Nonparametric Methods for Road Safety Analysis

by

Lalita Thakali

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Civil Engineering

Waterloo, Ontario, Canada, 2016

AUTHOR'S DECLARATION

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Crash models for predicting long-term crash risk at some specific components of a road network are fundamental to road safety analyses such as network screening and countermeasure studies. These models are often calibrated using historical crash data from the sites of interest, aiming at capturing the underlying relationship between crash risk and various risk factors. Based on how the relationships are determined, crash models can be classified into two types: parametric or nonparametric. Parametric models represent the state of the art and practice methodology for road safety analyses. While this approach provides an easy-to-implement and easy-to-interpret tool, they come at the cost of the need for pre-selection of model forms, which, without knowing the true relation of crash and risk factors, could easily lead to misspecifications and biased estimations. In contrast, a nonparametric approach does not pre-specify a model structure but instead determines the structure from data, thereby providing greater flexibility to capture underlying complex relations. Despite this advantage of being a specification free approach, nonparametric models have not yet been accepted as part of the mainstream methodologies for road safety analyses. Little were known about their relative performance in comparison to parametric models and the practical implications of their applications for the common road safety analysis tasks such as network screening and countermeasure effectiveness estimation. Furthermore, crash data for road safety analysis and modeling are growing steadily in size and completeness with the advancement in information and sensor technologies. It is, however, unclear what implications this increased data availability has for road safety analyses in general and crash modeling in specific. Will a data-driven nonparametric technique become a more attractive alternative for addressing the complex problem of crash modeling in this era of Big Data?

In this thesis, we have introduced one of the most popular nonparametric techniques - kernel regression (KR) - as an alternative for crash modeling. One of the uniqueness of this method is that it takes a fully data-driven approach in determining the relationship between crash frequency and risk factors. Compared to other nonparametric methods, it does not contain any hidden structures to train. Therefore, when a new crash dataset is available, it can be used directly in updating crash prediction without re-calibrating the underlying models. We made two methodological contributions to facilitate the application of a nonparametric model for road safety analyses. We first extended the KR method, similar to Empirical Bayesian (EB) method using parametric models, to account for the site-specific crash history in predicting risk. We then developed a bootstrap-based algorithm for identifying the important variables to be included in a nonparametric model.

The research also made significant knowledge contributions to the practice field related to applications of nonparametric models for road safety analyses. First, we benchmarked the crash prediction performance of the KR model against the mainstream model – Negative Binomial (NB) model. Using three large crash datasets, we investigated the performance of the KR and NB models as a function of the amount of training data. Through a rigorous bootstrapping validation process, we found that the two approaches exhibit strikingly different patterns, especially in terms of sensitivity to data size. While the performance of the KR method improved significantly with increase in data size, the NB model showed less sensitivity. Meanwhile, the KR method outperformed the NB model in terms of predictive performance, and that performance advantage increased noticeably by data size. Secondly, we compared the two approaches in their ability to capture the underlying complex relationships between crash frequency and predicting variables. The KR method was shown to yield more sensible results on the effects of various risk factors in both case studies as compared to the NB model.

Our other main contribution comes from the investigation on the practical implications of applying the KR models for two critical road safety analyses tasks – network screening and countermeasure study. Both KR method and NB model were employed in a case study under the two popular network screening frameworks, i.e., regression-based and EB-based. Their performances were compared in terms of site ranking and identification of crash hotspots. The two approaches were found to yield more similar rankings when applied in the EB-based framework, irrespective of the ranking measures (i.e., crash frequency or crash rate), than in the regression-based framework. Similar comparative results were obtained in locating the crash hotspots. Likewise, for countermeasure studies, the two popular approaches – the before-after EB study and the cross-sectional study – were considered in case studies using both KR and NB crash prediction models. As expected, the two different crash modeling techniques showed significant differences in their estimates on crash modification factors (CMF). Different from the NB model based approach, the KR-based method was able to capture the sensitivity of CMFs to traffic levels as well as combine the effect of multiple countermeasures without requiring any assumptions on the interaction between the countermeasures.

Acknowledgements

Foremost, I would like to express my deepest gratitude to my advisor, Dr. Liping Fu, who has been extremely supportive throughout my doctoral research. His great mentorship and immense passion towards research will always make me remember of him as a great professor. I could not have wished for a better mentor to accomplish this milestone. I am also very thankful to my co-advisor, Dr. Tao Chen, specifically for introducing me to the field of nonparametric statistics and providing invaluable guidance throughout the research.

I also would like to express my sincere appreciations to the members of my PhD examination committee: Dr. Changbao Wu, Dr. Frank F. Saccomanno, Dr. Mahesh Pandey and Dr. Bhagwant Persaud for their time and constructive comments and suggestions.

I want to thank Max Perchanock, Zoe Lam and Michael Pardo from the Ministry of Transportation Ontario for their help in providing some of the data used in this research. I would also like to acknowledge Dr. Taimur Usman and Shahram Heydari for sharing with me some of their datasets used in this thesis. I am also thankful to Matthew L. Muresan for his valuable time in proofreading my thesis. Also thanks to all my friends at iTSS lab, University of Waterloo and elsewhere for their great friendships. My special thanks goes out to my friend, Dr. Ambika Karkee, for always being there to talk to me at the time of some unproductive moments.

I am indebted to my dearest parents, Narayan P. Thakali and Santa K. Thakali, for their endless love, support and encouragement, which have meant a lot for completing my study**.** The value of honesty and compassion they have taught me has always helped me to live my life happily. My siblings, brother-in-law, sister-in-law, two grown up kids of my elder sister, uncles and aunts too deserve many thanks for their enormous support throughout my student life. I am also very thankful to all my in-laws (grandmother-in-law, parents-in-law, sister-in-law and brothers-in-law) for their support and blessings. Last but not least, my husband, Sawal Thakali, deserves a Big Thanks for his exceptional positive personality and greatest love, which meant a lot towards a successful completion of my thesis.

Dedication

To my parents Narayan Prasad Thakali and Santa Kumari Thakali

To my husband, Sawal Thakali

बुवा-आमा, नारायण प्रसाद थकाली र सान्ता कुमारी थकाली

श्रीमान्, सावल थकाली

Table of Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

AADT	Annual Average Daily Traffic
AASHTO	American Association of State Highway and Transportation Officials
AIS	Accident Information System
ANN	Artificial Neural Network
BV	Bootstrapping Validation
CMF	Crash Modification Factor
EB	Empirical Bayesian
FB	Full Bayesian
FWHA	Federal Highway Administration
GCIP	Grade Crossing Improvement Program
GIS	Geographical Information System
GIS	Geographical Information System
GNB	Generalized negative binomial
GNP	Gross National Product
HIMS	Highway Inventory Management System
HSM	Highway Safety Manual
IHSDM	Interactive Highway Safety Design Model
IRSI	Integrated Railway Information System
KR	Kernel Regression
KDE	Kernel Density estimate
KR-EB	Kernel Regression based Empirical Bayesian
LHRS	Linear Highway Referencing System
MAE	Mean Absolute Error
MARS	Multivariate Adaptive Regression Splines
MLE	Maximum Likelihood Estimation
MTO	Ministry of Transportation Ontario
MVK	Million Vehicle Kilometers
NB	Negative Binomial
NB-EB	Negative Binomial Model based Empirical Bayesian
PL	Poisson Lognormal
RMSE	Root Mean Square Error

RODS	Railway Occurrence Database System
RSA	Road Safety Analysis
RSI	Road Surface Index
RTM	Regression-to-Mean
SPF	Safety Performance Function
TVIS	Traffic Volume Inventory System
US	United States
VI	Variable Importance
VI’	Relative Variable Importance
VKT	Vehicle Kilometer Traveled
WHO	World Health Organization
ZINB	Zero-inflated Negative Binomial
ZIP	Zero-inflated Poisson

Chapter 1
Introduction

1.1 Background

Modern society runs on road transportation mainly due to the flexibility and convenience provided by affordable roads that move people and goods on a large scale. For this reason, governments spend a huge amount of resources on constructing and maintaining extensive road networks. However, the net result of the development of road network is like a double-edged sword. Extensive road transportation encourages individuals to own vehicles, allowing them to move and work farther away than what would have been possible without the roads. On the other hand, the increase in motorization rises traffic interaction among the road users, thereby causing serious road safety problems. In addition, many other factors, such as poor road design, adverse environmental conditions, human errors and vehicle defects could trigger road safety problems, leading to an increase in crashes related to property damages, injuries and fatalities. These effects, in turn, cause travel time delays that have substantial direct economic and social costs. Furthermore, travel time delays themselves create several additional indirect costs, such as an increase in fuel consumption, increase in air and noise pollution, and additional health treatment costs associated with the pollution. The fact that road transportation incurs lower infrastructure and maintenance costs compared to other modes of transportation (e.g., airway, railway) could be offset by all the significant economic and social losses incurred by road safety problems.

From a global perspective, every year, road crashes result in a large number of deaths and extensive property damage. The World Health Organization (WHO) has identified traffic crashes as one of the most critical public health issues around the world. According to the WHO's global status report on road safety, more than 1.2 million people die every year and as many as 50 million people suffer non-fatal injuries because of road crashes (WHO, 2015). Meanwhile, the majority of the people involved in traffic crashes are the economically active population. This study also shows that traffic crashes are the ninth leading cause of death, and they are projected to be the seventh in the year 2030 with an estimated annual fatality of 2.4 million people. In addition to the social costs due to deaths, there is a significant economic burden imposed due to the property damages and injuries. The estimated cost is 1% of the gross national product (GNP) for low-income countries, 1.5% for middle-income countries and 2% for high-income countries (Jacobs et al., 2000).

Similarly, other indirect costs of traffic crashes, such as traffic congestion and air/noise pollution caused by traffic crashes, are difficult to quantify in monetary terms but are likely to be substantially high. According to Garrison and Mannering (1990), each minute of traffic congestion resulting from crashes was associated with an equivalent loss of over 2000 dollars. Based on this rate, the estimated annual crash delay cost in the City of Seattle, U.S., alone was over 250 million dollars. Similarly, the cost mentioned above was estimated to be 501.9 million dollars for highways in Ontario, Canada (Vodden et al., 2007).

The sheer magnitude of road crash consequences has resulted in an increasing public demand for safer roads. Road agencies around the world have been expending significant resources on various programs to counteract road safety problems. The root cause of these problems mainly lies in the interactions of the four main components of the system: road users, roadways, environment and vehicles (HSM, 2010). Human factors include driver's characteristics such as age, judgment capacity, driving skills and experience, and physical state (e.g., fatigue, alcohol or drug usage level). Similarly, roadway and environmental factors include geometric alignment, cross-section elements, traffic control devices, weather factors and road surface conditions. Meanwhile, vehicle conditions, including the capacity to brake and steer smoothly, are equally important. The impacts of all these factors can be proactively reduced by implementing various road safety improvement programs that involve applying proper engineering treatments, educating road users, enforcing traffic laws, improving emergency response services, and improving vehicle safety technologies.

1.2 Road Safety Analysis

Prior to launching any road safety improvement programs, a systematic road safety analysis is necessary to investigate safety-related issues. This includes identification of crash hotspots by systematically screening a list of candidate locations (e.g., roadway segments or intersections) with high-risk levels. This is critical especially when the resources available to implement safety treatments on selected locations are limited. Therefore, network screening has been a standard procedure for launching cost-effective safety programs. Similarly, countermeasure studies are another important task which involve quantifying the effects of specific road safety treatments, such as signalizing intersections, converting a two-lane to a multiple-lane road and adding a median to an undivided road section. Both of the components of a systematic road safety analysis involve a detailed exploration of historical crash data and require appropriate modeling techniques to quantify risk levels from the given data. The following sections provide a brief discussion of these components.

1.2.1 Network Screening

The process of network screening involves ranking the sites of interest by a specific ranking measure related to crash risk level. For example, sites could be ranked on the basis of crash rate (crashes per vehicle-kilometers or per entering vehicles), crash frequency (crashes per km-year or crashes per year) or weighted crash severities (Laughland et al., 1975; Deacon et al., 1975; Mcguigan, 1981; Mcguigan, 1982; Stokes and Mutabazi, 1996; HSM, 2010). Sites could also be ranked by the probability that the crash frequency exceeds what is normal to reflect the potential benefit from applying safety treatments. The Highway Safety Manual (HSM) has listed 13 different ranking measures, suggesting that a wide choice of measures for ranking can be adopted for network screening. Broadly, the approach of determining these measures can be categorized into two groups. The first is the direct method where the risk level associated with each unit (section or intersection) can be measured by direct counting of observed crash frequency (or rate). The second is the regression-based approach, where risk levels are estimated in terms of expected long-term effects of given conditions by using some crash models.

In the past, when the use of regression-based approach was not common, transportation agencies frequently applied the direct method. The ranking measures in this approach are determined mainly based on the arithmetic means of the observed historical crash data. This method is very simple and easy to apply; however, there are few limitations from a statistical point of view. First, it lacks a probabilistic approach for determining the ranking measure, thereby ignoring the inherited randomness of crash occurrence. Moreover, it represents a short-term measure derived simply from the observed crashes and may not represent a reliable estimate of long-term safety effects. Such bias in the measure of safety effects is known as regression-to-mean (RTM) bias effect (HSM, 2010; Hauer, 1997). Furthermore, this approach cannot take into account site-specific factors, such as road geometric design features, weather conditions, traffic level and other factors, which may be useful indicators for measuring crash risk. A failure to account for all these issues may lead to a selection of a biased list of crash hotspots, and consequently, launching a safety program may result in a huge waste of resources.

Recently, the regression-based approach has been quite popular as it addresses the RTM problem and considers the effects of external factors causing crash risks by modeling crashes under a parametric framework. An extension of the regression-based approach, known as the Empirical Bayesian (EB) method, has been the state-of-art methodology for network screening and other road safety studies. This method provides a framework to combine estimates from a crash model and the site-specific crash history through some weighting schemes (Hauer, 1997). As a result, the crash model remains the most critical element of the EB approach (Hauer, 1997; Miranda-Moreno et al., 2005; Montella, 2010; HSM, 2010; Zou et al., 2013).

Statistically, there are two approaches to modeling crashes, namely, parametric and nonparametric approaches. To the best of our knowledge, all past network screening studies depended on the former technique for estimating the long-term effects of crash risk.

1.2.2 Countermeasure study

A countermeasure study involves evaluating the safety effects of one or more treatments, such as changing intersection's control type, adding a rumble stripe along the edge of a paved road section, and adding a median to an undivided road section. The effectiveness is commonly measured by a crash modification factor (CMF), which is obtained from a countermeasure study. CMFs can be used to select the best treatment option in terms of reducing crash risk in the identified hotspot sections. Our focus is on the methodological part of how CMFs can be obtained.

The CMF related to a treatment is determined by comparing the safety levels before and after the treatment conditions. The two main approaches are the before-after study and the cross-sectional study (Benekohal and Hashmi, 1992; Hauer, 1997; Gross et al, 2010; HSM, 2010). Before-after studies are further categorized into simple before-after, comparison before-after and EB-based before-after. Among these, the latter approach based on the EB method is the most popular as it reduces bias of RTM effects (Council and Stewart, 1999, Persaud et al., 2001; Srinivasan and Kockelman, 2002; Miaou and Song, 2005; Harkey et al., 2008; HSM, 2010; Li et al., 2008).

While the before-after EB-based study is the state-of-art approach in the study of countermeasures, it should be noted that some treatments may present a data restriction problem. This could include some extremely rare cases where treatments are applied to collect enough crash data. For example, when determining the safety effectiveness of widening shoulder or median widths, it is less practical for on-site modifications of these features to be made to allow for the collection of their before-after crash data. Meanwhile, such treatments have a wide range of possible design options to consider for determining their corresponding CMFs. These studies are common mainly in the context of determining the CMFs of roadway characteristics that extend along the road sections, such as altering shoulder, lane and median widths, and treating road shoulders with rumble strips (Lord & Bonneson, 2007; Fitzpatrick et al., 2008; Stamatiadis et al., 2009; Zeng & Schrock, 2013; Park et al., 2014; Park & Abdel-Aty, 2015). In such cases, a cross-sectional study is recommended whereby the data from similar sites are analysed using a crash model in a framework of with and without the treatment conditions (Gross et al., 2010). While the decisions on which study approach (i.e., before-after EB-based or the cross-sectional) to consider could be contextual, the improvement of crash models involved in both techniques remains the most critical issue.

As mentioned in the previous section, two approaches are usually used to modeling crashes: parametric and nonparametric. Among these categories, the parametric approach has been the mainstream technique to determine the CMFs of safety treatment measures. There have been very limited applications of the nonparametric approach for such studies (e.g., Park & Abdel-Aty, 2015).

1.3 Issues with a Parametric Approach

As previously discussed, crash models are required by the two most important components of road safety analyses. This significance of crash models has stimulated significant past efforts which have led to the development of a large number of statistical models, such as Poisson (Jovanis and Chang, 1986; Miao and Lum, 1993), Negative Binomial (NB) (Miaou, 1994; Persaud, 1994; Shankar et al., 1995; Council and Stewart, 1999), Poisson-Lognormal (PL) (Aguero-Valverde and Jovanis, 2008; Usman et al., 2012), Zero-inflated Poisson (ZIP) and Zero-inflated Negative Binomial (ZINB) (Lambert, 1992; Washington et al., 2003). These models are all parametric, posessing the following limitations: presumption of a specific probability distribution for crash data, and pre-specification of a functional form for the relationship between the expected crash frequency and the predicting variables.

For the probability distribution of crash occurrence, various distributions have been assumed in the crash models. For example, the Poisson model assumes that the frequency by which crashes occur follow a Poisson distribution where the mean and variance of the distribution are equal. However, crash data are often found to be over-dispersed, thereby resulting the variance to be greater than the mean (Miaou et al., 1993; Miaou and Lum, 1993; Shankar et al., 1997; Lord and Miranda-Moreno, 2008).

Therefore, a number of parametric models have been developed to deal with this limitation of Poisson model. Some of the examples include NB, PL, Generalized negative binomial (GNB) and Zero-inflated NB models. However, these models suffer from various issues. For example, NB model considers a constant dispersion parameter which may not reflect the actual heterogeneity condition in the crash data. Generalized negative binomial model has been developed to address the limitations of the NB model by specifying the over-dispersion parameter as a function of a set of covariates; however, this approach again has a problem of requiring an assumption on such specification. Similarly, Zero-inflated count models (ZIP and ZINB) have also been developed based on the assumption of the existence of the dual states, namely, safe and unsafe state. Although this particular form of models may increase the goodness-of-fit, they do not reflect the real data generating process due to the unrealistic assumption of absolute safe conditions in the road network.

The next common assumption in all parametric models is the specification of their model mean structures, i.e., the relationship between crash frequency and its predicting variables. This relationship is represented by an equation comprising a set of variables and its associated coefficients. The most common choice for the function (equation) that models the relationship between the expected crash frequency and various factors that affect the occurrence of crashes is an exponential function. While the model in a single equation form may be relatively easy to interpret and apply, the need for prior specification may limit its flexibility to improve estimation accuracy. That is, the functional form imposes a certain shape restriction without providing the full flexibility needed to reflect the actual crash data characteristics. Moreover, these specified functional forms are able to capture only the monotonic relation between crashes and the predicting variables. In other words, these relations represent either only increasing or decreasing trends without having enough flexibility to capture composite trends across the full range of values that variables could take. Such common practice of pre-specifying a functional form without any supporting theory may lead to erroneous and biased inferences. Meanwhile, the model coefficients associated with each predicting variables are estimated globally (e.g., maximum likelihood estimation method) using a given crash dataset. In the presence of outliers or some extreme cases, estimated model coefficients from such a global perspective can easily influence their magnitudes.

Once the model mean structure is defined, the most commonly used technique to estimate the model parameters (i.e., coefficients associated with predicting variables) is by using the maximum likelihood estimation (MLE) technique. In this technique, the model parameters are estimated by maximizing the probability for obtaining the observed crash data under a given distribution (e.g., Poisson, NB and others). Recently, the use of Bayesian techniques has also become quite popular. In this technique, the estimation of model parameters is improved through the use of prior information for the parameters of interest. However, choosing the right prior information could be as challenging as selecting the right functional forms. It is also noted that the Bayesian approach performs comparatively better than the MLE approach when the sample size and crash frequency is low (Lord and Miranda-Moreno, 2008). However, it is anticipated that the volume of crash-related data collected from the field will grow significantly due to the advancement in traffic-related technologies, thus providing the benefit of larger data size. This means that these estimation techniques for crash modeling would yield similar results when data become large enough.

Parametric models also lack the power to identify the interaction effect of multiple variables. For example, Shankar et al. (1995) explored the interaction effect of weather and geometric factors using the NB model with some assumptions about their interaction terms (e.g., snowfall-grade and snowfall-curve interactions). The problem with such an approach once again lies in pre-specifying the form of interaction with little basis. Because of this challenge to identify the interaction effects of multiple factors, the current version of the Highway Safety Manual (HSM) determines the joint CMF simply by multiplying the CMFs of individual countermeasures. The underlying assumption of this action is the strong assumption that their effects are independent from each other.

1.4 Potential of Data-driven Nonparametric Approach

The nonparametric approach is different from the parametric approach because it does not require specification of model functional form, especially in an equation structure, for the relation between dependent and independent variables. Therefore, the estimation is purely data-driven and is expected to be less biased as this approach avoids the misspecification issues of parametric models. Hauer (2015) also mentions, “Even when masterfully executed, the parametric fit will suffer from all the shortcoming of nonparametric one.”

Despite its advantages over parametric models, the data-driven nonparametric approach has not been accepted as a mainstream alternative due to some commonly cited challenges. The first challenge is that a nonparametric analysis is data hungry - it requires much larger sample sizes than a parametric method due to its lower convergence rate. However, recent advances in information and sensor technologies has increased the availability and completeness of crash data reducing the significance of this issue, especially in the context of future applications. Another commonly cited issue of a nonparametric method is the difficulty in direct interpretation of how each variable influences crash risk. However, such interpretations are not always necessary, especially when applied in network screening and countermeasure study. Furthermore, if required, we can easily generate the effect of each variable in a graphical form (Thakali et al., 2014).

In the past, a few studies have investigated the application of nonparametric methods, including artificial neural network (ANN), classification and regression tree (CART), multivariate adaptive regression splines (MARS). Karlaftis and Golias (2002) employed CART to explore the effects of rural road geometry and traffic volumes on crash rates. Similarly, Chang (2005) and Xie et al. (2007) applied ANN to model crash frequency based on highway geometric variables, traffic characteristics, and environmental factors. Likewise, Abdel-Aty & Haleem (2011) and Park et al. (2014) employed MARS in their road safety studies. However, these efforts are mostly limited to the effort of modeling crashes. Furthermore, these methods are often characterized as "Black Box" approach due to the involvement of some complex hidden model structures in their modeling frameworks, which also raises difficulty in interpreting their underlying relations between dependent and independent variables.

In this thesis, we propose alternative nonparametric methods to crash modeling that are fully data-driven. Apart from the motivation of reducing specification problems of traditionally used parametric models, this thesis also intends to explore some of the research gaps in implementing a nonparametric approach which have not been studied extensively in the past, as summarized in the following section:

- Most of the previous nonparametric methods (e.g., ANN, MARS, and CART) applied in road safety studies are relatively complex and require extensive effort for training due to the involvement of hidden model structures. There is a need for alternative nonparametric methods, especially with a full data-driven feature and relatively fewer model parameters or hidden structures.

- Compared to the parametric approach, the nonparametric approach is characterized as a data-hungry technique. It is believed that, with advancement in information technologies, crash data for road safety modeling will grow steadily in size. However, it is little known on the practical implication of data size on crash modeling and road safety analysis. It is also of interest to investigate how the relative performance of these two different approaches differs with growing data size.

- Nonparametric methods typically lack a variable selection process. Addressing this issue may not be as simple as in a parametric method where the significance of a variable can be easily tested statistically. This is another important issue that needs to be addressed for a newly introduced nonparametric method.

- The EB approach provides a framework to determine a long-term crash risk of a site by combining two different sources of evidence: site-specific observed crashes and the expected crash frequency. It has been one of the most popular and extensively used method by road agencies. For example, the Interactive Highway Safety Design Model (IHSDM) developed by the Federal Highway Administration (FHWA), US, and the SafetyAnalyst tool developed by the American Association of State Highway and Transportation Officials (AASTO) are both based on the EB approach. However, one of the issues with this approach is that it depends on a parametric crash model for estimating the "expected crash frequency". As previously discussed, the parametric models have specification problems, an issue that could be reduced by applying a data-driven nonparametric method; however, there is a need of a methodology to incorporate this alternative method within the popular EB framework.

- While some of the past studies have demonstrated the use of a few nonparametric methods, their efforts have been mostly limited to crash modeling. Without their applications in road safety analyses such as identification of crash hotspots and countermeasure studies, their significance may not be fully recognized.

1.5 Research Objectives

As discussed in the previous section, parametric models, the commonly applied methods for road safety analyses, have some issues mainly due to the need for model specifications. The primary goal of this research is to investigate the application of a nonparametric approach for its potential to address some of the limitations possessed by parametric approach. The particular objectives of this thesis are as follows:

1. Apply a nonparametric approach for modeling traffic crashes and investigate its applications and features for road safety analysis.
2. Perform a comparative study of parametric and nonparametric approaches from both theoretical and practical points of view.
3. Develop a framework to combine site-specific safety records and expected crash risk from a nonparametric model, similar to the EB framework using parametric models.
4. Develop a framework for the application of nonparametric methods (objective 1 and 3) to road safety analysis- network screening and countermeasure studies, including a few relevant case studies for each type.

1.6 Overview of Chapters

This thesis is organized into seven chapters. Chapter 1 provides an overview of the road safety problems including some of the research gaps in analysing the safety problems and the study objectives. Chapter 2 presents a brief literature review on various components of road safety analysis and their relation to the road safety modeling techniques. It also includes a comprehensive review of parametric and nonparametric methods that are common in the past road safety studies. Chapter 3 focuses on a proposed study methodology, describing the proposed crash estimation methods-both parametric and nonparametric approaches and the use of these models in an Empirical Baye's (EB) framework. Meanwhile, an algorithm is introduced for the nonparametric method to identify a list of relevant variables for the modeling purpose. Chapter 4 presents a comprehensive comparative study of parametric and nonparametric approaches for modeling crashes. In addition, this chapter also demonstrates the application of variable selection algorithm for the nonparametric method proposed in this thesis. Chapter 5 and 6 present the applications of proposed crash estimation methods in network screening and countermeasures studies, respectively. Finally, Chapter 7 highlights the main contribution of thesis and makes some suggestions for future research.

Chapter 2
Literature Review

Road safety studies involve analyzing various crash-related issues, identifying high crash risk sites, selecting effective countermeasures, and evaluating safety effects of treatment measures after their implementations. All these studies require crash models to estimate the expected crash risk of study sites. In road safety literature, parametric models have been proposed as dominant means for estimating crash risk as supported by a large body of literature and applications. However, these models possess various assumptions and specification issues which will be critically assessed in this chapter.

This chapter has three main parts. The first part, Section 2.1 to 2.3, provides a brief description of road safety analysis procedures with focus on its two main components, namely, network screening and countermeasure study. The second part, Section 2.4, discusses the concept of parametric approach and presents some of the commonly adopted models in road safety studies. Finally, the third part, Section 2.5, presents a brief review of past efforts on modeling crashes in a nonparametric approach.

2.1 Road Safety Analyses

Road traffic system consists of four basic components: road network, road users, vehicles and environment. Any adverse conditions in these four components, such as poor road designs, human errors, vehicle defects or adverse environmental conditions increase the likelihood of vehicle crashes. Traffic interactions between road users including other components of the road system also contribute to the safety problems, and these effects are expected to grow continuously as travelers increasingly depend on road transport. To counteract these increasing road safety problems, a comprehensive safety improvement programme plays a crucial role.

A safety improvement programme often consists of one or more of the five main safety strategies, including engineering, education about road safety, enforcement, improvement of emergency response service, and advancement of vehicle safety technologies (HSM, 2010). However, before launching a safety program, road agencies need to perform a systematic analysis to identify safety issues, quantify risk level, and identify suitable treatment measures. The Highway Safety Manual (HSM) provides six interrelated analytical steps in a framework of road safety management process that consists of network screening, diagnosis, countermeasure selection, economic appraisal, projects prioritization and countermeasure study. In our following review, we focus on the two main components, i.e., network screening and countermeasure study.

2.2 Network Screening

Networking screening is a systematic process of ranking sites that suffer from unacceptably high levels of crash risk. It provides a low-cost strategy in road safety management where a small group of sites is selected from a large population so that the available resources can be effectively deployed to relatively risk-prone areas thereby increasing the overall safety of the road network.

2.2.1 Network Screening Process

Figure 2-1 presents a framework for network screening as detailed in the HSM (2010). A brief discussion of each step is given below.

Establish focus: The first step in network screening is to establish the study focus which could be either to identify a list of hotspots in a network for safety improvement (applicable to this thesis) or to evaluate the network in terms of safety performance for formulating some specific policies.

Identify sites and establish reference population: This involves identification of a set of sites or facilities for screening. Normally, the facilities with similar characteristics are grouped together; for example, highway road sections and city roads are considered differently. Similarly, road sections and intersections are studied separately. This is important as the crash related data and the processing steps might vary depending on the nature of the study group.

Select performance (ranking)measures: Performance measures, also referred as the ranking measures, are used to gauge the relative risk levels of the study sites. Therefore, the methods used to estimate these risk measures are crucial as their accuracies vary accordingly. The HSM (2010) has identified 13 potential measures including crash frequency, crash rate and others. These measures can be determined either using crash counts directly or by employing crash models. The latter approach is, however, preferred as the crash models help to reduce the regression-to-mean (RTM) problem of the former approach. One of the objectives of this thesis is to apply alternative crash modeling techniques to improve the estimates of risk measures. We will provide a brief review on past practices in the next section.

Screen and evaluate result: Finally, the study sites are ranked based on the magnitude of estimated risk measure and a list of top high-risk sites, also known as crash hotspots, are selected for a further detailed investigation so that suitable countermeasures could be recommended for reducing their safety problems.

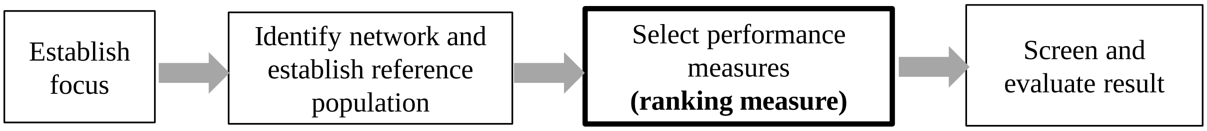

Establish focus
Identify network and establish reference population
Select performance measures
(ranking measure)
Screen and evaluate result

Figure 2-1: Framework for network screening (HSM, 2010)

2.2.2 Methods for Estimating Performance Measures

As mentioned in the earlier discussion, risk measure plays a key role in network screening. In the past, when the statistical techniques were not widely applied, road agencies simply used observed crash frequency (or rate) as the risk measure. However, this conventional approach does not account for the uncertainty in crash occurrence and thus suffers from the RTM effect. Recently, the regression-based and Empirical Bayesian (EB) approaches have been the popular techniques employed to estimate the risk measures needed in network screening as these address the RTM problem and consider the effects of external factors through the effort of modeling under a parametric framework.

A number of studies that involve in comparing the performance of these mentioned approaches are found in literature. For example, Cheng and Washington (2005) evaluated the performance of conventional approach of using simple crash count and the Empirical Baye's (EB) based approach in estimating risk measures. For the conventional approach, two measures were used. The first was the observed crash frequency where a set of sites was ranked in a descending order and the top most sites were selected as hotspots. The second was establishing a threshold value and comparing it with the observed crash counts. In the latter, the threshold value was calculated as a summation of the average observed crashes and the confidence interval. When the observed crashes exceeded the threshold value, then the sites were classified as hotspots. Similarly, for the EB-based approach, the risk measure was an EB estimated crash frequency. This study showed that the EB-based approach significantly outperformed the conventional method in identifying hotspots. The study further concluded that the importance of an EB-based approach is especially critical when there are high heterogeneities in crash data. Similar conclusions were also drawn in a study by Elvik (2008).

Recently, the EB-based and regression-based approaches have been the most extensively used techniques for network screening. The commonly used risk measure, i.e., expected crash frequency and crash rate, are obtained using parametric crash models such as Poisson and NB models calibrated from the maximum likelihood estimation (MLE) technique (Saccamanno et al., 2001; Greibe, 2003; Saccamanno et al., 2004; Mirinda-Moreno, 2005; Geedipally and Lord, 2010). Meanwhile, these measures can also be obtained from an EB approach where the NB model is extended in a framework of Bayesian approach (Higle & Witkowski 1988; Hauer, 1996; Montella, 2010; Persaud et al., 1999; AASHTO, 2010). Mathematically, the EB estimates are a combination of estimates from a crash model and site-specific observed crashes. Therefore, this approach is not appropriate in the absence of site-specific historical crash data (HSM, 2010). Other advanced forms of parametric models exist such as the full Bayesian approach (Miaou and Song, 2005; Miranda-Moreno et al., 2005; Miranda-Moreno et al., 2007; Huang et al., 2009; Miranda-Moreno et al., 2013; Wang et al., 2014). However, it is shown that in the case of a relatively large dataset and sample mean, the use of full Bayesian approach does not significantly contribute to the improvement of estimation results of the traditional MLE approach (Lord & Miranda-Moreno, 2008; Miranda-Moreno et al., 2013).

Some studies have compared regression-based and EB-based estimation techniques for network screening. For example, Saccamanno et al. (2001) applied Poisson model and EB method for identifying hotspots in a two-lane highway in Italy using crash frequency as the risk measure. They concluded that the numbers of hotspots identified by the EB estimate were less than that of the Poisson model. Furthermore, the authors mentioned that the results from the Poisson model may have been biased due to its inability to account for over-dispersion in crash data. Comparatively, the EB method has two main advantages. First, it includes NB model, thereby taking account of over-dispersion in the data structure, which would not be possible using a Poisson model. Second, the precision of crash estimation is improved by considering site-specific crash history under a Bayesian framework. Similarly, in another study by Miranda-Moreno et al. (2005), a significant difference was observed between the EB and regression-based approaches used in ranking of highway-railway grade crossings, thus underscoring the importance of method selection.

Similarly, Huang et al. (2009) compared EB and full Bayesian approach using NB and Poisson-lognormal model structure to identify crash hotspots of signalized intersections. First, the sites were ranked based on the average crash frequency estimated from individual methods using five years of crash data. Among them, a specific number of sites (e.g., 5%, 10% of total sites) were considered as hotspots. These hotspots were then compared with the "true" hotspots obtained from crash counts using ten years of crash data (1997-2006). These hotspots identified using observed crashes directly were considered as the true hotspots by following a logic that the crash data collected with a relatively longer duration is expected to capture both randomness in crashes and actual risk level of study sites. The study concluded that the full Bayesian approach showed better performance in identifying the actual hotspots.

In a nutshell, all these previous network screening studies employed parametric models to fulfill their need for crash modeling. One of the least explored approaches in this decision-making process is the use of the data-driven nonparametric approach as an alternative technique. The fact that this approach is specification free may provide a significant advantage in improving the accuracy of risk measure (e.g., crash frequency or crash rate) and eventually in the identification of crash hotspots (Persaud et al., 1999).

2.3 Countermeasure Study

Determining the effectiveness of countermeasures is crucial as this allows road agencies to conduct cost-benefit analysis such that the most cost effective treatment measures can be selected. In general, preference is given to the countermeasure with high safety benefits unless there is a significant cost associated to it.

2.3.1 Crash Modification Factor (CMF)

Typically, the effectiveness of a countermeasure is represented by a measure called the crash modification factor (CMF), which is defined on the basis of the safety status of two different conditions (illustrated in Figure 2.2). Mathematically, the CMF can be calculated as follows:

$$CMF = \frac{C_a}{C_b} \qquad (2\text{-}1)$$

where,

C_a= expected crash frequency for condition "a" i.e., after or with the treatment.

C_b = expected crash frequency for condition "b" i.e., before or without the treatment.

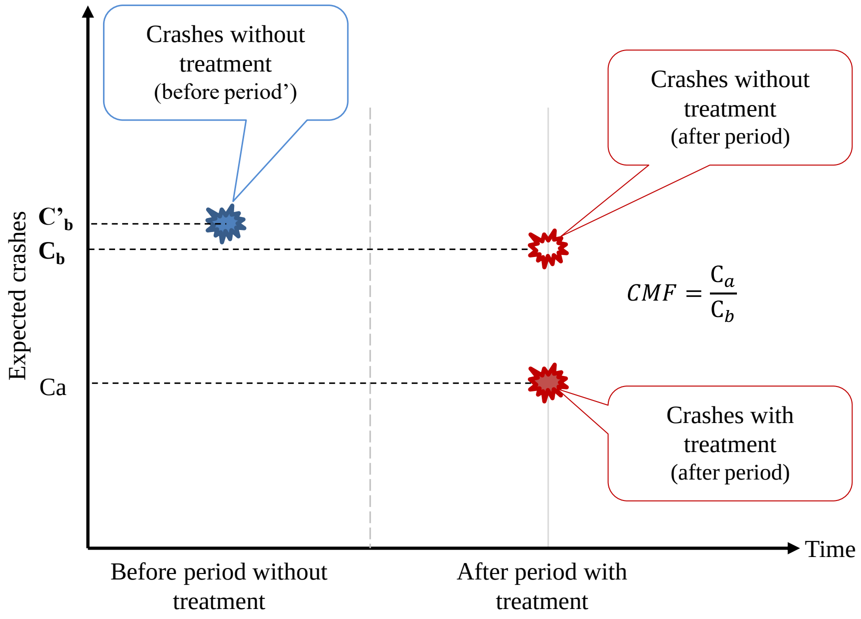

Crashes without treatment (before period')
Crashes without treatment (after period)
Crashes with treatment (after period)
Expected crashes
C'b
Cb
Ca
$CMF = \frac{C_a}{C_b}$
Time
Before period without treatment
After period with treatment

Figure 2-2: Determining CMF in a before-after or with-without study framework

CMFs appear as the multiplicative factors when computing the crash risk of implementing alternative treatments and/or designs in a given roadway section. For example, when a treatment has a CMF of 0.6 and the expected crashes without the treatment is 2 crashes per year, then expected crashes after the treatment becomes 1.2 crashes per year (i.e., CMF×2= 0.6×2). Most importantly, the magnitudes of CMFs can be used to interpret safety effectiveness of implementing the specific treatments. A CMF value below one indicates a reduction in expected crash frequency and vice versa for the value greater than one as compared to the before treatment condition. This factor could also be indirectly interpreted in terms of percentage decrease or increase in expected crash frequency. For the same example here, the treatment with CMF of 0.6 indicates that by implementing this countermeasure, the crash frequency is expected to reduce by 40 percent (i.e., (1-CMF) ×100 = 40%). Therefore, transportation planners and designers' interest lies in the countermeasures that have lower CMF values.

There are two popular approaches for determining the CMFs: before-after study and cross-sectional study (Benekohal and Hashmi, 1992; Hauer, 1997; Persaud et al., 1999; Harwood et al., 2002; Gross et al., 2010). The CMF measure in Equation (2-1) is either a single value or a functional form depending on the approach. A before-after study results in a single CMF value, whereas a cross-sectional study specifically using a crash model (parametric model) results CMF in an equation form. Therefore, often the CMFs from the latter approach are named as crash modification functions (Gross et al., 2010). An extensive list of CMFs obtained from both the approaches are documented in the HSM manual and the FHWA Clearinghouse web application (HSM, 2010; FHWA, 2015). The following section provides a brief review of each approach.

2.3.2 CMF: Before-After Study

The before-after study is commonly used approach to evaluate the safety effects of traffic controlling measures, such as adding left and/or right-turn lanes at intersections, converting an intersection to a roundabout (Hauer & Persaud, 1987; Harwood et al., 2002; Persaud et al., 1999; Persaud et al., 2001). It involves a direct comparison of site-specific risk levels of before and after the treatment conditions; therefore, in the case of enough observed before-after crash data, this approach is highly recommended (Gross et al., 2010). This approach is further categorized into three types: simple before-after, comparison before-after and EB-based before-after study.

In a simple before-after study, before-crash risk (C_b) is estimated using either the previous year's crash records or an arithmetic mean of crashes occurring in the past few years. However, obtaining estimates of crash risk from only the observed crashes are questionable because there are chances that some external factors could influence the safety of the treated sites. For example, there could be an increase in traffic volume, changes in weather conditions, modification in road design features and others. In such cases, the safety effect of a specific treatment is difficult to distinguish from those of the external factors.

Another before-after study type is using comparison (or control) sites where the crash data from sites with similar features are used to adjust the potential temporal change of before-crash risk (C_b) over the treatment period. One of the disadvantages of this method is the need of a relatively detailed crash data from multiple sites. Some research in the past have compared its performance with other alternative methods. For example, Benekohal & Hashmi (1992) conducted a before-after study in a two-lane highway to evaluate the highway improvement program that consisted of resurfacing, restoration and rehabilitation of road surface. Crash reduction factor (or 1- CMF) was determined using crash data from 51 treated sites and 31 control sites. The two approaches considered were model-based and comparison before-after approaches. In the model-based approach, two crash models were calibrated, each for before and after treated conditions using their respective crash data. Then, these models were used to estimate before (C_b) and after (C_a) crash risks and finally determined the reduction factor. For the second approach of using comparison sites, before and after crash data for the treatment were directly compared with an adjustment made from the crash data of the control sites. A slight deviation was observed between the crash reduction factors from these two selected approaches. The study recommended using before-after study with comparison sites whenever detailed data are available. Similarly, Griffith (1999) applied before-after comparison study to evaluate the safety benefits of adding shoulder rumble strips on freeways. The treatment sites were selected based on their sequence of surface improvement rather than from a list of hotspot sections. Thus, the study makes an argument that those selected sites do not have a selection bias, and therefore, the EB method is not required over the before-after comparison approach.

Before-after study based on EB technique is the most widely used approach compared to the two previously discussed study types. One of the main benefits of the EB method is the use of a crash model that helps to reduce the RTM problem. Harwood et al. (2002) applied before-after studies based on comparison and EB method to evaluate safety effects of providing left and right-turn lanes at the intersections. The CMFs from EB estimates were found relatively lower, and the fact that the EB method accounts for the RTM effect, the results from this method were considered more accurate. Similarly, there are a number of past countermeasure studies using the before-after EB approach (Hauer & Persaud, 1987; Al-Masaeid, 1997; Elvik et al., 2001; Persaud et al., 2001; Bahar et al., 2004; Lyon et al., 2005; Choi et al., 2015).

Recently, the full Bayesian (FB) technique has been applied as an alternative to the EB method for determining the safety effectiveness of countermeasures (Persaud et al., 2010; Lan et al., 2009). The main difference between these two techniques lies in the selection of priors for their model parameters. In the EB method, priors are commonly obtained from the parametric model (e.g., NB model) calibrated using MLE technique, whereas in the FB method, they are either selected from past studies or obtained by assuming some vague non-informative values, i.e., large variance for a typical prior distribution (Miranda-Moreno et al., 2013). Despite this, studies have shown significance of the FB technique over a non-Bayesian method (e.g., NB model calibrated using MLE technique) when the data size is relatively small (Lord & Miranda-Moreno, 2008; Persaud et al., 2010; Miranda-Moreno et al., 2013).

2.3.3 CMF: Cross-sectional Study

Cross-sectional studies using a crash model are among the most frequently used methods for estimating the CMFs (Wu et al., 2015). This approach involves establishing a relationship between crash frequency and predicting variables, which is then used to estimate safety effect of a countermeasure with and without applying it. These two conditions are compared to obtain the countermeasure specific CMF. The most widely used crash models are parametric models, and of the parametric models the NB model is most often used as it has an ability to account for over-dispersion of crash data (Council & Steward, 1999; Lord & Bonneson, 2007; Fitzpatrick et al., 2008; AASHTO, 2010; Zeng & Schrock, 2013; Park et al., 2014; Wu et al., 2015; Park & Abdel-Aty, 2015). In the study using parametric models, the model coefficients of the variables are directly used to estimate their CMFs. Note that a CMF can represent the safety effectiveness measure of a single treatment or combination of multiple treatments. The approach to determine CMFs of these two categories (single and multiple factors) may vary slightly. The following presents some of the past-related studies for each category.

1. CMF for single treatments

Council and Stewart (1999) adopted a cross-sectional study to evaluate safety effects of converting a two-lane highway to a four-lane highway, as typically for such conversions, before and after crash data are not easily available. Separate NB models were developed for each highway type using crash data from four different states in the U.S. Then, for the comparisons, the most typical sections were selected, i.e., for the two-lane section- shoulder width of 1.83 m and surface width of 7.32 m, and similarly, for the four-lane section- shoulder width of 3.05 m and surface width of 3.66 m. Meanwhile, same exposure levels were assumed for both the highway types (i.e., AADT and length). The results showed that converting two-lane to four-lane with divided section with varying condition of exposure levels, the crash reduction was expected to be in the range of 40 to 60 percent or in terms of CMF in the range of 0.6 to 0.4.

