *Journal of Transportation Engineering*. American Society of Civil Engineers, 138(6), pp. 786– 797.

Bijleveld, F. and Churchill, T. (2009) *The influence of weather conditions on road safety*. SWOV.

Brodsky, H. and Hakkert, A. S. (1988) 'Risk of a road accident in rainy weather', *Accident Analysis & Prevention*, 20(3), pp. 161–176.

Cafiso, S., D'Agostino, C. and Persaud, B. (2013) 'Investigating the influence of segmentation in estimating safety performance functions for roadway sections', in *92nd Annual Meeting of the Transportation Research Board, Washington, DC*.

Cafiso, S., Di Graziano, A., Di Silvestro, G. and La Cava, G. (2008) 'Safety performance indicators for local rural roads: comprehensive procedure from low-cost data survey to accident prediction model', in *Transportation Research Board 87th Annual Meeting*.

Cafiso, S. and Di Silvestro, G. (2011) 'Performance of safety indicators in identification of black spots on two-lane rural roads', *Transportation Research Record: Journal of the Transportation Research Board*. Transportation Research Board of the National Academies, (2237), pp. 78–87.

Cenek, P. D., Davies, R. B., McLarin, M. W., Griffith-Jones, G. and Locke, N. J. (1997) 'Road environment and traffic crashes', *Transfund New Zealand research report*, (79).

Chang, L.-Y. and Chen, W.-C. (2005) 'Data mining of tree-based models to analyze freeway accident frequency', *Journal of Safety Research*, 36(4), pp. 365–375.

Chen, Y. and Tjandra, S. (2014) 'Daily Collision Prediction with SARIMAX and Generalized Linear Models on the Basis of Temporal and Weather Variables', *Transportation Research Record: Journal of the Transportation Research Board*. Transportation Research Board of the National Academies, (2432), pp. 26–36.

Cheng, W. and Washington, S. (2008) 'New criteria for evaluating methods of identifying hot spots', *Transportation Research Record: Journal of the Transportation Research Board*. Transportation Research Board of the National Academies, (2083), pp. 76–85.

Cheng, W. and Washington, S. P. (2005) 'Experimental evaluation of hotspot identification methods', *Accident Analysis & Prevention*. Elsevier, 37(5), pp. 870–881.

Chin, H. C. and Quddus, M. A. (2003) 'Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections', *Accident Analysis & Prevention*. Elsevier, 35(2), pp. 253–259.

Coll, B., Moutari, S. and Marshall, A. H. (2013) 'Hotspots identification and ranking for road safety improvement: An alternative approach', *Accident Analysis & Prevention*. Elsevier, 59, pp. 604–617.

Daniels, S., Brijs, T., Nuyts, E. and Wets, G. (2010) 'Explaining variation in safety performance of roundabouts', *Accident Analysis & Prevention*. Elsevier, 42(2), pp. 393–402.

Deacon, J. A., Zegeer, C. V and Deen, R. C. (1974) 'Identification of hazardous rural highway locations'.

Donnell, E. T. and Mason, J. M. (2006) 'Predicting the frequency of median barrier crashes on Pennsylvania interstate highways', *Accident Analysis & Prevention*. Elsevier, 38(3), pp. 590– 599.

Duncan, C., Khattak, A. and Council, F. (1998) 'Applying the ordered probit model to injury severity in truck-passenger car rear-end collisions', *Transportation Research Record: Journal of the Transportation Research Board*, (1635), pp. 63–71.

Easa, S. and You, Q. (2009) 'Collision prediction models for three-dimensional two-lane highways: horizontal curves', *Transportation Research Record: Journal of the Transportation Research Board*, (2092), pp. 48–56.

El-Basyouny, K. and Sayed, T. (2006) 'Comparison of two negative binomial regression techniques in developing accident prediction models', *Transportation Research Record: Journal of the Transportation Research Board*. Transportation Research Board of the National Academies, (1950), pp. 9–16.

Farah, H., Bekhor, S. and Polus, A. (2009) 'Risk evaluation by modeling of passing behavior on two-lane rural highways', *Accident Analysis & Prevention*. Elsevier, 41(4), pp. 887–894.

Geedipally, S., Patil, S. and Lord, D. (2010) 'Examination of methods to estimate crash counts by collision type', *Transportation Research Record: Journal of the Transportation Research Board*. Transportation Research Board of the National Academies, (2165), pp. 12–20.