Lord & Bonneson (2007) developed CMFs for some road geometric elements such as changing lane and shoulder widths, extending edge markings for a rural frontage road and others. An exponential form of NB model was calibrated to compute the CMFs for these features. The results showed that increasing lane and shoulder widths were associated with lower CMF values, indicating that their safety effects are positive. Likewise, an increase in proportion of edge marking of a road section indicated a relatively safer conditions. Similarly, Fitzpatrick et al. (2008) applied NB model to quantify the effects of widening median width (with rigid barrier), widening left shoulder width including few other factors in freeways and rural multilane highways. As in other studies using NB model, the regression coefficients were used to determine the CMFs. The study showed that widening both the median and left shoulder widths resulted in reduction of crash risk.

Other similar applications of cross-sectional study in developing CMFs of road geometric elements include the works by Zeng and Schrock (2013) and Choi et al. (2015). Zeng and Schrock (2013) focused on developing CMF of varied shoulder width of rural two-lane highway using crash data from Kansas State. Four years (2003- 2007) of crash data were processed annually for the winter and non-winter seasons, and the datasets were used for calibrating the NB models (for each season) where shoulder width was considered as a categorical variable (total ten types). The result showed that widening of shoulder width resulted in increasing safety benefits. Meanwhile, the CMF of changing shoulder width for winter seasons was slightly larger with a variance of 13 to 25 percent. Similarly, Choi et al. (2015) developed two NB models, each for horizontal curve deflection and vertical grade, using a crash data from a Korean Expressway. Additional variables included in the models were length and AADT. The model coefficients were used to compute their CMFs. The result showed that sections with higher horizontal curve radius and lower vertical grades have a lower risk of crashes. However, this study using two separate models for each factor with limited variables is likely to have excluded the effects of important omitted variables.

As the method of cross-sectional study using parametric model is one of the most frequently used approaches to quantifying safety benefits of road geometric features, it is important to validate their results and at the mean time know their strengths and limitations. For this, we refer to the work of Wu et al. (2015) which presents a comprehensive simulation study for validating the CMFs. In this study, a crash dataset was generated by fixing following conditions: 1) assumed CMFs for three variables- lane width, curve density and pavement friction, 2) assumed a safety performance function (SPF) (or crash model) from the HSM manual. The SPF, which represented the base conditions, was multiplied by CMFs to form a complete model. Then this complete model was finally used to generate a simulated crash data. A number of NB models were calibrated using the simulated crash data with a variation in number of predicting variables included in the models. The CMFs were then back calculated using model coefficients for respective scenarios. The result showed that when all the variables are included, the estimated CMFs had much less deviation from their original values. However, when the variables were omitted in the model, the resulting CMFs were biased. The main finding from this study is that parametric crash models, such as NB model, can be effective in estimating CMFs only when a complete information is available including their model functional forms. However, in a real case study, a true model form between crashes and predicting variables is unknown. Therefore, the findings based on parametric model forms (SPFs) can hardly be generalized.

2. CMF for multiple treatments

Determining CMFs for multiple treatments is relatively a complex process as their simultaneous effects are difficult to capture mainly due to the practical difficulty of getting enough data. They are generally derived by an indirect approach by combining the CMFs of single treatments as discussed in NCHRP (2008). Among all, the method involving simple multiplication of individual CMFs is the most popular one (HSM, 2010). This is based on the assumption that the effects of individual factors are independent, which means that there are no interaction effects between the treatment measures. Similarly, other methods mentioned in NCHRP (2008) are designed to calculate CMF of multiple factors by combining their individual CMFs where less important factors are penalized by using some weighting schemes. Critical to this indirect approach is the CMF of a multiple treatment depends on the quality of CMF of individual treatments and the validity of the assumptions made to combine their effects.

Only a few studies have focused on developing CMFs for multiple treatments by considering their actual interaction effects. For example, Park et al., (2014) tried to estimate the CMF of combined treatments of adding shoulder rumble strip and widening shoulder width for rural multilane highway sections by applying the before-after and cross-sectional approaches. For the latter approach, a NB model with an exponential form was considered. The variables in the model were shoulder rumble strip (categorical form), shoulder width and their interaction term. The interaction term was not found significant in the model; however, it was still used to interpret their combined effects. As an alternative, only the interaction term could have been considered in the crash model, similar to the work of Bauer & Harwood (2012). However, the authors argue that such partial form of model may provide a biased result. The study showed that the wider shoulder widths with rumble strip on shoulder showed greater safety benefits and vice versa for the narrow shoulder widths. Another important finding from this study is the CMFs of single treatments obtained from before-after and cross-sectional studies only differ slightly (8%), suggesting that the latter method could be a viable alternative t when a before-after study is not feasible.

In contrast to the parametric approach, only one study had applied a nonparametric approach, i.e., by Park & Abdel-Aty (2015). They applied a nonparametric model called applied the multivariate adaptive regression spline (MARS) as well as NB model for estimating CMF using a crash dataset from a case study of multilane rural highways with the following roadside features: driveway density, pole density, distance to trees and others. The CMF obtained from the MARS method consists of a set of basis function (local parametric models) involving significant variables together with their corresponding model coefficients. Note that the model coefficients are obtained through a calibration process (i.e., training process) similar to other parametric models. CMFs from the two approaches are not comparable, as their true values are not known. Therefore, the study drawed a conclusion that since MARS outperformed the NB model in terms of model performance, the CMFs from MARS are expected to be more accurate.

2.4 Parametric Models

A crash model represents the conditional expectation of crash frequency as a function of a set of covariates. Consider Y as a random variable representing the number of crashes occurring during a specified time period (e.g., hour, month or year) and X, a vector of covariates, representing the potential factors such as traffic characteristics, weather conditions, and geometric features. The conditional expectation of the crash frequency is given by Eq. 2-2.

$$E(Y|X = x) = \mu(x;\beta) \tag{2-2}$$

where,
$\mu(.)$ = expected crash frequency, which is a function of *x* and β, with a known form.
β = a vector of regression coefficients associated with the covariates *x*.

In a parametric approach, the conditional probability of crash frequency is assumed to follow a specific distribution defined by its respective parameters. Its expected crash frequency, i.e., $\mu(.)$, which is a systematic component of a model, is then assumed to be a function of a given set of variables $x_1, x_2, \dots., x_D$ along with a set of regression coefficient $\beta_1, \beta_2 \dots \dots \beta_D$. The associated regression coefficients define the direction and magnitude of the effect of corresponding factors on crash frequency. This process of defining a shape of the relation between crashes and covariates is the basic approach in conventional parametric models. Again, the parameter estimation depends on the methods being applied. Two methods, namely, maximum likelihood method and Bayesian method are commonly used, which will be briefly discussed in Section 2.4.2 and 2.4.3, respectively.

2.4.1 Parametric Model Functional Forms

Specifying a functional form for the expected crash frequency, i.e., $\mu = \mu(x)$ for a given x, is one of the critical parts in the parametric approach. A wide spectrum of functional forms have been postulated in the past. All these forms can be generalized under a common structure given by Eq. 2-3 where the crash exposure and the crash risk, as a set of explanatory variables, appear in a multiplicative form.

$$\textit{Crash frequency} \sim \textit{crash exposure} \times \textit{crash risk} \tag{2-3}$$

Crash exposure represents the traffic level on the road entities of interest, representing the chances of exposing to crashes. If the entities of interest are road segment, it could be measured by traffic volume and the segment length. These factors appear in a model either as a product of individual effects (i.e., $(traffic)^{\beta_1} \times (length)^{\beta_2}$) or as a combined effect ($(traffic \times length)^{\beta_1}$), as shown in Table 2-1. This structuring of crash frequency model by specifying the crash exposure and crash risk in a multiplicative form supports the logic of "no traffic flows or no length" means no crash.

Table 2-1: Common functional forms of parametric crash models

S.N.	Model functional form	References
1	$\mu = traffic \times length \times \left(\beta_o + \sum_{d=1}^{D} \beta_d x_d\right)$	Jacobs and Sayer, 1983; Okamoto and Koshi, 1989; Zegeer et al., 1991; Miaou and Lum, 1993; Hong et al., 2005.
2	$\mu = traffic \times length \times \ e^{\beta_o + \sum_{d=1}^{D} \beta_d x_d}$	Miaou et al., 1992; Miaou, 1994; HSM, 2010; Ahmed et al., 2011.
3	$\mu = traffic^{\beta_1} \times length \times \ e^{\beta_o + \sum_{d=2}^{D} \beta_d x_d}$	HSM (2010) (undivided rural multilane and urban suburban arterial roads); Persaud et al., 1999; Montella, 2009.

S.N.	Model functional form	References
4	$\mu = (traffic \times length)^{\beta_1} \times e^{\beta_o + \Sigma_{d=2}^{D} \beta_d x_d}$	Hauer et al., 1996; Miaou, 1994; Fu et al., 2005; Usman et al., 2012; Wu et al., 2015.
5	$\mu = traffic^{\beta_1} \times length^{\beta_2} \times e^{\beta_o + \Sigma_{d=3}^{D} \beta_d x_d}$	Hadi et al., 1993; Washington et al., 2003; Qin et al., 2004; Miranda-Moreno, 2006; El-basyoung and Sayed 2009.

Note: μ = expected crash frequency; x_d = predicting variables, β_o= intercept; β_d= regression coefficient of x_d; β_1 and β_2 (in model 5) = regression coefficient of exposure variables, D is number of covariates.

Similarly, crash risk represents the effect of factors on crash such as road geometric features and weather variables. The effects of these factors are commonly defined by a simple functional form, a linear or an exponential function, as shown in Table 2-1. In a linear functional form, covariates appear in an additive form with its effect quantified by respective regression coefficients (β). This form is generally used in linear regression (Jacobs and Sayer, 1983; Miaou and Lum, 1993). However, the main limitation of a linear functional form is that it does not guarantee a non-negative outcome, which may easily violate the basic requirement of crashes as a count process.

To overcome the statistical constraint of a linear form, an exponential function has been widely used in crash models, where the linear form of covariates is linked by an exponential function (Table 2-1). This form is also known as "log-linear function" in literature as the same expression can be interpreted by the log of the dependent variable on the left side (crash frequency) linked to a linear form of covariates on the right side of the mathematical expression. Such a functional form ensures that the crash frequency always results a non-negative value (Miaou and Lum, 1993; Miaou, 1994). Due to exponential function in these forms, the elasticity of predicting variables on accident frequency can be easily expressed by the individual regression coefficients (Shankar et al., 1995; Milton and Mannering, 1998; Washington et al., 2003; Chang, 2005; Usman et al., 2012). Elasticity is interpreted as a measure of percentage change of effect of a certain factor on crashes occurrence provided that the other factors remain constant.

The fundamental problem with both the linear and exponential specifications and any other defined forms is that the relation between crash frequency and influencing factors is not known in advance. This parametric approach where the functional forms are arbitrarily selected only quantifies the magnitude of the coefficients for the assumed model and has no flexibility to detect the true shape of the underlying relation. As a result, the model outcomes obtained are limited to the general trend, and cannot detect a complex relation with potential irregularities, such as peak, valley, and point of inflection lying within the relation domain. Therefore, this potential problem in misspecification on which the whole estimation of regression coefficients depends on can easily run into a risk of biased estimates of model coefficients. Consequently, other derived measures such as model elasticity can easily lead to misinterpretations.

2.4.2 Maximum Likelihood Approach

As previously mentioned, there are two common parametric approaches for estimating the parameters of a model, namely, maximum likelihood estimation (MLE) and Bayesian approach. In the MLE technique, the most commonly used crash models are a group of parametric models namely Poisson, Negative Binomial (NB), Poisson-lognormal (PL), Zero-inflated Poisson (ZIP) and Zero-inflated Negative Binomial (ZINB), Generalized Negative Binomial (GNB), random-effect and random-parameter models. Fundamentally, they are all variants or extensions of the Poisson model.

Details on the MLE method used in various crash models can be found in McCullgh & Nelder (1989) and Washington et al. (2003). We have summarized the overall process into the following five steps:

Step 1: Specification of crash distribution

Consider Y represents crash frequency, a random variable, which are independently and identically distributed with an assumed probability distribution $f_Y\ (y;\theta)$ in which θ is the model parameter. We denote the distribution of Y as:

$$Y \sim f_Y\ (y;\theta) \tag{2-4}$$

where,

$f_Y\ (.)$ is the adapted distribution for Y

θ *is* distribution parameter which is a function of $\mu\ (.)$ and conditional on the given error term ϵ*;* here, $\mu(.)$ is expected value of Y and is conditioned on a set of covariates. Note that the parameter θ might have different forms, depending on the model type.

Step 2: Specification of functional form of model, i.e., $\mu(.)$

Pre-define model functional form for expected crash frequency, i.e., $\mu(.)$, expressed by a set of covariates (x) as shown in Eq. 2-5. This model form is the core output of the modeling. The role of coefficients of covariates (β) depend on how the function $\mu(.)$ is specified. Some of the functional forms common in road safety studies are discussed in Section 2.5.1.

$$E(Y|X = x) = \mu(x;\ \beta) \qquad (2\text{-}5)$$

where,

$\mu(.)$ is a function relating x on Y through the regression coefficients β.

Step 3: Specification of error term

In order to capture the variability of model, in most of the models, an error term (ϵ) is considered and specified with a specific probability distribution *f (ϵ; φ).*

$$\epsilon \sim f(\epsilon; \varphi) \qquad (2\text{-}6)$$

Step 4: Construction of likelihood function

A likelihood function *L*(.) is defined mathematically as:

$$L(\theta) = \prod_{i=1}^{n} f_Y(y_i; \theta) \qquad (2\text{-}7)$$

where, y_i is observed crashes, n is number of observations.

Step 5: Model calibration

The likelihood function *L(.)* can be transformed into sum of the probabilities of observed crash occurrences using a logarithmic functions *LL*(.) (Eq. 2-8).

$$LL(\theta) = \sum_{i=1}^{n} ln\ f_Y(y_i; \theta) \qquad (2\text{-}8)$$

The model coefficients β*,* which is part of θ parameter as explained in Eq. 2-4 and 2-5, are estimated by maximizing the *LL(.)* function. The main advantage of the MLE technique is that a closed function exists for the family of commonly used probability distributions (e.g. in Poisson, NB, GNB, PLN, ZIP, ZINB models). A simulation approach is required for some models where the LL function does not have a closed form (e.g., random effect model).

A number of parametric models are used in road safety analyses which are based on different assumptions made on a probability distribution of crash occurrence and its associated error structure. In the following section, some of the commonly used parametric crash models are reviewed.

1. Poisson Model

Poisson model is a single parameter model in which crash occurrence is assumed to follow a Poisson distribution. This model overcomes the statistical problem caused by discrete and non-negativity in a linear model (Jovanis and Chang, 1986; Jones et al., 1991; Miaou, 1994). From the statistical property of the Poisson distribution, the conditional expectation and variance are equal to the model parameter, i.e., $\mu\ (.)$ Mathematically,

$$\begin{gathered} Y \sim Poisson(\theta) \\ \theta = \mu(.) \\ E(Y|X = x) = \ \mu(x;\ \beta) \\ Var(Y|X = x) = \mu(x;\ \beta) \end{gathered} \tag{2-9}$$

where,

y is crash counts,

μ(.) is expected crash frequency as defined in Eq. 2-*5,*

x is a vector of covariates,

β is a vector of regression coefficients associated with covariates x.

One of the limitations of the Poisson model is that the crash data are most likely to be over-dispersed thereby resulting the variance to exceed the mean (Maher and Summersgill, 1996; Cameron and Trivedi 1998; Lord and Mannering, 2010). This may easily violate the mean-variance constraint imposed in the Poisson model which eventually misleads the asymptotic covariance estimate of regression coefficient i.e., β, in Eq. 2-5 affecting the standard error of model coefficients. Consequently, the biased estimate of standard error may invalidate the hypothesis testing on each coefficient (Jovanis and Chang, 1986; Miaou et al., 1992; Miaou and Lum, 1993; Fu et al., 2005; Miranda-Moreno, 2006). This may result in either inclusion or exclusion of variables failing to catch their effects. Therefore, given that the observed data are most likely to be over-dispersed, the use of Poisson model may lead to a biased inference.

2. Negative Binomial (NB) Model

The most popular model used in road safety analysis is the Negative Binomial (NB) model. NB model also known as Poisson-gamma model, is a derived from the Poisson model by including a gamma distributed error term (Lawless, 1987; Miranda-Moreno, 2006). Mathematically, it is given below in Eq 2-10.

$$\begin{gathered} Y \sim Poisson(\theta) \\ \theta = \mu(.)e^{\epsilon} \\ e^{\epsilon} \sim Gamma\ (\varphi,\ \delta) \\ Y \sim NB(\mu(.), \alpha) \\ E(Y|X = x) = \ \mu(x;\ \beta) \\ Var(Y|X = x) = \mu(x;\ \beta) + \ \alpha \times \mu^2(x;\ \beta) \end{gathered} \tag{2-10}$$

As seen above, the error term, i.e., e^{ϵ} is gamma distributed with parameters $\varphi > 0$ and $\delta > 0$. This distribution ensures that $\mu(.) > 0$ since $e^{\epsilon} > 0$. Furthermore, by specifying $\varphi = \delta$, crash occurrence will follow a NB distribution where $E(e^{\epsilon}) = 1$ and Var $(e^{\epsilon}) = 1/\varphi = \alpha$ (Lawless, 1987). The term α is usually defined as the over-dispersion parameter. If $\alpha \to 0$, then this model converges to a Poisson model. Thus α represents the difference between these two forms of parametric models. The expected crash frequency, i.e.,$\mu\ (.)$ is specified as a function of a set of covariates similar to a Poisson model. In this model, the inclusion of an error term provides the flexibility to permit the variance greater than the mean, which allows capturing any unmeasured heterogeneity in a dataset (Lord and Mannering, 2010).

Given this advantage of adjusting potential over-dispersed characteristics of crash data, NB model has been the most widely used model in road safety analysis (Hadi et al., 1993; Miaou, 1994; Shankar et al., 1995; Milton and Mannering, 1998; Harwood et al., 2000; Miaou and Song, 2005). NB model is a further extended by modeling dispersion parameter with all the other assumptions remaining unchanged. This form of model is sometimes referred as Generalized NB model. The over-dispersion parameter is generally specified as a linear combination of covariates linked by an exponential function. This has been proved to increase model fitness as compared to NB and PL models (El-Basyouny and Sayed, 2006; Usman et al., 2012).

3. Poisson-lognormal Model

Poisson-lognormal (PL) is also a variant of Poisson model, similar to the NB model. This model is obtained by replacing the Gamma distributed error term in NB model with Normal distribution with the mean equal to zero and variance σ^2. Note that if $\sigma^2 \rightarrow 0$, mean and variance are reduced to mean and variance of the Poisson model. PL was found to be advantageous for addressing spatial variation pattern of crashes (Milton et al., 2008; Li et al. 2008; Anastasopolos and Mannering 2009; El-basyoung and Sayed 2009).

4. Zero-inflated Models

Zero-inflated Poisson (ZIP) and zero-inflated Negative Binomial (ZINB) models are used when the over-dispersion of data are caused by excess zeroes (Miaou, 1994; Cameron and Trivedi, 1998). In addition to crashes being rare events, the problem of excess zeroes can appear when the roadway's geometric factors are considered by dividing roadway into short homogenous segments (Ahmed et al., 2011). To overcome such an inflated data structure, the ZIP and ZINB model structure defines a two-state regime (i.e., truly safe state and unsafe state) of crashes (Shankar et al., 1997; Qin et al., 2004).

Miaou (1994) employed Poisson, ZIP and NB models, and compared the model performance for truck crash frequency and suggested that the ZIP model is a good candidate model when the crash data exhibits excess zeros. However, these sets of models have limitations. The assumption of dual-state is not a true representation of an underlying crash occurrence process. There is no existence of the complete safe state. On the contrary, the reasons for excess zeros could be due to one of the following reasons: (1) sites with a combination of low exposure, high heterogeneity; (2) small time or spatial scales; (3) data with a relatively high percentage of missing or misreported crashes; and (4) crash models with omitted important variables (Lord et al., 2005). Therefore, such models may misrepresent the practical phenomena of crash occurrence.

5. Random Effect Count Models

In all the previously mentioned parametric models, there could be some unobserved site-specific factors, such as functional class of road, degree of side slopes, pavement surface conditions, users driving behavior, and temporal correlation which are not captured by the models (Shanker et al., 1997; Chin and Quddus, 2003). Such effects can be addressed by an alternative specification of random effects in a count model. Washington et al. (2003) considers NB as one of the examples of random effect model in a Poisson model setting, where a random effect parameter (i.e., error) is assumed to follow the Gamma distribution. This results in consideration of an over-dispersed nature of crashes as discussed in NB model. However, this model does not take into account of a location-specific variation and time correlation effect which are introduced through the random effect of models. This benefit of considering site-specific parameter was observed in a median crossover crash study by Shankar et al. (1997).

6. Random Parameter Count Model

A random parameter model is another variant of a Poisson model. In all the previous models, regression coefficients of covariates (β_i, known as parameter in this section) are considered fixed across all the observations which may however vary in reality. Therefore, to capture such potential heterogeneity in a data structure, random terms are introduced in the given parametric specification of the functional form of a count model. This provides flexibility for the parameters to vary over the observations (Milton et al., 2008; Anastasopoulos and Mannering, 2009).

Anastasopoulos and Mannering (2009) compared the statistical fit of a random parameter NB model (random parameter defined by a normal distribution) and the traditional NB model in establishing an empirical relation of crashes with various road geometric features and traffic exposure. The NB random parameter model was found to be superior to the traditional NB model. Unlike the previous parametric models, the log likelihood function for this model is computationally complicated due to the integration function of normal count model over the assumed normal distribution function of a random parameter. Therefore, a simulation based maximum likelihood method is commonly used for parameter estimation.

2.4.3 Bayesian Approach

The Bayesian approach is an alternative parameter estimation technique which overcomes some of the limitations imposed by the MLE technique. In this approach, both the response variable (here crash frequency) and the model parameters are considered as random variables defined by specific probability distributions. The probability structure of the response variable remains the same as in the MLE approach (e.g., Poisson and NB distribution). Additional flexibility is introduced by considering model parameters (mean crash frequency and over-dispersion parameter) following prior distribution defined by a set of parameters known as hyper-parameter (Tunaru, 2002; Miaou et al., 2003; Miaou and Song, 2005; Song et al., 2006).

Similar to the MLE technique, the Bayesian approach consists of various assumptions on probability distribution and model functional form at different levels. This is described in following four steps:

Step 1: Specification of crash frequency distribution

Consider crash frequency (Y) as a random variable with assumed probability distribution $f_Y(y;\theta)$ where, θ represents the model parameter. Note that the following probabilistic function of model at this step is similar to the MLE approach.

$$Y \sim f_Y(y;\theta) \tag{2-11}$$

where,

$f_Y\ (.)$ is the adapted distribution for Y

θ *is* distribution parameter which is a function of $\mu\ (.)$ and conditional on the given error term ϵ*;* here, $\mu(.)$ is expected value of Y and is conditioned on a set of covariates. Let's assume ϵ follows a probability distribution *f* $(\epsilon;\ \varphi)$. Note that the parameter θ might have different forms depending on the model type.

Step 2: Specification of functional form of model ($\mu(.)$)

Expected value of Y, i.e., $\mu(.)$, is expressed as a function of a set of covariates (x) with assumed functional structure (Eq. 2-12). The coefficients of covariates (β) depend on how the function $\mu(.)$ is specified as in the MLE method. Some of the functional forms common in crash models are discussed in Section 2.4.1.

$$E(Y|X = x) = \ \mu(x;\beta) \tag{2-12}$$

Step 3: Specification of prior distribution for model parameters

The model parameters that include model coefficients in μ(.) ($i.e.,\beta$ in Eq. 2-12) and ε in Eq. 2-11 are considered as random variables with assumed probability distributions $f_\beta(.)$ and $f_\varphi(.)$, respectively. Let's say their parameters (also known as hyper-parameter) are σ and γ, respectively (Eq. 2-13). Note that the number of hyper-parameters depends on the distribution type.

$$\beta \sim f_\beta\ (.,\sigma) \tag{2-13}$$

$$\varphi \sim f_{\varphi}\ (.\,,\gamma)$$

Prior distributions may be either informative or non-informative. Informative priors are those based on previous research. Often the prior information on dispersion parameter, the parameter describing error term specially for the NB model i.e., φ, can be drawn from the estimated result of the MLE models (Ma and Kockelman, 2006; Miranda-Moreno et al., 2013). Meanwhile, non-informative priors are used when there is a lack of past information (Ahmed et al., 2011; Miaou and Song, 2005). This is especially applicable for the model coefficient in μ(.), i.e., β. Studies have shown that the choice of prior information is especially critical for data with small sample sizes but is negligible for data with a large sample sizes (Kass and Wassermann, 1996). Similarly, a few studies have concluded that the use of prior information on the dispersion parameter in NB model increases the accuracy of the estimate when the sample size and crash mean are low (Nathan and Gary, 2006; Song et al., 2006; Lord and Miranda-Moreno, 2008; Miranda-Moreno et al., 2013). Therefore, in such cases, the use of non-informative priors can be problematic, resulting inaccuracy in parameter estimation.

In addition to the availability of information, the specification of prior distribution is based on the conjugal distribution that produces the full posterior distribution of the same form as the parent distributions, and can be computed easily (Miranda-Moreno, 2006). For example, in a Poisson lognormal Bayesian model, the inverse-gamma prior is selected for the hyperparameter as its combination with the normal distribution results in a conjugal distribution.

Step 4: Bayesian inference for model coefficient (β)

Based on Bayes' theorem, the posterior distribution of a parameter is constructed through prior information on the model coefficient β and the current information from the likelihood specification as shown in Eq. 2-14). The posterior distribution $f_{\beta_p}(.)$ is proportional to the product of likelihood function *L(.)* and the prior distribution f_{β}(.) as:

$$f_{\beta_p}(\,\beta) = \frac{L\ (\beta,\varphi)\ f_{\beta}\ (\sigma)}{\prod_{i=1}^{n} f_Y(y_i)} \propto L\ (\beta,\varphi)\ f_{\beta}(\sigma) \tag{2-14}$$

where, the likelihood function $L\ (.)$ is defined by:

$$L\,(\beta, \varphi) = \prod_{i=1}^{n} f_Y(\,y;\ \beta, \varphi) \tag{2-15}$$

where,

y_i is ith observed crash counts,

n is number of observations.

$f_Y(.)$ is the marginal probability distribution of Y.

Finally, the posterior inference for model coefficients (β) given by Eq. 2-16 is computed using Markov Chain Monte Carlo (MCMC) sampling method (Ntzoufras, 2008).

$$E(\beta) = \int \beta \times f_{\beta_p}(\,\beta)d\beta \tag{2-16}$$

Some of the commonly used Bayesian models in the road safety analysis are reviewed in the following section.

1. Poisson-Gamma (NB) Bayesian Model

The Negative Binomial distribution has been the most widely used probabilistic structure for the Bayesian crash model as it offers the simplest way to accommodate over-dispersion. In addition, its marginal distribution has a close form with many prior distributions (Gamma and Normal distribution) (Hauer, 1997; Lord and Miranda, 2008). The following specifications are assumed for the crash distribution and its associated priors.

$$Y{\sim}Poisson(\theta) \tag{2-17}$$

$$\theta = \mu(.)e^{\epsilon}$$

$$e^{\epsilon} \sim Gamma\ (\varphi,\ \varphi)$$

$$E(Y|X = x) = \ \mu(x;\ \beta)$$

Prior distribution $\varphi \sim f_{\varphi}(.)$ and $\beta{\sim}f_{\beta}\ (.)$

where, $\mu\ (.)$ is the model functional form (Section 2.4.1), x is a vector of covariates, β is a vector of regression coefficient, φ is a dispersion parameter. The specification for the main distribution (including error term) is same as the NB-MLE method mentioned in Section 2.4.2. The only difference with the NB-MLE method is the additional specification on prior distribution. Two different types of prior distributions (i.e., non-informative or informative) are specified for dispersion and model regression coefficients.

Ahmed et al. (2011) used non-informative priors to model crashes on mountainous freeways since the prior information for such types of road sections was lacking. Similarly, Miranda-Moreno et al. (2013) considered three different types of distributions for over-dispersion (Gamma distribution, Christiansen distribution, Uniform distribution), and investigated the performance of the model under these priors for four-lane rural highways and three-legged rural intersections. The model hyper-parameters for these respective distributions were derived statistically from several past studies using the MLE approach. In addition, they also considered non-informative gamma distribution priors to observe if any significant difference exists between the two sets of priors. In contrary to the over-dispersion parameter, the priors for the set of regression coefficient are generally considered non-informative and flat, defined by normal distribution with a large variance, e.g., $\beta \sim Normal\ (\beta', 10000)$ (Ahmed et al., 2011; Miranda-Moreno et al., 2013). Using different sets of data the results showed that for lower sample size and mean, the informative priors outperformed far above the non-informative flat priors, and specifically, Gamma and Uniform priors performed better. This shows that the selection of informative priors and distribution type is sensitive.

Similarly, in order to account for spatial variation (e.g., correlation between adjacent road segments), an additional parameter is often introduced in Bayesian framework described by a prior distribution. Normal distribution is commonly considered for computational convenience. This particular structure of model was proved to be a better fit compared to the zero inflation models that are designed to overcome dispersed data (Huang et al., 2010). Ahmed et al. (2011) compared the performance of spatial-effect model and NB model under a Bayesian approach, with non-informative priors and, concluded that the spatial effect was redundant while including well-defined geometric variables. Moreover, the author also concluded that these Bayesian-based models outperformed Poisson model by the MLE approach for the same set of data by accounting for over-dispersion.

2. Poisson-Lognormal Bayesian Model

In a Poisson-Lognormal Bayesian model, the error term follows a lognormal distribution. Since the conjugate distribution of Normal distribution (defining error term) is an inverse Gamma function, the hyperparameter of the error term is specified by the Gamma distribution (Lord and Miranda-Moreno 2008, Miranda-Moreno et al., 2013). These models are recommended over the Poisson–Gamma model when the prior information is missing and crash data characterized by low sample means (Lord and Miranda, 2008).

2.5 Nonparametric Models

Nonparametric methods provide a conditional expectation of crash frequency as a function of a set of predicting variables based on some defined data-driven rules. Examples of nonparametric methods employed in past studies for modeling crashes include Classification and Regression Tree (CART) (Karlaftis and Golias, 2002; Chang & Wang, 2006), Artificial Neural Network (Mussone et al.,1999; Abdelwahab and Abdel-Aty, 2001; Chang, 2005; Xie et al., 2007), Multivariate Adaptive Regression Splines (Abdel-Aty & Haleem, 2011; Park & Abdel-Aty, 2015), and Kernel Regression (Thakali et al., 2014[1]; Thakali et al., 2016[2]). These methods have varied data-driven rules and are briefly discussed in the following sections.

1. Classification and Regression Tree (CART)

The CART method applies a tree like structure for predicting a dependent variable from a given set of input variables. It involves recursively partitioning data points (i.e., training set) where each parent node is split into a binary node based on a selected predicting variable until it reaches to the terminal nodes (Karlaftis and Golias, 2002) (Figure 2-3). Critical to this method is making a choice of the variable at each split and its value to perform a binary split at each node. A numerical search algorithm is used such that the split at each node generates the greatest prediction accuracy, which is usually measured with node impurity measures, and in the meantime, to make sure that there is a greater relative homogeneity (the inverse of impurity) in the terminal nodes. Before application, CART requires training to determine this tree like structure with if-then splitting rules. Whenever a new prediction is to be made, we apply the if-then-else rule which will lead to one of the terminal nodes, and the average value of the terminal node provides the estimated value. However, when an updated training data set is available, the CART model structure needs re-training in order to update the if-then-else rule.

[1] Thakali, L., Fu, L., & Chen, T. (2014). A Comparison between Parametric and Nonparametric Approaches for Road Safety Analysis - A Case Study of Winter Road Safety. In Transportation Research Board Annual Meeting (Vol. 6, pp. 1–17).

[2] Thakali, L., Fu, L., & Chen, T. (2016). Model Based versus Data-driven Approach for Road Safety Analysis : Does More Data Help? Transportation Research Record: Journal of the Transportation Research Board, No. 2601.

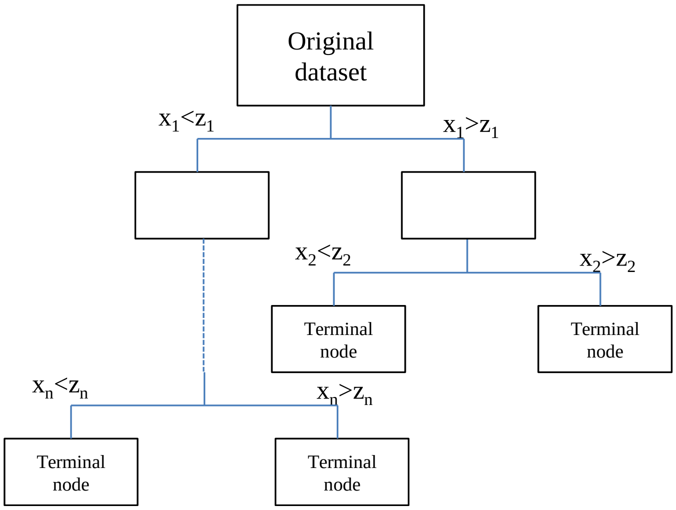


Figure 2-3: A typical structure of CART

2. Artificial Neural Network (ANN)

The architecture of an ANN model consists of an input layer, hidden layers and an output layer as shown in Figure 2-4. (Chang, 2005; Xie et al., 2007). The nodes in the input layer receive a set of predicting variables which are then processed through the hidden and output layers to obtain a final output. Each node in these layers (hidden and output) acts as a computational unit where the inputs coming to the node are first weighted, and then an activation function is used to transform the result. Finally, this becomes an input to the next layer as directed by the connections in the network. The ANN method requires a few pieces of prior information such as number of hidden layers, number of units (nodes) in each hidden layer, a network-learning rate (to controls size of weights) and activation functions. Some learning processes are used to determine the weights; for example, one of the most commonly used algorithm to train the model is the back-propagation algorithm (Rumelhart et al., 1986). Additionally, assumptions are needed on activation functions and a number of hidden layers are typically determined by trial-and-error process. This approach may have less control over negative outcomes, especially, when a dataset is dominated by a large number of zero crash counts. Note that whenever a new data set is available, the ANN approach requires training of the network model in order to update the weights assigned to each neuron that links predictors and the target variables.

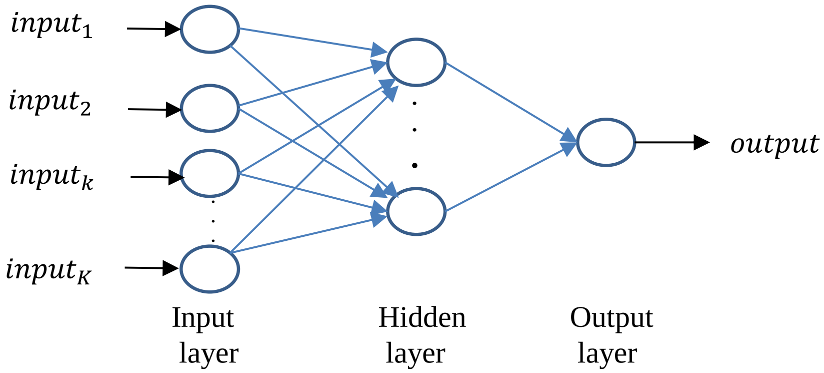


Figure 2-4: A typical structure of ANN with one hidden layer

3. Multivariate Adaptive Regression Splines (MARS)

This method consists of multivariate-segmented regressions which are defined by a set of basis functions for modeling a relationship (Friedman, 1991). Theoretically, this method is similar to a parametric model as there involve model coefficients associated with each basis function which are determined through a calibration process using a training dataset. The model coefficients are obtained through minimization of sum of residuals. However, compared to parametric models, there is a greater flexibility to capture any nonlinear and complex relationship by considering a large number of basis functions. Often the input variables to this method are identified using some alternative methods such as random forest technique (Abdel-Aty & Haleem, 2011).

4. Kernel Regression (KR)

This is a fully data-driven nonparametric method, and is relatively simple to understand as the parameters involved (i.e., bandwidths) are easily interpretable. The KR method requires a kernel function and bandwidths that are determined from training data set along with existing data points to make a new prediction. These parameters control the weight on each data point with closer data points getting larger weights compared to the farther ones. Due to its physically interpretable parameters, KR is often called a “grey-box” (Brown et al., 2011). Most importantly, unlike in previously discussed nonparametric methods, only a few parameters are required, and there are no hidden model structures to train. For example, ANN has layers of input, hidden and output layers with many different weights to train. Similarly, MARS has a set of basis functions whose corresponding coefficients need calibration and CART has a tree-like structure to generate if-then rules. These methods are often characterized as "Black Box" due to their complexity for direct interpretation.

Meanwhile, when it comes to updating of the results using a new dataset, the KR method is more adaptive. This is because the KR method is a fully data-driven technique, i.e., it uses the raw data points directly to make a prediction. Therefore, the newly collected data set can be easily combined with the existing set and update the results in a real-time. Of course, the bandwidths can be updated by learning from the updated new training set, but still without this step, the KR method can easily make use of new information. However, in other nonparametric methods, unless their hidden structures are re-trained using new training set, we cannot make use of new information. Details about the KR method are provided in next chapter.

2.6 Summary

The first part of this chapter (Section 2.1-2.3) discussed about road safety studies, including network screening and countermeasure studies, which are very popular for an effective management of road safety problems. Past studies showed that a large amount of research has contributed to developing and investigating alternative techniques, and have focused mainly on addressing the problem of data availability and improving risk estimation. In all these past efforts, whenever a crash model is involved, there is a skewed preference for parametric models. In particular, the NB model has been the most popular, whether it be in an EB framework or applied independently. While there is a continuous effort to improve the crash modeling component required for road safety studies, much less attention has been given to the use of a nonparametric approach.

The second and third part of this chapter (Section 2.4 to 2.5) briefly reviewed various statistical techniques (parametric and nonparametric) used for modeling crashes. Examples of parametric models include the standard Poisson, NB, Poisson-lognormal, zero-inflated Poisson, zero-inflated Negative Binomial models and a few others. While this approach provides an easy-to-apply tool due to its defined mathematical form (i.e., equation) and allows for convenient interpretation of the results, it comes at a cost: the need to pre-select a model form, which, without knowing the true relation is nothing but a guess. Also discussed were the two model calibration techniques commonly applied in the parametric approach: MLE method and the Bayesian technique. The latter approach is important to improve the models mainly when the sample size of crash data is small. Similarly, typical nonparametric models including CART, ANN, MARS, and kernel regression were discussed. Compared to parametric models, these methods do not make any assumptions about their model forms; but rather they are assumed to follow a certain data-driven rule to capture the relation of dependent and independent variables.

Chapter 3
Proposed Methodology

This chapter provides a detailed description of kernel regression (KR), a nonparametric method which is relatively less complex compared to some of the previously applied similar methods that require explicit training of their hidden model structures (e.g., tree-like structures in CART, basis functions in MARS, hidden layer components in ANN). While the application of the KR method is widely found in the field of social science and is growing in the field of engineering, its usage in road safety analyses has been relatively less explored. One of the limitations of this method is lack of systematic approaches to identify a list of important variables to feed into the process of crash prediction. Therefore, to overcome this issue of variable selection, an algorithm is developed which fully runs in a nonparametric framework. Furthermore, we propose an extended form of the KR method to account for the site-specific risk levels, similar to the Empirical Bayesian (EB) method based on parametric models. Lastly, a brief description of negative binomial (NB) model is also included, as this model is used for benchmarking the performance of the KR method in latter chapters.

3.1 Nonparametric Approach: Kernel Regression (KR)

Kernel regression (KR) is one of the most commonly used forms of nonparametric techniques in applied economics (Livanis et al., 2009). It was originally proposed by Nadaraya (1964) and Watson (1964); therefore, it is also known as the "Nadaraya–Watson" estimator. KR can be used to identify the functional relationship between a dependent variable and potential covariates without the need to pre-specify a functional form and probability distributions like in a parametric model. Apart from KR, there are other similar data-driven nonparametric methods available, such as spline and orthogonal polynomial. However, it is argued that all these methods are, in an asymptotic sense, equivalent to the KR method (Hardle and Mammen 1993; Silverman, 1984; Hardle, 1994); as a result, this research focuses on the KR method.

To briefly explain how the KR method works, we assume a dataset consisting of a set of Y and X variables, where Y is a dependent random variable representing the number of crashes per unit time and X is a vector of D-dimensional covariates, i.e., $X_1, \ldots X_D$. A regression function $m(.)$ is defined by a set of covariates X with error ϵ, where crash counts are assumed to be independent and identically distributed across road segments. Like any paramateric models, the KR method considers the conditional expectation of the dependent variables (here crash frequency) given a set of covariates (Eq. 3-1). This form of regression is estimated through the data-driven nonparametric approach of kernel density estimation (Eq. 3-2).

$$Y = m(x) + \epsilon \tag{3-1}$$

where,

$m(x) = E(Y|X = x)$,

x is realization of X covariates with D-dimensional vector form, i.e., $x = (x_1, x_2, \ldots . x_D)'$,

$E(\epsilon|X) = 0$, and

$Var(\epsilon|X = x) = \sigma^2(x)$, which is finite.

$$E(Y|X = x) = \int y \frac{f_{X,Y}(x, y)}{f_X(x)} dy \tag{3-2}$$

where,

y = realization of Y response variable,

$f_{X,Y}(x, y) =$ joint distribution function of covariates X and Y, and

$f_X(x) =$ marginal distribution function of covariates X.

The foundation of the KR method is based on the stochastic framework that begins with an estimation of the nonparametric density, thereby without considering any prior information. This approach can identify any irregularities in the density, which are difficult to capture by a parametric approach. A general form of the multivariate kernel density estimator for the case of D-dimensional covariates is defined by Eq. 3-3 (Hardle 1990; Li and Racine, 2003).

$$\hat{f}_k(x_1, x_2, \ldots . x_D) = \frac{\sum_{i=1}^{n} \prod_{d=1}^{D} k\left(\frac{x_d - x_{i,d}}{b_d}\right)}{n \prod_{d=1}^{D} b_d} \tag{3-3}$$

where,

$\hat{f}_k(.)$ = multivariate kernel density function,

$k(.)$ = kernel function — a continuous bounded and symmetric function i.e., *k (u) = k (-u)*, $\int k(u)du = 1$, $\int uk(u)du = 0$, $\int u^2 k(u)du > 0$,

x_d = point of evaluation for the *d*th variable (*d= 1…D*),
$x_{i,d}$ = observed value for the *d*th variable,
D = number of covariates,
n = sample size, and
b_d = bandwidth (a positive number) of dth variable, such that:$b_d^{(n)} \downarrow 0$ (goes down to 0 monotonically) and n$b_d^{(n)} \to \infty$ for all $d = 1, \ldots, D$.

The conditional expectation of crash frequency (the dependent variable) is given by Eq. 3-2. By substituting the kernel density estimate for the corresponding marginal and joint density given by the above-mentioned concept of density estimation process, the expression is deducted to Eq. 3-4.

$$\hat{m}(x_1, x_2, \ldots .x_D) = \frac{\sum_{i=1}^{n} \prod_{d=1}^{D} k\left(\frac{x_d - x_{i,d}}{b_d}\right) y_i}{\sum_{i=1}^{n} \prod_{d=1}^{D} k\left(\frac{x_d - x_{i,d}}{b_d}\right)} = \sum_{i=1}^{n} w_i(x_1, x_2, \ldots .x_D) y_i \qquad (3\text{-}4)$$

where:

$\hat{m}(.)$ = estimator of $m(.)$ in Eq. 3-1 (i.e., expected crash frequency),
y_i = observed crashes per unit time interval,
$w_i(x_1, x_2, \ldots .x_D) =$ a weighting factor which equals to $\frac{\prod_{d=1}^{D} k\left(\frac{x_d - x_{i,d}}{b_d}\right)}{\sum_{i=1}^{n} \prod_{d=1}^{D} k\left(\frac{x_d - x_{i,d}}{b_d}\right)}$, and
All other notations remain same as in Eq. 3-3.

Variance of estimate, i.e., $\hat{m}(.)$, is given by Eq. 3-5 (Silverman, 1986; Hardle, 1994; Hyfield & Rachin, 2008).

$$Var[\hat{m}(x_1, x_2, \ldots .x_D)] = \frac{\hat{\sigma}^2(x_1, x_2, \ldots .x_D) R(k)}{n \prod_{d=1}^{D} b_d \ \hat{f}(x_1, x_2, \ldots .x_D)} \qquad (3\text{-}5)$$

where,

$R(k) = \int_{-\infty}^{\infty} k^2(u) du$, also known as kernel roughness. Note that $R(k)$for the Gaussian kernel is $1/2\sqrt{\pi}$, i.e., 1.57;

$\hat{\sigma}^2(x_1, x_2, \ldots .x_D) = \frac{\sum_{i=1}^{n} \prod_{d=1}^{D} k\left(\frac{x_d - x_{i,d}}{b_d}\right) \epsilon_i^2}{\sum_{i=1}^{n} \prod_{d=1}^{D} k\left(\frac{x_d - x_{i,d}}{b_d}\right)}$, ϵ_i is error for the ith observation; and

All other notations remain same as in Eq. 3-3.

Hyfield & Rachin (2008) have also proposed a bootstrapping approach as an alternative technique to determine the variance.

As shown in Eq. 3-4, estimating dependent variable “Y” for a given condition is a weighted average of observed values, i.e., y_i’s, where the weights are determined jointly by a kernel function and bandwidths. It is easy to show that the weighting factor $\omega_i(.)$ has the following properties: $w_i(.) > 0$

and $\sum_{i=1}^{n} w_i(.) = 1$. Intuitively, for the fixed bandwidths, the weights are bigger for the observed points that are closer to evaluating point and smaller or possibly 0 when they are remote. This aspect makes the KR method straightforward and understandable unlike in other data-driven approaches, such as ANN and MARS, where interpretation of how a dependent variable is linked to a set of independent variables is relatively complex. Furthermore, the weighting approach of each observed y_i's to estimate the value for a given point of interest indicates that the KR method is a local fitting technique as opposed to a parametric method that selects a single curve of a certain shape to fit the given entire data points. By down weighting the observations that are further apart, the kernel nonparametric estimator uses more relevant information for estimation, hence it could capture variations that are overlooked by parametric models. However, one of the similarities between the two approaches is that both are determined based on a probabilistic framework (details in Pagan and Ullah, 1999).

In this method, we need to decide two things prior to the estimation: the kernel function and the variable bandwidths. The most common choice for the kernel function is the Gaussian kernel, but alternatives such as the Epanechnikog, triangular, and uniform kernels also exist. Note that for a large sample size, any kernel will be close to the optimum kernel (Pagan and Ullah, 1999). For this study, Gaussian kernel is selected as it has higher differentiability properties and less computation time compared to other kernel functions (Lavergne and Vuong, 2000).

Another important component of the KR method is the bandwidth of each variable. As discussed previously, bandwidths play a crucial role in KR estimates as they determine the size of the neighborhood for averaging. Though the kernel regression estimator is free from misspecification, i.e., it converges to the truth when sample size approaches infinity, it is biased for a finite sample. A smaller bandwidth reduces the bias but inflates the variance, while a bigger bandwidth reduces variance at the cost of bigger bias. This natural trade-off between the bias and variance helps to pin down the desirable bandwidth that minimizes the mean squared error of the estimator. A detailed discussion of various of bandwidth estimation methods could be found in Pagan and Ullah (1999). This includes methods from a simple rule of thumb to some advanced approaches such as cross- validation method (CV) (Racine, 2008; Hall et al., 2007; Parmeter et al., 2007; Zhang et al., 2009; Brown et al., 2011; Kohler et al., 2014). One of the limitations of the CV method is, for a large data size, a normal computer requires a substantial computation time to use the method. Therefore, in this thesis, we adopt a variation of the Silverman's rule of thumb to avoid this issue of computation (Silverman, 1986; Simonoff, 1996). A similar method is also applied in past studies (Lavergne and Vuong, 2000; Gu et al., 2007; Dudek, 2012). We calculate the bandwidth for the *j*th variable as:

$$b_d = \left(\frac{4}{2D+1}\right)^{\frac{1}{4+D}} \times \sigma_d \times n^{-\frac{1}{4+D}} \tag{3-6}$$

where,
σ_d = standard deviation of the corresponding d^{th} variable,
D = total number of variable, and
others are same as in Eq. 3-3.

3.2 Variable Selection

The data-driven and specification free nature of the KR method are appealing to modeling crashes and analyzing road safety problems. However, a few issues need to be addressed before it can become a potential tool to be applied by front-line practitioners. One of the important issues is the need for a systematic process to identify a set of input variables that feed into a modeling framework. Similar to regular parametric regression models, it is necessary to determine the variables that have a significant influence on crash predictions and filter out the unnecessary ones. Meanwhile, particularly for the KR method, fewer numbers of variables are favorable to avoid the curse-of-dimensionality effect, especially when the sample size of a dataset is relatively small (Silverman, 1986; Pagan & Ullah, 1999). An increasing number of variables reduces the number of data points available in the neighborhood of a point of interest, which may eventually affect the accuracy of predictions. In the past studies, the issue of variable selection for the KR method has not been explored extensively, which is also true for other nonparametric methods including artificial neural network (ANN), multivariate adaptive regression splines (MARS) and others applied for modeling crashes.

Developing a variable selection algorithm for a nonparametric method may not be as straightforward as in a parametric method. The normally used procedure in the latter method is to follow either backward or forward selection process where potential variables are successively included in the model, and the insignificant ones are excluded. This decision is made based on testing a certain hypothesis with an assumption of some parametric distributions for the test statistic. For example, the most commonly used is the t-test (or its equivalent Wald test) in a parametric model, which checks whether or not a calibrated variable coefficient is significant at a certain confidence level by stating a null hypothesis that the coefficient is of zero magnitude (Washington et al., 2003). Rejecting this null hypothesis means the variable is significant. For this, a test statistic is evaluated by assuming it follows a normal distribution with the mean equal to zero and that it has a certain variance. This value is then compared to a critical value from a standard normal distribution for a fixed confidence level. When the test statistic is larger or greater than the critical value, then the null hypothesis is rejected, thus suggesting that the variable is significant and should be included in the model. Thus, due to this convenience of the testing procedure, the approach of variable selection in a parametric method becomes a straightforward process.

In contrast, similar testing is not possible in a nonparametric method as there is no any definite model coefficient related to each variable by which a test can be conducted. Moreover, a situation may exist where a single parameter (i.e., for each variable) may not be sufficient to represent an underlying relation between a dependent and a predicting variable. For example, let us say that there is a non-linear, non-monotonic relation between a dependent variable “x” and an independent variable “y”. In the case of a linear model, we could simply test the significance of the variable by determining whether the slope of the proposed relation is zero or not. However, in a nonparametric approach, which is expected to capture a nonlinear relation, in some ranges of x values the relation may be flat indicating an insensitive relation, whereas it may be highly sensitive in other ranges of x values. Therefore, in a nonparametric method, testing whether a variable is significant or not requires special approach.

A few indirect approaches do exists. Examples include selecting variables that have been found to be significant in past-related studies or on the basis of the recommendations of road safety experts. However, these approaches could result in a subjective list of candidate variables that may not completely reflect the safety problems that are specific to study area. Another approach could be by making use of the findings from some parametric studies. That is, we could first, calibrate a certain parametric model following their standard procedure of testing variable significance levels, and then identify the final list of candidate variables to be included in the KR method. This was the approach taken in our previous study, which was reasonable as the main objective was to compare the NB model and the KR method (Thakali et al., 2014). Similarly, Li et al. (2008) applied support vector machine to predict crashes with variables selected using traditional NB model. Meanwhile, in some past safety-related studies that applied nonparametric methods, such ANN, this issue of variable selection was not explicitly explored (Xie et al., 2007; Chang 2005). Another alternative approach could be by conducting an exhaustive search for a subset of variables that has the optimum performance (Goethals et al., 2001; Cateni et al., 2011; Cateni et al., 2012; Dudek, 2012). However, such an approach lacks detailed information and insights of variable effects at their individual levels.

3.2.1 Proposed Methods

In this section, we propose a two-step approach to select the variables in the KR method. The first step is to apply a bootstrap algorithm, which determines the relative variable importance (VI') of each potential factor. The VI' measure, as detailed in the following section, is used to decide whether a specific variable should be included or excluded from the model. A detailed explanation is given in Section 3.2.2. Note that Prinzie & Poel (2008) and Hossain & Muromachi (2012) applied a slightly similar method in their studies. However, their focus was to solve some classification problems, such as classifying consumer preferences and classifying real-time crash risk level of a freeway, and therefore, do not directly fulfill the need of a regression problem.

The second step is to measure the overall performance of the model. This is a supplementary step to the previous algorithm. After knowing importance level of each variable, the final decision of what variables to include or exclude, especially those at marginal VIs', are made by measuring the overall performance of the KR method. This is the essence of a regression model where optimum performance is desired. The indicator used is a mean absolute error (MAE) which is given by Eq. (3-7). We adopt a backward selection approach where the least important variable is excluded step-by-step until we reach to optimum performance.

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \tag{3-7}$$

where,

y_i= i^th observed crash frequency,

$\hat{y}_i$= estimated crash frequency for i^{th} observation, and

n = total number of observations.

3.2.2 Algorithm for Determining VI'

We propose a bootstrap-based algorithm to quantify impact level of each variable. The idea behind the bootstrapping is to create a large number of sample datasets by resampling the original dataset. These generated datasets provide an opportunity to produce a number of possible outcomes, thereby providing a better representation of the imaginary population. The measure of outcome, for example, could be the model goodness-of-fit. Finally, the average statistic obtained from the bootstrapping process can be used for making decisions.

This algorithm follows the idea of bootstrapping approach by generating some random samples (also termed as bootstrap samples) a part of which are then used as a training set for the KR method. Note that the KR method uses these sampled datasets to determine the bandwidth of variables. At first, all the potential variables are included in the training set and the final decision to select variables are made based on their individual average performance level measured as relative variable importance (VI'). The entire process involved in determining the VI' is presented in Figure 3-1, and a detailed explanation is given below.

1. Select all the potential variables (D) from a given dataset that could have effects on the dependent variable.
2. Generate B random datasets- bootstrap samples, from a given dataset of N sample size and D number of variables. For each bootstrap sample (b), split the dataset into a training set (T^b) and a validation set (V^b). This means, a total of B training sets and corresponding B validation sets are generated. The percentage split between T^b and V^b is 80/20.

   In the case of a case study that consisted of crash data measured over a period for the same unit (sections), in that case, we preferred to split the initial training (T^b) and validation set (V^b) by a certain time horizon. For example, consider a situation where the first few years are assigned to $T^{b'}$ and last few years to V^b. Now, the final training set (T^b) for each tree is selected by randomly considering 90% of the initial training set ($T^{b'}$). However, note that the validation set is fixed. This approach is taken to have a more representative validation set to test the performance of the KR method.

3. For each bootstrap sample, calculate the bandwidth of each variable using T^b set and estimate the crash frequency for its corresponding V^b set using KR method. Then, calculate prediction error of each tree on its V^b set as:

$$SAD^b = \sum_{i=1}^{n} |y_i - \widehat{y_i}|$$

   where,

SAD^b = sum of absolute deviation for bootstrap "b",

y_i= observed crash frequency,

$\hat{y_i}$= estimated crash frequency,

n= number of observations in V^b.

4. Permute each variable one at time and recalculate the percentage increase in prediction error as below:

$$PE_d^b = \frac{SAD^{\mathrm{b}} - SAD_d^{p_{\mathrm{b}}}}{SAD^{\mathrm{b}}}$$

   where,

   PE_d^b= prediction error for bootstrap sample "b" and "d" variable,

   $SAD_d^{p_{\mathrm{b}}}$ = performance measure after permuting the variable "d" in bootstrap sample "b".

   The idea of permuting a variable is fundamental for measuring its variable importance (VI). A similar concept can be found in a random forest method (Breiman, 2001). When a variable is important for a model, then permuting its values is expected to increase the model prediction error (or decrease model accuracy) and vice versa when the variable is less important (Prinzie & Poel, 2008; Hossain & Muromachi, 2012). Therefore, there is a positive correlation between the magnitude of VI and the impact level of a variable.

5. Repeat step 4 for each variable.
6. Repeat step 3 to 5 for each bootstrap sample "b".
7. Calculate VI of each variable as:

$$VI_d = \frac{\sum_{b=1}^{B} PE_d^b}{B}$$

   where,

   VI_d= variable importance of variable d,

   B = total number of bootstrap samples.

8. Rank the variables based on their magnitude of VI.
9. Find the relative importance of each variable (VI') by comparing to the highest VI.

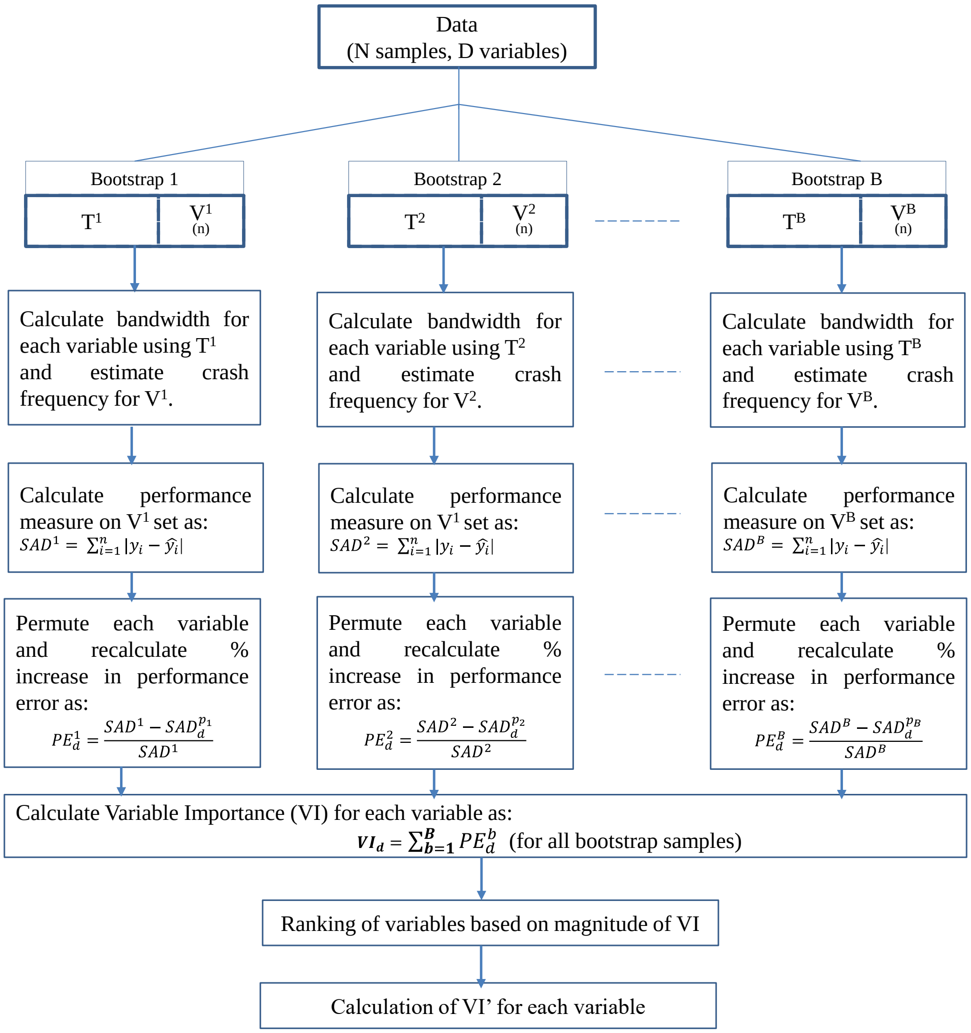

Data
(N samples, D variables)
Bootstrap 1
T^1
$V^1_{(n)}$
Bootstrap 2
T^2
$V^2_{(n)}$
Bootstrap B
T^B
$V^B_{(n)}$
Calculate bandwidth for each variable using T^1 and estimate crash frequency for V^1.
Calculate bandwidth for each variable using T^2 and estimate crash frequency for V^2.
Calculate bandwidth for each variable using T^B and estimate crash frequency for V^B.
Calculate performance measure on V^1 set as: $SAD^1 = \sum_{i=1}^{n} |y_i - \hat{y}_i|$
Calculate performance measure on V^1 set as: $SAD^2 = \sum_{i=1}^{n} |y_i - \hat{y}_i|$
Calculate performance measure on V^B set as: $SAD^B = \sum_{i=1}^{n} |y_i - \hat{y}_i|$
Permute each variable and recalculate % increase in performance error as: $PE_d^1 = \frac{SAD^1 - SAD_d^{p_1}}{SAD^1}$
Permute each variable and recalculate % increase in performance error as: $PE_d^2 = \frac{SAD^2 - SAD_d^{p_2}}{SAD^2}$
Permute each variable and recalculate % increase in performance error as: $PE_d^B = \frac{SAD^B - SAD_d^{p_B}}{SAD^B}$
Calculate Variable Importance (VI) for each variable as:
$\boldsymbol{VI_d} = \sum_{\boldsymbol{b=1}}^{\boldsymbol{B}} PE_d^b$ (for all bootstrap samples)
Ranking of variables based on magnitude of VI
Calculation of VI' for each variable

Figure 3-1: A bootstrap algorithm for variable importance

3.3 Parametric Approach: Negative Binomial Model

The Negative Binomial (NB) model has been the model most extensively used by transportation agencies for crash modeling and road safety analyses (Hadi et al., 1993; Miaou, 1994; Shankar et al., 1995; Persaud et al., 1997; Milton and Mannering, 1998; Harwood et al., 2000; Persaud, 2001; Miaou and Song, 2005; Lord and Mannering, 2010). It has also been recognized as the mainstream model and documented in the highway safety manual.

The NB model, also known as Poisson-gamma model, is derived from the Poisson model that includes a gamma-distributed error term (Maher and Summersgill, 1996; Lord and Mannering, 2010; Cameron and Trivedi 2013). As we explore the conceptual background of NB model, we can see that there are a few pre-specification requirements prior to the model calibration. Let Y be a number of crashes occurring at a certain site for a specified time period (here year). Assuming Y follows a Poisson distribution i.e., mathematically, $\mathrm{Y}{\sim}Poisson(\theta)$, where $\theta = \mu(.) * e^{\varepsilon}$ and e^{ε} follows a Gamma distribution (a two-parameter distribution), by specifying both parameters of the gamma distribution equal and greater than zero results in the NB model. Then the conditional expected crash frequency over the specified time period, $E(Y|X = x)$ or μ (.), is written as in Eq. 3-8.

$$E(Y|X = x) = \mu(x;\ \beta) \tag{3-8}$$

where,

$\mu(.)$ = a parametric function representing the relationship between crash frequency and a set of covariates (x), and

β = a set of model coefficients to be calibrated

Generally, an "exponential function", as shown in Eq 3-9, is used for $\mu(.)$ as it ensures that the crash frequency is always non-negative value (Miou and Lum, 1993; Miaou, 1994 and more references given in Section 2.5.1).

$$\mu(exposure, x_{2\ldots.}x_D; \beta_0, \beta_1 \ldots\ldots \beta_D) = (exposure)^{\beta_1} e^{(\beta_0 + \sum_{d=2}^{D} \beta_d * x_d)} \tag{3-9}$$

where,

$\mu(.)$ = expected crash frequency,

exposure = generally defined as the product of total traffic volume and the road section length,

β_1 = exponent of the exposure variable,

β_0 = intercept,

β_d = coefficient of explanatory variable $x_{d,}$

x_d = d^{th} explanatory variable, and

D = number of covariates.

The model coefficients in Eq. 3-9 can be estimated with the maximum likelihood method using a crash dataset. Meanwhile, a backward selection approach could be employed to determine those variables that are significant at a given level of significance (e.g., 5%).

3.4 Empirical Bayesian (EB) Approach

As mentioned in Chapter 2, EB method has been the state-of-art approach for estimating crash risk both in network screening and countermeasure studies. This section provides an overview of the EB method including its current practices and potential extension for the nonparametric analysis

3.4.1 Concept of EB Approach

EB method provides a framework to determine the long-term crash risk of a site by combining risk measures from two clues- site-specific observed crashes and the expected crash frequency. The latter represents the risk of a site estimated from reference sites with similar features which is achieved through an effort of crash modeling. The use of the first measure, site-specific observed crashes, alone is not sufficient to capture uncertainty in crash occurrence as it is expected to fluctuate over time. This uncertainty phenomenon of crash occurrence in road safety studies is also known as regression-to-mean (RTM) effect (Figure 3-2). Consequently, this value must be supplemented by an estimate from a crash model, which forms the basis of the EB approach.

Following Hauer's (1997) notations, the expected crash frequency at a site is expressed as a weighted average of crash frequency obtained from a reference population "$E(k)$" and the observed crash counts on that site (K). Mathematically,

$$E(k/K) = wE(k) + (1-w)K \tag{3-10}$$

where,

$E(k/K)$ = expected number of crashes at a given site (e.g., a road segment, or an intersection) given that K crashes occurred,

$E(k)$= expected value for mean crash frequency (i.e., k) as referenced to similar units,

K = observed number of crashes at that given site,

w = weight.

$$w = \frac{1}{1+\frac{Var(k)}{E(k)}} \tag{3-11}$$

where,

$Var(k) =$variance of mean crash frequency.

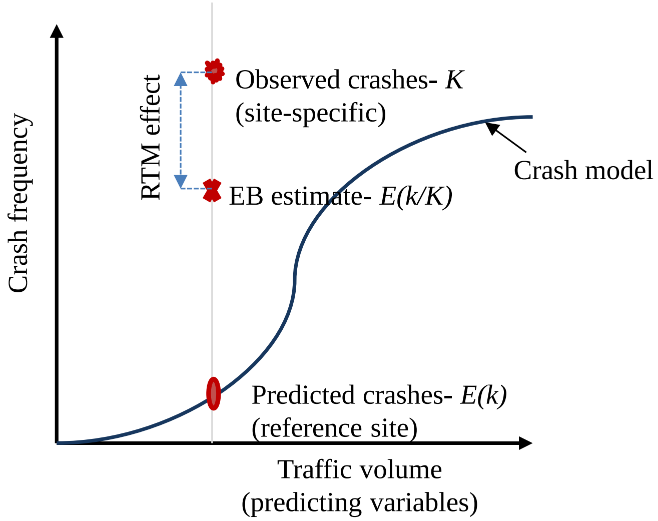


Figure 3-2: Regression-to-mean effect and EB estimate

As seen in Eq. 3-10, an EB estimate is a weighted average of observed crashes and the expected crash frequency obtained from reference sites. In this equation, the weighing factor (w) is the key input which can be derived from two different approaches. The first method is using the Bayesian approach of combining priors and data likelihood following the Baye's rule (see Appendix A. 1). Note that the EB is different from the full Bayesian approach in selecting the priors. In the EB method, the priors come from crash data that are used to calibrate the crash models; whereas in the full Bayesian approach, some distributions are assumed on the bases of previous relevant studies. The second method is using the concept of combining two random variables aimed to minimize their resultant variance (see Appendix A. 2). Hauer (1997) showed that these two approaches are equivalent, yielding the same final expression, i.e., Eq. 3-11, for the weighing factor. The detailed steps involved for the derivations are included in Appendix A. Note that in this thesis we aim to incorporate the KR method into the EB framework by considering the second approach.

As seen in Eq. 3-10 and 3-11, the EB framework requires estimates of two measures to calculate the weight and expected crash estimate from reference sites: $E(k)$ and $Var(k)$. Hauer (1997) proposed the following two methods to estimate $E(k)$: (1) method of sample moment, and (2) method of multivariate regression. In the first method, a simple sample mean is used to estimate the value of E(k). While this method is simple to apply with only a few assumptions, it, however, does not account for possible effects of safety related factors. This issue is addressed by using second method, i.e., regression approach. This approach takes into account of effects due to crash causing factors by developing their relation with crash frequency. In the past, use of multivariate regression from a family of parametric models have been the standard process.

Similarly, the estimate of site-specific variance (i.e., Var(k)) is needed to determine the weight. Hauer (1997) and Hauer (2015) define this as a function of the variance of crash counts and the model estimates (Eq. 3-12). The further details related to this are presented in Appendix A.

$$Var(k) = Var(K) - E(k) \tag{3-12}$$

where,

$Var(K)$ = variance of crash counts and other terms same as in previous definitions.

Again, Hauer (1997) proposed two methods to estimate the variance: (1) method of sample moment, and (2) regression method. In the first method, the variance is directly estimated as sample variance minus the sample mean of crash counts. Note that, in this method, no safety impacts of crash-related factors are considered. In the second approach of using regression, the variance is determined based on two measures: crash counts and estimated mean crashes. By imagining a population where each unit (row in the dataset) is a sample of one, we estimate the sample variance of crash counts by the square difference (SD) between the observed crash counts and fitted values from a crash models. Then replacing SD in Equation (3-12), we obtain:

$$Var(k) = SD - E(k) \tag{3-13}$$

where,

SD is defined as the square difference between the observed crash counts and crash estimates from a model. Details are presented in Appendix A.

In the past, the use of parametric models has been the common tradition to estimate all the parameters required for an EB estimate. This include Poisson-gamma model, also known as NB, (Hauer, 1997; Persaud et al., 1997, Miranda-Moreno et al., 2005; HSM, 2010; Zou et al., 2013), Poisson-lognormal model, Sichel model (Zou et al., 2013) and Poisson-lognormal model (Miranda-Moreno et al., 2005). As discussed in Chapter 2, all these models use a pre-specified mathematical equation to specify a relation of crashes and predicting variable. However, if an improper function is specified, the resulting risk estimates may be biased. To address this specification problem, we propose to employ the KR method.

3.4.2 NB-based EB Method

Referring to Eq. 3-10, an EB estimate requires three main inputs: crash counts (K), expected crash frequency $E(k)$ and a weighing factor (w). Obtaining first input is straightforward as we can directly extract it from a historical crash dataset. For the latter two inputs, they depend on the choice of crash modeling approach. Note that the weighing factor "w" depends on the precision of estimate from a crash model. The larger the variance less is the weight given to the model estimate and more on site specific crash counts; and vice versa for the case of smaller variance. Similarly, it has an inverse relation to the model estimate.

In the road safety field, it is well known that among the various options available for modeling crashes, the NB model is the most extensively used method. In this section, we obtain the two inputs (i.e., E(k) and w) as follows:

- **E(k):** obtain from NB model ($\text{i.e.}, \mu(.)$), as given in Eq. 3-9.
- **Weight (*w*):** As shown in Eq. 3-11, this is a function of the mean (E(k)) and variance (Var(k)). Hauer & Persaud (1987), in their study using NB-based EB method, observed a systematic relation between these two measures. A quadratic function, given by Eq. 3-14, was used to fit the relation. After substituting the values of mean and variance back in Eq. 3-11, the final expression obtained for weight is represented by Eq. 3-15. Since then, in NB-based EB method, it has been a standard procedure to apply this proposed relation of mean-variance to calculate the weights (Persaud et al., 1999; HSM, 2010). It is also noted that there is an inverse relationship between the weight and the over-dispersion parameter and weight and NB estimate.

$$Var(k) = [E(k)]^2 * \alpha \tag{3-14}$$

$$w = \frac{1}{1 + \frac{[E(k)]^2 * \alpha}{E(k)}} = \frac{1}{1 + \mathrm{E(k)} * \alpha} = \frac{1}{1 + \mu(.) * \alpha} \tag{3-15}$$

where, α is over-dispersion parameter

3.4.3 KR-based EB Method

In this section, we propose a KR-based EB method adopting a specification free and data-driven approach of the KR method for estimating crash frequency. Although the fundamental derivation of this proposed nonparametric approach is not based on Bayes' rule; instead, it follows the approach of combining two random variables, we still use the term "KR based EB method" to reflect its similarity with the original NB-based EB framework. The following explains how E(k) and weight (w) are obtained.

- **E(k):** Obtained from an estimate of $\hat{m}(.)$ using from Eq. 3-4.
- **Weights (w):** The steps involved for determining the weights (w) are not as straight forward as in previously discussed NB-based EB method. For this particular method, we trace back to its initial form in Eq. 3-11, where it is represented as a function of variance and mean of crash estimates. The following three steps are needed to determine the weights:
    1. Estimate the variance associated with each site using Eq. 3-13.
    2. Use KR approach to establish a relation of mean and variance. The detailed process is described in Section 3.2. Note that, establishing a mean-variance relation here is a univariate case.
    3. Use Eq. 3-11 to calculate weights associated with each site.
    4. Finally, use Eq. 3-10 to obtain the final KR-based EB estimates by substituting the values of *E(k)* and *w*. Figure 3-3 provides an illustration of steps from 1 to 4 which is also applicable to NB-based EB method.

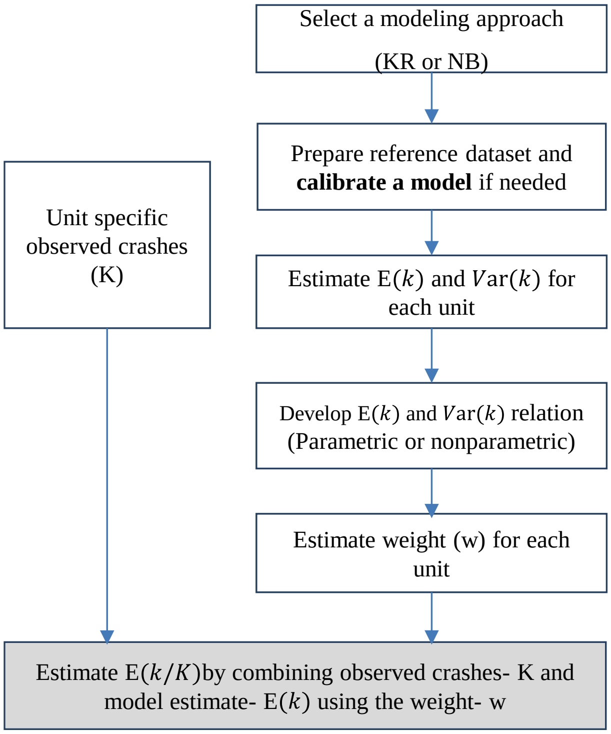

Select a modeling approach
(KR or NB)
Prepare reference dataset and **calibrate a model** if needed
Estimate E(k) and Var(k) for each unit
Develop E(k) and Var(k) relation (Parametric or nonparametric)
Estimate weight (w) for each unit
Unit specific observed crashes (K)
Estimate E(k/K)by combining observed crashes- K and model estimate- E(k) using the weight- w

Figure 3-3: Steps to determine EB estimate

3.5 Summary

In this chapter, we proposed kernel regression (KR), a data-driven nonparametric approach, for crash modeling. The problem of how to select the appropriate variables to be included in a KR model has not been explored extensively. We proposed a variable selection algorithm which is able to detect the importance of variables at their individual levels. We also discussed the negative binomial (NB) model, one of the most commonly used parametric models, as it is employed in later chapters for comparisons with the KR method. Another extensively used estimation method in road safety studies is the Empirical Bayesian (EB) method for which the NB model has been the most extensively used model. We introduced an EB extension of the KR method so that two important measures of crash risk (site-specific crash history and estimates from a crash model) can be combined in the final estimate.. The main motivation was to improve crash modeling through a data-driven technique. It should be noted that, similar to the NB-based EB model, the proposed KR-based EB method still subject to the issue of using the crash data twice - one for regression modeling and the other for EB adjustment.

Chapter 4
Performance Comparison of Parametric and Nonparametric Crash Modeling Techniques

In road safety studies, parametric models have been the most popular choice for predicting crash risk. While parametric models are relatively convenient to apply and comprehend due to their defined mathematical functions between crashes and potential explanatory variables (e.g., traffic exposure, geometric features), pre-specifying such relations is one of their critical issues. This could easily be overcome by implementing a nonparametric approach where no prior specification of a model form is required. However, this approach is often characterized as a data-hungry technique, thereby demanding a larger data set. The good news is that crash data for road safety analysis and modeling are growing steadily in size due to recent advancements in information and sensor technologies, thereby motivating us to explore the nonparametric methods.

In this chapter, we apply two popular techniques from the two approaches: negative binomial models (NB) for the parametric approach and kernel regression (KR) for the nonparametric counterpart. Our main objective is to compare performance of these two methods and investigate how their relative performance varies with the data size. This helps answer our research question that whether or not we could benefit from adopting a nonparametric approach to road safety analysis in a scenario of growing crash data. For this, case studies consisting of three large crash datasets: hourly winter road crash dataset from 31 patrol routes in the province of Ontario in Canada, yearly crash dataset from Highway 401 of Ontario in Canada and yearly crash dataset from the rural highways of Colorado State in the U.S. were used. Prior to this comparative study, we present results of variable selection for the KR method, which is based on the proposed algorithm described in Chapter 3. Furthermore, we extended our analysis to compare how the KR and NB methods perform for modeling the relationship of crashes and predicting variables by studying their individual effects on crash frequency.

4.1 Description of Crash Datasets

This section briefly describes the crash datasets, including the data sources and the steps involved in data processing. These datasets are used in different case studies for performance comparison of the KR and NB methods.

4.1.1 Crash Dataset 1: Highway 401 (Multilane Access Controlled Highway)

This dataset consists of historical crash data from 2000-2008, along with traffic and road geometric data from Highway 401 in Ontario, Canada, one of the busiest highway in North America (map in Appendix B.1). The highway extends across the Southern, Central and Eastern regions of Ontario from Quebec in the east to the Windsor-Detroit international border in the west. According to the 2008's traffic volume data, the annual average daily traffic (AADT) ranges from 14,500 to 442,900, indicating an extremely busy road corridor. Its total length is 817.9 km of which approximately 800 km was selected for this study. The details on the data sources and processing steps are explained below.

Data Sources

The databases used in this dataset include: 1) historical crash records from 2000 to 2008 extracted from MTO's Accident Information System (AIS); 2) historical AADT data for the same years from MTO's Traffic Volume Inventory System (TVIS); and 3) road geometric features from MTO's Highway Inventory Management System (HIMS) database. Note that each record in these databases is referenced to MTO's linear highway reference system (LHRS), a one-dimensional spatial referencing system with a unique five-digit number representing a node/link on a particular highway. LHRS can be used to locate the position of features on a map using a Geographical Information System (GIS) tool.

Road Segmentation and Geocoding

Crashes are aggregated on an annual basis over the individual homogenous sections (HSs), each of which represents a segment with similar characteristics such as number of lanes, shoulder width, the presence of median, curvatures, and other roadway features. As previously mentioned, all the data (crash, road geometry and traffic data), are spatially referenced to MTO's LHRS system. All features were geocoded in the GIS map through a multi-step procedure (see Figure 4-1). First, the geometric features from the HIMS database, which was in a spreadsheet format, were geocoded in the GIS platform. There were a total of 244 records in the HIMS layer, each representing a road section with a set of uniform road geometry features. However, as the road curvature was missing in the database, further geoprocessing was needed to obtain the final set of HS sections. For this, curve sections were first demarcated on a map using a GIS tool, thus generating a curve layer. This tool automatically created an attribute table for the curve layer with detailed information such as LHRS number, start point, length and radius related to each curve. For a refined set of HS sections, the initial HIMS layer was split at the intersection of the curve layer, and the segmented HIMS layer was spatially joined to the curve layer in order to transfer all the curvature related information. As each road section's initial HIMS may have one or more curvature sections, these were disaggregated into smaller subsections thereby including an additional level of homogeneity. Note that the shortest length of HS section was 0.2 km. This selection of a certain lower threshold value complies with literature as it had been suggested that very short road segments might have higher uncertainty and lower exposure problems (Council and Stewart, 1999; Begum et al., 2009; HSM, 2010; Ahmed et al., 2011). Finally, these HSs, assigned with unique IDs, were used as the spatial unit for integrating crash and traffic volume data. There are a total of 418 unique HS sections covering 800.03 km, or 97.9%, of Highway 401.

Data Aggregation and Integration

Crash data are stored in two different databases: one on property damage only (PDO) crashes and the other on fatality & injury (FI) crashes. The PDO records are managed at the vehicle level, whereas the FI crashes are organized at the person level. The summary result presented in Appendix B.2 showed that the total PDO crashes were approximately three times that of total FI crashes. These two datasets were then combined into one dataset with only crash level information extracted. Note that each crash is represented by a unique ID. The extracted crashes were geocoded using the linear referencing tool in a GIS platform, resulting in a crash layer (Appendix B.3). This crash layer was spatially mapped to the previously generated HS layer, thereby associating each crash to a specific HS section. Finally, the crashes were aggregated by individual HSs on annual basis. The distribution of crash counts is presented in Appendix B.4.

Traffic count data consist of AADT and average annual commercial vehicle counts for the period 2000-2008. As each observation recorded LHRS and offset information, the traffic counts were spatially located using the linear referencing GIS tool. Each HS was then assigned the nearest traffic observation. Note that a total of 170 traffic counting stations were available for the 418 HSs. Approximately 85% of the HSs have traffic values assigned from its nearest count station within 2 km distance indicating that the coverage of traffic monitoring was quite extensive on this particular highway (Appendix B.5). Finally, the processed crash and traffic data were integrated into a single dataset with HS and year as the mapping fields that resulted in a total of 3762 records. Table 4-1 shows the summary statistics of final processed dataset, and the distributions for individual factors included in the dataset are given in Appendix B.6.

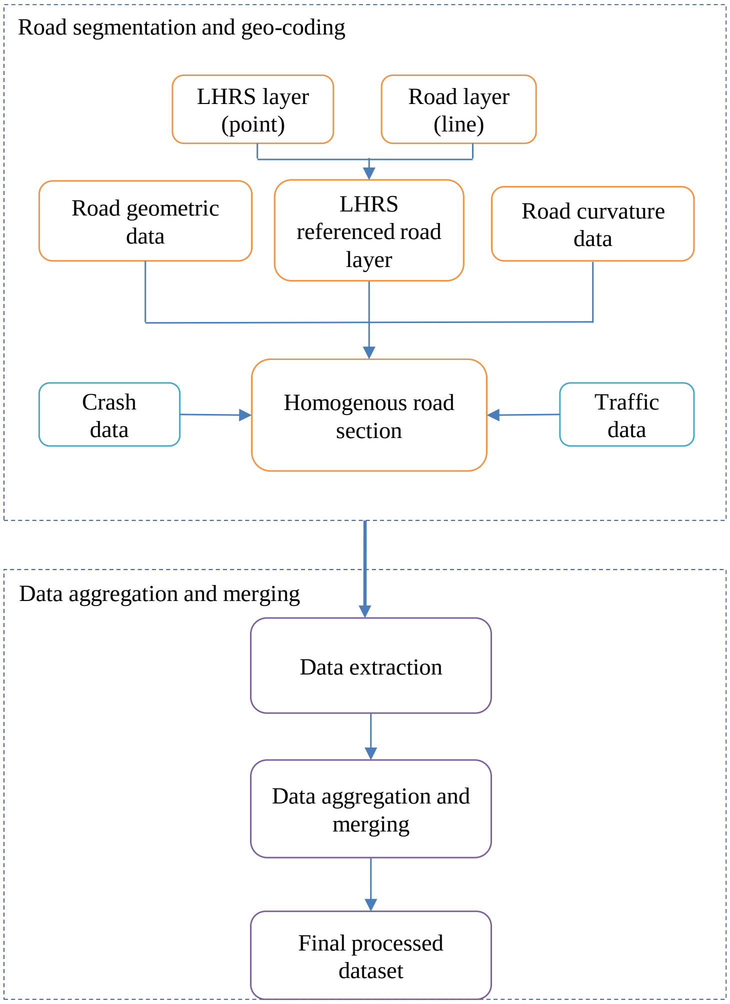


Figure 4-1: A framework for data processing and integration

Table 4-1: Descriptive Statistics- Dataset 1

	Total Crash (per year)	AADT (veh/day)	Exposure (million vehicle-kilometer)	Commercial AADT (veh/day)
Mean	23.81	76633	41.79	13993
St.dev.	50.02	91476	54.05	6719
Min	0	12000	1.66	0
Max	468	442900	611.41	42076

Sample size	3762	3762	3762	3762
	Median width (m)	Shoulder width-right (m)	Curve deflection (1/km)	shoulder width-left (m)
Mean	11.11	3.14	0.19	1.60
St.dev.	6.14	0.28	0.35	1.19
Min	0.60	2.60	0.00	0.00
Max	30.50	4.00	1.86	5.19
Sample size	3762			

4.1.2 Crash Dataset 2: Ontario Multilane Highways

This dataset was originally prepared by Usman et al. (2012) for winter road safety analysis. The dataset consists of historical crash data for six winter seasons (2000-2006), along with traffic count, weather, and road surface condition data from different sources for 31 highway patrol routes in Ontario, Canada (map in Appendix B.1). These selected patrol routes belong to either Highway 1 or 2 and are used as the spatial analysis unit for processing the data. A brief description on data sources and processing steps is given below.

Crash data come from Ministry of Transportation, Ontario (MTO), and are originally collected by the Ontario Provincial Police. This database includes information about each crash at personal level including crash time, crash location, crash type and severity, weather and road surface conditions. Hourly traffic count data was extracted from loop detector data obtained from MTO’s COMPASS system and permanent data count stations. The average value was taken for highway sections with multiple count data. Similarly, hourly weather data such as temperature, precipitation, visibility, wind speed were collected from nearby Road Weather Information System and Environment of Canada weather stations. The Road Surface Index (RSI) variable was constructed as a surrogate measure based on MTO’s road surface condition weather information system. It measures the frictional levels of road sections. All these data sets were merged into a single hourly data set using date, time and location as the basis of merging for each selected highway section. Finally, only the hours defined by snow storm events for the given six winter seasons were considered. A summary of the dataset is presented in Table 4-2, and the distribution of each factor included in the dataset is given in Appendix B.7.

Table 4-2: Descriptive Statistics- Dataset 2

	Crashes	Temp (C)	Wind speed (km/hr)	Visibility (Km)	Precipitation (cm/hr)	RSI	Exposure (*10000) VKT)	Length (Km)
Mean	0.02	-5.12	16.28	11.16	0.24	0.745	5.7	58.08
St.Dev	0.18	5.56	9.62	7.91	0.37	0.197	8.08	33.2
Min	0	-33.55	0	0	0	0.05	0.004	12.9
Max	7	28	69	40.2	13.8	1	154.59	139.5
Sample size: 122058								

VKT is vehicle kilometer traveled, RSI is road surface index

4.1.3 Crash Dataset 3: Colorado Two-lane Rural Highways

This dataset contains crash data from rural two-lane highways in the Colorado State, U.S., and it was downloaded from http://extras.springer.com (Hauer, 2015). The reasons for using this dataset are as follows. First, it represents a case of two-lane rural roads with an average AADT of approximately 2200, which is significantly lower than the two multiplane highway cases described previously. Second, this dataset covers a total of 4593 unique sections (section length larger than 200 m) with observations from 1991 to 1998, which can be considered to be large in sample size. A summary of the dataset is given in Table 4-3, and the histograms showing the distribution of included factors are given in Appendix B.8.

Table 4-3: Descriptive Statistics- Dataset 3

	Total Crash (per year)	AADT (veh/day)	Length (km)
Mean	0.9	2217	2.1
St.dev.	2.2	2534	2.4
Min	0.0	40	0.2
Max	54.0	21720	31.8
Sample size	36743		

4.2 Comparing Crash Models

This section provides the modeling results for the previously presented three datasets including model coefficients for NB model, bandwidths for KR method and their goodness-of-fit measures. Eq. 4-3 to 4-5 present the summary results of crash models with their details included in Appendix B.9. Note that the results of the KR method do not have any reportable model forms like in the NB models, as this method follows a fully data-driven approach for predicting crashes. At this initial stage of modeling, the model variables are selected based on some prior evidence on their safety effects from the past road safety studies. However, we will discuss more about their individual effects and the selection process in next section. Note that in this thesis, we used a statistical software platform "R" wherever required by the methods (http://www.r-project.org/).

To compare the performance of the two approaches, two goodness-of-fit measures, namely, mean absolute error (MAE) and root mean square error (RMSE), are used. As given by Eq. 4-1 and 4-2, these measures quantify the average deviation of the estimated crash frequencies from the observed values. Therefore, smaller the magnitude of these measures better is their performance level. Note that the difference between these two performance measures is how the residuals are weighted. In MAE, equal weights are given to the residuals from the observed points, whereas in RMSE larger residuals are given greater weights by squaring the deviation. For example, an estimation that is two units off the observed data produces a weight four in RMSE compared to two in MAE. As a result, RMSE is always greater than MAE. When the values of MAE and RMSE are close, we can also conclude that the residuals are more concentrated to unit values.

$$MAE = \frac{\sum_{i=1}^{n}|\hat{y}_i - y_i|}{n} \qquad (4\text{-}1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \qquad (4\text{-}2)$$

where,

y_i= ith observed crash frequency,

$\hat{y}_i$ = estimated crash frequency for i^{th} observation, and

n = total number of observations.

Case 1: Highway 401

- $\mu^{NB_1} = e^{-1.04+0.82\ \ln(exposure)+0.001AADT_C-0.02MW-0.09SW_l+0.16SW_R-0.17CD}$ (4-3)

where,

μ^{NB_1} = crash frequency (per year),

exposure= million vehicle kilometer travelled,

$AADT_C$= commercial AADT (veh/day),

MW= median Width (m),

SW_l= shoulder width- left (m),

SW_R= shoulder width – right (m),

CD= curve deflection (1/km).

- Bandwidths for the KR method: 21.08, 2621, 2.39, 0.465, 0.11, 0.13 (variables are in the same order as in the NB model)

  (MAE^{NB}= 11.86; $RMSE^{NB}$= 26.64, MAE^{KR}= 7.34, $RMSE^{KR}$= 14.81)

Case 2: Ontario multilane highways

- $\mu^{NB_2} = e^{-2.58+0.72\ln(exposure)-2.83RSI-0.01P-0.04V+0.01WS-0.0001T}$ (4-4)

where,

μ^{NB_2} = crash frequency (per hour),

exposure = vehicle kilometer travelled (‘0000),

RSI = road surface index,

P = precipitation (cm/hr),

V = visibility (km),

WS = wind speed (km/hr),

T = temperature (C).