Geedipally, S. R., Lord, D. and Dhavala, S. S. (2012) 'The negative binomial-Lindley generalized linear model: Characteristics and application using crash data', *Accident Analysis & Prevention*, 45, pp. 258–265.

Golob, T. F. and Recker, W. W. (2003) 'Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions', *Journal of Transportation Engineering*, 129(4), pp. 342– 353.

Gregoriades, A. (2007) 'Road safety assessment using bayesian belief networks and agent-based simulation', in *2007 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, pp. 615–620.

Hadi, M. A., Aruldhas, J., Chow, L.-F. and Wattleworth, J. A. (1995) 'Estimating safety effects of cross-section design for various highway types using negative binomial regression', *Transportation Research Record*, 1500, p. 169.

Haleem, K. and Gan, A. (2015) 'Contributing factors of crash injury severity at public highwayrailroad grade crossings in the US', *Journal of safety research*. Elsevier, 53, pp. 23–29.

Harb, R., Yan, X., Radwan, E. and Su, X. (2009) 'Exploring precrash maneuvers using classification trees and random forests', *Accident Analysis & Prevention*, 41(1), pp. 98–107.

Hauer, E. (1980) 'Bias-by-selection: Overestimation of the effectiveness of safety countermeasures caused by the process of selection for treatment', *Accident Analysis & Prevention*. Elsevier, 12(2), pp. 113–117.

Hauer, E., Ng, J. C. N. and Lovell, J. (1988) *Estimation of safety at signalized intersections (with discussion and closure)*.

Hauer, E., Persaud, B. N., Smiley, A. and Duncan, D. (1991) 'Estimating the accident potential of an Ontario driver', *Accident Analysis & Prevention*. Elsevier, 23(2), pp. 133–152.

Hensher, D. A. and Mannering, F. L. (1994) 'Hazard-based duration models and their application to transport analysis', *Transport Reviews*. Taylor & Francis, 14(1), pp. 63–82.

Heydari, S., Miranda-Moreno, L. F. and Liping, F. (2014) 'Speed limit reduction in urban areas: A before–after study using Bayesian generalized mixed linear models', *Accident Analysis & Prevention*, 73, pp. 252–261.

Hogema, J. H. and Van der Horst, A. R. A. (1994) 'Driving behaviour in fog: analysis of inductive loop data'.

Hojati, A. T., Ferreira, L., Washington, S. and Charles, P. (2013) 'Hazard based models for freeway traffic incident duration', *Accident Analysis & Prevention*. Elsevier, 52, pp. 171–181.

Hongguo, X., Huiyong, Z. and Fang, Z. (2010) 'Bayesian network-based road traffic accident causality analysis', in *Information Engineering (ICIE), 2010 WASE International Conference on*. IEEE, pp. 413–417.

Ivan, J. N., Gårder, P., Bindra, S., Jonsson, T., Shin, H.-S. and Deng, Z. (2007) 'Network-Based
Highway Crash Prediction Using Geographic Information Systems'. New England Transportation Consortium.

Johansson, P. (1996) ‘Speed limitation and motorway casualties: a time series count data regression approach’, *Accident Analysis & Prevention*, 28(1), pp. 73–87.

Jones, B., Janssen, L. and Mannering, F. (1991) ‘Analysis of the frequency and duration of freeway accidents in Seattle’, *Accident Analysis & Prevention*, 23(4), pp. 239–255.

Joshua, S. C. and Garber, N. J. (1990) ‘Estimating truck accident rate and involvements using linear and Poisson regression models’, *Transportation planning and Technology*, 15(1), pp. 41– 58.

Jovanis, P. P. and Chang, H.-L. (1986) ‘Modeling the relationship of accidents to miles traveled’, *Transportation Research Record*, 1068, pp. 42–51.

Jung, S., Jang, K., Yoon, Y. and Kang, S. (2014) ‘Contributing factors to vehicle to vehicle crash frequency and severity under rainfall’, *Journal of safety research*. Elsevier, 50, pp. 1–10.

Junhua, W., Haozhe, C. and Shi, Q. (2013) ‘Estimating freeway incident duration using accelerated failure time modeling’, *Safety science*. Elsevier, 54, pp. 43–50.