- Bandwidths for the KR method: 2.23, 0.054, 1.53, 2.65, 2.17, 1.53 (variables are in the same order as in the NB model)

  (MAE^{NB}= 0.046; $RMSE^{NB}$= 0.178, MAE^{KR}= 0.031, $RMSE^{KR}$= 0.137)

Case 3: Colorado two-lane rural highways

- $\mu^{NB_3} = e^{-8.03+0.95ln(AADT)+1.07ln(L)}$ (4-5)

where,

μ^{NB_3}= crash frequency (per year),

AADT= vehicle per day,

L= length of highway section (km).

- Bandwidths for the KR method: 423.48, 0.4 (variables are in the same order as in the NB model) *(MAE^{NB}= 0.781; $RMSE^{NB}$= 1.529, MAE^{KR}= 0.752, $RMSE^{KR}$= 1.333)*

As seen in Eq. 4-1 to 4-3, the KR method performed better in comparison to the NB model (measured in terms of MAE and RMSE) in all the case studies. This could be due to the fact that the KR method does not impose any functional form on the relationship between the expected crash frequency and the predicting variables, thereby allowing its data-driven process to capture the underlying shape of the relation. On the other hand, this flexibility may have been restricted in the NB model due to its need for the pre-specification of model form (here the exponential form). However, the performance comparison in this section is based on the entire datasets, i.e., without holding a validation set. Therefore, to address this issue, we apply a bootstrap validation approach, which is discussed in Section 4.4. Prior to that, we will first discuss the variable selection process of the KR method based on the algorithm we proposed in Chapter 3.

4.3 Variable Selection for KR Method

As discussed in Chapter 3, a bootstrap-based algorithm was proposed to solve the problem of variable selection in the KR method. This algorithm determines the impact level of each variable as it feeds into the model, using an indicator called relative Variable Importance (VI'). A larger value of VI' means that the given variable is relatively more important or influential in predicting crash risk. We applied the algorithm to all the previously presented crash datasets, and meanwhile, compared the results to their corresponding variable selection process in the parametric counterpart, i.e., the NB model.

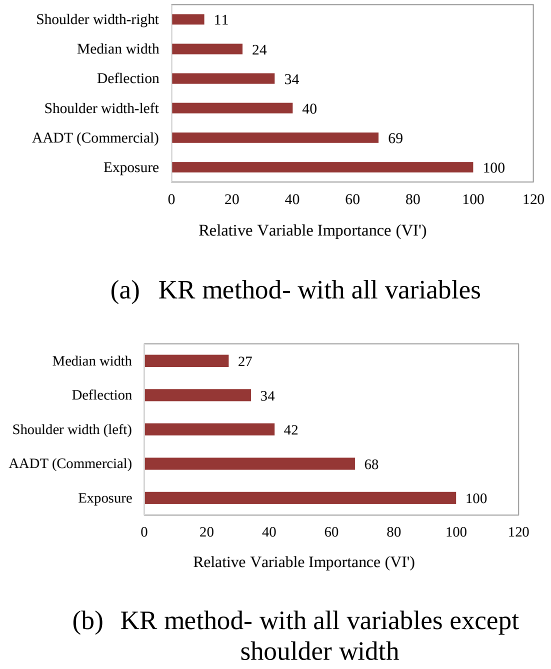
4.3.1 Variable Selection: Case Studies

Case study 1: The data split for the algorithm is as follows: (1) training set – crash data from 2000-2006 (90% of the data is randomly selected for each bootstrap training set), (2) validation set – crash data from 2007-2008. There are a total of six potential variables that are crash-related whose impact levels need to be determined. Figure 4-2 (a) presents relative variable importance (VI') of each variable after applying the bootstrap-based algorithm. As seen, the exposure factor appears to be the most influential variable whereas the shoulder width the least influential one.

Furthermore, we compare the magnitude of VI' of each variable to its corresponding t-value obtained from the parametric NB model. The t-value in the NB model is used to test a null hypothesis to infer whether or not a given variable has a significant effect on the dependent variable - crash frequency.

The larger t-value (absolute) implies that the given variable has a higher impact level, whereas the value close to zero implies that its effect is negligible. The results from the NB model show that all the variables are significant at 5% significance level. Meanwhile, comparing the impact levels of the variables (VI's) in the KR method with the t-values in the NB model, we observed a similar trend (Figure 4-2(c)). For example, the exposure variable in both methods has the highest explanatory power as indicated by their respective measures, i.e., t-value of 49.91 in the NB model and VI' value of 100 in the KR method. Similarly, both methods showed the safety effect of shoulder width to be minimal, i.e., t-value of 3.32 in the NB model and VI's of 11 in the KR method. By referring to a relatively high VI's values in Figure 4-2 (a), we make an intuitive decision to select all the variables for the KR method. We also confirmed this based on the findings that it has the highest performance compared to those using other subsets of variables.

We also conducted a hypothetical analysis on the effect of excluding the least important variable. As in this particular case study, shoulder width has the least VI', therefore, this factor was excluded. This is similar to the backward elimination process of a parametric modeling technique. The updated result for this hypothetical scenario is shown in Figure 4-2 (b) where their relative effects are found similar to the previous result in Figure 4-2 (a) with a slight change in their magnitudes.



(a) KR method- with all variables

(b) KR method- with all variables except shoulder width

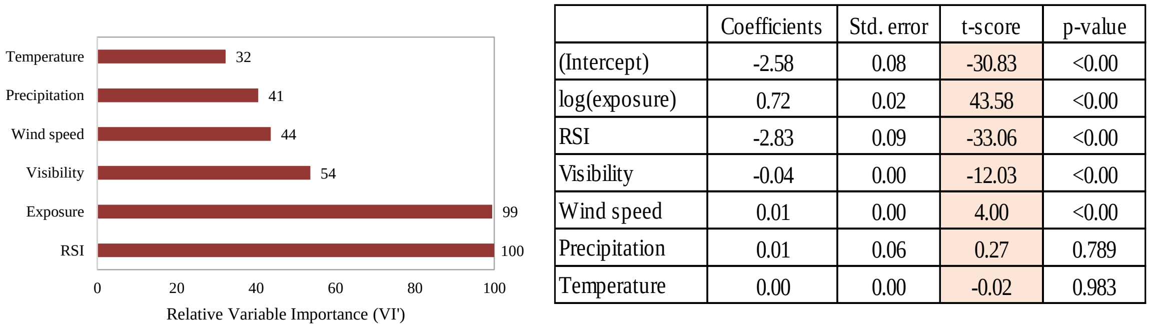
Coefficients	Estimate	Std. error	t-value	P-value
(Intercept)	-1.04	0.16	-6.52	<0.001
log(Exposure)	0.82	0.02	49.91	<0.001
AADT (Commercial)	0.00	0.00	20.57	<0.001
Median Width	-0.02	0.00	-6.16	<0.001
Shoulder width (left)	-0.09	0.01	-8.44	<0.001
Shoulder width (right)	0.16	0.05	3.32	<0.001

(c) NB model- with all variables

Figure 4-2: Variable selection: (a) and (b) KR method, (c) NB model**.**

Case Study 2: The data split for the algorithm is as follows: (1) training set – crash data from 2000-2004 (90% of the data is randomly selected for each bootstrap training set), (2) validation set – crash data from 2004-2006. The proposed bootstrap-based algorithm was applied using these training and validation sets. There are six variables that could potentially cause traffic crashes. Note that this particular case study represents road safety in winter conditions where poor road surface condition is expected to impose a relatively high crash risk. Figure 4-3 (a) presents the result of relative variable importance (VI') of each variable, where, as expected, the exposure and RSI factors seem to have a relatively large effect on crash risks. Meanwhile, other variables, such as temperature, precipitation, visibility and wind speed, also showed high VI's (Figure 4-3 (a)). Also, the optimum performance of the KR method appeared when all the variables were included.

Similar to the previous case study, we also compared the VI' of each variable to its corresponding t-value from the NB model. As shown in Figure 4-4(b), the orders of the variables' influence on crash risk are overall similar. Regarding the significance of variables tested in the NB model, temperature and precipitation are insignificant at 5% significance level. This might be due to the existence of some nonlinear relations between these variables and the crash frequency as detailed in Section 4.5. Note again that the KR method is capable of capturing nonlinear relations, which might explain why the method had yielded high VI's for these two variables.



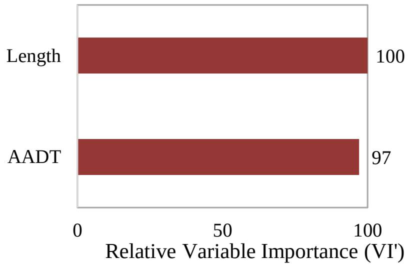
	Coefficients	Std. error	t-score	p-value
(Intercept)	-2.58	0.08	-30.83	<0.00
log(exposure)	0.72	0.02	43.58	<0.00
RSI	-2.83	0.09	-33.06	<0.00
Visibility	-0.04	0.00	-12.03	<0.00
Wind speed	0.01	0.00	4.00	<0.00
Precipitation	0.01	0.06	0.27	0.789
Temperature	0.00	0.00	-0.02	0.983

(b) NB model with all variables

(a) KR with all variables

Figure 4-3: Variable selection- (a) KR method; (b) NB model.

Case Study 3: The data split for this case study is as follows: (1) training set – crash data from 1991-1996 (90% of the data is randomly selected for each bootstrap training set), (2) validation set – crash data from 1997-1998. Similar to previous case studies, the proposed bootstrap-based algorithm was applied using these training and validation sets to determine the significance of two variables – Traffic and length. Figure 4-4 (a) presents the result of relative variable importance (VI’) of each variable, where length appears to have a slightly larger effect than the AADT on crash risks. Similar trend is observed from the NB model with both variables appearing significant at 5% significance level (Figure 4-4(b)).



(a)

	Coefficients	Std. error	t-score	p-value
Intercept	-8.03	0.07	-121.00	<0.001
ln(AADT) (veh/day)	0.95	0.01	115.60	<0.001
ln(Length) (km)	1.07	0.01	119.70	<0.001

(b)

Figure 4-4: Variable selection- (a) KR method; (b) NB model

4.3.2 Variable Selection: Simulation Study

We also conducted a simulation study to test the robustness of the proposed algorithm as the simulated results can be easily validated by comparing them to their true values. For this, we first generated two datasets by assuming two different parametric model forms. Then, the previously described algorithm was applied to quantify effect of each variable on the dependent variable and the results obtained were compared to the parametric models by re-calibrating their individual models using the original model specifications. Note that a re-calibrated parametric model represents the true relation between dependent and independent variables. The paragraphs below provide a description of the two simulated datasets.

Linear model: A linear model consisting of three predicting variables- *X* (i.e., x_1, x_2 *and* x_3) - of varying magnitude of impacts on a dependent variable (*y)* was assumed (Eq. 4-6). *X* variables were randomly generated using a normal distribution with mean of 10 and standard deviation (sd) of 5. Meanwhile, some noise was added to the model by assuming a normal distribution (mean= 0, sd= 5). Finally, a new external variable (x_4) was introduced (mean= 10, sd= 5). A sample of 1000 observations were simulated from the assumed model setting. A summary of the dataset is provided in Appendix B.10.

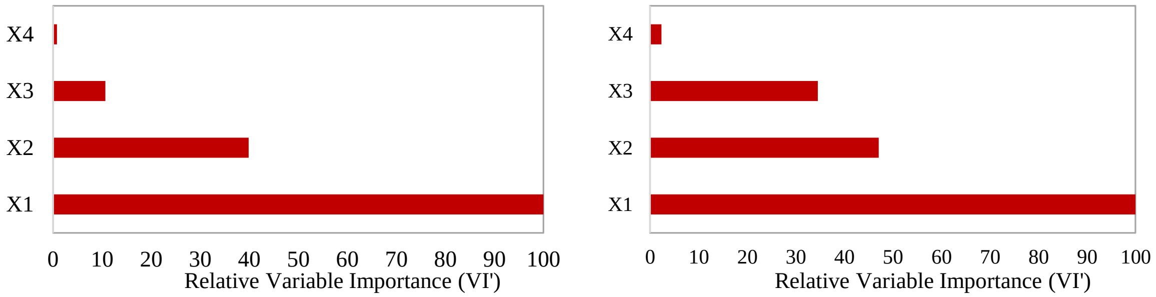
$$y = 10x_1 + 5x_2 + 2x_3 + error \tag{4-6}$$

Nonlinear model: An exponential form of model given by Eq. 4-7 was considered. Predicting variables were randomly generated using a normal distribution (mean= 20, sd= 5). Meanwhile, some noise was added in the model by assuming it to follow a normal distribution (mean= 0 and sd= 0.5). Similar to the previous dataset, a new external variable (*x_4*) was introduced by assuming a normal distribution (mean= 20, sd= 5). A total of 1000 observations were simulated from the assumed model setting. A summary of the dataset is provided in Appendix B.11.

$$y = e^{0.05x_1 + 0.03x_2 - 0.005x_3 + error} \tag{4-7}$$

The training and validation sets required in the algorithm were obtained by randomly splitting the given dataset into two groups containing 80 and 20 percent of the data, respectively. Figure 4-5 (a) and (b) present the results of VI' of each variable for the above mentioned linear and nonlinear models, respectively (VI is presented in Appendix B.9 (b) and (d)). As seen, the relative magnitude of VI' of variables are such that the *x_1* has the highest influence, followed by x_2, x_3 and x_4 variables. This shows a strong correlation between the impact levels of variables indicated by VI's and their corresponding coefficients in their original models. It is also noted that the variable "x_4", which was not a part of the original models, shows a negligible effect.

Furthermore, we validate the performance of the algorithm by comparing VI's of variables to their corresponding t-values obtained from the parametric models. For this, models were re-calibrated following the same specifications as in their original forms i.e., linear and nonlinear (Eq. 4-6 and 4-9 respectively). Table 4-4 shows the model coefficient associated with each variable and their corresponding t-value. As seen, there is a strong correlation between these measures- VI's and t-values. For example, *x_1* appears to be the most influential variable in both methods as indicated by its highest VI' and t-value. Similarly, the *x_4* variable appears to be insignificant in 95% confidence interval in the parametric models and its corresponding VI' appears to be close to the zero value, suggesting its exclusion from the final model set.



(a)
Linear model

(b)
Non-linear model

Figure 4-5: VI's of variables- linear and nonlinear models

Table 4-4: Summary results of calibrated models

Variable	coefficient	Std. error	t-value	p-value	Variable	coefficient	Std. error	t-value	p-value
1) Linear model					**2) Nonlinear model**				
X1	10.03	0.03	367.19	<0.001	X1	0.05	0.00	75.77	<0.001
X2	5.00	0.03	185.30	<0.001	X2	0.03	0.00	43.82	<0.001
X3	1.98	0.03	72.09	<0.001	X3	-0.01	0.00	-9.52	<0.001
X4	0.00	0.03	0.11	0.917	X4	0.00	0.00	-0.23	0.822

4.3.3 Summary: Variable Selection

While the data-driven nonparametric KR method can be considered as an alternative approach to crash modeling, the issue of selecting what variables to include has been relatively less explored. To address this issue, we proposed a bootstrap-based algorithm (in Chapter 3) in which an indicator– relative Variable Importance (VI') is used to measure the impact level of each variable. Most importantly, this indicator in the algorithm is estimated by applying the KR method itself.

The previous sections presented the results of studies comparing the proposed algorithm with its parametric counterpart. First, three real case studies were considered to test the algorithm. Second, simulated datasets were used to validate the results as the parametric approach represented the true state. In all these cases, strong correlations were observed between the VI' measures from the algorithm and the t-values generated from their corresponding parametric models. From this, we can make the following two conclusions. First, the proposed algorithm appears quite robust in capturing impact levels of variables at their individual levels. Second, we could employ a parametric model for selecting important variables in a nonparametric method. However, the result from this short-cut approach is expected to be less biased only when the model specification of a selected parametric model is relatively accurate.

4.4 Bootstrap-based Performance Comparison[3]

In this section, we propose a bootstrap-based validation approach to complement the performance comparison of the nonparametric and parametric models presented in Section 4.2. In a commonly applied validation approach, the original dataset is split into two sets: the first, known as a training set, is used for model calibration, and the second, known as a validation set, is used for testing. However, we extend this traditional approach in two aspects. First, we adopt a bootstrap-based validation (BV) approach in which the standard validation process is designed to repeat a large number of times (here 100 for small dataset and 25 for large dataset) with training set being selected randomly in each step. Second, we also evaluate the sensitivity of data size to the model performance (both KR and NB methods), by sub-setting the original training set into different sample sizes. The following eight steps describe the proposed validation approach and the flow chart related to these steps is given in Appendix B.13.

1. Split a given dataset into two subsets, one for calibration (i.e., training set), and other for validation (i.e., validation set). The following are the case study specific splits.
    - Case study 1– Highway 401, Ontario: First seven years of crash data (2000-2006) as a training set and last two years (2007- 2008) as a validation set. The sample size for the training set is 2926 and the validation set is 836.
    - Case study 2– Patrol routes, Ontario: First four years of crash data (2000-2004) as a training set and last 2 years (2004- 2006) as a validation set. The sample size for the training set is 85183 and the validation set is 36875.
    - Case study 3– Two-lane rural highway, Colorado State: First six years of crash data (1991-1996) as a training set and last 2 years (1997-1998) as a validation set. The sample size for the training set is 27558 and the validation set is 9186.

[3] Thakali, L., Fu, L., & Chen, T. (2016). Model Based versus Data-driven Approach for Road Safety Analysis : Does More Data Help? Transportation Research Record: Journal of the Transportation Research Board, No. 2601.

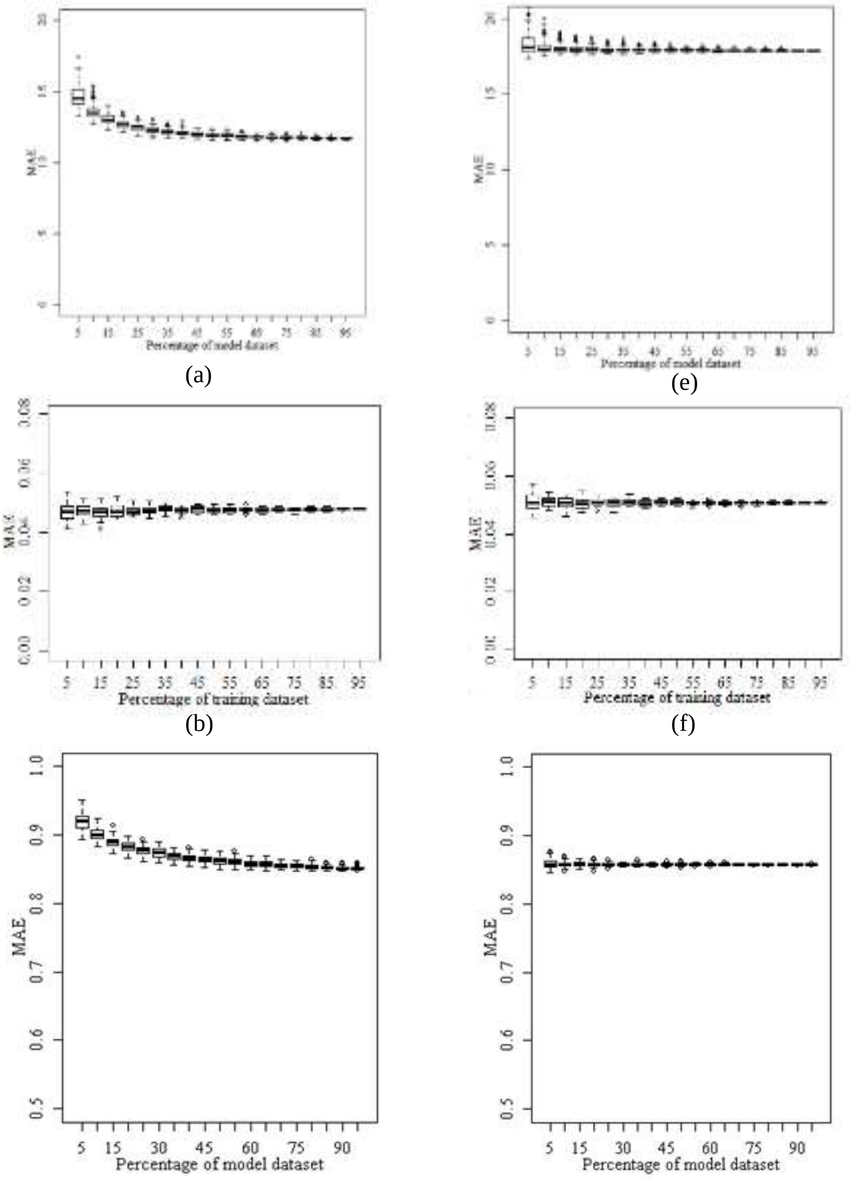
2. For the model set, specify split percentage (s) starting at 5%. This results in a certain sample size.
3. Select the final model set randomly as "s” % of the original training set defined in step 1.
4. Using the model set in step 3, estimate performance measures (MAE and RMSE) for the validation set.
5. Repeat steps 3 to 4 100 times to generate bootstrapping samples. As we randomly select the final model set in Step 3, MAE and RMSE are expected to vary.
6. Increase the split percent “s” by 5% and go to step 2.
7. Repeat steps 2 to 6 until “s” is 95%.
8. Finally, for each split, calculate the percentage of times that the KR method outperformed NB (i.e., out of 100 samples).

Figure 4-6 shows the results of bootstrap-based validation using boxplots for all three case studies discussed in Section 4.2. Figure 4-6 (a), (b) and (c) are the boxplots of performance measure MAE using KR method for case studies of Highway 401, the 31 highway patrol routes and the two-lane rural highways, respectively, whereas, Figure 4-6 (d), (e) and (f) are the boxplots of MAE for their corresponding NB models. Results for the RMSE performance measure are presented in Appendix B.14.

From the first case study (Figure 4-6 (a) and (d), we can draw two important findings. First, on average, regardless of the sample size, the KR method has higher estimation accuracy compared to the NB model. Secondly, the performance of the KR method increases with increasing sample size. In contrast, the average performance of the NB model varied little with change in the sample size (although its reliability did improve as in the KR method). Similar trends were observed in the case of Colorado rural highways except that at the lower data size (<30% split) the KR method showed lower performance. Meanwhile, in the case of the 31 patrol routes, the performance of both methods were less sensitive to the data size. One of the reasons could be due to its relatively large data size when compared to the two other crash datasets. The overall findings presented in these boxplots suggest that the nonparametric KR approach sensitive to the data sample size as compared to its parametric counterpart.

The boxplots that presented results of 100 BVs for each model split[4] (i.e., each data size) are further summarized by calculating a new performance indicator measured as the percentage of times the KR method outperformed the NB model. Figure 4-7 illustrates this comparison results for all the case studies. The results show that in case study first and second, at all the sample sizes, the KR method outperformed the NB model in all bootstrapping instances. Similarly, Figure 4-7 illustrates the result for the two-lane rural road where the comparison indicators are found to increase in the direction of increasing sample size, suggesting that KR performance is correlated with sample size. Therefore, this result further confirms the sensitivity of the KR method to data sample size. The larger the data size available, the higher is its estimation accuracy.

[4] 25 BV in case study 2



(a)

(e)

(b)

(f)

(c)

(g)

Figure 4-6: Boxplots- (a), (b) and (c) represent MAE of KR for case study 1, 2 and 3, respectively; (d), (e) and (f) represent MAE of NB for case study 1, 2 and 3, respectively

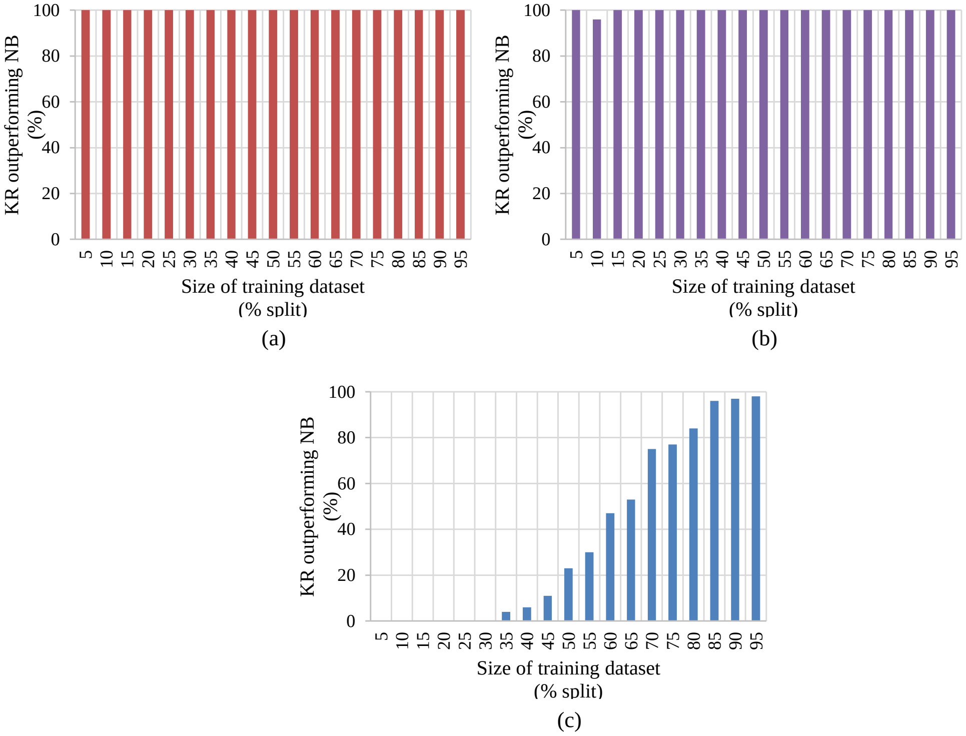


Figure 4-7: Bootstrap validation (BV) results: (a) Case study1- Highway 401; (b) Case study 2- Patrol routes; (c) Case study 3- Two-lane rural highway

In summary, this section conducted a comprehensive comparative study of nonparametric (KR) and parametric (NB) approaches by using three relatively large datasets, all related to road safety. Bootstrapping validation results showed that the KR method has comparatively better performance compared to the NB model. This could be due to the advantage of adopting a data-driven nonparametric approach used in the KR method. Furthermore, this section also investigated the question of how the relative performance of these two alternative approaches changes as a function of data size. The findings suggested that the KR method performance increases significantly with the growing sample size, unlike the NB model. Based on this finding, a spectrum of crash estimation methods could be recommended that varies according to data size. If the spectrum were arranged according to data size, with the left side having smaller data sizes and the right side having larger data sizes, then the NB model calibrated from the maximum likelihood (MLE) approach would be located towards the left side of the spectrum while the KR method would be on the right side. Accordingly, NB and similar models calibrated using Bayesian approach would appear at the far left end of the spectrum, as they have been shown to be able to address the problem of relatively small data size in the past studies.

4.5 Comparisons of Factors Effecting Road Safety[5]

This section describes two case studies to demonstrate the differences between the two approaches (i.e., KR and NB) in modeling the effects of various factors on crash risk. For this, we use two crash models from previous case studies presented in Section 4.2. The first case study is Highway 401 with annual crash data. This dataset is used to illustrate the safety effects of various road geometric elements. The second case study is highway patrol routes with hourly winter crash data and this study illustrates safety effects of various weather and road surface related factors that have direct implications on winter road maintenance. Note that in an analysis of a given factor, all the remaining factors are fixed at their mean levels and the result is presented using a regression plot. The following two sections provide case specific findings. It is important to note that our interpretation of the regression plots will focus on the general trends and the regions within which there are sufficient data for reliable estimates from the models. This is especially relevant for the KR method regression curves which may yield local unsmoothed waving and unreliable estimates at the boundary region of the variables.

4.5.1 Case Study 1: Highway 401

Effects of exposures: Figure 4-8 (a) and (b) illustrate the effects of traffic exposures, including total traffic in million-vehicle-kilometers traveled and commercial traffic in AADT, on the expected frequency of crashes on individual sections of Highway 401. As expected, both models show a positive correlation with the traffic level. However, there are differences in capturing the underlying nonlinearity patterns between the two models. For example, in the range of exposure 100 MVK to 200 MVK, the safety effect is relatively constant based on the KR model, while it shows a continuous increasing trend according to the NB model. Likewise, the crash frequency estimated by the KR method is relatively higher at the upper ranges of the exposure as compared to the NB method. Similarly, the KR model shows that the relative effect of commercial traffic is negligible when its volume is below 22500 and it then increases sharply until the AADT reaches 32500 veh per day. However, the NB model

[5] A part of the results is based on “Thakali, L., Fu, L., & Chen, T. (2014). A Comparison between Parametric and Nonparametric Approaches for Road Safety Analysis - A Case Study of Winter Road Safety. In *Transportation Research Board Annual Meeting* (Vol. 6, pp. 1–17)”.

has a smooth and continuous increasing trend with increasing commercial traffic volume. These peculiar nonlinearity patterns that were captured by the KR model appear to make physical sense as the interactions between vehicles are often minimal under low traffic and then increase quickly as traffic reaches to a certain level. This reason has also been used to explain the fact that traffic speed is usually insensitive to traffic volume when the traffic volume is low but decreases quickly when it approaches to the capacity of the road. Overall, the findings from both methods are expected and consistent with those from the literature (Miaou, 1994; Hauer et al., 1996; Usman et al., 2012).

Effects of Road Geometric Features: The road geometric features included in the analysis are three cross-sectional elements: median width, median shoulder width (left) and shoulder width (right), and an alignment element – horizontal curvature. Both methods show a smooth linearly decreasing trend with respect to median width, with the NB model indicating slightly larger effects, which is consistent with those from the past studies (HSM, 2010). However, the effects of median shoulder width (i.e., on left) from the two models are different. The NB model shows a slight negative correlation between crash risk and shoulder width, suggesting that larger shoulder widths are favorable in reducing crash risk. The KR model shows a non-monotonic relationship: adding a one meter shoulder is beneficial in reducing the risk but wider shoulders have a negative effect on safety. For the right shoulder width, results from the KR and NB methods are somehow inconsistent. While the NB model shows an increasing trend of crash risk with widening of shoulder width, an almost constant trend could be observed from the KR model suggesting the insensitivity of crash risk to right-side shoulder width (beyond the commonly used width of three meters).

Horizontal curve deflection (CD) is measured as the reciprocal of curve radius, where small values indicate relatively flat sections and vice versa for the large ones. A straight road section has zero curve deflection. As shown in Figure 4-8 (f), the NB method suggests a decreasing crash risk as curve deflection increases (or decreasing radius). The KR model, however, shows a clear non-linear and non-monotonic relation between crash risk and curve deflection. While overall there is a general trend of decreasing risk with CD (similar to the finding from the NB model), there is a range of CD values (CD = 0.50~0.75) over which curves have in fact a negative effect on safety. The general trend from both models are inconsistent with those from the safety literature which have generally concluded that crash risk should increase as the curve radius increases (Hauer, 1999). This could be due to the fact that the highway being analysed – Highway 401– is a freeway system with a minimum radius greater than 500 meters, which is beyond the sensitive range that has been identified in literature. The non-monotonic patterns identified from the KR model appear to make intuitive senses from a driver's behaviour point of view. For example, it has long been recognized that straight sections are prone to causing driving fatigues and higher speeds. Furthermore, drivers tend to be more alert and cautious when driving on a curved section. Note that these findings from the KR model could have significant implications to the geometric design of highways such as Highway 401.

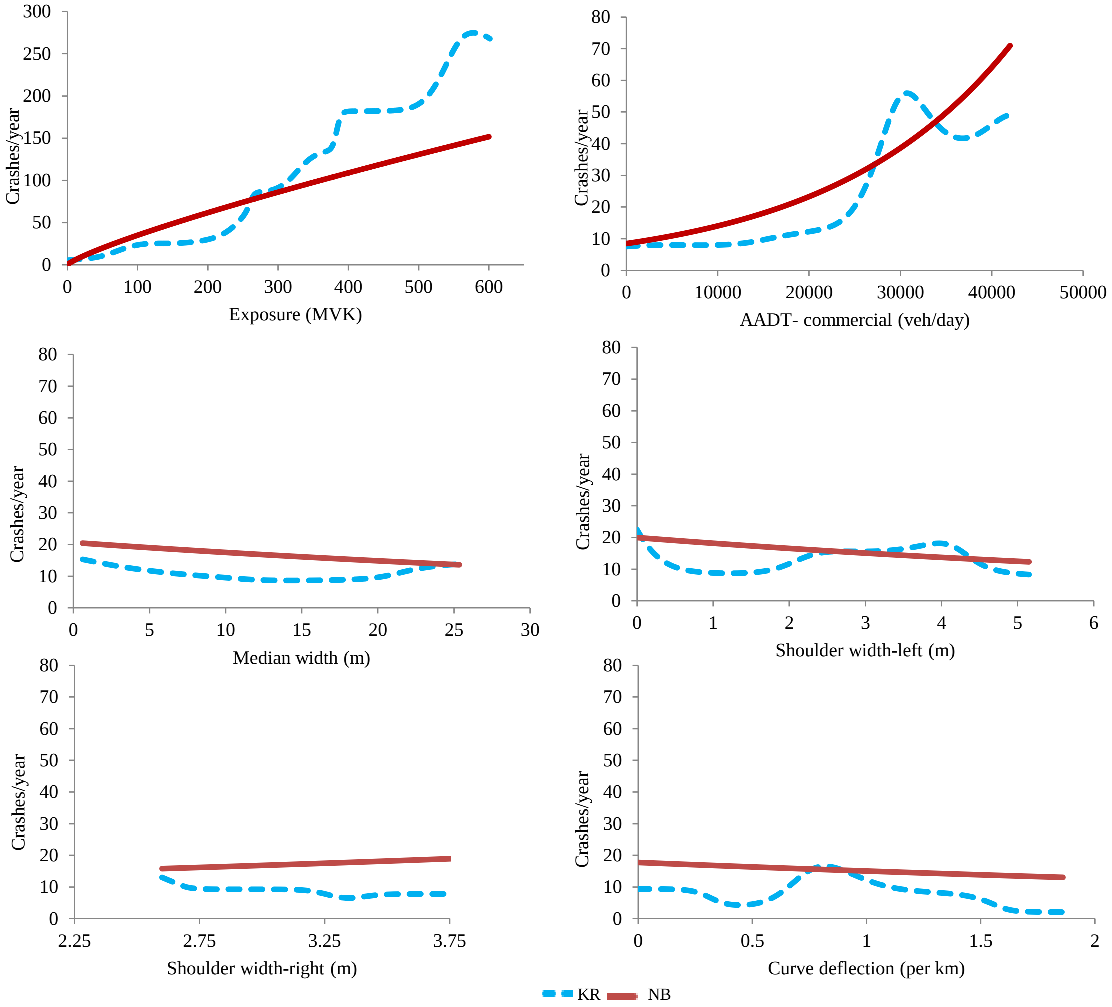


Figure 4-8: Factors affecting crash frequency (Case study 1)

4.5.2 Case Study 2: Highway Patrol Routes

Effect of Road Surface Conditions (RSI): Figure 4-9 (a) shows the effect of RSI on the expected crash frequency from the two modeling approaches. According to the KR method, the average crash frequency is fairly constant for RSI ranging from 0.4 to 1.0. The crash risk starts to increase drastically as the RSI drops below 0.4. Meanwhile, the expected number of crashes estimated by the NB model is much lower than those from the KR method, especially for the low range RSI values. This is mainly due to the fact that the NB being a parametric approach focuses on a global statistical fit to the assumed functional relationship while the KR method places more weights on the local information. Another noticeable difference is that the KR method result shows a clear two-regime relationship with a turning point located around 0.4. Interestingly, literature aimed at determining a relation between crash risk and the frictional level of road pavements has shown a similar turning point. For example, Wallman and Astrom (2001) identified a threshold friction value of 0.45, and suggest that frictions below this value increases the crash risk exponentially (see Appendix B.15). Overall, the nonlinear result revealed by the KR model is important as it could have a significant implication to the establishment of optimal maintenance policy.

Effect of Exposure: Traffic exposure is defined as the total Vehicle-Kilometers Travelled (VKT) as in most past road safety studies (Jovanis & Chang, 1986; Miaou & Lum, 1993; Usman et al., 2012). As expected, both approaches show a general increasing trend in the expected crash frequency with respect to the exposure, as shown in Figure 4-9 (b). This result is consistent with past road safety studies (Jovanis & Chang, 1986; Miaou & Lum, 1993; Miaou, 1994; Hauer et al., 1996; Usman et al., 2012). Additionally, the KR method shows significant nonlinearity. Initially, the crash risk increases linearly until it reaches to the point of 0.015 million VKT, then the trend remains constant between 0.015 to 0.5 million VKT, and finally, rises again. This nonlinearity in relationship could possibility be the reflection of driver's behavior at different exposure level. In contrast, the parametric NB model shows a smooth, uniform increasing trend with respect to exposure, which is likely due to the pre-defined functional form.

Effects of Weather Conditions: Figures 4-9 (c)–(e) shows the effect of weather conditions on the crash risk. Both approaches show a negative correlation of visibility to crash risk (Figure 4-9 (c)), which also confirms from past studies (e.g., Al-Ghamdi, 2007). Comparatively, the effect of visibility is slightly underestimated by the NB model. Similar to the effect of other factors, the NB model shows a smooth linearly decreasing trend with respect to visibility. The KR modeling result, however, shows that the crash risk is sensitive to visibility only at low range values (< 15 km), which makes an intuitive sense. When the visibility reaches to a certain high level, it no longer imposes any effect on driving and thus safety. This nonlinear effect of visibility on safety is not captured by the NB model.

A significant nonlinear relationship between crash frequency and precipitations is captured by the KR method, as illustrated in Figure 4-9 (d). When the precipitation intensity is low (< 0.5 cm/hr), its effect on road safety is fairly minor and constant. After the precipitation rate passes this value, it starts to have a negative effect on crash risk. This trend starts to reverse after the precipitation intensity reaches 1.3 cm/hr. When the precipitation intensity increases passing 2 cm/hr, its effect becomes relatively small. The later patterns may be attributed to the behavior response of the drivers who are likely to drive more cautiously and slowly under heavy snowfalls. In contrast, the parametric NB model shows that the effect of precipitation is negligible throughout the whole range of precipitation intensity, which does not make intuitive sense and contradicts with the results from past studies (Knapp et al., 2000; Andreay et al., 2001).

According to both modeling approaches, the effect of temperature on crash risk is minimal (Figure 4-9 (e)). While past studies have shown mixed results regarding its direction of influence on crash risk, both decreasing (Scott, 1986) and increasing (Antoniou et al., 2013; Karlaftis and Yannis, 2010), no such notable relations are observed in this case study. Similarly, the effect of wind speed on crash frequency is illustrated in Figure 4-9 (f), where both KR and NB methods show a slightly increasing trend, though with low effects. This effect of wind speed on crash risk in snow-storm conditions makes intuitive sense, and is also confirmed from the literature (e.g., Baker and Reynolds, 1992; Knapp et al., 2000).

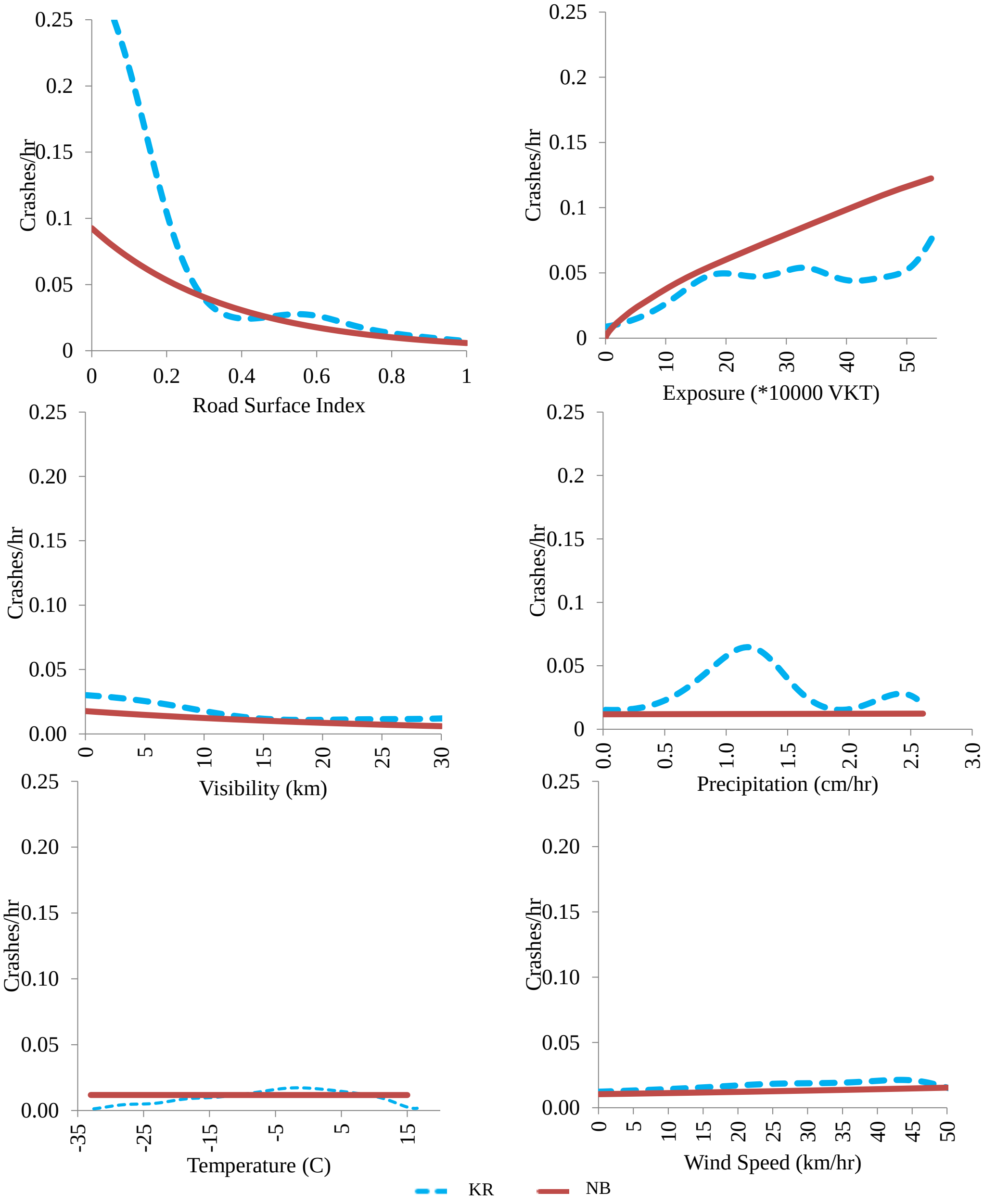


Figure 4-9: Factors affecting crash frequency (Case study 2)

4.5.3 Summary: Comparing Effect of Variables

This section compared the crash modeling results from the nonparametric (KR) and parametric (NB) approaches. As the KR method does not contain any variable specific interpretable parameters to quantify their effects, a direct comparison to the NB model was not possible. Therefore, we presented the results in a graphical form using partial regression plots. The results from both case studies showed that the KR method was able to capture some nonlinear and non-monotonic effects of some risk-related factors, whereas the NB model failed to do so due to its pre-specified model form. This could be the main reason that in our previous study of performance comparison, the KR method showed better results.

Chapter 5
Network Screening: Nonparametric and Parametric Approaches

Network screening is one the most important components of road safety analysis and involves selecting a list of crash hotspots so that a countermeasure program can be launched effectively. A hotspot is the site with a relatively high level of crash risk as determined by a crash prediction technique. Broadly, there are two commonly used statistical approaches for estimating the crash risk: regression-based and Empirical Bayesian (EB) based methods. One of the reasons for their popularity is that they help reduce the regression-to-mean bias (RTM) problem of a simple crash history-based method. In particular, the EB method is known for its robustness as it accounts for site-specific crash history while still incorporating the risk estimates from a regression model (i.e., crash model).

Central to both approaches is the crash prediction model that is used to estimate the risk levels of study sites. As discussed in the literature review and Chapter 4, the most popular crash modeling technique in road safety studies is the parametric approach. This technique, however, needs a prior specification of the relation between crashes and the potential explanatory variables. Therefore, it has a potential risk of misspecification due to the complex, unknown relation of crashes and crash-related factors. Consequently, any misspecification of the crash model may result in an inaccurate list of crash hotspots in network screening, thereby leading to improper allocation of resources for safety improvements. In this chapter, we introduce kernel regression (KR) as an alternative to the parametric model applied under both crash estimation frameworks (i.e., regression-based and EB-based) to network screening. We compare its performance with the parametric counterpart- negative binomial (NB) model (both in regression and EB frameworks) with a case study.

5.1 Framework for Network Screening

Networking screening is a systematic process of ranking sites that suffer from unacceptably high levels of crash risk. This process consists of five main components: 1) preparation of the dataset; 2) selection of ranking measure; 3) selection of a method for estimating crash risk; 4) ranking of sites, and 5) selection of high-risk sites (or crash hotspots) (see Figure 5-1). We describe each of these components by considering a case study of Highway 401 in Ontario, Canada.

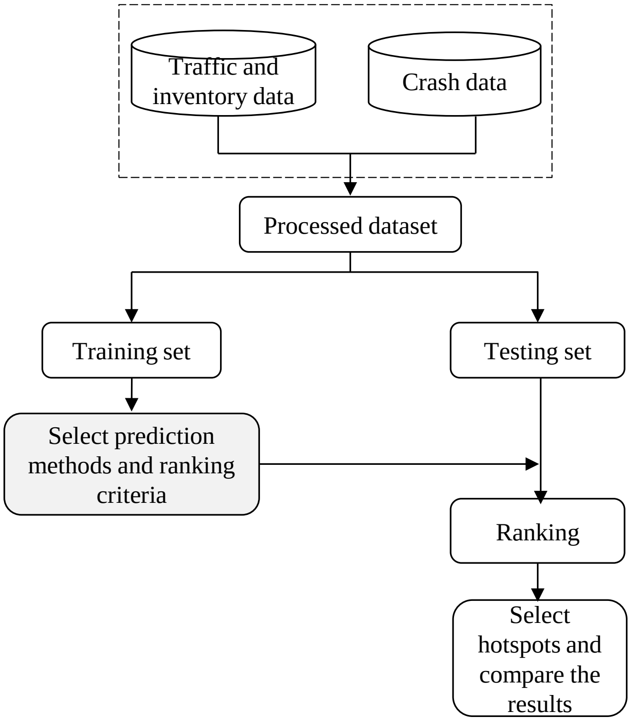

Traffic and inventory data
Crash data
Processed dataset
Training set
Testing set
Select prediction methods and ranking criteria
Ranking
Select hotspots and compare the results

Figure 5-1: A framework for network screening

Preparation of dataset: The Highway 401 case study used in Chapter 4, which provides a detailed description of data sources and processing steps, is also used in this analysis. Note that the data processed was yearly-based, a time-span large enough for analyzing safety problems from a planning perspective, as is required in network screening. We split the original dataset, which contains nine years of crash data (2000-2008), into a model set (2000-2006) and an evaluation set (2007-2008).

Selection of ranking measure: A variety of risk measures can be used as the ranking measure for network screening, such as average crash frequency (crash per km-year), average crash rate (crash per vehicle-kilometers) or weighted crash frequency based on crash severities. Oher measures that can be used are listed in the HSM manual. In our case study, we considered the first two measures: crash frequency/km and crash rate (crash frequency/exposure) as determined by normalizing the estimated crash frequency by length and exposure, respectively. Note that the frequency-based measure emphasizes maximizing the system-wide benefits of safety intervention targeted to the selected set of treatment sites, while the rate-based measure underscores the importance of individual road users' safety perspective (Tarko & Kanodia, 2003). While the choice depends on the interest and priority set identified by the stakeholders or road agencies, we consider both measures in order to explore their implications in identifying the crash hotspots.

Selection of methods to estimate crash risks: Two approaches are considered: Regression-based (KR and NB) and Empirical Baye's-based (both KR and NB) methods. The following paragraphs explain the steps involved in estimating crash risk using each approach.

- For the regressions-based approach, the model dataset was first used to estimate model coefficients for the NB model and calculate the variable bandwidths for the KR method. To relax the constant over-dispersion parameter of the NB model, we considered using its extended form – a generalized negative binomial (GNB) model, where the dispersion parameter is modeled as a function of a set of covariates (Hauer, 2001; Miaou and Lord 2003; Miranda-Moreno et al., 2005; Usman et al., 2010). This may be significant in the EB approach as it uses dispersion parameter of the NB model to determine the weights. Then, the evaluation set was used for estimating the crash risk of highway sections using their respective crash model. Note that, in this particular case study, there are a total of 418 highway sections.

- For the EB-based approach, the steps are slightly different. We refer 2005-2006 as the base-year and 2007-2008 as the ranking-year. First, the EB estimates (both KR and NB) were obtained for the base-year. Then the base-year estimates were extrapolated to the ranking-year by multiplying them by a factor $r = \mu(.)^{ranking-year} / \mu(.)^{base-year}$, where $\mu(.)$ represents crash frequency estimated by respective regression techniques. This two-step process presented here is same as the EB estimates used in before-after countermeasure studies where crashes estimated for the before treatment period (here base period) are projected to the after treatment period (here ranking period) with adjustment made through the crash models to account for changes in the variables (HSM, 2010; Choi et al., 2015). This approach was taken to match the prediction framework and is similar to that of a regression model where a separate evaluation set is considered.

Ranking of sites: Two ranking measures, crash frequency (crashes/km) and rate (crashes/MVK), are determined as follows. First, the number of crashes occurring in each of 418 HS sections was estimated using two models (i.e., KR, NB) including their respective EB methods (i.e., KR-based EB and NB-based EB). The estimates are then normalized using section length and exposure to obtain normalized crash frequencies and rates, respectively. All sections are then sorted in descending order based on the risk estimates obtained from each approach.

To compare the ranking of sites between any pairs of estimation techniques, we calculated the Spearman's rank correlation coefficients (SC) for each ranking measure. SC is a measure obtained by a nonparametric method to quantify the linear association between any two independent ranking variables, as given by Eq. 5-1. SCs can vary from -1 to +1 with values close to 1 indicating that the results from the two estimation techniques are highly similar.

$$\rho = 1 - \frac{6\sum_{i=1}^{m} d_i^2}{m(m^2-1)} \qquad (5\text{-}1)$$

where,

ρ = Spearman's rank correlation coefficient,

m = total number of sites (here 418),

d_i = difference between the two ranks of site *i*.

Selection of hotspots: The final step involved in network screening is to select hotspot sites, i.e., a subset of sites with relatively high crash risk that warrant safety interventions. The threshold risk levels used to determine hotspots depend on the amount of resources available for the safety improvement program. In the case of this case study, resource availability was not a concern; therefore, the top *x* percent of sites (e.g., *x* could be 10%, 20%, etc.) were selected as the crash hotspots.

5.2 Crash Models

Eq. 5-2 shows the NB model calibrated using model dataset (2000-2006) applying the maximum likelihood estimation technique. The summary of the model results is presented in Appendix C.1

$$\mu = e^{-1.16+0.84\ \ln(exposure)+5\text{E}-05AADT_C-0.01MW-0.01SW_l+0.16SW_R-0.09CD} \qquad (5\text{-}2)$$

$$\alpha = e^{-51-0.83L} \text{ where, } \alpha \text{ is over-dispersion parameter} \qquad (5\text{-}3)$$

where,

μ = crash frequency (per year),

exposure= million-vehicle-kilometer travelled,

$AADT_C$= commercial AADT (veh/day),

MW= median Width (m),

SW_l= shoulder width- left (m),
SW_R= shoulder width - right(m),
CD= curve deflection (1/km),
L= length of road segment (km).

For the KR method, the bandwidths identified are 21.08, 2621, 2.39, 0.46, 0.11, 0.135 for *exposure*, $AADT_c$, *MW*, SW_l, SW_R*, and CD* variables, respectively.

As discussed in Chapter 3, the EB-based approach combines crashes occurred at the specific site and expected crash frequency from the reference sites (using crash model) through a weighting scheme. Determination of the weight factors (w) depends on the types of regression model used in the EB framework. For the NB-based EB method, we first calculate the value of dispersion parameter (α) using Eq. 5-3 which is then used to obtain the weight factor as $w = 1/(1 + \mu \times \alpha)$. This is based on the use of parametric modeling approach for defining the relation of variance-mean (details in Chapter 3). Note that larger the value of "*w*" more weight is given to model estimates and lesser to the observed crashes.

In contrast, for the EB-based approach using the KR method (hereafter KR-based EB approach), we do not specify such relation for mean-variance to calculate the weight; rather, a data-driven approach (KR method) is used to establish the underlying relation (detailed steps explained in Section 3.4.3 of Chapter 3). Figure 5-2 illustrates a nonlinear relation of mean-variance using the model dataset. The result indicates that, in general, the variance has a positive correlation with the mean crash estimate. This implies that the weight decreases with increasing variance as the relation between these two measures is inversely proportional. This is similar to the NB-based EB method where the larger model estimates (i.e., larger variance) result in the lower magnitude of weights.

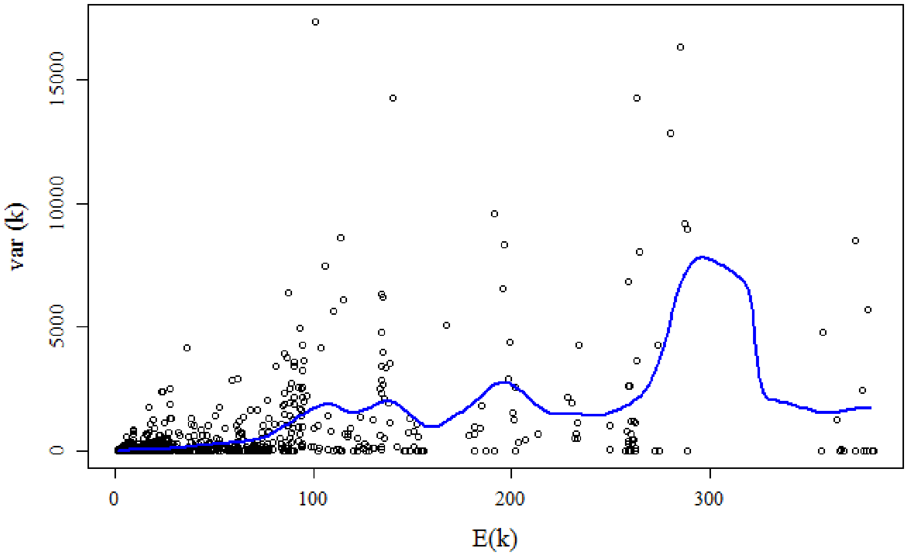


Figure 5-2: Mean-variance relation for EB-based KR method (bandwidth of E(k) is 10.13)

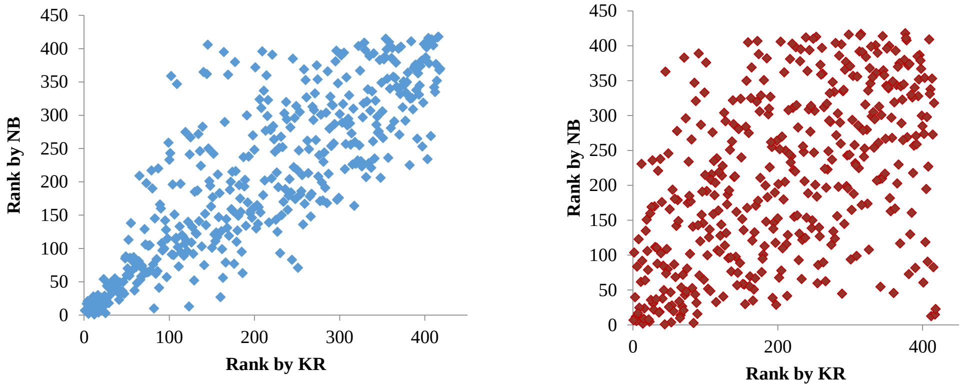
5.3 Comparing Ranking [6]

Figures 5-3 and 5-4 present scatter plots of one to one ranking of highway sections (total 418 sections) based on a pair of crash models (KR and NB) applied under the two frameworks: regression-based and EB-based, respectively. These figures also include ranking comparisons for the two measures- crash frequency and crash rate. We can visually observe some deviations in ranking as some of the sites are found significantly off the diagonal line, and this varies based on estimation approach and ranking criteria. Between the crash estimation techniques, i.e., regression-based and EB-based, the ranking correlation of the KR and NB methods in the latter approach is comparatively high under both ranking criteria. One of the reasons could be due to the involvement of site-specific crash history in the EB-based framework, thereby resulting in similar magnitude of risk measures. While the effects of ranking criteria appears less visible in the EB-based approach as seen in Figure 5-4 (a and b), this is quite different in the regression-based approach. Comparatively, the frequency criterion (Figure 5-3 (a)) shows a high correlation in ranking than by the rate criterion (Figure 5-3 (b)).

As mentioned earlier, Spearman's correlation (SC) coefficients are used to determine the correlation between the two methods, where large SC values represent high correlation and vice versa for the low

[6] Thakali et al. (2016). Comparing crash estimation techniques for ranking of sites in a network screening process, CSCE conference, June 1-4, 2016.

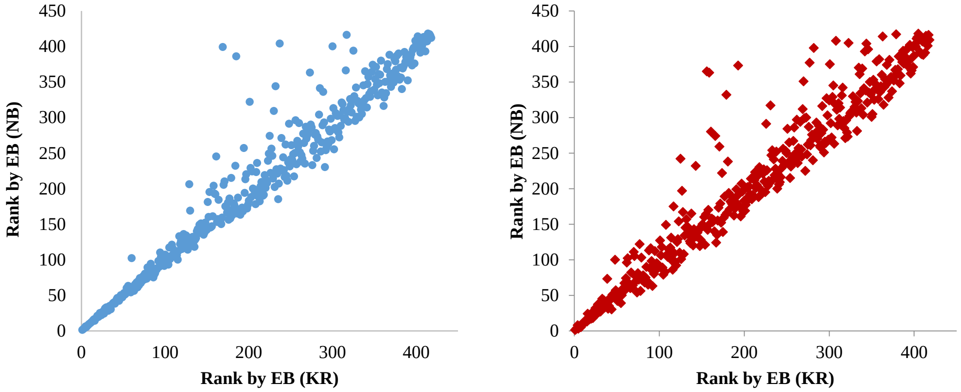
values. As summarized in Table 5-1, the KR and NB models in a regression-based approach showed relatively lower correlation (SC ranging from 0.526 to 0.826) compared to the EB-based approach (SC ranging from 0.965 to 0.973). Meanwhile, as previously discussed, ranking correlations are higher with the frequency measure (SC ranging from 0.826 to 0.973) compared to the rate measure (SC ranging from 0.526 to 0.965). The results from this example suggest that the choice of risk measures for ranking may have a significant impact on the relative performance of the parametric and nonparametric methods, especially when adopting the regression-based approach.



(a)

(b)

Figure 5-3: Ranking comparison based on regression modeling approach- (a) crash frequency/km and (b) crash frequency/million vehicle km



(a)

(b)

Figure 5-4: Ranking comparison based on EB approach- (a) crash frequency/km and (b) crash frequency/veh-km

Table 5-1: Spearman's correlation (SC) coefficients

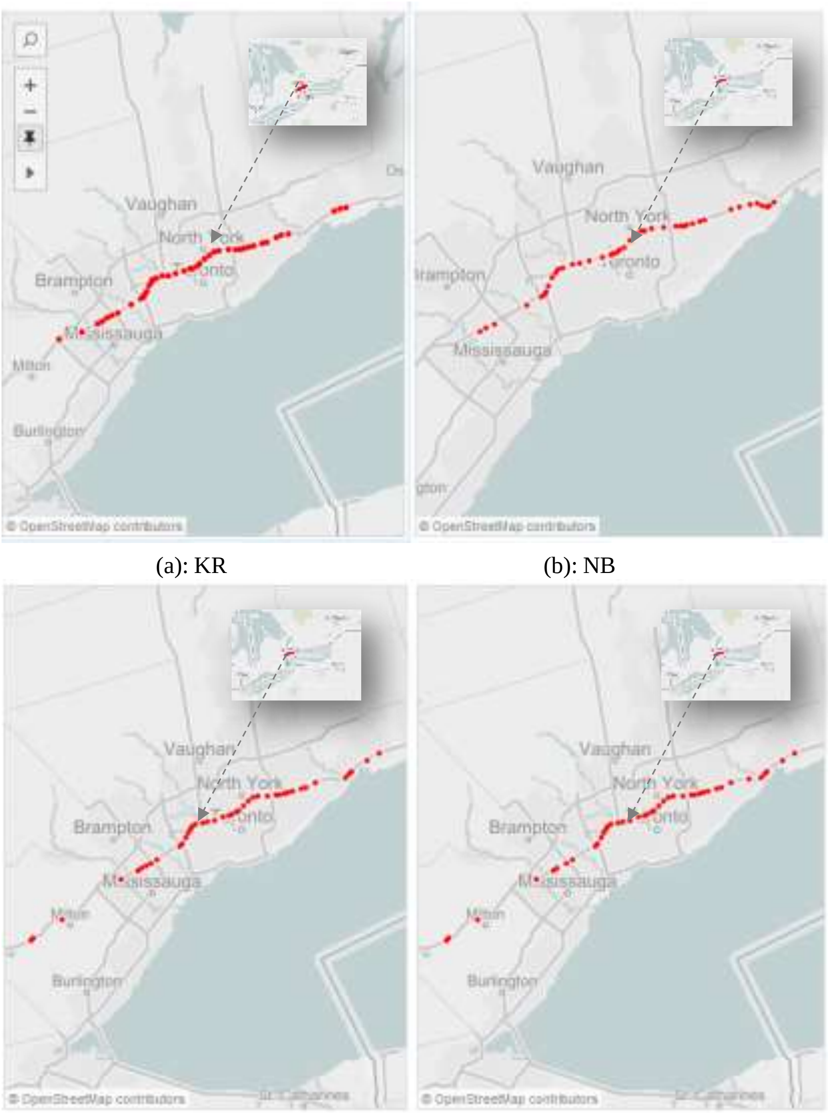
Methods	Ranking measure	
	Crash frequency	Crash rate
Regression-based approach	0.826	0.526
EB-based approach	0.973	0.965

5.4 Comparing Hotspots

Crash hotspot sites are selected by considering the top 10% of the total sites with highest risk levels. However, it is noted that in the real field, this number depends on the amount of resources available for a safety program. Figure 5-5 and 5-6 illustrate crash hotspots based on ranking according to crash frequency and crash rate, respectively. The main implication of selecting a specific ranking measure for identifying hotspots is visible on the maps. As expected, hotspots with frequency indicator are located in the vicinity of the urbanized section of the City of Toronto where the traffic levels are high, causing higher risks. In contrast, when crash rate is used, the hotspots are little scattered as we normalize the crash risk by their respective traffic levels.

As seen Figure 5-5, there is a slight variation in the hotspot sections of frequency risk measure depending on the approach– regression-based or EB-based – and the type of crash model (parametric-NB or nonparametric-KR) involved in each approach. Comparatively, the EB-based approach shows more similarity in the location of hotspots. Similarly, Figure 5-6 shows different locations of hotspots for the rate risk measure and draws a similar conclusion. Furthermore, to quantify their differences in hotspots locations, we calculated the percentage-matching rate. In addition, we also explored how this matching rate varies as more hotspot sites are considered (e.g., 10%, 20%, etc.) and the results are shown in Figure 5-7. Matching rate is relatively constant in the EB-based approach between the KR and NB methods as it is already in the higher end. However, for the regression-based approach, the matching rate between the crash models increases as more sites are considered, and this rate is comparatively high for the frequency measure.

Overall, regarding the identification of hotspots using different techniques, we summarize the findings as follows. First, the choice of crash models (KR or NB) had less influence when applying the EB-based approach. Second, this choice of modeling approach made quite a difference in the regression-based approach, however, their differences were reduced as the number of sites selected as hotspots increased.



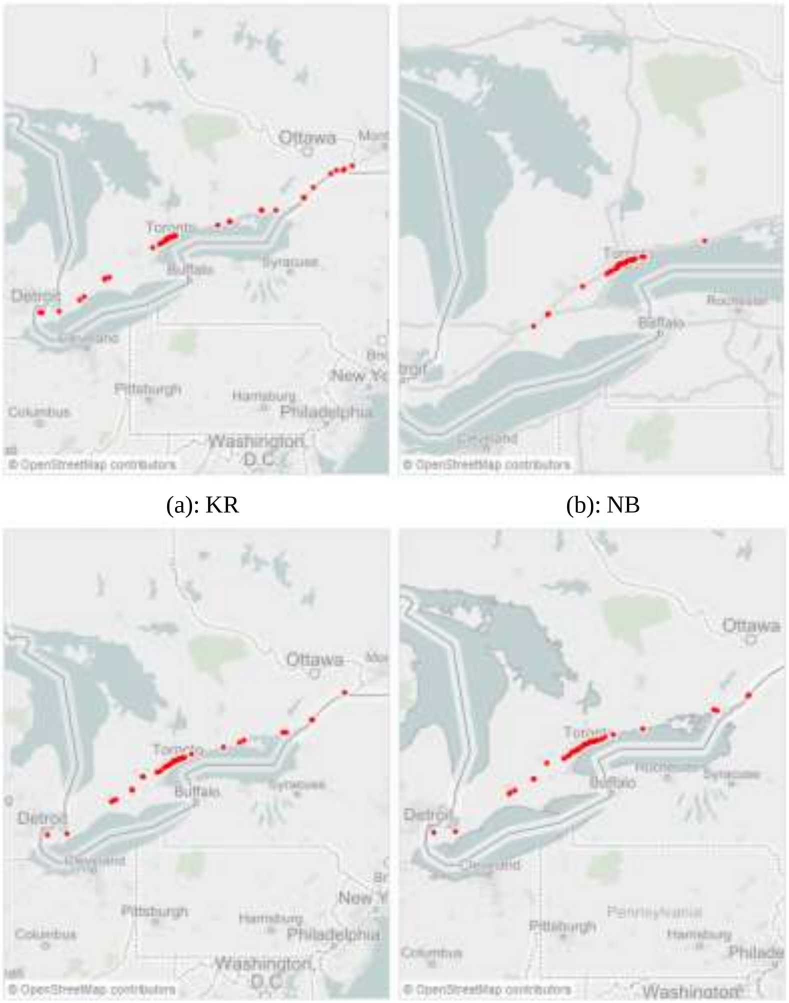
(a): KR

(b): NB

(c): EB (KR)

(d): EB (NB)

Figure 5-5: Locations of 42 hotspot sections (10% of total sites) based on crash frequency



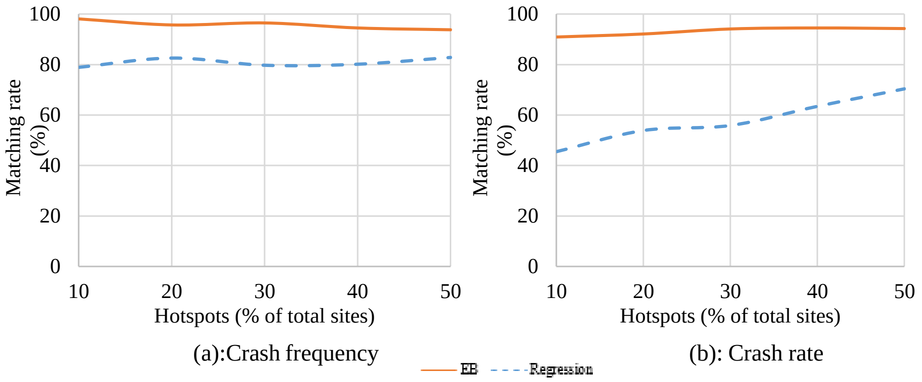
(a): KR

(b): NB

(c): EB (KR)

(d): EB (NB)

Figure 5-6: Locations of 42 hotspot sections (10% of total sites) based on crash rate



(a):Crash frequency

(b): Crash rate

Figure 5-7: Comparisons of hotspots in terms of percentage matching (total sites= 418)

5.5 Summary Conclusions

In this chapter, we presented a case study of network screening using regression-based and EB-based approaches with the main objective of evaluating the practical implications of adopting a nonparametric cash model within these two crash estimation frameworks. For this, we considered the kernel regression (KR), a data-driven nonparametric approach, for estimating the crash risk. We benchmarked its performance for network screening against the parametric counterpart, the negative binomial (NB) model.

The comparative results from network screening in terms of ranking of sites and identification of hotspots showed that the nonparametric and parametric approaches have more similarities when applied in the EB-based framework than in the regression-based framework. In the EB-based framework, the ranking results based on the KR and NB models were highly similar regardless of the choice of ranking measures– crash frequency or crash rate. In contrast, their differences were more visible when used in the regression-based approach, with the rate measure showing a relatively high variation compared to the frequency measure. The findings also showed that differences in locating hotspot sections based on different approaches reduced as the number of selected sites increased, thereby providing greater flexibility to select either nonparametric or parametric methods. We also make a note that in our previous comparative study, the KR method performed better than the parametric NB model, the ranking of sites and the selection of hotspots based on the nonparametric method (regression-based or EB-based) is expected to be more reliable.

Chapter 6
Countermeasure Study: Nonparametric and Parametric Approaches

The study of countermeasures is another important component of a road safety analysis aiming at quantifying the effect of safety treatment measures through crash modification factors (CMFs). As discussed in Literature Review, there are two popular approaches to a countermeasure study, namely, before-after study and cross-sectional study. Both methods are popular as they reduces regression-to-mean (RTM) problem of a naïve crash count-based approach. When enough crash data related to before-after treatments are available, the use of before-after Empirical Baye's (EB) method is recommended. However, when before-after data are limited, an alternative viable method is using the cross-sectional study approach. In this approach, CMFs are obtained by comparing with and without crash risk conditions using the cross-sectional data from similar sites. These studies are common mainly in the context of determining CMF of roadway characteristics, such as altering shoulder, lane and median width, and treating road shoulders with rumble strips.

Parametric models have been commonly used in both types of studies. In this chapter, we employ kernel regression (KR) as an alternative to the traditionally used parametric count models. We also compare its performance with the parametric counterpart- negative binomial (NB) model.

6.1 Framework for Countermeasure Study

Figure 6-1 provides an overview of a framework for countermeasure study. The main components include preparation of dataset for model development, selection of treatment sites, collection of detailed information about before and after treatment conditions, selection of crash-modeling techniques, calculation of the CMFs for the selected treatments, and finally comparison of the CMFs obtained from different techniques. The CMFs are computed by comparing the crashes before and after the period of treatments as given by Eq. 6-1. Depending on the study approach selected (i.e., before-after study or cross-sectional study), there may be a slight variation in final CMF calculations. Detailed explanations are given in the following sections (6.2 and 6.3).

$$CMF = \frac{C_a}{C_b} \tag{6-1}$$

where,

C_a= expected crashes for condition “a” i.e. after[7] or with[8] treatment,

C_b = expected crashes for condition “b” i.e. before or without treatment.

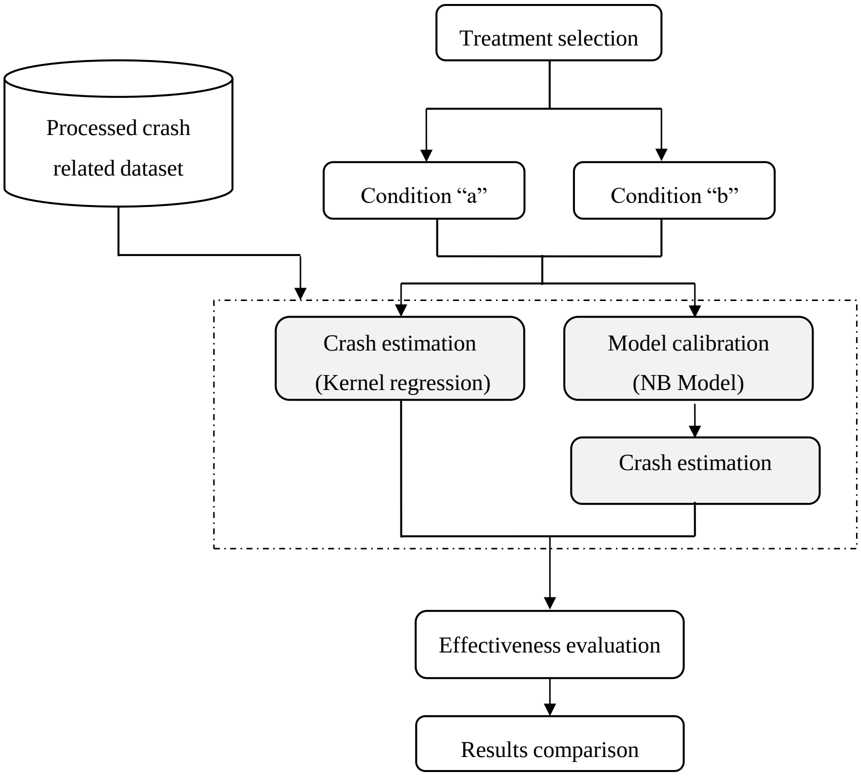


Figure 6-1: Framework for a countermeasure study

6.2 Approach 1: Before-After EB Study

The EB method combines safety from two measures: the observed numbers of crashes obtained from individual sites and the expected number of crashes estimated from the reference population. The latter is achieved using a regression model developed from crash data of the reference sites. The EB method is given its name depending on the type of model applied. For example, the KR-based EB method involves KR method, and similarly, the NB-based EB method involves NB model. Detailed descriptions have been given previously in Chapter 3.

[7] Applicable for before-after study

[8] Applicable for cross-sectional study

Determining the CMF of a treatment in a before-after study requires selection of a number of sites that have implemented the specific treatment and is based on the combined safety effects before and after the treatment periods as given by Eq. 6-2.

$$CMF' = \frac{\sum_i^n C_{a_i}}{\sum_i^n C'_{b_i}} \tag{6-2}$$

where,

C_{a_i} = observed number of crashes after the treatment at the site i,

C'_{b_i}=expected number of crashes before the treatment at the site i (obtained from EB estimate for before the treatment period),

n = total numbers of sites, and

CMF' = unadjusted CMF.

We normalize the before treatment crashes in Eq. 6-2 as follows:

$$C_{b_i} = C'_{b_i} r_i$$

$$r_i = \frac{\mu_{a_i}}{\mu_{b_i}} \tag{6-3}$$

where,
r_i = adjustment factor for a change in site conditions,
μ_{a_i}= predicted crashes (by a model) at the site i after the treatment period, and
μ_{b_i}= predicted crashes (by a model) at site i before the treatment period.

The factor "r_i" is used to adjust for changes in actual safety in the treatment sites due to the change of traffic volumes and other engineering interventions. This step makes sure that the crashes of before and after the treatment conditions are compared in the same time horizon, i.e., after treatment period.

After replacing normalized before -treatment crashes C_{b_i}in Eq. 6-2, we obtain the final expression of CMF given in Eq. 6-4 (or Eq. 6-5 for an unbiased estimate).

$$CMF = \frac{\sum_i^n \mathrm{C}_{a_i}}{\sum_i^n \mathrm{C}_{b_i}} \tag{6-4}$$

To adjust for an unbiased estimate (Hauer, 1997):

$$CMF = \frac{\frac{\sum_i^n c_{a_i}}{\sum_i^n c_{b_i}}}{1+\frac{var(\sum_i^n c_{b_i})}{\sum_i^n {c_{b_i}}^2}} \tag{6-5}$$

The variance of CMF is estimated by:

$$Var(CMF) = \frac{CMF^2(\frac{Var(\sum_i^n \mathrm{C}_{a_i})}{\left(\sum_i^n \mathrm{C}_{a_i}\right)^2} + \frac{Var(\sum_i^n \mathrm{C}_{b_i})}{\left(\sum_i^n \mathrm{C}_{b_i}\right)^2}}{\left[1+\frac{Var\ (\sum_i^n \mathrm{C}_{b_i})}{\sum_i^n {\mathrm{C}_{b_i}}^2}\right]^2} \tag{6-6}$$

where,

$Var(\sum_i^n \mathrm{C}_{a_i}) = \sum_i^n \mathrm{C}_{a_i} = \sum_i^n \mathrm{K}_i$ (Assuming crashes follow Poisson distribution and K_i is observed crashes at site i; Persaud et al., 2001; Hauer, 1997),

$Var(\sum_i^n \mathrm{C}_{b_i}) = \sum_i^n Var\ (\mathrm{C}_{b_i})$ (Assuming individual variances are mutually independent, and for individual variance, we use $Var\ (\mathrm{C}_{b_i}) = (1-w)E(k/K)$ (Hauer, 1997).

6.3 Approach 2: Cross-sectional Study

The CMF in a cross-sectional study is determined by comparing the crash risk of with and without the treatment. Depending on the type of crash model involved, (e.g., nonparametric or parametric) Eq. 6-1 for CMF varies slightly. We provide a brief explanation of each type in the following sections.

6.3.1 Nonparametric: KR method

The KR method, being nonparametric in nature, does not contain any model parameters to relate predicting variables to crash risk (i.e., model coefficients). Instead, it adopts a data-driven approach to crash estimation by weighing all the observed crash data points. The weights are determined jointly by a kernel function and the bandwidth of the covariates. The weights vary depending on the distance between the covariates of observed crash data points and the evaluation point. Therefore, the KR method requires an explicitly defined condition of each covariate, both with and without treatment conditions, for predicting the crash risk. We represent CMF of a covariate “x_d " as:

$$CMF_d = \frac{\mathrm{C}_a}{\mathrm{C}_b} = \frac{m_a(x_1^b, x_2^b, \ldots\ x_d^a\ \ldots x_D^b\ )}{m_b(x_1^b, x_2^b, \ldots\ x_d^b\ \ldots x_D^b\ )} \tag{6-7}$$

where,

C_a and C_b (have same definitions as in Eq. 6-1) are functions of covariates,

$m_a(.)$ is expected crash frequency for after (with) treatment case,

$m_b(.)$ is expected crash frequency for before (without) treatment case,

x_d = the covariate whose CMF is to be calculated,

x_d^a = after (with) condition of x_d ,

x_d^b= before (without) or base condition of x_d ,

$x_1^b, x_2^b \dots x_D^b$= are remaining covariates at their base conditions (excluding x_d), and

D = number of covariates.

As shown in Eq. 6-7, the CMF of a covariate involves a comparison of crash risk with (C_a) and without (C_b) the treatment conditions. In summary, there are three main inputs required for CMF calculation:

- A base case (or before case) for the covariate whose CMF is to be determined. This represents the “without the treatment condition”.
- A treatment case (or after case) for the same covariate. This represents the “with the treatment condition”
- Base cases for remaining covariates. These represent controlling variables.

Standard error[9] of CMF is given by (Kendall, 1998):

$$SE_{CMF} = \left[(\frac{\mathrm{C}_a}{\mathrm{C}_b})^2 (\frac{Var(\mathrm{C}_a)}{\mathrm{C}_a{}^2} + \frac{Var(\mathrm{C}_b)}{\mathrm{C}_b{}^2}) \right]^{0.5} \qquad (6\text{-}8)$$

where, $Var(\mathrm{C}_a)$ and $Var(\mathrm{C}_b)$ are determined by a bootstrap approach adopted from Hyfield & Rachin (2008).

6.3.2 Parametric: NB model

As the NB model is represented by an equation, the CMF calculation becomes relatively easy. For the exponential form of a NB model, the CMF of a covariate "x_d " is represented as:

$$CMF_d = \frac{\mathrm{C}_a}{\mathrm{C}_b} = \boldsymbol{e}^{\boldsymbol{\beta}_d (x_d^a - x_d^b)} \qquad (6\text{-}9)$$

where,

β_d= estimate regression coefficient associated to covariate d, and

[9] Standard error is the standard deviation of a sample mean (Gross et al., 2010).

Others notations same as in Eq. (6-7)

The exponential form of the NB model is the most popular specification used in the past studies (Council & Steward, 1999; Lord & Bonneson, 2007; Fitzpatrick et al., 2008; Stamatiadis et al., 2009; Carter et al., 2012; Zeng & Schrock, 2013; Park et al., 2014; Choi et al., 2015; Park & Abdel-Aty, 2015; Wu et al., 2015). However, we should note that depending on the choice of a functional form for the NB model, the expression for the CMF in Eq. 6-9 changes.

As seen, there is a fundamental difference between the CMF of nonparametric approach (Eq. 6-7) and parametric approach (Eq. 6-9). In the latter approach, there is no need for defining base cases for the controlling variables, i.e., variables other than "x_d " whose CMF is to be determined. This is because the effects of remaining variables get canceled out as they appear in divisional forms.

The standard error of the CMF can be calculated by using two equations: 1) Eq. 6-10.a as adopted by Park & Abdel-Aty (2015) and Eq. 6-10.b as recommended by Bahar (2010) and HSM (2010, part D).

$$SE_{CMF} = \frac{\exp\left(\beta_k + SE_{\beta_d}\right) - \exp(\beta_k - SE_{\beta_d})}{2} \quad \text{(6-10.a)}$$

$$SE_{CMF} = \frac{SE_{\beta_d}}{t_{\beta_d}}\, CF \quad \text{(6-10.b)}$$

where,
SE_{CMF}= standard error of CMF,
SE_{β_d}= standard error of coefficient β_d,
t_{β_d} = t-statistic of coefficient β_d, and
CF= correction factor (2) obtained from Bahar (2010).

Note that the standard error provides the precision of an estimate, and does not say anything about the accuracy of the estimate. Therefore, which CMF estimates appear closer to the true value may not be concluded simply from a direct comparison of the CMFs based on their standard errors.

6.4 Case Study: Before-After Study

Safety at railway-highway grade crossings is a serious concern to transportation agencies, and various traffic control devices, either passive controls such as stop sign and yield sign or active controls such as flashing light and bell (FLB), FLB with Gates (FLBG) and others are often deployed to reduce the potential risk of crashes. These control types are expected to have different levels of safety effects (or

CMFs) depending on their degree of protection. In this section, we present a case study to determine the CMFs of three sets of controls in relation to their specific base conditions using crash data from grade crossings in Canada. These include adding FLB to passively controlled crossings, adding gates to FLB crossings and adding a constant warning time device to crossings with FLB.

6.4.1 Data Description

The before-after EB study requires two different sets of crash data. The first one is a model dataset which is collected from a reference population for calibrating a crash model, and the second is a before-after observed dataset which is obtained from a set of sites that have implemented a specific treatment. Below provides a brief description of these two datasets used in this study.

Model dataset: The data for the model set are obtained from two different sources: 1) inventory data from Integrated Railway Information System (IRSI) database and 2) observed crash data from Railway Occurrence Database System (RODS) database. The IRIS database contains information related to characteristics of crossings, such as control type (e.g., passive, FLB, FLBG and others), location, traffic volume (both vehicle and train). Similarly, the RODS database records information related to individual crashes, such as date of occurrence, type of trains and vehicles involved, crash severities, average traffic volume. For each crossing (only the public crossing), crashes occurring from 2008- 2013 were aggregated. This was then integrated with the inventory data based on their unique crossing IDs. Those crossings with missing inventory information such as road speed, traffic volumes and train volume were excluded from the processed dataset. Meanwhile, only the crossings with the following three control types- passive crossing, flashing light and bell (FLB) and flashing light and bell with gates (FLBG) were considered. Finally, the dataset was split for these control types. Appendix D.1 provides a summary of each dataset.

Before-after dataset: The data for the before-after set are obtained from three different sources: the Grade Crossing Improvement Program (GCIP), IRIS and RODS. The GCIP database contains information related to safety projects implemented across Canada under Transport Canada's funding program – GCIP. It includes information such as date of project completion, types of intervention, crossing conditions at the time of implementation and crossing ID. For this case study, we selected projects related to the following treatment types: converting passive controls to FLB, adding gates to FLB crossings and adding constant warning time device to FLBG crossings. For each project site, crash data for the five years before and after the project completion were extracted from RODS. The GCIP database lacks information about the traffic conditions before and after the treatment implementations; therefore, this missing information was filled in using the following procedure. For the before treatment condition, crossing related information (e.g., traffic volume) was extracted from the RODS database by referring to the crashes of that period. Similarly, for the after treatment condition, the IRIS database was used as it contained the most current information on crossings.

6.4.2 Crash Models

The crash model used in a before-after EB study has two main roles. First, as discussed in Chapter 3, the EB method used to predict crash frequency requires a crash model to incorporate the risk levels of similar sites. Second, as mentioned in Section 6.3 (Eq. 6-3), we use a crash model to obtain an adjustment factor to account for the change in risk levels that could have resulted from the changes in traffic volume including other interventions.

NB models were calibrated for each control type: passive, FLB and FLBG. Initially, their full models were calibrated by considering a set of significant variables that included vehicle volume, train volume, train maximum speed, road speed and number of tracks (see Appendix D.2). However, as the RODS database used for extracting the before treatment conditions contained only details on traffic volume, we calibrated traffic-only models as presented in Eq. 6-11 to 6-13 (see Appendix D.3 for more information). When comparing the performance of full and traffic-only models for all control types, only marginal differences were observed. For example, for the passive control, the full model has Akaike information criterion (AIC) (NB): 2431.7, MAE (NB): 0.074 and MAE (KR): 0.045, whereas the traffic-only model has AIC (NB): 2448, MAE (NB): 0.075 and MAE (KR): 0.048. Similarly, for FLB and FLGB types, these differences were minimal. Meanwhile, for all the datasets, the KR method outperformed the NB model as indicated by its lower MAE values. Note that AIC is one of the commonly used goodness-of-fit measures in a parametric model calibrated using the MLE technique. Lower the value of AIC, better is the model performance.

As discussed in Chapter 3, the EB method using KR method requires a mean-variance relation for determining a weight factor. We present this relation specific to each control type in Appendix D.4.

Crash model for Passive control

$\mu^{p} = TV^{0.68}VV^{0.45}e^{-6.03}$ (6-11)

MAE^{NB}= 0.075; MAE^{KR}= 0.048

Bandwidths for KR method: *TV*=1.74 and *VV*=243. 44

where,

VV= vehicle volume (per day), and

TV= train volume (per day).

Crash model for FLB control

$\mu^{FLB} = TV^{0.64}VV^{0.52}e^{-7.34}$ (6-12)

MAE^{NB}= 0.13; MAE^{KR}= 0.082

Bandwidths for KR method: *TV*=1.61 and *VV*=1240. 96

Crash model for FLBG control

$\mu^{FLB} = TV^{0.56}VV^{0.32}e^{-6.04}$ (6-13)

MAE^{NB}= 0.217; MAE^{KR}= 0.139

Bandwidths for KR method: *TV*=4.21 and *VV*= 1706. 19

6.4.3 CMF Results

This section presents the results of CMFs determined by before-after EB study as described in Section 6.3. Two different approaches, the nonparametric (KR method) and the parametric (NB model), are employed under the EB framework. It is noted that a few sites were excluded prior to the calculation of CMFs due to the issue of sparse data points in the KR method resulting in very low crash estimates. This low values of estimates, if included, would greatly influence the calculation of an adjustment factor "r", i.e., ratio of expected crashes of without treatment conditions for after and before the treatment period. For example, when not enough neighborhood data points are available to estimate before crashes (*Cb*), very low estimates are expected (close to zero), thereby causing "r" to be very large (Eq. 6-1). This problem could have been due to not having enough representative data points in the model dataset (i.e., training set) to reflect before and/or after without treatment conditions of the selected sites. Therefore, the sites with such issues were excluded prior to the calculation of CMFs.

Table 6-1 summarizes the results of CMFs for all three countermeasures: converting passive controls to FLB, converting FLB to FLBG and adding a constant warning time device to FLBG. As seen, the differences in the results obtained from the two approaches (i.e., KR and NB) vary across the countermeasures. For example, adding gates to FLB crossing reveals a comparatively small deviation in their safety effects, and both methods show a reduction of crash risk. This countermeasure is expected to result in approximately 80% reduction in crash frequency (approximate CMF is 0.2) with a marginal difference between the estimates of the KR and the NB methods. Meanwhile, this difference is much higher in the case of converting a passive control to FLB crossing. The KR method shows an approximately 80% reduction in crashes, whereas the NB method indicates a 65% reduction. In contrast to the previous two countermeasures where both approaches agreed showing a reduction of crash risk, they showed opposite effects from adding a constant warning time device to FLBG crossing. As seen, the result from the KR method shows a reduction in crash frequency whereas the NB method shows a slight increase in crash risk. Intuitively, the result from the KR method is more meaningful as providing a constant warning is expected to increase safety level of a crossing.

In Table 6-2, we also present the CMFs of two countermeasures- passive to FLB and passive to FLBG, which were obtained from past studies. As seen, there is a wide range of values within the same treatment measure. This variation, including the differences in results we presented in Table 6-1, could be due to a number of factors, such as local conditions of the treatment sites, numbers of treatment sites, and the methods applied in determining the CMFs.

Table 6-1: CMFs obtained from the before-after EB study

Countermeasures (or treatments)	KR-based EB method	NB-based EB method	Number of sites
Passive to FLB	0.184 (0.09)	0.35 (0.18)	52
Adding gates to FLB	0.178 (0.1)	0.225 (0.12)	67
FLBG to FLBG + CWD	0.597 (0.26)	1.1 (0.51)	21

Values in parenthesis indicate standard error; CWD stands constant warning time device

Table 6-2: CMFs of similar treatments from past studies

Study references	Passive to FLB	Passive to FLBG
Park (2007)	-	0.35

Saccomanno and Lai (2005)	0.42	0.37
U.S. DOT (1980)*	0.30	0.17
California (1974)*	0.36	0.12
Hedley (1952)*	0.37	0.04

*FHWA (2015)

6.5 Case Study: Cross-sectional Study

Highway safety improvement programs often focus on changing the geometric design elements of highway sections, such as shoulder width, the degree of curvatures and others, for improving their safety. In this section, we present a case study of Highway 401 in Ontario, Canada with the objective of determining the safety effectiveness (or CMFs) of some of these design features by employing a cross-sectional study.

6.5.1 Data Description

For the data sources and processing, we refer to Section 4.2 of Chapter 4. This dataset, hereafter referred to as a model set, consists of nine years of crash data (2000-2008) from Highway 401, Ontario. It was previously used for comparing the performance of parametric and nonparametric crash modeling techniques.