Karlaftis, M. G. and Golias, I. (2002) ‘Effects of road geometry and traffic volumes on rural roadway accident rates’, *Accident Analysis & Prevention*. Elsevier, 34(3), pp. 357–365.

Khattak, A. (1999) ‘Effect of information and other factors on multi-vehicle rear-end crashes: crash propagation and injury severity’, in *78th Annual Meeting of the Transportation Research Board, Washington, DC*.

Khorashadi, A., Niemeier, D., Shankar, V. and Mannering, F. (2005) ‘Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis’, *Accident Analysis & Prevention*, 37(5), pp. 910–921.

Kim, D.-G. and Washington, S. (2006) ‘The significance of endogeneity problems in crash models: an examination of left-turn lanes in intersection crash models’, *Accident Analysis & Prevention*, 38(6), pp. 1094–1100.

Kim, J.-K., Wang, Y. and Ulfarsson, G. F. (2007) ‘Modeling the probability of freeway rear-end crash occurrence’, *Journal of transportation engineering*. American Society of Civil Engineers, 133(1), pp. 11–19.

Kockelman, K. M. and Kweon, Y.-J. (2002) ‘Driver injury severity: an application of ordered probit models’, *Accident Analysis & Prevention*, 34(3), pp. 313–321.

Kononov, J., Bailey, B. and Allery, B. (2008) ‘Relationships between safety and both congestion and number of lanes on urban freeways’, *Transportation Research Record: Journal of the Transportation Research Board*, (2083), pp. 26–39.

Koorey, G. (2009) ‘Road data aggregation and sectioning considerations for crash analysis’, *Transportation Research Record: Journal of the Transportation Research Board*. Transportation Research Board of the National Academies, (2103), pp. 61–68.

Krull, K., Khattak, A. and Council, F. (2000) 'Injury effects of rollovers and events sequence in single-vehicle crashes', *Transportation Research Record: Journal of the Transportation Research Board*, (1717), pp. 46–54.

Kumara, S. S. P. and Chin, H. C. (2003) 'Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros', *Traffic injury prevention*, 4(1), pp. 53–57.

Kweon, Y.-J. and Kockelman, K. M. (2004) 'Spatially disaggregate panel models of crash and injury counts: the effect of speed limits and design', in *Annual meeting of the Transportation Research Board*. Citeseer.

Lao, Y., Wu, Y.-J., Corey, J. and Wang, Y. (2011) 'Modeling animal-vehicle collisions using diagonal inflated bivariate Poisson regression', *Accident Analysis & Prevention*. Elsevier, 43(1), pp. 220–227.

Lao, Y., Zhang, G., Wang, Y. and Milton, J. (2014) 'Generalized nonlinear models for rear-end crash risk analysis', *Accident Analysis & Prevention*. Elsevier, 62, pp. 9–16.

Lee, C., Abdel-Aty, M., Park, J. and Wang, J.-H. (2015) 'Development of crash modification factors for changing lane width on roadway segments using generalized nonlinear models', *Accident Analysis & Prevention*. Elsevier, 76, pp. 83–91.

Lee, C. and Li, X. (2014) 'Analysis of injury severity of drivers involved in single-and twovehicle crashes on highways in Ontario', *Accident Analysis & Prevention*. Elsevier, 71, pp. 286– 295.

Li, R. and Shang, P. (2014) 'Incident duration modeling using flexible parametric hazard-based models', *Computational intelligence and neuroscience*. Hindawi Publishing Corp., 2014, p. 33.

Li, X., Lord, D., Zhang, Y. and Xie, Y. (2008) 'Predicting motor vehicle crashes using support vector machine models', *Accident Analysis & Prevention*, 40(4), pp. 1611–1618.

Li, Z., Liu, P., Wang, W. and Xu, C. (2012) 'Using support vector machine models for crash injury severity analysis', *Accident Analysis & Prevention*, 45, pp. 478–486.

Lin, L., Wang, Q. and Sadek, A. W. (2016) 'A combined M5P tree and hazard-based duration model for predicting urban freeway traffic accident durations', *Accident Analysis & Prevention*. Elsevier, 91, pp. 114–126.

Lord, D., Guikema, S. D. and Geedipally, S. R. (2008) 'Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes', *Accident Analysis & Prevention*, 40(3), pp. 1123–1134.

Lord, D. and Mannering, F. (2010) 'The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives', *Transportation Research Part A: Policy and Practice*. Elsevier, 44(5), pp. 291–305.

Lord, D., Washington, S. and Ivan, J. N. (2007) 'Further notes on the application of zero-inflated models in highway safety', *Accident Analysis & Prevention*, 39(1), pp. 53–57.

Lord, D., Washington, S. P. and Ivan, J. N. (2005) 'Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory', *Accident Analysis & Prevention*, 37(1), pp. 35–46.

Ma, J. and Kockelman, K. (2006) 'Bayesian multivariate Poisson regression for models of injury count, by severity', *Transportation Research Record: Journal of the Transportation Research Board*, (1950), pp. 24–34.

Ma, J., Kockelman, K. M. and Damien, P. (2008) 'A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods', *Accident Analysis & Prevention*, 40(3), pp. 964–975.

Ma, X., Wu, Y.-J. and Wang, Y. (2011) 'DRIVE Net: E-science transportation platform for data sharing, visualization, modeling, and analysis', *Transportation Research Record: Journal of the Transportation Research Board*. Transportation Research Board of the National Academies, (2215), pp. 37–49.

Ma, Z., Zhao, W., Steven, I., Chien, J. and Dong, C. (2015) 'Exploring factors contributing to crash injury severity on rural two-lane highways', *Journal of safety research*. Elsevier, 55, pp. 171–176.

Maher, M. J. and Summersgill, I. (1996) 'A comprehensive methodology for the fitting of predictive accident models', *Accident Analysis & Prevention*. Elsevier, 28(3), pp. 281–296.

Malyshkina, N. and Mannering, F. (2008) 'Effect of increases in speed limits on severities of injuries in accidents', *Transportation Research Record: Journal of the Transportation Research Board*, (2083), pp. 122–127.

Malyshkina, N. V and Mannering, F. L. (2010a) 'Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents', *Accident Analysis & Prevention*. Elsevier, 42(1), pp. 131–139.

Malyshkina, N. V and Mannering, F. L. (2010b) 'Zero-state Markov switching count-data models: an empirical assessment', *Accident Analysis & Prevention*, 42(1), pp. 122–130.

Manual, H. S. (2010) 'American association of state highway and transportation officials (AASHTO)', *Washington, DC*, 10.

McCullagh, P. (1984) 'Generalized linear models', *European Journal of Operational Research*. Elsevier, 16(3), pp. 285–292.

Mergia, W. Y., Eustace, D., Chimba, D. and Qumsiyeh, M. (2013) 'Exploring factors contributing to injury severity at freeway merging and diverging locations in Ohio', *Accident Analysis & Prevention*. Elsevier, 55, pp. 202–210.

Miaou, S.-P. (1994) 'The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions', *Accident Analysis & Prevention*. Elsevier, 26(4), pp. 471–482.

Miaou, S.-P. and Lum, H. (1993) 'Modeling vehicle accidents and highway geometric design relationships', *Accident Analysis & Prevention*. Elsevier, 25(6), pp. 689–709.

Miaou, S.-P., Song, J. J. and Mallick, B. K. (2003) 'Roadway traffic crash mapping: a space-time modeling approach', *Journal of Transportation and Statistics*, 6, pp. 33–58.

Milton, J. C., Shankar, V. N. and Mannering, F. L. (2008) 'Highway accident severities and the mixed logit model: an exploratory empirical analysis', *Accident Analysis & Prevention*. Elsevier, 40(1), pp. 260–266.

Milton, J. and Mannering, F. (1998) 'The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies', *Transportation*, 25(4), pp. 395–413.

Montella, A. (2010) 'A comparative analysis of hotspot identification methods', *Accident Analysis & Prevention*. Elsevier, 42(2), pp. 571–581.

Mujalli, R. O. and de Oña, J. (2013) 'Injury severity models for motor vehicle accidents: a review'. Thomas Telford.

Nam, D. and Mannering, F. (2000) 'An exploratory hazard-based analysis of highway incident duration', *Transportation Research Part A: Policy and Practice*. Elsevier, 34(2), pp. 85–102.

Nassar, S. A., Saccomanno, F. F. and Shortreed, J. H. (1994) 'Road accident severity analysis: a micro level approach', *Canadian Journal of Civil Engineering*, 21(5), pp. 847–855.

NHTSA (2014) *State Traffic Safety Information for Year 2014*.