6.5.2 Selecting Typical Treatment Cases

To compute CMFs, we first identify the typical highway condition combinations in terms of geometric features and traffic from the dataset. This is particularly necessary for the KR method as it requires an explicitly defined base and treatment conditions for the covariate whose CMF is to be determined, including the base cases for the remaining variables that act as controlling factors (Eq. 6-7). Because the KR method is a local estimator, selecting typical features (i.e., most common) from the dataset will ensure that enough near data points are available to obtain relatively accurate crash estimates. However, in determining CMFs using the NB model, this selection is not a requirement as the model coefficients of respective variables are directly used to determine their CMFs (see Eq. 6-9). Table 6-3 presents a list of typical values for traffic volume, shoulder width, median width and horizontal curve deflection (see Appendix B.6 and D.6 for histograms). Note that, for the traffic-related variables, we first aggregated the records into bins of uniform width and then retained their mid values.

Table 6-3: Typical road geometric sections and traffic levels for developing CMFs

AADT (all vehicles) (veh/day)	AADT (Commercial) (veh/day)	Median width (m)	Shoulder width- left (m)	Shoulder width- right (m)	Curve deflection (per km)
12000	5250	5	1	3	0
17000	8750	10	2	3.5	0.4
22000	12250	15	3	4	0.5
27000	15750	20	4	Total= 3	0.6
32000	19250	25	5		0.7
37000	22750	30	Total= 5		0.8
42000	26250	Total= 6			0.9
47000	29750				1
52000	Total= 8				1.25
57000					1.5
62000					Total= 10
67000					
Total= 12					

6.5.3 Preliminary Setups for the KR method

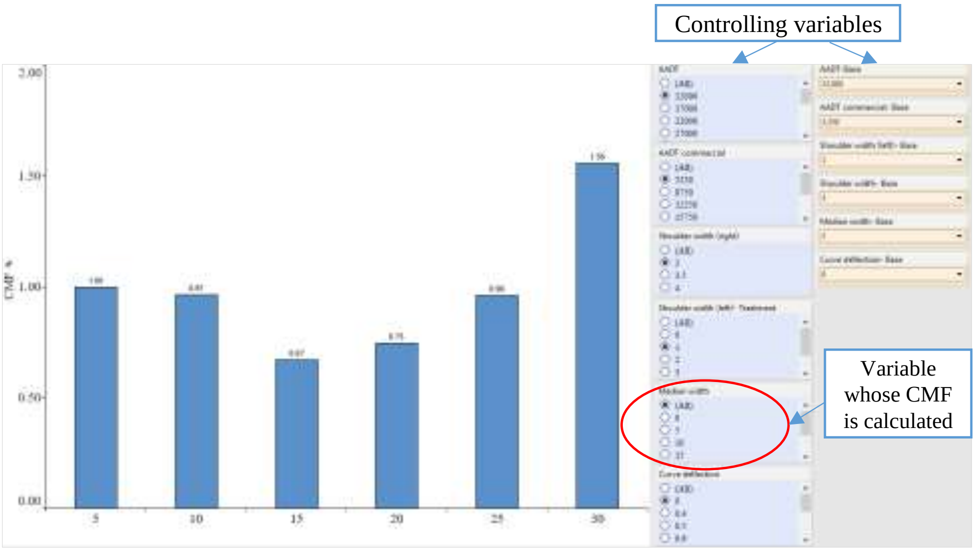
For computing the CMF of a treatment, as discussed in Section 6.3 (Eq. 6-7), we first need to estimate expected number of crashes for with (C_a) and without (C_b) the treatment conditions. Prior to this, we normalize the crash frequency by section length, that is, the dependent variable is crash rate (crash per year per unit length) rather than the crash frequency (crash per year). Note that the CMF is a unitless factor; therefore, this normalizing step has no effect on its estimate. The list of predicting variables includes AADT, commercial AADT, median width, shoulder width (left), shoulder width (right) and horizontal curve deflection. Bandwidths of these variables determined from the model dataset are given in Appendix D**.7**

A new dataset (also known as the evaluation set) consisting of all the combinations of typical values of the predicting variables was created (This consisted of total 86400 records**)**. Then, using the model set

(2000-2008 data), crash rates were estimated for the evaluation set. A visualization tool was used to interactively select expected crashes for two conditions- with (C_a) and without (C_b) the treatment conditions- thereby allowing automatic generation of CMF for the selected countermeasure (see Figure 6-2).

To compare the CMFs from the KR and NB method, we fix the following common base cases: median width of 5 m, shoulder width (right) of 3 m, shoulder width left of 1 m, and curve deflection of zero. Note that, in calculating the CMF of a given variable, we set all the remaining predicting variables to their base values. To account for the effects of changing traffic levels, we consider three different scenarios:

- Scenario 1: Low traffic level- AADT 12000, commercial AADT 5250
- Scenario 2: Medium traffic level- AADT 37000, commercial AADT 15750
- Scenario 3: High traffic level- AADT 67000, commercial AADT 22750



Use filters to select values

Figure 6-2: Visualization tool used for CMF calculation, an example of median width in a low traffic level scenario

6.5.4 Results of CMFs for Single Treatments: KR Method vs NB Model

Figure 6-3 to 6-6 illustrate the results of CMFs obtained from the KR and NB models for four different road geometric features. As discussed previously in Section 6.3, the CMFs in a parametric model are determined directly from its estimated model coefficients, unlike in a nonparametric method which does not contain such easy-to-use parameters due to its data-driven estimating approach. All the CMFs for the NB model are derived from the previously calibrated model (see Section 4.2) as shown in Eq. 6-14:

$$\mu^{NB} = exposure^{0.82} e^{-1.04+0.001AADT_C-0.02MW-0.09SW_L+0.16SW_R-0.17\,CD} \qquad (6\text{-}14)$$

where,

μ^{NB} = expected crash frequency (per year),

$exposure$= million vehicle kilometer travelled,

$AADT_C$ = AADT of commercial vehicle (veh/day),

MW = median width (m),

SW_L= shoulder width on left (m),

SW_R = shoulder width on right (m),

CD = curve deflection or reciprocal of radius (per km).

We briefly discuss each CMF from both the KR method and the NB model in the following paragraphs. As the CMFs from the KR method vary by traffic level, we present the results in three different scenarios: low, medium and high traffic. Detailed results, including their standard errors, are given in Appendix D.8 and D.9.

CMFs for changing median width: Figure 6-3 illustrates CMFs for changing median widths. The KR method shows that, in a scenario of high and medium traffic volumes, widening the median width (except 10 m width) results in a decrease in CMF magnitude, suggesting a lowering of crash risk level compared to the base case of 5 m median width. However, this decreasing trend gradually reverses in a low traffic scenario, particularly for the widths larger than 25 m. This indicates that in a relatively low traffic volume section, widening of shoulder widths does not improve highway safety. In other words, this result from the KR method suggests that increasing the design standard in regard to median width may require caution. Similarly, CMF results from the NB model with the same base case (i.e., 5 m) are illustrated in the same figure (Figure 6-3) generated from the CMF function, i.e., Eq. 6-15. As shown, there is a smooth decreasing trend in CMFs with an indication that the wide median widths have lower crash risk compared to the median of narrow widths. This overall trend from the NB model is similar to the case of high and lower traffic volumes in the KR method with latter having a relatively high reduction in crash risk.

$$CMF_{MW} = e^{-0.02(MW-5)} \qquad (6\text{-}15)$$

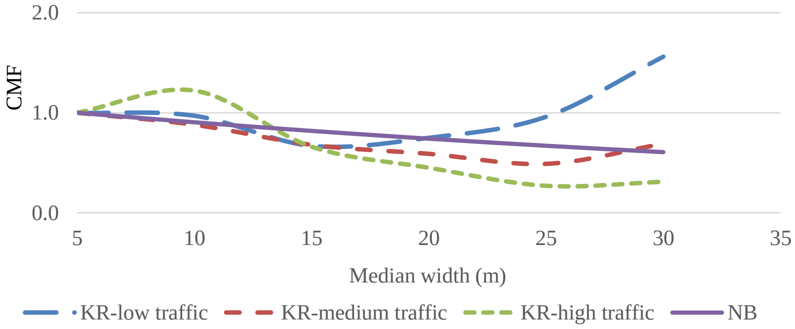


Figure 6-3: CMFs of single factors- median width

CMFs for changing right shoulder width: Figure 6-4 illustrates CMFs for changing shoulder widths (SW_R), i.e., shoulder on the right side of traffic flow. Note that there is less variation in the dimension of SW_R. As shown, the effect of widening shoulder widths on CMFs in medium traffic volume using the KR method reveals a similar decreasing trend as in the NB model (also see Eq 6-16) with the former having relatively higher effects. Meanwhile, the KR method at high traffic volume shows an opposite trend, where widening the shoulder width increases the risk level significantly. The conventional notion that the widening of shoulder width increases safety may not always be true in a highway section with a relatively high traffic volume. The width of 3.5 m in low traffic volumes from the KR method shows a relatively high-risk level compared to the base case of 3 m.

$$CMF_{SW_R} = e^{0.16(SW_R-3)} \qquad (6\text{-}16)$$

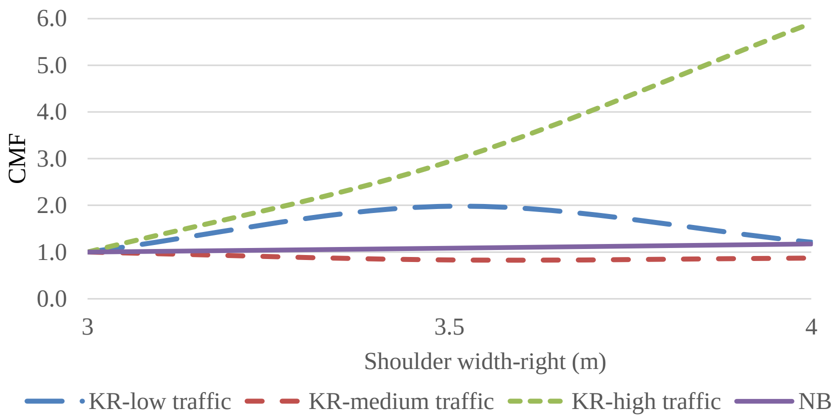


Figure 6-4: CMFs of single factors- shoulder width on right

CMFs for changing left shoulder width: Figure 6-5 presents the CMFs for changing shoulder widths (SW_L), i.e., shoulder on the left side of traffic flow, with the base case of 1 m width. Overall, the trend of CMFs across the SW_L widths in the NB model shows an opposite result compared to the KR method. The NB model shows a reduction in relative crash risk with increasing SW_L (Eq. 6-17), whereas this relation in the KR method for low and high traffic levels is the opposite. Meanwhile, CMFs from the KR method in the medium traffic scenario show slightly different results. Initially, CMF increases with widening shoulder widths, which, after reaching 4 m, starts to decrease.

$$CMF_{SW_L} = e^{-0.09(SW_L-1)} \tag{6-17}$$

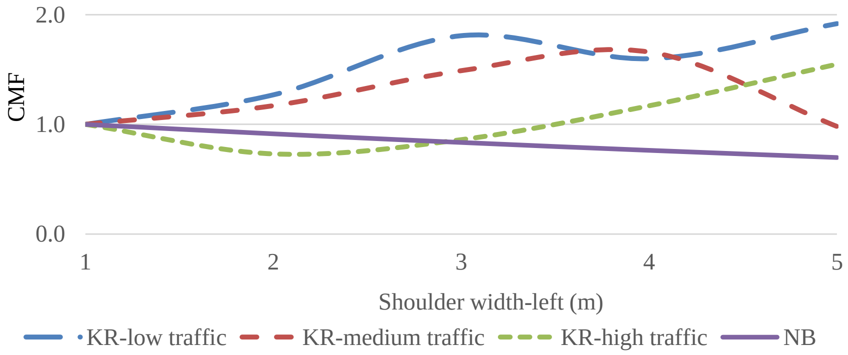


Figure 6-5: CMFs of single factors- shoulder width on left

CMFs for changing curve deflection: Figure 6-6 presents the safety effects of changing the horizontal curve deflection (CD) of a highway section with the base case of zero CD, i.e., straight section. Note that the larger the CD values, the sharper the curve turnings. The results from the KR method show a highly nonlinear trend of CMFs across all the traffic levels with a relatively high magnitude in the medium traffic level, followed by high and low traffic volume scenarios. The results can also be interpreted stating that the risk of crash occurrence in a relatively straight road section is higher than in a curved section. Comparing the two methods, overall, the results from the NB and KR method have similar results at the lower range of CD. One of the main differences between the two methods is that the result from the NB model shows a smooth decreasing trend of CMFs for the increasing values of CDs whereas this trend appears highly nonlinear in the KR method.

$$CMF_{CD} = e^{-0.17CD} \qquad (6\text{-}18)$$

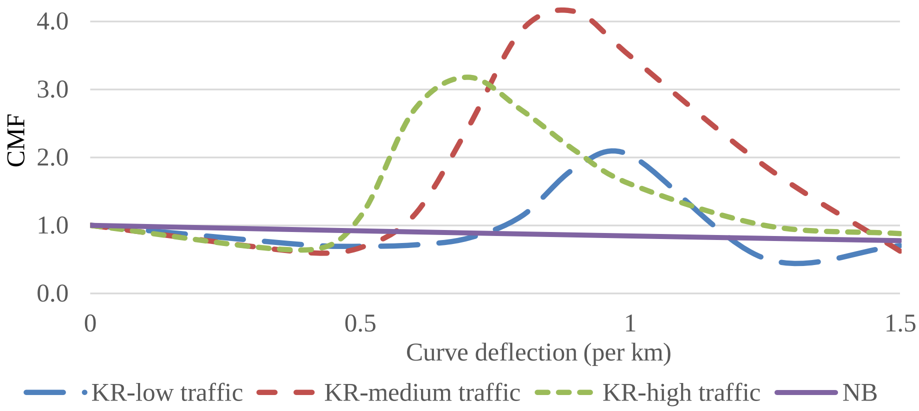
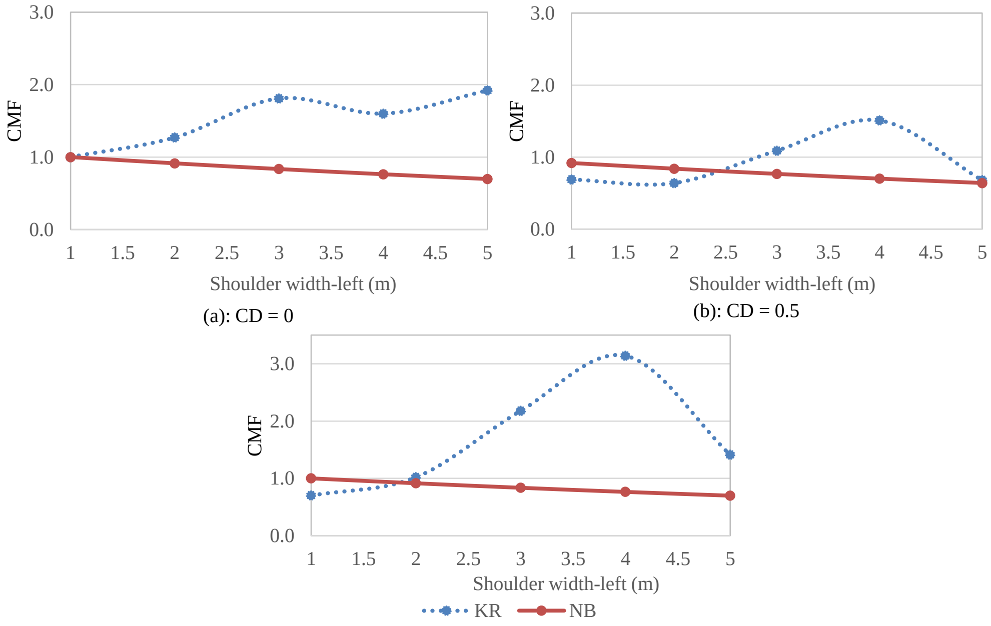


Figure 6-6: CMFs of single factors- curve defection

6.5.5 Results of CMFs for Multiple Treatments: KR Method vs NB Model

Multiple CMFs, hereafter called M-CMFs, are important for evaluating the safety benefits when more than one treatment is considered. In a parametric approach, this is obtained by: 1) including an interaction term in the regression model and using its coefficient directly to determine the M-CMF (Bauer & Harwood, 2013), or 2) multiplying the CMFs of individual treatment measures assuming their effects are independent (i.e., $CMF = CMF_1 \times CMF_2$ for two treatments). The latter approach is the most popular and is also suggested by the Highway Safety Manual. We follow this second approach for the M-CMFs from the NB model. However, for the KR method, as it follows a nonparametric approach, such assumptions are not necessary. The data-driven process of KR method automatically considers the joint effects of multiple treatments.

We present an example of M-CMF for the changing shoulder width (left) and horizontal curve deflection (CD). Figures 6-7 (a), (b) and (c) show the results from the KR method and the NB model for three dimensions of CD, i.e., 0, 0.5 and 1.5, respectively, arranged in the order of increasing curve sharpness. A section with 5 m shoulder width and zero CD represents the base case in determining M-CMFs. The results presented from the KR method represent the scenario of low traffic volume. As shown, the KR method indicates that the combined crash risk of shoulder width and curve deflection increases with the widening of shoulder width, except for a dimension greater than 4 m for the CD above 0.5. Meanwhile, the M-CMF appears larger in the curved sections (i.e., CD = 1.5).   In contrast, the results from the NB model show a reversed trend with a relatively lower effect.



(a): CD = 0

(b): CD = 0.5

(c): CD = 1.5

Figure 6-7: Multiple CMFs: curve deflection (CD) and shoulder width

6.6 Summary Conclusions

This chapter presented countermeasure studies under the two commonly used frameworks, namely, before-after EB and cross-sectional approaches. A summary of findings specific to each study framework is given below.

Before-after EB study: CMFs were developed for three countermeasures: converting passive control to FLB, converting FLB to FLBG and adding a constant warning time device to FLBG, using before-after crash data of railway-highway grade crossings in Canada. While the parametric models, especially the NB model, have been extensively used under the before-after EB framework, no study has attempted to propose nonparametric models under this framework. We therefore introduced the KR method as an alternative to the NB model with the main motive to take advantage of its data-driven approach to crash estimation. As expected, the two different crash modeling techniques showed some discrepancies in the results of effectiveness measures.

Cross-sectional study: CMFs of four highway geometric features were developed using crash data of Highway 401, Ontario, Canada. We applied both nonparametric (KR method) and parametric (NB model) crash-modeling techniques. The fundamental difference between the results from these two approaches were such that the CMFs from the KR method showed sensitivity to traffic levels unlike those from the NB model. For example, in the case of widening of median width, the results from KR and NB model had a similar trend that showed decreasing crash risk with increasing median width. However, the KR method in high traffic volume indicated a reverse trend. This could be the result of complex nonlinear relation of median width and traffic interaction with crashes. In contrast, in the NB model, only the model coefficient of shoulder width and its associated value play a role. We also explored the applications of the KR and NB models to determine the joint effect of multiple countermeasures. An example of changing shoulder width (left) and horizontal curve deflection was presented and the results from these two approaches were quite different.

Our analysis on the performance of KR and NB methods in both countermeasures studies has revealed the significant differences in the resulting CMFs, but it did not point out which method is relatively better since the true values are unknown. In such cases, it is reasonable to consider the method that has the highest prediction performance as the favourable one. Following this logic, we can conclude that the results from the KR method are more reliable as our previous study in Chapter 4 have shown strong evidence that it performs better than the NB model in terms of model fitting and prediction accuracy. However, the NB model has been widely used in research and practice with a large body of knowledge being accumulated. A meta-heuristic approach could be taken to combine the estimates from the parametric and nonparametric methods.

Chapter 7
Conclusions and Future Research

In road safety studies, such as in identification of crash hotspots and analysis of safety countermeasures, the most popular approach to crash modeling is of parametric nature, in which, crash frequency is assumed to follow a certain distribution. One of the reasons for its popularity is the adoption of relatively simple forms of model structures, making it easy to comprehend and convenient to apply for analyzing road safety problems. However, there is a risk of modeling bias as the model form that relates crashes and risk factors requires prior specification. In addition, a simple parametric form for crash modeling provides limited flexibility in capturing the underlying complex relations. An alternative to this could be a nonparametric approach, which relaxes restriction of parametric model pre-specification and allows the data to speak for themselves. However, the nonparametric approach has not been explored extensively in past road safety studies and its potential and limitations have not been fully understood. The primary objectives of this thesis are therefore to introduce alternative data-driven nonparametric methods to crash modeling, investigate their potential applications for various road safety studies, and compare their performances with their parametric counterparts. This chapter highlights the main contributions of this thesis followed by direction for future research.

7.1 Contributions

The main contributions of this thesis are as follows:

- **Conducted an in-depth investigation of different approaches to crash modeling techniques**

  This research conducted a detailed literature review of various crash modeling techniques, broadly categorized as parametric and nonparametric approaches. The review suggested that parametric models are the most popular form adopted by both frontline researchers and the practitioners. Examples of these models include the standard Poisson, negative binomial (NB), Poisson-lognormal, zero-inflated Poisson, zero-inflated Negative Binomial models. While each of these models provides an easy-to-apply tool due to an involvement of simple mathematical construct relating crash risk and a set of risk factors, they come at a cost of need for pre-selection of the model form, which could easily lead to biased outcomes. Studied also indicated that these parametric crash models are determined by two popular calibration techniques: the maximum likelihood estimation (MLE) and the Bayesian methods. The latter calibration technique is known to have a significant role in improving the accuracy of models based on a relatively smaller crash dataset. Meanwhile, the nonparametric approach, a specification free crash modeling method, has been relatively less explored in past road safety studies because it is often perceived as a “Black Box” technique. Among the few nonparametric methods that were previously employed are Classification and Regression Tree (CART), Artificial Neural Network (ANN), kernel regression and Multivariate Adaptive Regression Splines (MARS) methods.

- **Introduced a nonparametric method to crash modeling including its extension to an EB-based framework**

  In this thesis, we introduced a nonparametric method called kernel regression (KR) for road safety studies. The KR method is a fully data-driven method without any hidden model structure. It is relatively simple to understand as the parameters involved (i.e., bandwidths) are easily interpretable; therefore, this method is often characterized as a “Grey-Box” technique. In contrast, some other nonparametric methods (e.g., ANN, MARS), involving complex hidden structures with difficulty in their interpretation, are characterized as “Black Box” techniques. Another added benefit of the KR method is that it is highly adaptive to changes in system conditions. This is because the KR method can use all the new data directly in making a prediction, unlike other nonparametric methods that require calibration of their hidden model structures prior to their applications. Whenever a new dataset is available, the data can be easily pooled into the original dataset, and the results can be updated using the KR method with a minimal effort. In contrast, in other nonparametric methods, unless their hidden structures are re-trained using an updated training set, they cannot make use of the newly collected information.

  Similarly, another modeling contribution made in this thesis involves extending the KR method in an Empirical Baye’s (EB) framework. As it is commonly recognized, the attractiveness of EB-based framework is that it combines both the site-specific crash history and expected crashes from a crash model to estimate the risk of a study site. We developed a similar EB approach where the role of a parametric model (e.g., NB, PLN) was substituted by the nonparametric KR method. Note that this approach can be adopted in any nonparametric method.

- **Developed a variable selection algorithm for a nonparametric approach**

  While the KR method proposed in this thesis provides an alternative data-driven nonparametric technique to crash modeling, this method lacks a systematic process of selecting a list of relevant explanatory variables. To address this issue, we developed a bootstrap-based algorithm designed to measure the relative safety effects of each potential risk factor. We performed a simulation study to validate the algorithm, and also conducted a few case studies to explore its practical implications. Meanwhile, the performance of the algorithm was benchmarked to its parametric counterpart of the variables selection process.

  Overall, the comparison results indicated a strong correlation between the variable importance measure from the algorithm and its corresponding indicator from the parametric models. Furthermore, the key findings are as follows. First, the proposed algorithm was shown quite robust in capturing the impact of variables at their individual levels. Whenever a selected variable appears less significant in terms of its magnitude of the effect, it is recommended to exclude it. However, the final decision is made based on the optimum performance of the model. Second, we may also employ a parametric model for selecting important variables in a nonparametric method. However, the result of this short-cut approach is expected to be less biased when the model specification of a selected parametric model is relatively accurate. Finally, this developed algorithm can also be applied to other nonparametric methods that lack a variable selection process.

- **Conducted a comprehensive performance comparison of crash models using parametric and nonparametric approaches**

  This thesis conducted a systematic comparison of crash models using parametric and nonparametric approaches to identify differences in their performance. For this, we focused on comparing two popular techniques from the two approaches: the KR method for the nonparametric approach and the NB model, the most extensively used parametric method in road safety studies, for the parametric counterpart. A validation approach was adopted in which the original dataset was split into training and testing sets, the training set being used for model calibration or bandwidth calculation and the testing set for computing goodness-of-fit measures. Three case studies consisting of large crash datasets showed that the KR method has relatively better performance than its parametric counterpart. This could be due to the KR method potentially reducing the modeling bias of the NB model by imposing no specific model structure on the expected crash frequency other than considering of a smoothing parameter, i.e., bandwidth.

Next, we compared the performance of these two methods (KR and NB) in extracting the underlying relationship between crashes and various risk factors. As the KR method does not contain any variable specific interpretable parameters to quantify their effects, we generated partial regression plot for each factor. The nonparametric method was shown to be successful in capturing some sensible nonlinear effects of various factors on crashes. This could be the main reason that in the comparison of goodness-of-fit measures, the KR method showed better results.

- **Examined the relative performance of crash models using parametric and nonparametric techniques for varying data size**

  The nonparametric approach is often characterized as a data-hungry technique as it requires a relatively large dataset to exhibit performance advantage. However, no study was found in road safety literature that involved comparing the performance of this approach with the parametric counterpart in relation to changing data size. Despite the data-hungry nature of the nonparametric method, studying its relative performance could provide insights into the selection of an appropriate crash modeling technique. Therefore, this motivated us to develop a bootstrap-based validation algorithm to investigate their relative performance. The algorithm was designed such that the original dataset was repetitively resampled to obtain its subsets with varying sample sizes, which were subsequently used for performance comparison of the KR and the NB methods. Through a rigorous bootstrapping validation process, we found that the two approaches exhibit strikingly different patterns in terms of sensitivity to data size. The performance of the KR method improved significantly as the data size grew, which was not the case for the NB model. This finding is a good indication for the future application of the data-hungry nonparametric approaches as an alternative to the traditional parametric models since high-quality crash data are growing steadily in size due to latest advancement in information technologies.

- **Developed a framework for network screening using nonparametric approach and compared it to its parametric counterpart**

  In this thesis, we demonstrated the practical application of the KR method as an alternative data-driven nonparametric method for network screening, including ranking of highway sections based on their relative risk and selection of crash hotspots. The nonparametric method was employed under the two popular network screening frameworks, i.e., regression-based and EB-based. For comparison purposes, we also considered the traditional NB model for the same analysis. A case study was conducted using crash data from the busiest highway in Canada - Highway 401.

  The comparative results in terms of ranking of sites and identification of hotspots showed that the nonparametric and parametric approaches have more similarities when applied in the EB-based framework, irrespective of the ranking measures, than in the regression-based framework. Meanwhile, their differences under the regression-based framework were relatively high for the crash rate ranking measure. Similar results were obtained while comparing their crash hotspots. One of the reasons for obtaining similar results using nonparametric and parametric methods under the EB-based framework could be the inclusion of site-specific crash counts while estimating the crash risk. It was also noted that the difference in the list of crash hotspots from the two methods decreased as the percentile of site selection increased, thereby suggesting that the choice of crash modeling approach could be of less importance while considering a relatively large number of sites. While the true ranking results and crash hotspots were not known in the comparisons, those from the nonparametric method (regression-based or EB-based) are expected to be relatively unbiased due to their higher performance as concluded in our previous findings.

- **Developed a framework for countermeasure study using nonparametric approach and compared it to its parametric counterpart**

  This thesis demonstrated the application of the KR method as an alternative to the traditionally used parametric models to countermeasure study, which involves determining the safety effectiveness of treatment measures. The two popular approaches, the before-after EB study and the cross-sectional study, were considered using the parametric (i.e., NB model) and nonparametric (i.e., KR method) methods.

A case study using crash data of railway-highway grade crossing in Canada under the before-after study framework was presented to determine the CMFs of a set of selected countermeasures. While the parametric models, especially the NB model, has been extensively used under this framework, no study has attempted applying nonparametric models for similar study. As expected, the two different crash modeling techniques showed a slight variation in their results. Similarly, we also performed a case study of cross-sectional study using crash data of a highway in Ontario, Canada. The fundamental difference between the results from these two approaches were such that the CMFs from the KR method were able to capture sensitivity to traffic levels whereas the NB model showed no such effect. Furthermore, for determining the CMF of multiple countermeasures, unlike the NB model, the KR method did not require any assumptions to combine the effect of multiple countermeasures. In all these studies, it is expected that the performance of the selected crash model, nonparametric or parametric, has a direct influence on the values of CMFs.

7.2 Future Works

The following are some of the recommendations for future studies on extending this research.

- **Develop a data-driven system to road safety analyses**

  In this thesis, we demonstrated the potential applications of kernel regression including its extended form in an EB approach to road safety analyses, particularly for network screening and countermeasure study. Future works could involve developing a data-driven automated system that performs all the steps involved in these analyses on a single platform. For this, we could divide the system into two main modules: the first related to data handling, such as connecting the system to a continuous flow of data from different sources followed by data processing and data integration; and the second related to the modeling and application part by following the frameworks presented in this thesis. By connecting these two modules, we could automate the entire process and most importantly, take a unique advantage of its high adaptive property to newly collected information. For example, in network screening, the crash hotspots list is expected to change as site-specific crash history and/or site characteristics change. Through this proposed data-driven system, soon after we have new crash data, which could be collected in a yearly basis, the new crash hotspots list can be easily updated. However, one of the challenges while developing this system, especially for a large data size as desired for better performance, is the need for a relatively powerful computation environment. This is because the application of the KR method, being fully data-driven, involves the use of all the data points when making a prediction.

- **Explore alternative methods to improve the performance of kernel regression**

  The "Nadaraya-Watson" kernel regression proposed in this thesis could be improved or extended in several aspects. First, we can investigate the use of the cross-validation approach to determine bandwidths as it is known to provide relatively less biased results. However, this approach of bandwidth calculation involves a large number of computations given its direct correlation to the variable dimension and the data size. Therefore, this approach may demand a high computation environment. Second, we could apply the locally weighted local polynomial regression (LWLPR) introduced by Fan (1993), which is an extended version of the "Nadaraya-Watson" KR method. Comparatively, the LWLPR method is known to have better performance at the boundary of the regression space; however, it could be interesting to explore if their difference in overall performance is significant, especially in the case of big data size.

- **Compare performance of methods related to kernel approach (spatial and non-spatial) in road safety studies**

  As in any parametric count models, the KR method also has a limitation in accounting the spatial correlations of crashes in the road network while estimating their crash risk. By contrast, the kernel density estimate (KDE) method when applied in a spatial framework does take into account of their spatial correlations. This KDE method is simple, and therefore, quite popular in network screening for determining the crash hotspots. However, as it is a univariate technique, this method does not consider the effects of external factors in its risk calculations (Thakali et al., 2015)[10]. With these three alternative nonparametric methods of kernel type, i.e., KDE, KR and KR-based EB methods, it would be interesting to compare their performances and explore how the results vary across the methods.

[10] Thakali, L., Kwon, T. J., & Fu, L. (2015). Identification of crash hotspots using kernel density estimation and kriging methods: a comparison. *Journal of Modern Transportation*, *23*, 93–106.

Bibliography

Abbess, C., D. Jarret, & C. C. Wright. (1981). Accidents at Blackspots: Estimating the Effectiveness of Remedial Treatment, with Special Reference to the Regression-to-the-Mean Effect. *Traffic Engineering and Control*, Vol. 22, No. 10, pp. 535–542.

Abdel-Aty, M., & Haleem, K., (2011). Analyzing Angle Crashes at Unsignalized Intersections Using Machine Learning Techniques. *Accident Analysis & Prevention*, 43(1), 461-470.

Aguero-Valverde, J., & Jovanis, P. P. Analysis of Road Crash Frequency with Spatial Models. *Transportation Research Record: Journal of the Transportation Research Board*, 2061(1), 2008, pp.55-63.

Ahmed, M., Huang, H., Abdel-Aty, M., & Guevara, B. (2011). Exploring a Bayesian Hierarchical Approach for Developing Safety Performance Functions for a Mountainous Freeway. *Accident Analysis and Prevention*, 43(4), 1581-1589.

Al-Ghamdi, A.S., (2007). Experimental evaluation of fog warning system. *Accident Analysis & Prevention*. 39, 1065–1072.

Al-Masaeid, H.R., Performance of Safety Evaluation Methods, *Journal of Transportation Engineering*, Vol. 123, No. 5, 1997, pp. 364-369.

Anastasopoulos, P. C., & Mannering, F. L. (2009). A Note on Modeling Vehicle Accident Frequencies with Random-Parameters Count Models. *Accident Analysis and Prevention*, 41(1), 153-159.

Andrey, J., Mills, B., Vandermolen, J. (2001). Weather Information and Road Safety. Institute for Catastrophic Loss Reduction, Toronto, Ontario, Canada. Paper Series – No. 15.

Antoniou, C., Yannis, G., Katsohis, D. (2013). Impact of meteorological factors on the number of injury accidents. In: Proceedings of the 13th World Conference on Transportation Research, 15–18 July, 2013, COPPE—Federal University of Rio de Janeiro at Rio de Janeiro, Brazil.

Bahar, G. (2010). Methodology for the Development and Inclusion of Crash Modification Factors in the First Edition of the Highway Safety Manual. *Transportation Research E-Circular*, (E-C142).

Bahar, G., Mollett, C., Persaud, B., Lyon, C., Smiley, A., Smahel, T., and H. McGee. (2004). Safety Evaluation of Permanent Raised Pavement Markers. National Cooperative Highway Research Program, NCHRP Report 518, TRB, National Research Council, Washington, D.C.

Baker, C.J., Reynolds, S. (1992). Wind-induced accidents of road vehicles. *Accident Analysis & Prevention*. 24 (6), 559–575.

Begum, M., Persaud, B., Nichol, S., and Lyon, C. (2009). Safety Performance of Ontario Road Segments. *In Proc., 19th Canadian Multidisciplinary Road Safety Conference*, pp. 8-10.

Begum, M., Safety Performance Assessment of Ontario Highway Sections, M.Sc. Thesis, Ryerson University, 2008.

Benekohal, R. F., & Hashmi, A. M. (1992). Procedures for Estimating Accident Reductions on Two-Lane Highways. *Journal of Transportation Engineering*, 118, 111–129.

Bissantz, N., and Munk, A. (2001). New Statistical Goodness of Fit Techniques in Noisy Inhomogeneous Inverse Problems with Application to the Recovering of the Luminosity Distribution of the Milky Way. *Astronomy and Astrophysics*, 376, pp. 735–744.

Bissantz, N., Munk, A. and Scholz, A. (2003), Parametric Versus Non-Parametric Modelling? Statistical Evidence Based On P-Value Curves. Monthly Notices of the Royal Astronomical Society, 340: 1190–1198.

Blundell, R., and A. Duncan.(1995). Kernel Regressions in Empirical Micro- economics. *Journal of Human Resources*, Vol. 33, No. 1, pp. 62–87.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Brown, M., Barrington-Leigh, C., & Brown, Z. (2012). Kernel Regression for Real-Time Building Energy Analysis. *Journal of Building Performance Simulation*, 5(4), 263–276.

Cameron, A. C., & Trivedi, P. K. (2013). Regression Analysis of Count Data (Vol. 53). Cambridge university press.

Carter, D., Srinivasan, R., Gross, F., Council, F., (2012). Recommended Protocols for Developing Crash Modification Factors. NCHRP 20-7(314) Final Report.

Cateni, S., Colla, V., & Vannucci, M. (2011). A Genetic Algorithm-Based Approach for Selecting Input Variables and Setting Relevant Network Parameters of a SOM-Based Classifier. In International *Journal of Simulation Systems, Science & Technology*. UKSim 4th European Modelling Symposium on Mathematical modelling and computer simulation (Vol. 12, No. 2).

Cateni, S., Vannucci, M., Vannocci, M., & Colla, V. (2012). Variable Selection and Feature Extraction through Artificial Intelligence Techniques. *Multivariate Analysis in Management, Engineering and the Science*, 103-118.

Chang, L. Y. (2005). Analysis of Freeway Accident Frequencies: Negative Binomial Regression versus Artificial Neural Network. *Safety science*, 43(8), 541-557.

Chen, E., & Tarko, A. P. (2014). Modeling Safety of Highway Work Zones with Random Parameters and Random Effects Models. *Analytic methods in Accident Research*, 1, pp. 86-95.

Cheng, L., Geedipally, S. R., & Lord, D (2013). The Poisson–Weibull Generalized Linear Model For Analyzing Motor Vehicle Crash Data. *Safety science*, 54, pp. 38-42.

Cheng, W., & Washington, S. P. (2005). Experimental Evaluation of Hotspot Identification Methods. *Accident Analysis and Prevention*, 37(5), 870–81.

Cheng, W., & Washington, S., (2008). New Criteria for Evaluating Methods of Identifying Hot Spots. *Transportation Research Record: Journal of the Transportation Research Board*, 2083, p 76-85.

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 and Prevention*, 35(2), 253.

Choi, Y. Y., Kho, S. Y., Lee, C., & Kim, D. K. (2015). Development of Crash Modification factors of Alignment Elements and Safety Countermeasures for Korean Freeways. *In Transportation Research Board 94th Annual Meeting (*No. 15-0503).

Connors, R.D., Maher, M., Wood, A., Mountain, L., Ropkins, K. Methodology for Fitting and Updating Predictive Accident Models with Trend. *Accident Analysis and Prevention* 56, 2013, pp.82–94.

Council, F. M., and Stewart, J. R. (1999). Safety Effects Of The Conversion Of Rural Two-Lane To Four-Lane Roadways Based On Cross-Sectional Models. *Transportation Research Record: Journal of the Transportation Research Board*, 1665(1), 35-43.

Daniels, S., Brijs, T., Nuyts, E., & Wets, G. (2010). Explaining Variation in Safety Performance of Roundabouts. *Accident Analysis & Prevention*, 42(2), pp.393-402.

Deacon, J. A., C. V. Zegeer, and R. C. Deen. (1975). Identification of Hazardous Rural Highway Locations. In Transportation Research Record 543, TRB, National Research Council, Washington, D.C., pp. 16–33.

Dudek, G. (2012). Variable Selection In The Kernel Regression Based Short-Term Load Forecasting Model. *In Artificial Intelligence and Soft Computing*, pp. 557-563.

El-Basyouny, K., & Sayed, T. (2009). Collision Prediction Models Using Multivariate Poisson-Lognormal Regression. *Accident Analysis & Prevention*, 41(4), pp.820-828.

El-Basyouny, K., &Sayed, T. (2006). Comparison of Two Negative Binomial Regression Techniques in Developing Accident Prediction Models. *Transportation Research Record: Journal of the Transportation Research Board*, 1950(1), 9-16.

Elvik, R. (2008). Comparative Analysis of Techniques for Identifying Locations of Hazardous Roads. *Transportation Research Record: Journal of the Transportation Research Board*, 2083(1), 72-75.

Elvik, R., Amundsen, F.H., and F. Hofset. (2001). Road Safety Effects of Bypasses, Journal of the Transportation Research Record 1758, TRB, National Research Council, Washington, D.C., pp. 13-20.

Eubank, R.L. (1999). Nonparametric Regression and Spline Smoothing, Statistics: Textbooks and Monographs, Vol. 157 (2nd edition) Marcel Dekker, New York.

Federal Highway Administration (FHWA). (2015). http://safety.fhwa.dot.gov/xings/com_roaduser/07010/sec04b.cfm#i. Federal highway Administration. Accessed 02/11/2015.

Federal Highway Administration (FHWA). (2015). U.S. Department of Transportation. Crash Modification Factor Clearinghouse. http://www.cmfclearinghouse.org/.

Fitzpatrick, K., Lord, D., & Park, B.-J. (2008). Accident Modification Factor for Medians on Freeways and Multi Lane Rural Highway in Texas. *Transportation Research Record: Journal of the Transportation Research Board*, 136(9), 827–835.

Fu, L., Moreno, L. M., Shah, Q. A., & Lee, C., (2005). Effects of Winter Weather and Maintenance Treatments on Highway Safety, Ministry of Transportation, Ontario.

Gallo, M., Marzano, V., & Simonelli, F. (2012). Empirical Comparison of Parametric and Nonparametric Trade Gravity Models. *Transportation Research Record: Journal of the Transportation Research Board*, 2269(1), 29-41.;

Garrison, D., Mannering, F. (1990). Assessing the Traffic Impacts of Freeway Incidents and Driver Information. *ITE J*. 60 (8), 19–23.

Geedipally, S., & Lord, D. (2010). Identifying Hot Spots by Modeling Single-Vehicle and Multivehicle Crashes Separately. *Transportation Research Record: Journal of the Transportation Research Board*, 2147, 97–104.

Goethals, D. T., P. L. M., & Pauw, N. De. (2001). Use of Genetic Algorithms to select Input Variables in Artificial Neural Network Models for the Prediction of Benthic Macroinvertebrates. *Environmental Toxicology*, 2, 136–141.

Greibe, P. (2003). Accident Prediction Models For Urban Roads. *Accident Analysis and Prevention*, 35 (2), 273−285.

Griffith, M. S. (1999). Safety Evaluation Of Rolled-In Continuous Shoulder Rumble Strips Installed On Freeways. *Transportation Research Record: Journal of the Transportation Research Board*, 1665(1), 28-34.

Gross, F., Persaud, B., & Lyon, C. (2010). A Guide to Developing Quality Crash Modification Factors-FHWA Report.

Gu, J., Li, D., & Liu, D. (2007). Bootstrap Non-Parametric Significance Test. *Journal of Nonparametric Statistics*, 19(6-8), 215-230.

Hadi, M. A., Aruldhas, J., Chow, L. F., & Wattleworth, J. A. (1995). Estimating Safety Effects of Cross-Section Design for Various Highway Types Using Negative Binomial Regression. *Transportation Research Record: Journal of the Transportation Research Board*, 1500, 169.

Haleem, K., Abdel-Aty, M., & Santos, J. (2010). Multiple Applications of Multivariate Adaptive Regression Splines Technique to Predict Rear-End Crashes at Unsignalized Intersections. *Transportation Research Record: Journal of the Transportation Research Board*, 2165(-1), 33–41.

Hall, P., Li, Q., & Racine, J. S. (2007). Nonparametric Estimation of Regression Functions In The Presence Of Irrelevant Regrssionros. *Review of Economics and Statistics*, 89, 784-789.

Hardle, W., & Mammen, E., (1993). Comparing Nonparametric versus Parametric Regression Fits. *The annals of Statistics*, 21, p 1926–1947.

Harkey, D. L., Srinivasan, R., Baek, J., Council, F. M., Eccles, K., Lefler, N., Gross, F., Hauer, E., Bonneson, J. A. (2008). Accident Modification Factors for Traffic Engineering and ITS Improvements. NCHRP Report 633, Transportation Research Board, Washington, D.C.

Harwood, D., Bauer, K., Potts, I., Torbic, D., Richard, K., Rabbani, E., Griffith, M. (2003). Safety Effectiveness of Intersection Left- and Right-Turn Lanes. *Transportation Research Record: Journal of the Transportation Research Board*, 1840(July), 131–139.

Hauer E., (2004). Statistical Road safety modeling. Transportation Research Record: Journal of the Transportation Research Board, No. 1897, TRB, National Research Council, Washington, D.C., pp. 81-87.

Hauer, E, and B.N. Persaud. (1987). How to Estimate the Safety of Rail-Highway Grade Crossings and the Safety Effects of Warning Devices. *Transportation Research Record: Journal of the Transportation Research Board*, 1114, pp. 131-140.

Hauer, E. (1986). On the Estimation of the Expected Number of Accidents. *Accident Analysis & Prevention,* Vol. 18, No. 1, pp. 1–12.

Hauer, E. (1996). Identification of Sites with Promise. *Transportation Research Record: Journal of the Transportation Research Board*, 1542(1), 54–60.

Hauer, E. (1997). Observational Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety, Pergamon.

Hauer, E. (1999). Safety and the Choice of Degree of Curve. In Transportation Research Record: Journal of the Transportation Research Board, No. 1665, TRB, National Research Council Washington, D.C., pp. 22-27.

Hauer, E. (2001). Overdispersion in Modelling Accidents on Road Sections and In Empirical Bayes Estimation. *Accident Analysis and Prevention*, 33(6), 799–808.

Hauer, E., (2015). The Art of Regression Modeling in Road Safety. Cham: Springer International Publishing.

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

Higle, J. L., and J. M. Witkowski (1988). Bayesian Identification of Hazardous Locations. *Transportation Research Record: Journal of the Transportation Research Board*, 1185, pp. 24–36.

Hong, D., Lee, Y., Kim, J., Yang, H. C., & Kim, W. (2005). Development of Traffic Accident Prediction Models by Traffic and Road Characteristics in Urban Areas. *In Proceedings of the Eastern Asia Society for Transportation Studies*, Vol. 5, p 2046-2061.