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

Ogle, J. H., Alluri, P. and Sarasua, W. (2011) 'MMUCC and MIRE: The role of segmentation in safety analysis', in *Proceedings of the Paper Presented at the 90th Annual Meeting of the Transportation Research Board*.

Oh, J., Washington, S. P. and Nam, D. (2006) 'Accident prediction model for railway-highway interfaces', *Accident Analysis & Prevention*. Elsevier, 38(2), pp. 346–356.

Okamoto, H. and Koshi, M. (1989) 'A method to cope with the random errors of observed accident rates in regression analysis', *Accident Analysis & Prevention*, 21(4), pp. 317–332.

de Oña, J., López, G., Mujalli, R. and Calvo, F. J. (2013) 'Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks', *Accident Analysis & Prevention*, 51, pp. 1–10.

de Oña, J., Mujalli, R. O. and Calvo, F. J. (2011) 'Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks', *Accident Analysis & Prevention*. Elsevier, 43(1), pp. 402–411.

Park, B.-J., Lord, D. and Lee, C. (2014) 'Finite mixture modeling for vehicle crash data with application to hotspot identification', *Accident Analysis & Prevention*. Elsevier, 71, pp. 319–326.

Park, E. and Lord, D. (2007) 'Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity', *Transportation Research Record: Journal of the Transportation Research Board*. Transportation Research Board of the National Academies, (2019), pp. 1–6.

Patil, S., Geedipally, S. R. and Lord, D. (2012) 'Analysis of crash severities using nested logit model—accounting for the underreporting of crashes', *Accident Analysis & Prevention*, 45, pp. 646–653.

Pereira, F. C., Rodrigues, F. and Ben-Akiva, M. (2013) 'Text analysis in incident duration prediction', *Transportation Research Part C: Emerging Technologies*. Elsevier, 37, pp. 177–192.

Poch, M. and Mannering, F. (1996) 'Negative binomial analysis of intersection-accident frequencies', *Journal of Transportation Engineering*, 122(2), pp. 105–113.

Psarros, I., Kepaptsoglou, K. and Karlaftis, M. G. (2011) 'An empirical investigation of passenger wait time perceptions using hazard-based duration models', *Journal of Public Transportation*, 14(3), p. 6.

Qin, X., Ivan, J. N. and Ravishanker, N. (2004) 'Selecting exposure measures in crash rate prediction for two-lane highway segments', *Accident Analysis & Prevention*, 36(2), pp. 183–191.

Qin, X. and Wellner, A. (2012) 'Segment length impact on highway safety screening analysis', in *Transportation Research Board 91st Annual Meeting*.

Qu, X. and Meng, Q. (2014) 'A note on hotspot identification for urban expressways', *Safety Science*. Elsevier, 66, pp. 87–91.

Resende, P. T. V and Benekohal, R. F. (1997) 'Effects of roadway section length on accident modeling', in *Traffic congestion and traffic safety in the 21st century: Challenges, innovations, and opportunities*.

Robin, P. (2014) 'Use on multinomial logistic regression in work zone crash analysis for Missouri work zones'.

Savolainen, Mannering et al. (2011)

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

Savolainen, P. and Tarko, A. (2005) 'Safety impacts at intersections on curved segments', *Transportation Research Record: Journal of the Transportation Research Board*, (1908), pp. 130–140.

Sen, B., Smith, J. D. and Najm, W. G. (2003) *Analysis of lane change crashes*.

Shankar, V., Albin, R., Milton, J. and Mannering, F. (1998) 'Evaluating median crossover likelihoods with clustered accident counts: an empirical inquiry using the random effects negative binomial model', *Transportation Research Record: Journal of the Transportation Research Board*, (1635), pp. 44–48.

Shankar, V., Mannering, F. and Barfield, W. (1995) 'Effect of roadway geometrics and environmental factors on rural freeway accident frequencies', *Accident Analysis & Prevention*. Elsevier, 27(3), pp. 371–389.

Shankar, V. N., Ulfarsson, G. F., Pendyala, R. M. and Nebergall, M. B. (2003) 'Modeling crashes involving pedestrians and motorized traffic', *Safety Science*, 41(7), pp. 627–640.

Stamatiadis, N. (2009) *Impact of shoulder width and median width on safety*. National Academies Press.

Stewart, J. (1996) 'Applications of classification and regression tree methods in roadway safety studies', *Transportation Research Record: Journal of the Transportation Research Board*, (1542), pp. 1–5.