Hossain, M., & Muromachi, Y. (2012). A Bayesian Network Based Framework for Real-Time Crash Prediction on the Basic Freeway Segments of Urban Expressways. *Accident Analysis & Prevention*, 45, 373-381.

Huang, H., Chin, H., & Haque, M. (2009). Empirical Evaluation of Alternative Approaches in Identifying Crash Hot Spots: Naive Ranking, Empirical Bayes, and Full Bayes Methods. *Transportation Research Record: Journal of the Transportation Research Board*, (2103), 32-41.

Jacobs G, Aeron-Thomas A and Astrop A. (2000). Estimating Global Road Fatalities. Crowthorne, Transport Research Laboratory (TRL Report, No. 445).

Jacobs, G. D., & Sayer, I., (1983). Road Accidents in Developing Countries. *Accident Analysis and Prevention*, 15, p 337-353.

Jones, B., Janssen, L., & Mannering, F. (1991). Analysis of the Frequency and Duration of Freeway Accidents in Seattle. *Accident Analysis & Prevention*, 23(4), 239-255.

Jovanis, P. P., and Chang, H. (1986). Modeling The Relationship Of Accidents To Miles Traveled. *Transportation Research Record: Journal of the Transportation Research Board*, 1068, 42-51.

Karlaftis, M. G., & Golias, I. (2002). Effects of Road Geometry and Traffic Volumes on Rural Roadway Accident Rates. *Accident, Analysis and Prevention*, 34, p 357–65.

Karlaftis, M., Yannis, G. (2010). Weather effects on daily traffic accidents and fatalities: a time series count data approach. In: Proceedings of the 89th Annual Meeting of the Transportation Research Board, January 10–14, 2010, Washington, D.C.

Kass, R.E., Wassermann, L. (996). Selecting prior distributions by formal rules. *Journal of the American Statistical Association* 91, 1343– 1370.

Kendall’s Advanced Theory of Statistics, Arnold, London. (1998). 6th Edition, Volume 1, by Stuart & Ord, p. 351.

Knapp, K. K., Smithson, D.L., Khattak, A. J. (2000). The Mobility and Safety Impacts of Winter Storm Events in a Freeway Environment. Mid-Continent Transportation Symposium, May 15-16, Iowa State University, Ames, Iowa.

Köhler, M., Schindler, A., & Sperlich, S. (2014). A Review and Comparison of Bandwidth Selection Methods for Kernel Regression. *International Statistical Review*.

Lambert, D. (1992). Zero-Inflated Poisson Regression, With an Application to Defects in Manufacturing. *Technometrics* 34, p 1–14.

Lan, B., Persaud, B., Lyon, C., & Bhim, R. (2009). Validation of a Full Bayes Methodology for Observational Before-After Road Safety Studies and Application to Evaluation of Rural Signal Conversions. *Accident Analysis and Prevention*, 41(3), 574–80.

Laughland, J. C., Haefner, L. E., Hall, J. W., and Clough, D. R. (1975). Methods for Evaluating Highway Safety Improvements (No. HS-018 724).

Lavergne, P., & Vuong, Q. (2000). Nonparametric Significance Testing. *Econometric Theory*, 16(04), 576-601.

Li, X., Lord, D., Zhang, Y., & Xie, Y. (2008). Predicting Motor Vehicle Crashes Using Support Vector Machine Models. *Accident Analysis & Prevention*, 40 (4), 1611-1618.

Livanis, G., Salois, M., & Moss, C. (2009). A Nonparametric Kernel Representation of the Agricultural Production Function: Implications for Economic Measures of Technology. *In 83rd Annual Conference of the Agricultural Economics Society*, Dublin.

Lord, D., & Bonneson, J. A. (2008). Development of Accident Modification Factors for Rural Frontage Road Segments in Texas. *Transportation Research Record: Journal of the Transportation Research Board*, 2023(-1), 20–27.

Lord, D., & Mannering, F., (2010). The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. *Transportation Research Part A: Policy and Practice*, 44, p 291–305.

Lord, D., & Miranda-Moreno, L. F. (2008). Effects Of Low Sample Mean Values And Small Sample Size On The Estimation Of The Fixed Dispersion Parameter Of Poisson-Gamma Models For Modeling Motor Vehicle Crashes: A Bayesian Perspective. *Safety Science*, 46, 751–770.

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

Lyon, C., Haq, A., Persaud, B.N., and S.T. Kodama. (2005). Safety Performance Functions for Signalized Intersections in Large Urban Areas: Development and Application to Evaluation of Left-Turn Priority Treatment. *Transportation Research Record: Journal of the Transportation Research Board*, 1908, pp. 165-171.

Ma, J. and K. M. Kockelman. (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.

Maher, M. J., & Summersgill, I. (1996). A Comprehensive Methodology for the Fitting of Predictive Accident Models. *Accident Analysis & Prevention*, 28(3), 281-296.

McCullagh, Peter, and John A. Nelder (1989). Generalized linear models. Vol. 37. CRC press.

McGuigan, D. R. D (1981). The Use of Relationships between Road Accidents and Traffic Flow in ‘Black-Spot’ Identification. *Traffic Engineering and Control*, Aug.–Sept. 1981, pp. 448–453.

McGuigan, D. R. D. (1982). Nonjunction Accident Rates and Their Use in ‘Black-Spot’ Identification. *Traffic Engineering and Control*, Feb, pp. 60–65.

Miaou, S. P (1994). The Relationship between Truck Accidents and Geometric Design of Road Sections: Poisson versus Negative Binomial Regressions. *Accident Analysis & Prevention*, 26(4), pp. 471-482.

Miaou, S. P., & Lord, D. (2003). Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes versus Empirical Bayes Methods. *Transportation Research Record: Journal of the Transportation Research Board*, (1840), pp. 31-40.

Miaou, S. P., & Song, J. J. (2005). Bayesian Ranking of Sites for Engineering Safety Improvements: Decision Parameter, Treatability Concept, Statistical Criterion, and Spatial Dependence. *Accident Analysis and Prevention*, 37(4), 699-720.

Miaou, S. P., and Lum, H. (1993). Modeling Vehicle Accidents and Highway Geometric Design Relationships. *Accident Analysis & Prevention*, 25(6), 689-709.

Miaou, S.-P., Hu, P.S., Wright, T., Rathi, A.K., Davis, S.C. (1992). Relationship between Truck Accidents and Highway Geometric Design: A Poisson Regression Approach. *Transportation Research Record: Journal of the Transportation Research Board*, 1376.

Milton, J. and Mannering, F. (1998). The Relationship among Highway Geometrics, Traffic-Related Elements and Motor Vehicle Accident Frequencies. *Transportation* 25, 395-413.

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

Miranda-Moreno, L. F., Fu, L., Saccomanno, F. F., & Labbe, A. (2005). Alternative Risk Models for Ranking Locations for Safety Improvement. *Transportation Research Record: Journal of the Transportation Research Board*, 1908, p 1-8.

Miranda-Moreno, L. F., Heydari, S., Lord, D., & Fu, L. (2013). Bayesian Road Safety Analysis: Incorporation of Past Evidence and Effect of Hyper-Prior Choice. *Journal of safety research*, 46, 31-40.

Miranda-Moreno, L. F., Labbe, A., & Fu, L. (2007). Bayesian Multiple Testing Procedures for Hotspot Identification. *Accident Analysis & Prevention*, 39(6), 1192-1201.

Miranda-Moreno, L., Fu, L., Saccomanno, F., & Labbe, A. (2005). Alternative Risk Models for Ranking Locations for Safety Improvement. *Transportation Research Record: Journal of the Transportation Research Board*, 1908, 1-8.

Miranda-Moreno, L.F. (2006). Statistical Models and Methods for Identifying Hazardous Locations for Safety Improvements. PhD thesis report, University of Waterloo.

Mitra, S., & Washington, S. (2012). On The Significance of Omitted Variables in Intersection Crash Modeling*. Accident Analysis & Prevention*, 49, pp.439-448.

Mohamed A. Abdel-Aty, A.Essam Radwan. (2000). Modeling Traffic Accident Occurrence and Involvement, *Accident Analysis & Prevention*, Vol 32, pp 633-642.

Montella, A. (2010). A Comparative Analysis of Hotspot Identification Methods. *Accident Analysis & Prevention*, 42(2), 571-581.

Nadaraya, E. A. (1964). On Estimating Regression. Theory of Probability & Its Applications, 9(1), pp.141-142.

National Cooperative Highway Research Program (NCHRP). (2008). Crash Reduction Factors for Traffic Engineering and ITS Improvements. Final Report NCHRP Project 17-25 (to appear as NCHRP Report 572).

Ntzoufras, I. (2008). Bayesian Modeling Using WinBUGS. Wiley.

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

Pagan A., Ullah A. (1999). Nonparametric Econometrics, Cambridge University Press.

Park, J., Abdel-Aty, M., & Lee, C. (2014). Exploration and Comparison of Crash Modification Factors for Multiple Treatments on Rural Multilane Roadways. *Accident Analysis and Prevention*, 70, 167–177.

Park, Y.-J. (2007). Estimating Effectiveness of Countermeasures Based on Multiple Sources: Application to Highway-Railway Grade Crossings, University of Waterloo.

Parmeter, C. F., Zheng, Z., & McCann, P. (2009). Cross-Validated Bandwidths and Significance Testing. *Advances in Econometrics*, 25, 71.

Persaud, B. N. (1990). Blackspot Identification and Treatment Evaluation. Ontario Ministry of Transport, TDS-90-04.

Persaud, B. N. (1994). Accident Prediction Models For Rural Roads. *Canadian Journal of Civil Engineering*, 21 (4), pp. 547–554.

Persaud, B. N. (2001). NCHRP Synthesis of Highway Practice 295: Statistical Methods in Highway Safety Analysis. TRB, National Research Council, Washington, D.C.

Persaud, B. N., Lan, B., Lyon, C., & Bhim, R. (2010). Comparison Of Empirical Bayes And Full Bayes Approaches For Before-After Road Safety Evaluations. *Accident; Analysis and Prevention*, 42(1), 38–43.

Persaud, B. N., Lyon, C., & Nguyen, T. (1999). Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Safety Improvement. *Transportation Research Record: Journal of the Transportation Research Board,* 1665, p 7-12.

Persaud, B. N., Retting, R. A., Garder, P. E., & Lord, D. (2001). Observational Before-After Study of the Safety Effect of U. S. Roundabout Conversions Using the Empirical Bayes Method. *Transportation Research Record: Journal of the Transportation Research Board*, (1751), 1–8.

Prinzie, A., Poel, D.V. (2008). Random Forests for Multiclass Classification: Random Multinomial Logit. *Expert Syst*. Appl. 34 (3), 1721–1732.

Qin, X., Ivan, J. N., & Ravishanker, N. (2004). Selecting Exposure Measures in Crash Rate Prediction for Two-Lane Highway Segments. *Accident Analysis and Prevention*, 36, p 183–191.

Racine, J. (1997). Consistent Significance Testing For Nonparametric Regression. *Journal of Business & Economic Statistics*, 15(3), 369-378.

Racine, J. S. (2008). Nonparametric Econometrics: A primer. Now Publishers Inc.

Saccomanno, F. F., Fu, L., & Miranda-Moreno, L. F. (2004). Risk-Based Model for Identifying Highway-Rail Grade Crossing Blackspots. *Transportation Research Record: Journal of the Transportation Research Board*, 1862(1), 127-135.

Saccomanno, F. F., Grossi, R., Greco, D., and Mehmood, A. (2001). Identifying Black Spots along Highway SS107 in Southern Italy Using Two Models. *Journal of transportation engineering*, 127(6), 515-522.

Saccomanno, F., & Lai, X. (2005). A Model for Evaluating Countermeasures at Highway-Railway Grade Crossings. *Transportation Research Record: Journal of the Transportation Research Board*, 1918(1).

Saccomanno, F., Fu, L., & Roy, R. (2001). Geographic Information System—Based Integrated Model for Analysis and Prediction of Road Accidents. *Transportation Research Record: Journal of the Transportation Research Board*, 1768(-1), 193–202.

Scott, P.P. (1986). Modelling time-series of British road accident data. *Accident Analysis & Prevention*. 18 (2), 109–117.

Shankar, V., Mannering, F., & Barfield, W. (1995). Effect of Roadway Geometrics and Environmental Factors on Rural Freeway Accident Frequencies. *Accident Analysis and Prevention*, 27(3), 371-389.

Shankar, V., Milton, J., and Mannering, F. (1997). Modeling Accident Frequencies as Zero-Altered Probability Processes: An Empirical Inquiry. *Accident Analysis & Prevention*, 29(6), 829-837.

Silverman, B. W. (1984). Spline Smoothing: The Equivalent Variable Kernel Method. *The annals of Statistics*, 12(3), pp. 898–916.

Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis (Vol. 26). CRC press.

Simonoff, J. S. (1996). Smoothing Methods in Statistics. Springer.

Song, J., Ghosh, M., Miaou, S., & Mallick, B. (2006). Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis*, 97, 246 – 273.

Srinivasan, S., Kockelman, K.M. (2002). The Impacts of Bypasses in Small and Medium Sized Communities: An Econometric Analysis. *Journal of Transportation and Statistics* 5, 57–69.

Stamatiadis, N., Pigman, J., Sacksteder, J., Ruff, W., Lord, D. (2009). Impact of Shoulder Width and Median Width on Safety. NCHRP Report 633, Transportation Research Board, Washington, D.C.

Stokes, R. W., and Mutabazi, M. I. (1996). Rate-Quality Control Method of Identifying Hazardous Road Locations. *Transportation Research Record: Journal of the Transportation Research Board*, 1542(1), 44-48.

Tarko, A.P., Kanodia, M. (2003). Hazard Elimination Program- Manual on improving safety of Indiana Road Intersections and Sections.

Thakali, L., Fu, L., & Chen, T. (2014). A Comparison between Parametric and Nonparametric Approaches for Road Safety Analysis - A Case Study of Winter Road Safety. *In Transportation Research Board Annual Meeting* (Vol. 6, pp. 1–17).

Thakali, L., Fu, L., & Chen, T. (2016). Model Based versus Data-driven Approach for Road Safety Analysis : Does More Data Help? *Transportation Research Record*: *Journal of the Transportation Research Board,* No. 2601.

Thakali, L., Kwon, T. J., & Fu, L. (2015). Identification of Crash Hotspots Using Kernel Density Estimation and Kriging Methods: A Comparison. *Journal of Modern Transportation*, 23, 93–106.

Tristen Hayfield and Jeffrey S. Racine. (2008). Nonparametric Econometrics: The np Package. Journal of Statistical Software 27(5)

Tunaru, R., 2002. Hierarchical Bayesian models for multiple count data. *Austrian Journal of Statistics* 31 (3), 221–229.

Ukkusuri, S., Hasan, S., & Aziz, H. M. A. (2011). Random Parameter Model Used To Explain Effects Of Built-Environment Characteristics On Pedestrian Crash Frequency. *Transportation Research Record: Journal of the Transportation Research Board*, 2237(1), pp.98-106.

Usman, T., Fu, L., & Miranda-Moreno, L. F. (2012). A Disaggregate Model For Quantifying The Safety Effects Of Winter Road Maintenance Activities At An Operational Level. *Accident Analysis and Prevention*, 48, pp. 368–78.

Usman, Taimur. (2011). Models For Quantifying Safety Benefit Of Winter Road Maintenance. Ph.D. thesis report, University of Waterloo.

Vodden, K., Smith, D., Eaton, F., & Mayhew, D. (2007). Analysis and Estimation of the Social Cost of Motor Vehicle Collisions in Ontario. Ministry of Transportation Ontario, Canada.

Wallman, C.G., and Astrom, H. (2001). Friction Measurement Methods and the Correlation between Road Friction and Traffic Safety — A Literature Review. VTI Report, M911A.

Wang, X., Xie, K., Abdel-Aty, M., Chen, X., and Tremont, P. (2014). Systematic Approach to Hazardous-Intersection Identification and Countermeasure Development. *Journal of Transportation Engineering*, 140(6).

Washington, S.P., Karlaftis, M.G., Mannering, F.L. (2003). Statistical and Econometric Methods for Transportation Data Analysis. Chapman & Hall/CRC.

Watson, G. S. (1964). Smooth Regression Analysis. Sankhya: *The Indian Journal of Statistics*, Series A, 26 (4), pp.359–372.

World Health Organization (WHO). (2015). Global Status Report on Road Safety, WHO Press, World Health Organization, Geneva, Switzerland

Wu, L., Lord, D., & Zou, Y. (2015). Validation of CMFs Derived from Cross - Sectional Studies Using Regression Models. *In Transportation Research Board*.

Wu, Z., Sharma, A., Mannering, F. L., & Wang, S. (2013). Safety Impacts Of Signal-Warning Flashers And Speed Control At High-Speed Signalized Intersections. *Accident Analysis & Prevention*, 54, pp. 90-98.

Xie, Y., Lord, D., & Zhang, Y. (2007). Predicting Motor Vehicle Collisions Using Bayesian Neural Network Models: An Empirical Analysis. *Accident Analysis and Prevention*, 39(5), 922–33.

Zegeer, C. V., Stewart, J. R., Council, F. M., Reinfurt, D. W., & Hamilton, E. (1991). Safety Effects of Geometric Improvements on Horizontal Curves (No. 1356).

Zeng, H., Schrock, D.W. (2013). Safety-Effectiveness of Various Types of Shoulders on Rural Two-Lane Roads in Winter and Non-Winter Periods. In 92nd Annual Meeting of the Transportation Research Board, Washington, D.C.

Zhang, P. (1991). Variable Selection in Nonparametric Regression with Continuous Covariates. *The Annals of Statistics*, pp.1869-1882.

Zhang, X., King, M. L., & Hyndman, R. J. (2006). A Bayesian Approach to Bandwidth Selection for Multivariate Kernel Density Estimation. *Computational Statistics & Data Analysis*, 50(11), 3009-3031.

Zou, Y., Lord, D., Zhang, Y., & Peng, Y. (2013). Comparison Of Sichel And Negative Binomial Models In Estimating Empirical Bayes Estimates. *Transportation Research Record: Journal of the Transportation Research Board*, (2392), 11-21.

Appendix A

A.1 EB Estimate based on Bayesian approach

Applying Bayes' rule with crash count (K) Poisson distributed and mean crash frequency (k) gamma distributed with parameter a and b, we obtain the following expressions:

Property of a gamma distributed random variable (here k) with parameters a and b

$E(k) = b/a$

$Var(k) = b/a^2$

Then,

$a = \frac{E(k)}{Var(k)}; b = \frac{E(k)^2}{Var(k)}$

Applying Baye's rule

$\boldsymbol{E(k/K) = \frac{K+b}{1+a}}$ (EB estimate)

Re-arranging above expression for ***E(k/K),*** the results are same as in the derivation in **A.2**

$$E(k/K) = \frac{K + \frac{E(k)^2}{Var(k)}}{1 + \frac{E(k)}{Var(k)}}$$

$$E(k/K) = \frac{K \times Var(k) + E(k)^2}{Var(k) + E(k)}$$

$$E(k/K) = \frac{K \times Var(k)}{Var(k) + E(k)} + \frac{E(k)^2}{Var(k) + E(k)}$$

$$E(k/K) = \frac{K \times Var(k)}{Var(k) + E(k)} + \frac{E(k) \times E(k)}{Var(k) + E(k)}$$

$$E(k/K) = \frac{K \times Var(k)}{Var(k) + E(k)} + \frac{E(k)}{Var(k)/E(k) + 1}$$

$\underline{\boldsymbol{E(k/K) = K(1-w) + E(k)w}}$

where,

$\boldsymbol{w} = \frac{1}{1+\frac{Var(k)}{E(k)}}; \boldsymbol{1-w} = 1 - \frac{1}{\frac{Var(k)}{E(k)}+1} = \frac{Var(k)}{Var(k)+E(k)}$

Note: Inputs to the expression of "*w*" is the *E(k)* and *Var(k).*

A.2 EB estimate based on approach of combining two random variables

Adding two random variables of different precision

Let X and Y be two independent random variables with variances VAR(X) and VAR (Y) and w a constant. Let us define a new variable Z as:

$Z = wX + (1-w)Y$; then

$$Var(Z) = w^2 Var(X) + (1-w)^2 Var(Y)$$

Determining weight

The value of w is determined by minimizing sum of square deviance (i.e., Var(Z)) as follows

$\frac{dVar(Z)}{dw} = 2 \times w \times Var(X) - 2 \times (1-w) \times Var(Y) = 0$; Therefore

$$w = \frac{Var(Y)}{Var(X) + Var(Y)}, \text{ or}$$

$$w = \frac{\frac{1}{Var(X)}}{\frac{1}{Var(X)} + \frac{1}{Var(Y)}}$$

$$1 - w = \frac{\frac{1}{Var(Y)}}{\frac{1}{Var(X)} + \frac{1}{Var(Y)}}$$

This suggests that the weights w and $1-w$ are inversely proportional to the variance of the two random variables.

Now replacing the above notation with two independent estimates i.e., estimate from a crash model and observed crash counts:

X: corresponds to *E(k)* (model estimate) and *Var(X)* to *Var (k) (in the reference population of k, mean is E(k) and variance is Var(k) as all k's may not be same.*

Y: corresponds to *K* (crash count)

$$w = \frac{\frac{1}{Var(k)}}{\frac{1}{Var(k)} + \frac{1}{Var(K)}}$$

Assuming crash count, "*K*", follows a Poisson distribution with a mean "*k",* from its equal mean variance relation, the *Var(K)= k*. Therefore,

$$w = \frac{1}{1 + \frac{Var(k)}{E(k)}}$$

Hauer (1997) mentioned that “The merit of derivation is in that it does not require any assumptions about the distribution of the k’s in the reference population and agrees with the result of the derivation in which one assumes that the k’s are gamma distributed”.

A.3 Variance-mean relation

We refer to Hauer’s (2015) and Hauer’s (1997) derivation to estimate a relation between variance “*Var (k)”* and mean “*E(k)”* of an estimate mentioned in previous relation to determine the weight (w).

Let’s say we have two random variables- X and Y. From the law of total variance, we get following relation:

$$Var(Y) = E[Var(Y|X)] + Var[E(Y|X)] \qquad \text{(A.3.1)}$$

This fundamental concept is applied in following derivation by replacing *k*- mean crash frequency from a model for X and *K*- observed crashes for Y.

The logic here is “**k” and “K”** are considered **random variables** for the following reasons: k obtained as mean crash frequency from a crash model provides an estimate by relating to some safety related factors. However, the mean crashes (i.e., “k”), across units (road sections or intersections) belonging to same populations could vary as there may be many other excluded unit specific factors. Therefore, k of units from the same population are expected to vary. Meanwhile, we also know that crash counts (i.e., “K”) are random and we normally use Poisson distribution to describe the process. Now, by replacing *k* for the value of X and *K* of Y in above expression we get:

$$Var(K) = E[Var(K|k)] + Var[E(K|k)] \qquad \text{(A.3.2)}$$

Using the property of Poisson distribution, we have following relation for a unit case “i”:

$Var(K_i|k_i) = k_i$ and therefore first summand equals to $E(k)$

$E(K_i|k_i) = k_i$ and therefore, second summand equals to $Var(k)$

For each unit, it follows that

$Var(K) = E(k) + Var(k)$; Or $Var(k) = Var(K) - E(k)$ (A.3.3)

where,

$Var(K)$ is variance of crash counts and

$E(k)$ is mean of crash frequency obtained from the crash model

Imagining population where each road segment (row in dataset) is a sample of *one*, we can estimate the variance of crash counts, i.e., *Var (K),* by the square difference (SD) between the observed crash counts (K) and fitted values (k) *(References: Hauer (2015) p. 207 and Hauer (1997) p. 202).*

$Var(K) = SD$ (A.3.4)

where, SD = $(K - k)^2$

From Eq. A.3.3 and Eq. A.3.4, we obtain final expression for $Var(k)$as:

$Var(k) = SD - E(k)$ (A.3.5)

Appendix B

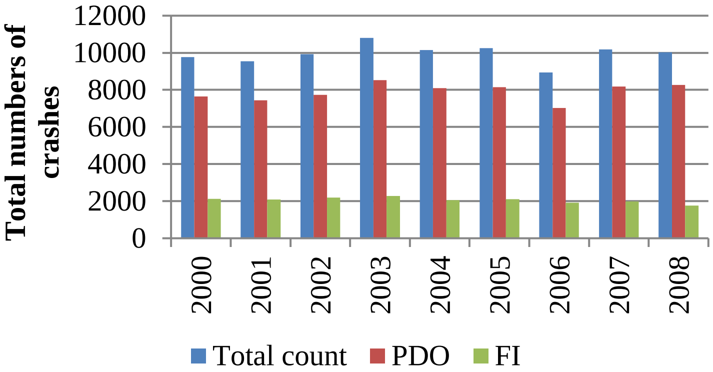
B.1: Geographical locations of case studies.

Hwy 401
Toronto
Buffalo
Detroit
Chicago
Pittsburgh
New York
Philadelphia
Washington

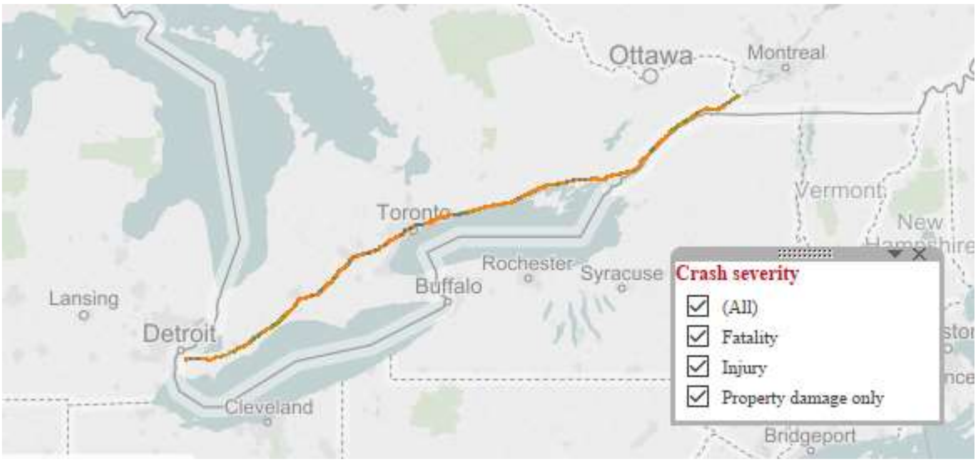
Case 1: Highway 401, Ontario

Case 2: 31 highway patrol routes, Ontario

B.2: Total annual crash counts in Highway 401 based on crash severities



B.3: Crashes after geocoding-Highway 401



Crash severity

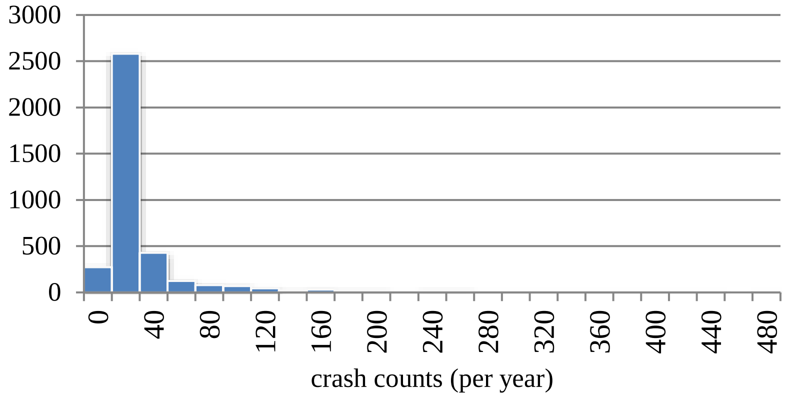
☑ (All)

☑ Fatality

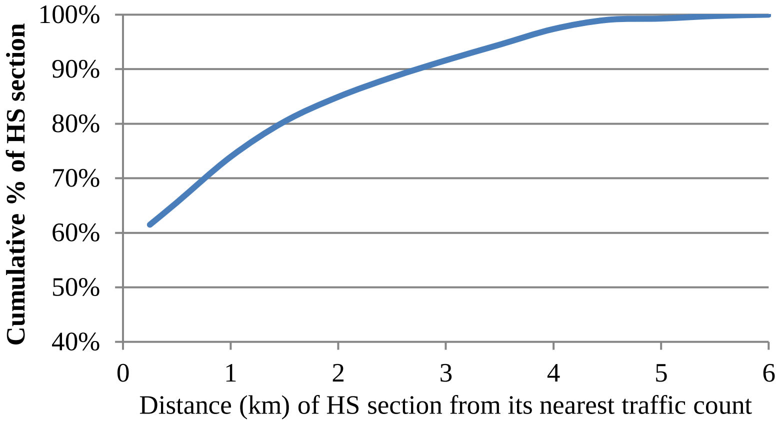
☑ Injury

☑ Property damage only

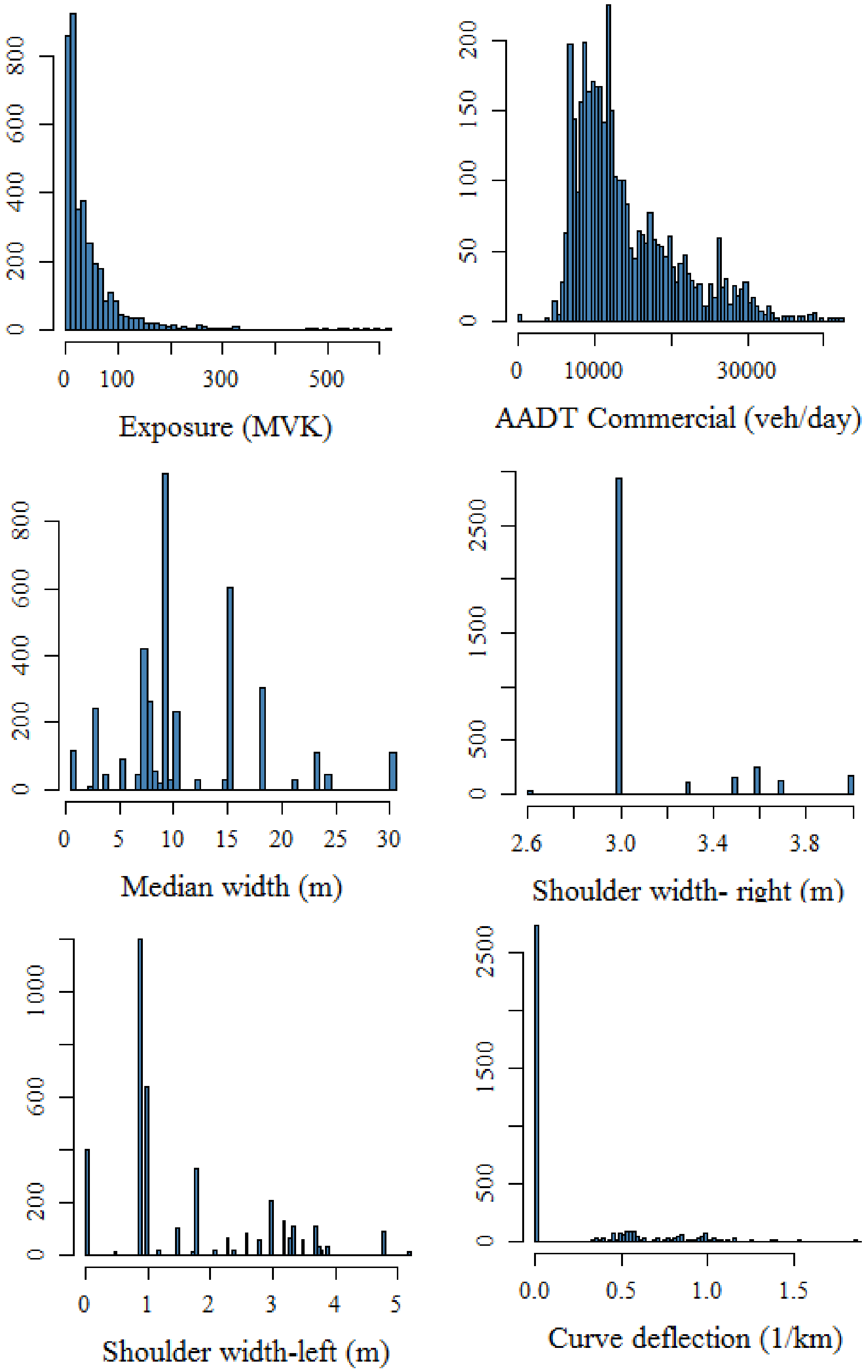
B.4: Distribution of annual crash counts



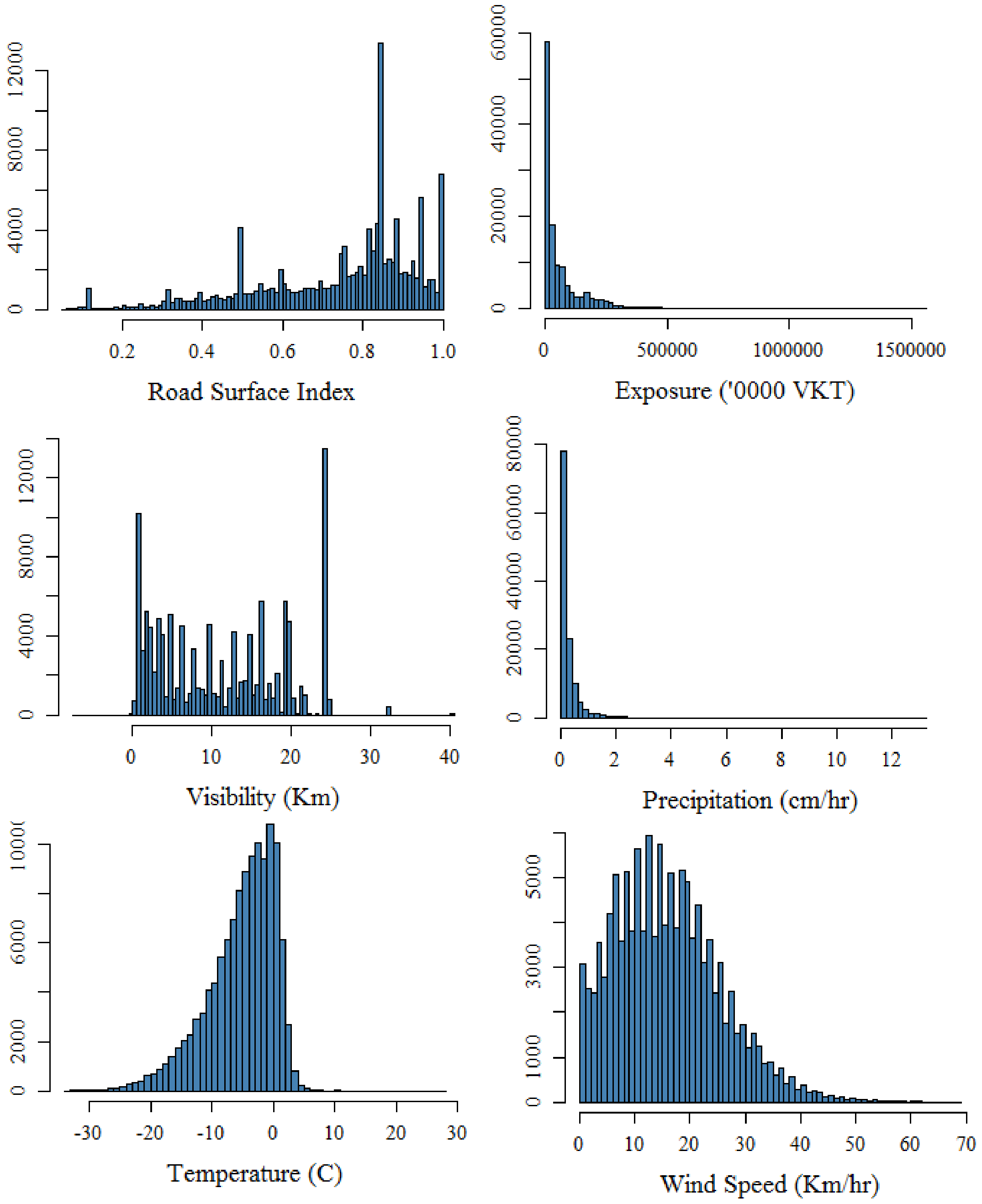
B.5: Distribution of distance between HS sections and the nearest traffic count locations



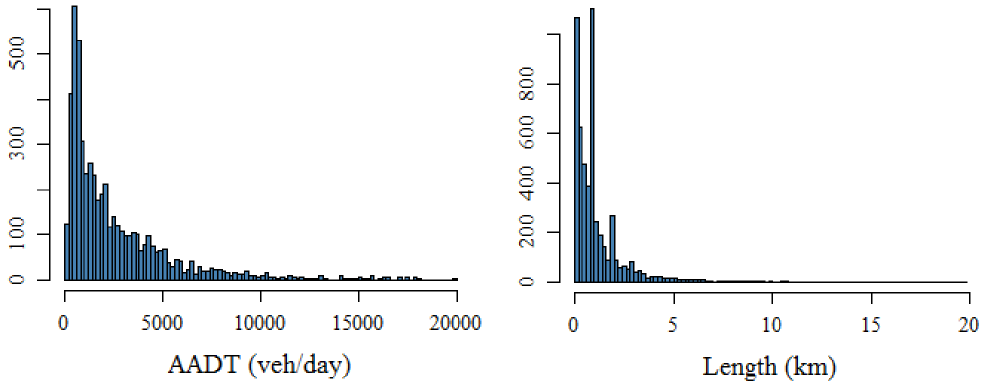
B. 6: Histogram of factors included in case 1 dataset: Highway 401



B. 7: Histogram of factors included in Case 2 dataset: 31 patrol routes, Ontario



B. 8: Histogram of factors included in Case 3 dataset: Two-lane rural roads, Colorado



B.9: Summary of crash models

Variables	NB				KR
	Coefficient estimate	Std. error	t-value*	p-value	Bandwidth
(a) Case study 1: Highway 401, Ontario (2000-2008)					
Intercept	-1.04	0.16	-6.52	<0.001	
ln(Exposure) (MVK)	0.82	0.02	49.91	<0.001	21.08
AADT (Commercial) (veh/day)	0.0001	0.00	20.57	<0.001	2621
Median Width (m)	-0.02	0.003	-6.16	<0.001	2.397
Shoulder width- left (m)	-0.09	0.01	-8.44	<0.001	0.465
Shoulder width - right(m)	0.16	0.05	3.32	<0.001	0.111
Curve deflection (1/km)	-0.17	0.04	-3.78	<0.001	0.135
theta	1.94	0.06			
AIC	25461				
MAE	*11.86*				*7.34*
RMSE	*26.64*				*14.81*
(b) Case study 2: 31 patrol routes, Ontario (2000-2006)					
(Intercept)	-2.58	0.08	-30.83	<0.00	
log(exposure) in '0000 veh km	0.72	0.02	43.58	<0.00	2.227
RSI	-2.83	0.09	-33.06	<0.00	0.054

Variables	NB				KR
	Coefficient estimate	Std. error	t-value*	p-value	Bandwidth
Precipitation (cm/hr)	0.01	0.06	0.27	0.789	1.532
Visibility (km)	-0.04	0.00	-12.03	<0.00	2.649
Wind speed (km/hr)	0.01	0.00	4.00	<0.00	2.170
Temperature ©	-0.0001	0.00	-0.02	0.983	1.530
theta	0.27	0.02			
AIC	24372				
MAE	*0.046*				*0.031*
RMSE	*0.137*				*0.178*
(c) Case study 3: Two-lane rural roads, Colorado State (1991-1998)					
Intercept	-8.03	0.07	-121	<0.001	
ln(AADT) (veh/day)	0.95	0.01	115.6	<0.001	423.48
ln(Length) (km)	1.07	0.01	119.7	<0.001	0.4
theta	2.1597	0.0631			
AIC	74377				
MAE	*0.781*				*0.752*
RMSE	*1.529*				*1.333*

Note: log function is not applicable for the KR bandwidths; MVK is million-vehicle-kilometer travelled, theta is $1/\alpha$, t-value* is equivalent to z-value when sample size is large.

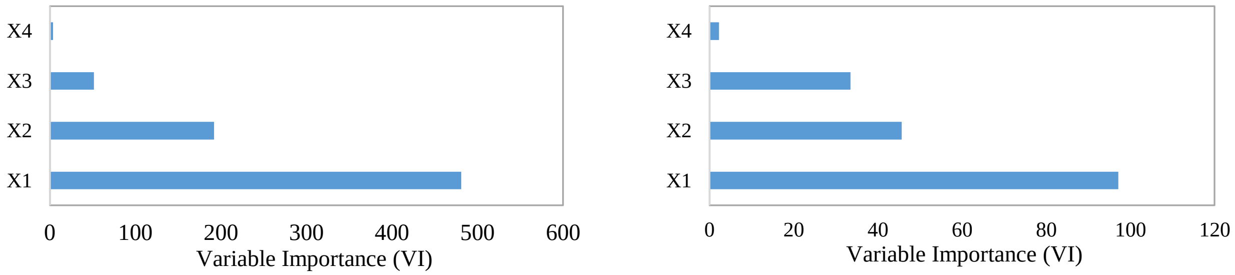
B.10: Dataset summary- nonlinear model

	y	x_1	x_2	x_3	x_4	*error*
Min.	1.00	4.94	2.88	3.73	2.94	1.98
Mean	4.77	20.22	20.11	19.88	20.06	0.01
Max.	13.87	35.86	37.08	35.86	36.67	1.62

B.11: Dataset summary- nonlinear model

	y	x_1	x_2	x_3	x_4	*error*
Min.	1.00	4.94	2.88	3.73	2.94	1.98
Mean	4.77	20.22	20.11	19.88	20.06	0.01
Max.	13.87	35.86	37.08	35.86	36.67	1.62

B.12: Variable Importance (VIs) of simulated datasets



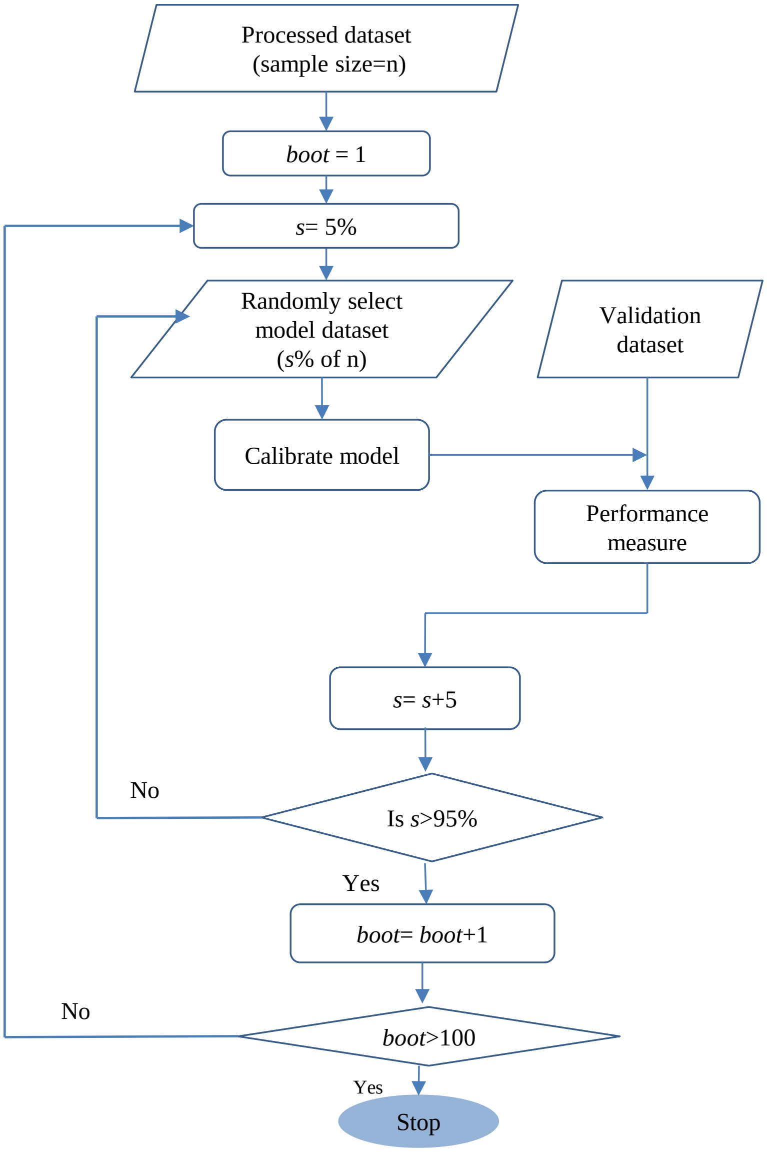
(a)

Linear model

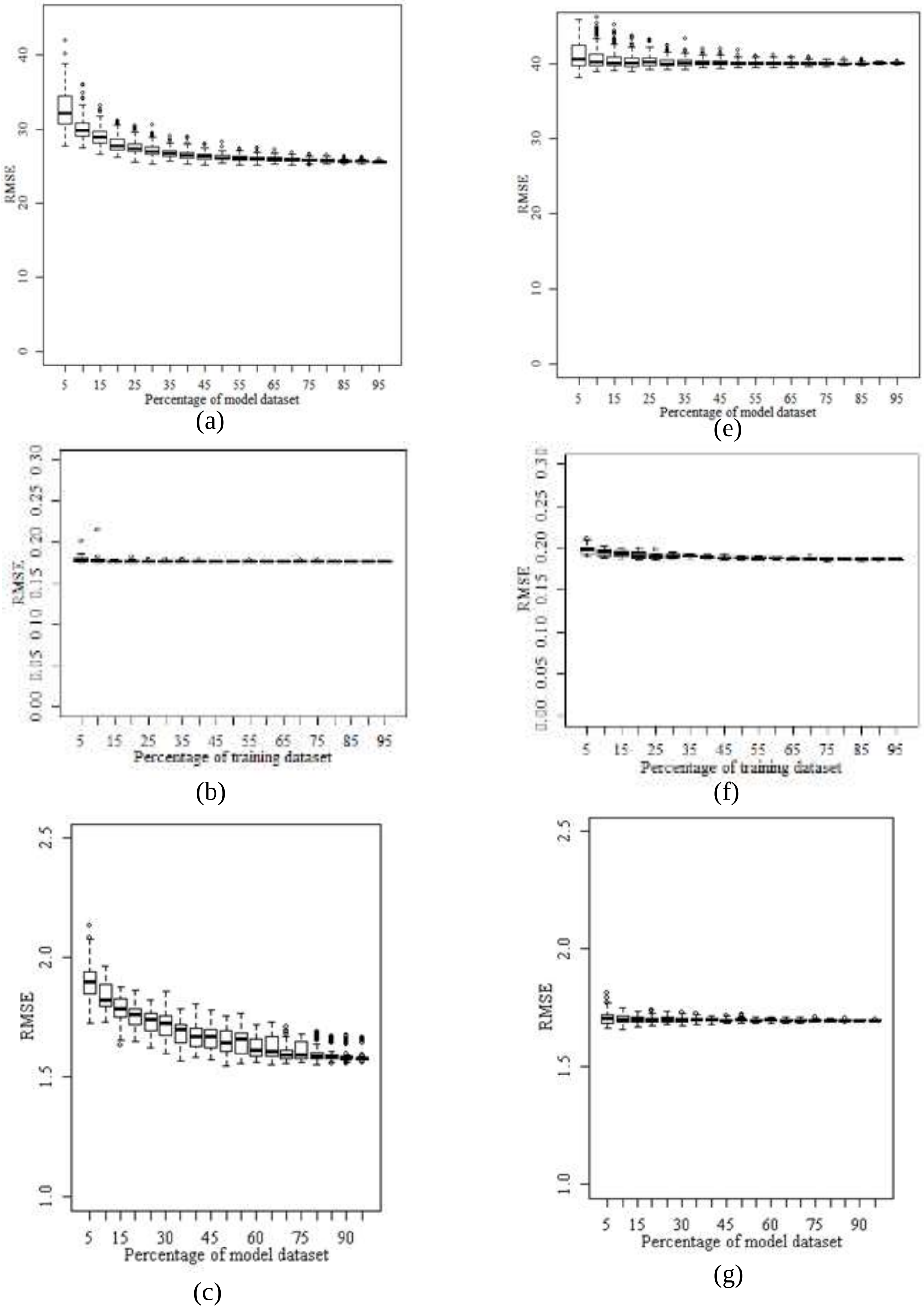
(c)

Non-linear model

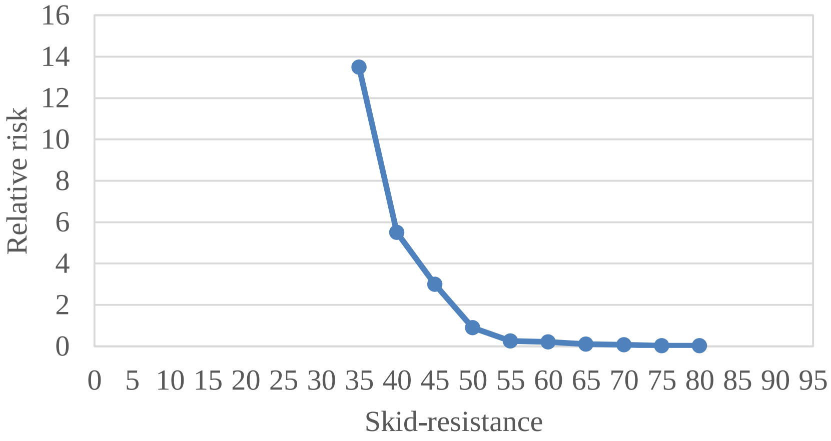
B.13: Framework for bootstrap-based validation



B.14 Boxplots- (a), (b) and (c) represent RMSE of KR method for case study 1, 2 and 3, respectively; (d), (e) and (f) represent RMSE of NB model for case study 1, 2 and 3, respectively.



B.15 Relation of relative crash risk and skid-resistance (i.e., friction) (Wallman & Astrom, 2001)



Appendix C

C. 1: Summary results of model

Variables	NB model				KR
	Coefficients estimate	Std. error	t-value	p-value	Bandwidth
Intercept	-1.16	0.15	-7.528	<0.001	
ln(exposure) (MVK)	0.84	0.02	50.75	<0.001	21.08
AADT (Commercial) (veh/day)	5E-05	0.00	18.69	<0.001	2621
Median width (m)	-0.01	0.00	-5.23	<0.001	2.397
Shoulder width- left (m)	-0.10	0.01	-9.14	<0.001	0.465
Shoulder width- right (m)	0.16	0.04	3.51	<0.001	0.111
Curve deflection (1/km)	-0.09	0.05	-1.69	0.052	0.135
Dispersion parameter (α)					
Intercept	-0.51	0.03	-15.10	<0.001	
Length (km)	-0.83	0.04	-20.11	<0.001	
AIC	19303				

Appendix D

Before-after study

D.1: Summary of datasets based on control types (2009-2013)

	Train Volume (train/day)	Traffic Volume (veh/day)	Total Exposure	Tracks	Lanes	Road Speed (km/hr)	Train Max Speed (km/hr)	Crashes
Passive crossings								
Mean	6.14	259.53	698.03	1.12	1.81	65.51	54.64	0.04
Std.dev	8.27	1130.37	4551.51	0.41	0.44	21.66	27.82	0.22
Min	0.01	1	0.01	1	1	5	1.609	0
Max	55	24990	313200	9	6	100	160.9	4
Sample Size: 8018								
FLB								
Mean	5.89	2634.50	9830.47	1.12	2.14	62.74	51.67	0.07
Std.dev	6.68	5139.69	22538.01	0.43	0.62	18.37	26.73	0.31
Min	0.01	5	0.25	1	1	5	8.045	0
Max	46	71500	432900	6	6	110	128.72	4
Sample Size: 4038								
FLBG								
Mean	20.45	4088.81	67053.90	1.58	2.22	60.43	87.93	0.13
Std.dev	15.90	6444.92	149445.21	0.76	0.75	16.64	39.54	0.39
Min	0.01	5	25	1	1	5	8.045	0
Max	162	51000	3000000	7	7	100	160.9	4
Sample Size: 2324								

D.2: Summary results of full crash models

Variables	NB model				KR method
	Coefficient estimate	Std. error	t-value	p-value	Bandwidth
Passive crossing					
(Intercept)	-7.24	0.37	-19.71	<0.001	2.43
log(Train volume)	0.52	0.07	7.04	<0.001	332.04
log(Vehicle volume)	0.52	0.04	13.39	<0.001	8.17
Train speed	0.01	0.00	2.49	0.01	6.36
Road speed	0.01	0.00	3.02	0.00	
theta	0.54	0.147			
AIC	2431.7				
MAE	0.074				0.045
Sample size	8018				
FLB crossing					
(Intercept)	-7.84	0.47	-16.56	<0.001	
log(Train volume)	0.60	0.08	7.95	<0.001	2.14
log(Vehicle volume)	0.49	0.05	9.01	<0.001	1644.92
Train speed	0.00	0.00	1.64	0.100	8.56
Lanes	0.23	0.08	2.91	0.004	0.20
theta	0.85	0.26			
AIC	1975				
MAE	0.129				0.073
Sample size	4038				
FLBG crossing					
(Intercept)	-5.78	0.45	-12.73	<0.001	
log(Train volume)	0.63	0.09	7.02	<0.001	4.85
log(Vehicle volume)	0.30	0.04	7.02	<0.001	1966.11
Train speed	0.00	0.00	-1.67	0.0948	12.06
theta	1.22	0.44			

Variables	NB model				KR method
	Coefficient estimate	Std. error	t-value	p-value	Bandwidth
AIC	1763				
MAE	0.217				0.13
Sample size	2324				

Note: for KR bandwidth, ignore the log function.

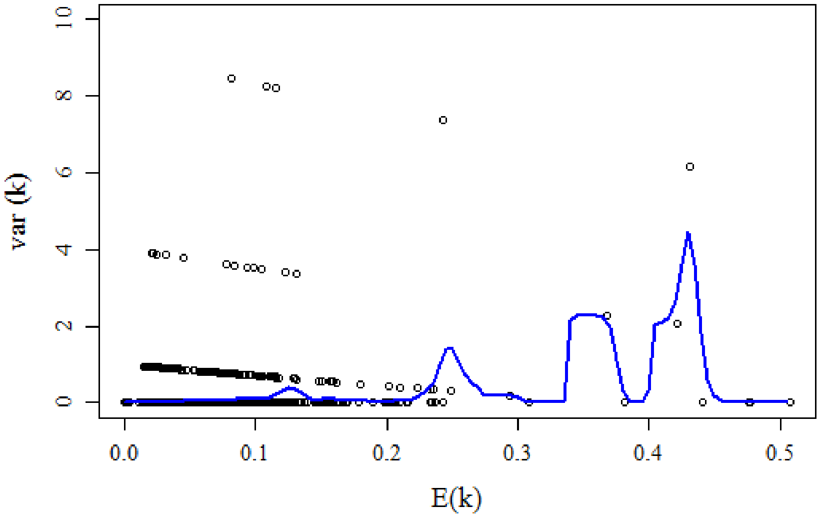
D.3: Summary results of Traffic-only crash models

Variables	NB model				KR method
	Coefficient estimate	Std. error	t-value	p-value	Bandwidth
Passive crossing					
(Intercept)	-6.03	0.22	-27.45	<0.001	
log(Train volume)	0.68	0.05	12.48	<0.001	1.74
log(Vehicle volume)	0.45	0.03	13.46	<0.001	243.44
theta	0.48	0.123			
AIC	2448				
MAE	0.075				0.048
Sample size	8018				
FLB crossing					
(Intercept)	-7.34	0.40	-18.32	<0.001	
log(Train volume)	0.64	0.07	9.53	<0.001	1.61
log(Vehicle volume)	0.52	0.05	11.60	<0.000	1240.96
theta	0.81	0.24			
AIC	1982				
MAE	0.13				0.082
Sample size	4038				
FLBG crossing					
(Intercept)	-6.04	0.43	-14.10	<0.000	

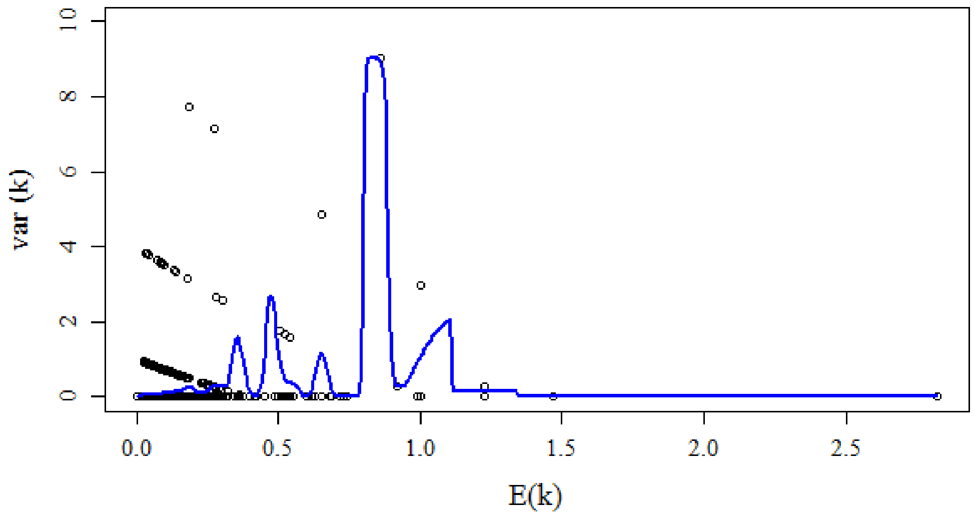
Variables	NB model				KR method
	Coefficient estimate	Std. error	t-value	p-value	Bandwidth
log(Train volume)	0.56	0.08	7.08	<0.000	4.21
log(Vehicle volume)	0.32	0.04	7.97	<0.000	1706.19
theta	1.206	0.436			
AIC	1763				
MAE	0.217				0.139
Sample size	2324				

Note: for KR bandwidth, ignore the log function.

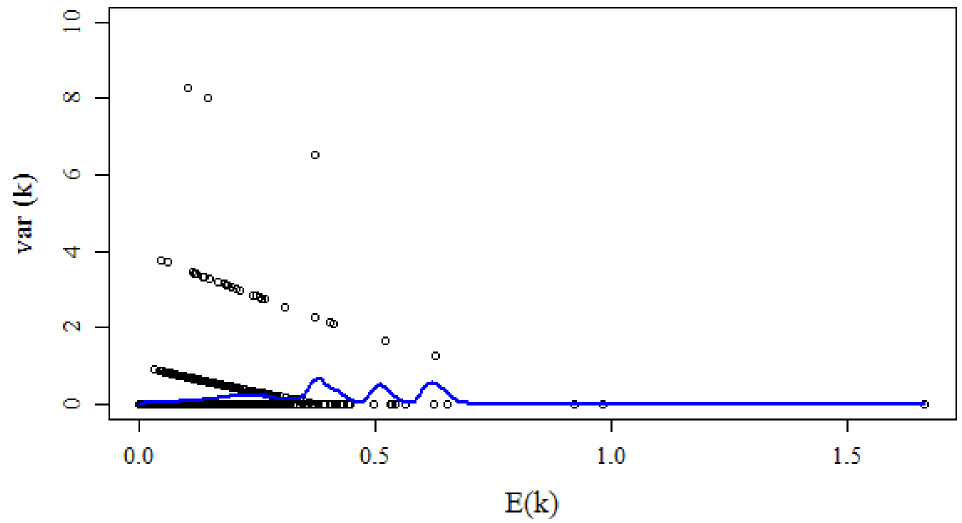
D. 4: mean- variance relation for all crossing types (use in EB estimate)



Mean-variance relation- with passive control (bandwidth= 0.006)



Mean-variance relation- FLB (bandwidth= 0.018)



Mean-variance relation- FLBG (bandwidth= 0.019)

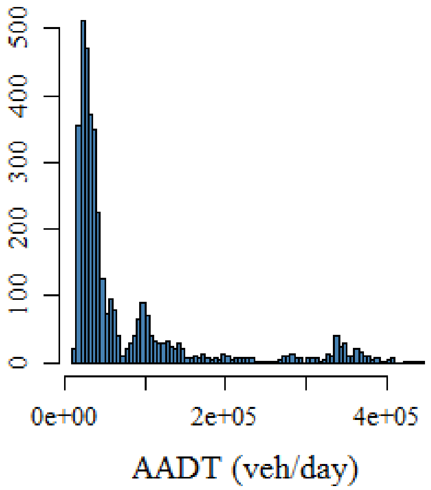
D.5: Crash modification (CMF) and crash reduction factors (CRF) based on different studies

Study reference	Passive to FLB	Passive to FLBG
Park (2007)	-	0.35 (65)
Saccomanno and Lai (2005)	0.42(58)	0.37 (63)
U.S. DOT (1980)*	0.30(70)	0.17 (83)
California (1974)*	0.36(64)	0.12 (88)
Hedley (1952)*	0.37(63)	0.04 (96)

*FHWA (2015); values in parenthesis indicates percentage of crash reduction

Cross-sectional study

D.6: Histogram of traffic volume (AADT) in Highway 401 dataset, for other factors refer to B.6.



D.7: Bandwidths for KR method

Variable	Bandwidth
AADT (veh/day)	35700
AADT (Commercial) (veh/day)	2620
Median width (m)	2.4
Shoulder width- left (m)	0.466
Shoulder width- right (m)	0.11
Curve deflection (per km)	0.126

D.8: CMFs from KR method

Variable	Low Traffic (AADT- 12000; Commercial AADT- 5250		Medium Traffic (AADT- 37000; Commercial AADT- 15750		High Traffic (AADT- 67000; Commercial AADT- 22750		Average CMF
	CMF	Std. error	CMF	Std. error	CMF	Std. error	
Median width							
5	1.00	0.087	1.00	0.13	1.00	0.334	1.00

Variable	Low Traffic (AADT- 12000; Commercial AADT- 5250		Medium Traffic (AADT- 37000; Commercial AADT- 15750		High Traffic (AADT- 67000; Commercial AADT- 22750		Average CMF
	CMF	Std. error	CMF	Std. error	CMF	Std. error	
10	0.97	0.067	0.88	0.09	1.22	0.408	1.02
15	0.67	0.045	0.68	0.06	0.66	0.586*	0.67
20	0.75	0.050	0.59	0.07	0.45	0.385	0.60
25	0.96	0.075	0.49	0.05	0.27	0.358	0.57
30	1.56	0.130	0.69	0.18	0.31	3.14*	0.85
Shoulder width- left							
1	1.00	0.080	1.00	0.12	1.00	0.334	1.00
2	1.27	0.189	1.17	0.12	0.73	0.189	1.06
3	1.81	0.261	1.49	0.14	0.86	0.209	1.39
4	1.60	0.550	1.66	0.2	1.17	0.329	1.48
5	1.92	0.860	0.98	0.49	1.55	2.244*	1.48
Shoulder width- right							
3	1	0.087	1	0.12	1	0.33	1.00
3.5	1.98	0.880*	0.83	0.14	2.94	0.96*	1.92
4	1.21	0.139	0.87	0.39	5.90	29.44*	2.66
Curve deflection							
0	1.00	0.087	1.00	0.13	1	0.33	1.00
0.4	0.71	0.100	0.60	0.15	0.64	0.88*	0.65
0.5	0.69	0.080	0.67	0.25	1.13	1.71*	0.83
0.6	0.71	0.080	1.15	0.51*	2.7	2.34*	1.52
0.7	0.81	0.115	2.44	0.95*	3.18	1.79*	2.14
0.8	1.14	0.214	3.88	1.13*	2.69	1.28*	2.57
0.9	1.85	0.370	4.14	1.5*	2.09	1*	2.69
1	2.03	0.448	3.49	1.67*	1.61	0.88*	2.38
1.25	0.51	0.220	1.87	3.16*	1.00	1.36*	1.13

Variable	Low Traffic (AADT- 12000; Commercial AADT- 5250		Medium Traffic (AADT- 37000; Commercial AADT- 15750		High Traffic (AADT- 67000; Commercial AADT- 22750		Average CMF
	CMF	Std. error	CMF	Std. error	CMF	Std. error	
1.5	0.70	0.80*	0.62	10.51*	0.88	14.07*	0.73

Note: * represents CMF with standard error >0.5

D.9: CMFs from NB model

Variable	CMF	Std. error	
		Park & Abdel-Aty 2015	Bahar (2010)
Median width			
5	1.00	0.003	0.005
10	0.90		
15	0.82		
20	0.74		
25	0.67		
30	0.61		
Shoulder width- left			
1	1.00	0.01	0.022
2	0.91		
3	0.84		
4	0.76		
5	0.70		
Shoulder width- right			
3	1	0.055	0.095
3.5	1.08		
4	1.17		
Curve deflection			

Variable	CMF	Std. error	
		Park & Abdel-Aty 2015	Bahar (2010)
0	1.00	0.037	0.088
0.4	0.93		
0.5	0.92		
0.6	0.90		
0.7	0.89		
0.8	0.87		
0.9	0.86		
1	0.84		
1.25	0.81		
1.5	0.77		

Appendix E

<u>Codes in R</u>

E.1. Bootstrap-based variable selection approach

1. *Generate a simulated dataset*

```
# providing the location of a folder to save outputs.
setwd("C:\\Users\\Lalita\\Dropbox\\4. WorkingFolder\\6. Simulation\\VariableSelection")
library("np")
n= 1000
X1<- rnorm(n, mean= 20, sd= 5)
X2<- rnorm(n, mean= 20, sd= 5)
X3<- rnorm(n, mean= 20, sd= 5)
X4<- rnorm(n, mean= 20, sd= 5)
error<- rnorm(n, mean= 0, sd= 0.5)
# assume a nonlinear model form
Y<- exp(0.05* X1+0.03*X2-0.025*X3+0.00*X4)+error
dataB<- data.frame(Y1, X1, X2, X3, X4)
summary(dataB)
```

2. *Estimate performance indicators following the variable selection algorithm described in Chapter 3*

```
# Select m independent variables randomly
m<- 4 # vary this value depending on the numbers of potential variables in a given dataset
#randomize columns of predicting variables
dataC<-dataB[,sample(2:ncol(dataB), m, replace=FALSE)].
dataD <- cbind(dataB[, 1],dataC) # combine fields
# rename field "Y" as "accident" to be consistent with other crash related datasets
names(dataD)[1]<- "accident"
SAD.B <- as.data.frame(matrix(0, ncol = m+1, nrow = ))
# create an empty dataframe for storing the results; +1 is for test statistics
SSD.B <- as.data.frame(matrix(0, ncol = m+1, nrow = ))
nn<- length(dataD$accident)
for (boot in 1:100) {
        train_id<- sample(1:nrow(dataD), round(0.8*nn ,0),replace=FALSE)
```

```
      dataTrainBoot<- dataD[train_id, ] # 80% of training sets
      dataTestBoot<- dataD[-train_id, ]
      X<- dataTrainBoot[,-1]
      Y<- dataTrainBoot[, 1]
     dataTestX<- dataTestBoot[, -1] # select only the predicting variables to permute
     dataTestY<- dataTestBoot[, 1]
      # bandwidths
      di= m
      n<- length(dataTrainBoot$accident)
      c2= (4/((2*di+1)*n))^(1/(di+4))
      bww<- c()
         for (i in 1:m){
         bww[i]<- sd(X[, i])}
      bw<- bww*c2
      Model<- npreg(txdat= X, tydat= Y, bws= bw, bandwidth.compute= FALSE)
# Estimate percentage of error using test dataset
PredictTest<- predict(Model, exdat= dataTestX)
              function.SAD<- function(predict, actual) { # sum of absolute deviation
              SAD<- sum(abs(predict-actual))
              return(SAD) }
SAD.Test<- function.SAD(PredictTest, dataTestY)
               function.SSD<- function(predict, actual) {# sum of square deviation
               SSD<- sum((predict-actual)^2)
              return(SSD) }
SSD.Test<- function.SSD(PredictTest, dataTestY)
SAD.Perm<- c() # created to sort the results after permuting each variable in the steps
described below
SSD.Perm<- c()
# Permute each variable, one at a time.
for (VarPer in 1:m) {
        Z1<- as.vector(dataTestX[[VarPer]]) #create a vector
        Z2<- data.frame(sample(Z1, replace = FALSE)) #permute a selected variable
```

```
            names(Z2)[1]<-names(dataTestX)[VarPer] #making variable name consistent
            # Arrange the dataset such that the position of Z2 goes to its original position
              if (VarPer==1){
              Z3<- data.frame(c(Z2, dataTestX[(VarPer+1):m]))} # VarPer=1
              else if (VarPer< m){
              Z3<- data.frame(c(dataTestX[1:(VarPer-1)],Z2, dataTestX[(VarPer+1):m]))}
              else {
              Z3<- data.frame(c(dataTestX[1:(VarPer-1)],Z2))} # VapPer= m
            PredictPermu<- predict(Model, xdat= X, ydat= Y, exdat= Z3)
            SAD.Perm[VarPer]<- function.SAD(PredictPermu, dataTestY)
            SSD.Perm[VarPer]<- function.SSD(PredictPermu, dataTestY)
         }
     SAD.B[boot, ]   <- c(SAD.Perm, SAD.Test)
    # combine outputs—permutation and test statistics— for each bootstrap sample
     for (name in 1:m){
     names(SAD.B)[name]<-names(Z3)[name] }
     SSD.B[boot, ]   <- c(SSD.Perm, SSD.Test)
     for (name in 1:m){
     names(SSD.B)[name]<-names(Z3)[name]}
     print(paste("bootstrap", boot))
   }
 # Extract output files
write.csv(SAD.B, file= "TestResultSAD.csv")
write.csv(SSD.B, file= "TestResultSSD.csv")
```

3. *Calculate variable importance (VI)*

   # Use excel spreadsheet to calculate VIs using output files from step 2. Note that results are presented based on SAD measure. Similar trends were observed using the SSD.

E.2. Bootstrap-based validation algorithm

```
setwd("C:\\Users\\Lalita\\Dropbox\\4.WorkingFolder\\2.RegressionYearly\\1.Regression\\Bootstraping-MAE")
datanb= read.csv("401C_Y_Model_00-06.csv", header= TRUE)
testing<- read.csv("401C_Y_07-08.csv", header= TRUE)
MAE.np.v<- c()
RMSE.np.v<- c()
MAE.p.v<- c()
RMSE.p.v<- c()
# create four dataframes for storing the values of goodness-of-fit measures (MAE and RMSE) for the KR (np) and NB (p) models.
# MAE dataframe for KR method
Data.MAE.np.v <- data.frame(Split5= numeric(0), Split10= numeric(0), Split15= numeric(0), Split20= numeric(0), Split25= numeric(0), Split30= numeric(0), Split35= numeric(0), Split40= numeric(0), Split45= numeric(0), Split50= numeric(0), Split55= numeric(0),Split60= numeric(0), Split65= numeric(0), Split70= numeric(0), Split75= numeric(0), Split80= numeric(0), Split85= numeric(0), Split90= numeric(0), Split95= numeric(0))

# RMSE dataframe for KR method
Data.RMSE.np.v <- data.frame(Split5= numeric(0), Split10= numeric(0), Split15= numeric(0), Split20= numeric(0), Split25= numeric(0), Split30= numeric(0), Split35= numeric(0), Split40= numeric(0), Split45= numeric(0), Split50= numeric(0), Split55= numeric(0),Split60= numeric(0), Split65= numeric(0), Split70= numeric(0), Split75= numeric(0), Split80= numeric(0), Split85= numeric(0), Split90= numeric(0), Split95= numeric(0))

# MAE dataframe for NB model
Data.MAE.p.v <-data.frame(Split5= numeric(0), Split10= numeric(0), Split15= numeric(0), Split20= numeric(0), Split25= numeric(0), Split30= numeric(0), Split35= numeric(0), Split40= numeric(0), Split45= numeric(0), Split50= numeric(0), Split55= numeric(0),Split60= numeric(0), Split65= numeric(0), Split70= numeric(0), Split75= numeric(0), Split80= numeric(0), Split85= numeric(0), Split90= numeric(0), Split95= numeric(0))
```

```
# RMSE dataframe for NB model
Data.RMSE.p.v <- data.frame(Split5= numeric(0), Split10= numeric(0), Split15= numeric(0), Split20=
numeric(0), Split25= numeric(0), Split30= numeric(0), Split35= numeric(0), Split40= numeric(0),
Split45= numeric(0), Split50= numeric(0), Split55= numeric(0),Split60= numeric(0), Split65=
numeric(0), Split70= numeric(0), Split75= numeric(0), Split80= numeric(0), Split85= numeric(0),
Split90= numeric(0), Split95= numeric(0))

for (boot in 1: 100){
    s<- 0.05 # initialize percentage split as 5%
    for (i in 1:19) {    # a total of 19 splits between 5% to 95%
        training <- datanb[sample(1:nrow(datanb),size= trunc(2927*s), replace= FALSE),]
        # total sample size of model dataset (2000-2006) is 2927
        #Read data
         attach (training)
        n= length (A_count) # of training set
             b<- Exposure
             ci<- AADT_Comm
             d<- MEDIAN_WID
              e<- MED_SHLDWI
              f<- SHLD_WIDTH
               g<- Deflection
         # Bandwidths
        di= 6
        c2= (4/((2*di+1)*n))^(1/(di+4))
        bw<- c(c2*sd(b), c2*sd(ci), c2*sd(d), c2*sd(e), c2*sd(f),c2*sd(g))
        detach (training)
        # KR method
        library(np)
        model.np<-
        npreg(A_count~Exposure+AADT_Comm+MEDIAN_WID+MED_SHLDWI+SHLD_
        WIDTH+Deflection,bws= bw, data= training, bandwidth.compute= FALSE)
        MAE.np.m[i]<- model.np$MAE
    RMSE.np.m[i]<- sqrt(model.np$MSE)
     # Validation-  KR method
    predict.np<- predict(model.np, data= training, newdata=testing)
    error.np<- testing$A_count-predict.np
    nt<- length(testing$A_count)
    MAE.np.v[i]<- sum(abs(error.np))/nt
    RMSE.np.v[i]<- sqrt(sum(error.np^2)/nt)
    # NB model
     library(MASS)
     model.p<-
    glm.nb(A_count~log(Exposure)+AADT_Comm+MEDIAN_WID+MED_SHLDWI+SH
    LD_WIDTH+Deflection, link=log, data=training)
    summary(model.p)
    n<- length(training$A_count)
    model.fit<- fitted(model.p)
    #Goodness-of-fit- model set
    E1<- (model.fit-training$A_count)
    MAE.p.m[i]<- sum(abs(E1))/n
    RMSE.p.m[i]<-sqrt(sum(E1^2)/n)

     # validation- NB model
    predict.p<- predict(model.p, type="response", newdata=testing)
    error.p<- testing$A_count-predict.p
    nt<- length(testing$A_count)
    MAE.p.v[i]<- sum(abs(error.p))/nt
    RMSE.p.v[i]<- sqrt(sum(error.p^2)/nt)
    s<- s+0.05 # increase the split (s) by 5%
    }
 # Appending calculated values from each bootstrap step to corresponding dataframes
Data.MAE.np.v[boot, ]   <- MAE.np.v
Data.RMSE.np.v[boot, ]  <- RMSE.np.v
Data.MAE.p.v[boot, ]  <- MAE.p.v
```

```
        Data.RMSE.p.v[boot, ] <- RMSE.p.v
        print(paste("bootstrap", boot))
}
write.csv(Data.MAE.np.v, file= "Data.MAE.np.v.csv")
write.csv(Data.RMSE.np.v, file= "Data.RMSE.np.v.csv")
write.csv(Data.MAE.p.v, file= "Data.MAE.p.v.csv")
write.csv(Data.RMSE.p.v, file= "Data.RMSE.p.v.csv")
```

E.3. Network screening: regression-based and EB-based methods

```
library(np)
library(gamlss)
setwd("C:\\Users\\Lalita\\Dropbox\\4.WorkingFolder\\2.RegressionYearly\\2.NetworkScreening\\NS-
4Methods")
dataAll= read.csv("1. 401C_Y_M_All.csv", header= TRUE)
# Splitting the  dataset
        dataT<- subset(dataAll, Year<2007) # model set
        data_07= subset(dataAll, Year==2007)
        data_08= subset(dataAll, Year==2008)
# KR estimates
        bw<- c(21.08, 2621.70, 2.397, 0.465, 0.111, 0.135) # same as in E.2.
        model.KR<-
        npreg(A_count~Exposure+AADT_Comm+MEDIAN_WID+MED_SHLDWI+SHLD_WIDTH
        +Deflection,bws= bw, data= dataT,bandwidth.compute= FALSE)
        summary(model.KR)
        MAE.KR<- model.KR$MAE
        RMSE.KR<- sqrt(model.KR$MSE)
        # data from year 2000-2006 is used for modeling
        # Predict using KR method
        predict.KR_07<- predict(model.KR, data= dataT, newdata=data_07)
        predict.KR_08<-predict(model.KR, data= dataT, newdata=data_08)
        predict.KR_T<- model.KR$mean
        predict.KR<- (predict.KR_07+predict.KR_08) # add for two years (output)
```

```
# EB estimate based on KR method
     #Step 1: Get estimates of variance for the training set; Reference Hauer (2015)
    variance<-function(observed, estimated){
            var<- (observed-estimated)^2
            return(var)
            }
    var.k.T<- variance(dataT$A_count, predict.KR_T)
    var.KR.T<- var.k.T-predict.KR_T # page 25 Hauer(2015)
    plot(predict.KR_T, var.KR.T)
    # var(mu) vs fitted values, some of the values of var.mu are negative which is replaced by zero (Hauer, 2015)
    var.KR.T<- replace(var.KR.T, var.KR.T<0, 0) #  to replace negative values by zeros
    #Step 2: Establish a relationship between mean and variance and estimate the variance
    #Use KR method to estimate var.mu
    n<- length(var.KR.T)
    di= 1
    c2= (4/((2*di+1)*n))^(1/(di+4))
    bwv<- c(c2*sd(predict.KR_T))
    # var.mu vs E.mu; note: there is change in name of variables to make same as in prediction dataset
    dataV<- data.frame(var.KR.T, predict.KR_T) # dataframe created though the values could be taken from the environment; use same name to match later in calculating var.mu.B
    model.var.KR<- npreg(var.KR.T~ predict.KR_T, bws= bwv, data= dataV, bandwidth.compute= FALSE) # fixed bandwidths, use updated dataV
    summary(model.var.KR)
    MAE.var.KR<- model.var.KR$MSE
    var.KR.T<- model.var.KR$mean # no need
    dataV_07<- data.frame(predict.KR_07) # estimated value for year 07
    names(dataV_07)[1]<-"predict.KR_T" # to make variable name same as in training dataset "dataV"
    var.KR_07<-  predict(model.var.KR, data=dataV, newdata=dataV_07)
    dataV_08<- data.frame(predict.KR_08)
```

```
     names(dataV_08)[1]<-"predict.KR_T" # to make variable name same as in training dataset "dataV"
      var.KR_08<-  predict(model.var.KR, data=dataV, newdata=dataV_08)
     #Step 3: Compute weights
         weight<-function(var, mu){
         w<- 1/(1+var/mu)
         return(w)
         }
        w_07<- weight(var.KR_07, predict.KR_07)
        w_08<- weight(var.KR_08, predict.KR_08)
     #Step 4: Use EB approach
        EB.KR.estimate<- function (observed, mu, w){
         EB<- w*mu+(1-w)*observed
         return(EB)
        }
        predict.EB.KR_07<- EB.KR.estimate(data_07$A_count, predict.KR_07, w_07)
        predict.EB.KR_08<- EB.KR.estimate(data_08$A_count, predict.KR_08, w_08)
        predict.EB.KR<- (predict.EB.KR_07+predict.EB.KR_08) # add for two years
# NB model estimates
    Model.NB<-
    gamlss(A_count~log(Exposure)+AADT_Comm+MEDIAN_WID+MED_SHLDWI+SHLD_WIDTH+Deflection,sigma.fo= ~log(Length), family=NBI, data=dataT)
     summary(Model.NB)
     model.fit<- predict(object= Model.NB, what= "mu",newdata= dataT, type= "response" )
    # predict for NB
    predict.NB_07<- predict(object= Model.NB, what= "mu",newdata=data_07, type="response")
    predict.NB_08<- predict(object= Model.NB, what= "mu",newdata=data_08, type="response")
    predict.NB<- (predict.NB_07+predict.NB_08) # add for two years

# EB estimates based on NB model
    EB.KR.estimate<- function(fitted, alpha, crash) {
       w<- 1/(1+fitted*alpha) # alpha is 1/ theta
     estimate.EB<-w*fitted+(1-w)*crash
     return(estimate.EB)
    }
fitted.sigma_07<- predict(object= Model.NB, what= "sigma",newdata= data_07, type=
"response" ) # sigma (or alpha) reciprocal of theta in MASS package
fitted.sigma_08<- predict(object= Model.NB, what= "sigma",newdata= data_08, type=
"response" )
predict.EB.NB_07<- EB.KR.estimate(predict.NB_07, fitted.sigma_07, data_07$A_count)
predict.EB.NB_08<- EB.KR.estimate(predict.NB_08, fitted.sigma_08, data_08$A_count)
predict.EB.NB<- (predict.EB.NB_07+predict.EB.NB_08) # add for two years (output)
observed<- data_07$A_count+data_08$A_count

#Combine all outputs in a single dataframe
   Output<-data.frame(data_08$HS_Section,data_08$Length,data_08$Exposure,predict.NB,
   predict.EB.NB, predict.KR, predict.EB.KR, observed, data_07$A_count, data_08$A_count,
   predict.KR_07,predict.KR_08,predict.EB.KR_07,predict.EB.KR_08,predict.NB_07,predict.N
   B_08, predict.EB.NB_07, predict.EB.NB_08 )
   write.csv(Output, file= "Outputs.csv")
   # Use Excel spreadsheet for ranking and calculating spearman's correlation coefficients.
```

E.4. Countermeasure study: regression-based approach using KR method

```
library(np)
setwd("C:\\Users\\Lalita\\Dropbox\\4. WorkingFolder\\2.RegressionYearly\\3. CMStudies")
dataA<- read.csv("1. CMF-data.csv", header= TRUE) # Case study of Highway 401, Ontario, Canada
X<- dataA[,-1] # includes year as well
Y<- dataA[, 1]
# KR method
    #Bandwidths
        di= 6
        n<- length(dataA$CrashPerKm)
        c2= (4/((2*di+1)*n))^(1/(di+4))
        bww<- c()
        for (i in 1:di){
          bww[i]<- sd(X[, i])
        }
        bw<- bww*c2
model.np<-
npreg(CrashPerKm~AADT+AADT_Comm+MEDIAN_WID+MED_SHLDWI+SHLD_WIDTH+Deflection,bws= bw, data= dataA,bandwidth.compute= FALSE) # fixed bandwidths
 summary(model.np)
# Create a new dataframe for determining the CMFs
      aadt<- c(12000, 17000, 22000, 27000, 32000, 37000, 42000, 47000, 52000, 57000, 62000, 67000)
       aadt_comm<- c(5250, 8750, 12250, 15750, 19250, 22750, 26250, 29750)
      median<- c(0, 5, 10, 15, 20, 25, 30)
      shld_left<- c(0, 1, 2, 3, 4, 5)
      shld_right<- c(3, 3.5, 4)
      deflection<- c(0, 0.4, 0.5,0.6, 0.7, 0.8, 0.9, 1, 1.25, 1.5)

 # Create a dataframe from all possible combinations of values
   data.a<- expand.grid(aadt, aadt_comm, median, shld_left, shld_right, deflection)
    colnames(data.a)<-c("AADT","AADT_Comm","MEDIAN_WID","MED_SHLDWI",
    SHLD_WIDTH", "Deflection")
```

```
Predict<- predict(model.np, newdata=data.a, se.fit=TRUE)
result<-data.frame(data.a$AADT,\data.a$AADT_Comm,data.a$MEDIAN_WID,
data.a$MED_SHLDW, data.a$SHLD_WIDTH, data.a$Deflection, Predict$fit, Predict$se.fit)
colnames(result)<-c("AADT","AADT_Comm","MEDIAN_WID","MED_SHLDWI",
"SHLD_WIDTH", "Deflection", "CrashPerKm", "error")
write.csv(result, file= "Predict_for_CMF.csv")
# Finally, import the output file in the “Tableau software” and determine CMFs interactively.
```

E.5. Countermeasure study: EB-based KR approach

```
# Case study of rail-highway grade crossing, Canada; converting passive controls to FLB
setwd("C:\\Users\\Lalita\\Dropbox\\4.WorkingFolder\\7.GradeCrossing\\Before-After\\Passive-FLB")
dataT= read.csv("3.4.Passive.csv", header= TRUE) #training set (reference group)
dataB= read.csv("Passive-FLB_B.csv", header=TRUE) # before case
dataA= read.csv("Passive-FLB_A.csv", header=TRUE) # after case
#Step 1: Obtain estimates of mean crashes for before and after cases
         #Bandwidths
        n<- length(dataT$accident)
        di= 2
        c2= (4/((2*di+1)*n))^(1/(di+4))
        bw<- c(c2*sd(dataT$trainflow), c2*sd(dataT$vehflow))
         library(np)
         model.np<- npreg(accident~ trainflow+vehflow,bws= bw, data= dataT,bandwidth.compute=
         FALSE)
         summary(model.np)
         MAE<- model.np$MSE
         E.mu.T<- model.np$mean
         E.mu.B<- predict(model.np,data=dataT, newdata= dataB) # for before period
         E.mu.A<- predict(model.np,data=dataT, newdata= dataA) # for after period

#Step 2: Obtain estimates of variance for training set (Hauer, 2015)
       variance<-function(observed, estimated){
       var<- (observed-estimated)^2
```

```
    return(var)
    }
    var.k.T<- variance(dataT$accident, E.mu.T)
    var.mu.T<- var.k.T-E.mu.T # page 25 Hauer (2015)
    plot(var.mu.T, var.k.T)
 plot(E.mu.T, var.mu.T)
var.mu.T<- replace(var.mu.T, var.mu.T<0, 0) #  to replace negative values by zeros

#Step 3: Establish a relation of mean and variance and estimate variance for before case
    #Use KR method
    n<- length(var.mu.T)
    di= 1
    c2= (4/((2*di+1)*n))^(1/(di+4))
    bwv<- c(c2*sd(E.mu.T))

  # var.mu vs E.mu; note: there is change in name of variables to make it consistent as in prediction
dataset
    dataV<- data.frame(var.mu.T, E.mu.T) # dataframe created eventhough the values could be
    taken from the environment; make same name to match later in calculating var.mu.B
    names(dataV)[2]<-"E.mu.B"
     model.var.mu<- npreg(var.mu.T~ E.mu.B, bws= bwv, data= dataV, bandwidth.compute=
    FALSE) # fixed bandwidths, use updated dataV
     summary(model.var.mu)
     MAE.var.mu<- model.var.mu$MSE
     var.mu.T<- model.var.mu$mean # no need (updated in march 14, because of this estimated
    value, we need to recalculate to find the mean-variance relation)
     dataV.new<- data.frame(E.mu.B) # estimated value for befor period in step 1
     var.mu.B<-  predict(model.var.mu, data=dataV, newdata=dataV.new)

#Step 4: Calculate weights
  weight<-function(var, mu){
   w<- 1/(1+var/mu)
```

```
      return(w)
     } # Hauer (1997) optimizing variance of weighted two random variables
     w.T<- weight(var.mu.T, E.mu.T) # for training dataset, not needed
     w.B<- weight(var.mu.B, E.mu.B) # no need of w for after

#Step 5: EB estimates
        EB.estimate<- function (observed, mu, w){
        EB<- w*mu*5/6+(1-w)*observed
        return(EB)
        } #5/6 to adjust for unit of model estimates (six-year) and observed value (five- year)
        EB.mu.T<- EB.estimate(dataT$accident, E.mu.T, w.T)
        plot(E.mu.T, EB.mu.T)

     # for before dataset
        EB.mu.B<- EB.estimate(dataB$accident, E.mu.B, w.B)
        # Ratio (R), is used for adjustment to change in traffic levels; similar to Persaud et al. (2009)
        R<- E.mu.A/E.mu.B
        EB.mu.BB<- R*EB.mu.B
        # Unit conversion is no need as the number of years for before observation and after is same i.e.,
        5 years
        var.EB.mu<- (1-w.B)*EB.mu.B

# Step 6: Outputs
         write.csv(data.frame(dataB$xng.no.,dataB$accident,dataB$trainflow,dataB$vehflow,
        dataA$accident, dataA$trainflow, dataA$vehflow, E.mu.B, w.B, EB.mu.B,E.mu.A, R,
        EB.mu.BB, var.EB.mu), file= "outputBA.csv")
        # Use Excel spreadsheet to summarize the results and determine the CMFs
```

E.6. Countermeasure study: EB-based NB approach

```
 # Same case study as in E.5.
 #Step 1: Obtain estimates of mean for before and after cases
       library(MASS)
       model.NB <- glm.nb(accident~ log(trainflow)+ log(vehflow), link=log, data=dataT)
       summary(model.NB)
       MAE<- model.NB$MSE
       E.mu.T<- fitted(model.NB)
       E.mu.B<- predict(model.NB,type="response", newdata= dataB) # for before period
       E.mu.A<- predict(model.NB,type="response", newdata= dataA) # for after period
#Step 2: Compute weights and determine EB estimates
      weight<-function(mu, theta){
       w<- 1/(1+mu/theta)
       return(w)
      }
     w.B<- weight( E.mu.B, model.NB$theta) # no need of w for after
      EB.estimate<- function (observed, mu, w){
       EB<- w*mu*5/6+(1-w)*observed
      return(EB)
      }
      # for before dataset
      EB.mu.B<- EB.estimate(dataB$accident, E.mu.B, w.B)
      R<- E.mu.A/E.mu.B
      EB.mu.BB<- R*EB.mu.B
      var.EB.mu<- (1-w.B)*EB.mu.B
# Step 3: Outputs
     write.csv(data.frame(dataB$xng.no.,dataB$accident,dataB$trainflow,dataB$vehflow,
     dataA$accident, dataA$trainflow, dataA$vehflow, E.mu.B, w.B, EB.mu.B,E.mu.A, R, EB.mu.BB,
     var.EB.mu), file= "outputBA_NB.csv")
     # Use Excel spreadsheet to summarize the results and determine the CMFs.
```