Stokes, R. and Mutabazi, M. (1996) 'Rate-quality control method of identifying hazardous road locations', *Transportation Research Record: Journal of the Transportation Research Board*. Transportation Research Board of the National Academies, (1542), pp. 44–48.

Tay, R. and Rifaat, S. M. (2007) 'Factors contributing to the severity of intersection crashes', *Journal of Advanced Transportation*. Wiley Online Library, 41(3), pp. 245–265.

Wang, Y., Ieda, H. and Mannering, F. (2003) 'Estimating rear-end accident probabilities at signalized intersections: occurrence-mechanism approach', *Journal of Transportation engineering*. American Society of Civil Engineers, 129(4), pp. 377–384.

Wang, Y., Lao, Y., Wu, Y. and Corey, J. (2010) 'Identifying high risk locations of animalvehicle collisions on Washington state highways', *Transportation Northwest (TransNow) and Washington State Department of Transportation (WSDOT) Research Report WA-RD*, 752, pp. 2004–2010.

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

Wang, Z., Chen, H. and Lu, J. (2009) 'Exploring impacts of factors contributing to injury severity at freeway diverge areas', *Transportation Research Record: Journal of the Transportation Research Board*. Transportation Research Board of the National Academies, (2102), pp. 43–52.

Washington, S., Haque, M. M., Oh, J. and Lee, D. (2014) 'Applying quantile regression for modeling equivalent property damage only crashes to identify accident blackspots', *Accident Analysis & Prevention*. Elsevier, 66, pp. 136–146.

Wegman, F., Commandeur, J., Doveh, E., Eksler, V., Gitelman, V., Hakkert, S., Lynam, D. and Oppe, S. (2008) 'SUNflowerNext: Towards a composite road safety performance index', *Deliverable D6*, 16.

Wei, C.-H. and Lee, Y. (2007) 'Sequential forecast of incident duration using Artificial Neural Network models', *Accident Analysis & Prevention*. Elsevier, 39(5), pp. 944–954.

Weng, J., Zheng, Y., Qu, X. and Yan, X. (2015) 'Development of a maximum likelihood regression tree-based model for predicting subway incident delay', *Transportation Research Part C: Emerging Technologies*, 57, pp. 30–41.

Weng, J., Zheng, Y., Yan, X. and Meng, Q. (2014) 'Development of a subway operation incident delay model using accelerated failure time approaches', *Accident Analysis & Prevention*. Elsevier, 73, pp. 12–19. Winkelmann, R. and Zimmermann, K. F. (1995) 'Recent developments in count data modelling: theory and application', *Journal of economic surveys*. Wiley Online Library, 9(1), pp. 1–24.

Wong, S. C., Sze, N.-N. and Li, Y.-C. (2007) 'Contributory factors to traffic crashes at signalized intersections in Hong Kong', *Accident Analysis & Prevention*. Elsevier, 39(6), pp. 1107–1113.

Wu, J.-D. and Chen, T.-R. (2008) 'Development of a drowsiness warning system based on the fuzzy logic images analysis', *Expert Systems with Applications*. Elsevier, 34(2), pp. 1556–1561.

Yaacob, W. F. W., Lazim, M. A. and Wah, Y. B. (2012) 'Modeling Road Accidents using Fixed Effects Model: Conditional versus Unconditional Model', in *Proceedings of the World Congress on Engineering*.

Yan, X. and Radwan, E. (2006) 'Analyses of rear-end crashes based on classification tree models', *Traffic injury prevention*, 7(3), pp. 276–282.

Yang, X., Abdel-Aty, M., Huan, M., Peng, Y. and Gao, Z. (2015) 'An accelerated failure time model for investigating pedestrian crossing behavior and waiting times at

signalized intersections', *Accident Analysis & Prevention*. Elsevier, 82, pp. 154–162.

Zhang, H. (2010) *Identifying and quantifying factors affecting traffic crash severity in louisiana*. University of Louisiana at Lafayette.

Zhang, X., Liu, P., Chen, Y., Bai, L. and Wang, W. (2014) 'Modeling the Frequency of Opposing Left-Turn Conflicts at Signalized Intersections Using Generalized Linear Regression Models', *Traffic injury prevention*. Taylor & Francis, 15(6), pp. 645–651.