May 2019

Identifying and Incorporating Driver Behavior Variables into Crash Prediction Models

Mohammad Razaur Rahman Shaon
University of Wisconsin-Milwaukee

IDENTIFYING AND INCORPORATING DRIVER BEHAVIOR VARIABLES INTO CRASH PREDICTION MODELS

by

Mohammad Razaur Rahman Shaon

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Engineering

at

The University of Wisconsin-Milwaukee

May 2019

ABSTRACT

IDENTIFYING AND INCORPORATING DRIVER BEHAVIOR VARIABLES INTO CRASH PREDICTION MODELS
by

Mohammad Razaur Rahman Shaon

The University of Wisconsin-Milwaukee, 2019
Under the Supervision of Professor Xiao Qin

All travelers are exposed to the risk for crashes on the road, as none of the roadways are entirely safe. Under Vision Zero, improving traffic safety on our nation's highways is and will continue to be one of the most pivotal tasks on the national transportation agenda. For decades, researchers and transportation professionals have strived to identify causal relationships between crash occurrence and roadway geometry, and traffic-related variables on the mission of creating a safe environment for the traveling public. Although great achievements have been witnessed such as the publication of the Highway Safety Manual (HSM), research is rather limited in the area of incorporating driver behavior variables into safety modeling. As driver errors are responsible for more than 90 percent of crashes occurred, excluding such important information could cause ineffective, inaccurate, and incorrect prediction results and parameter inferences.

The primary reasons for this research void are the lack of driver information and methods for integrating driver data with roadway and traffic characteristics. Standard procedures for collecting and archiving driver behavior data do not exist, as highway agencies are not obligated to collect them. The most relevant source for driver behavior information is perhaps the crash report where police officers may record driver conditions and the possible driver factors contributing to the crash. However, such information is not available to near misses, traffic conflicts and non-crash traffic events where good behaviors prevail. As a result, unobserved data

heterogeneity will induce data overdispersion issues which are a significant limiting factor to safety modeling. Furthermore, the conventional approach to treating crashes as originated from a single risk source also induces heterogeneity in crash data and yields biased parameter estimates. Thus, a statistically rigorous methodology is in urgent need to consider the consequence of missing critical driver information in a crash model as well as to distinguish between distinct risk generating sources of a crash event when the driver information is available.

This dissertation contributes to the prediction of crash frequency and severity by explicitly considering human factors and driver behaviors in the modeling process. This endeavor began with a comprehensive literature review that identified and addressed data needs, technical issues, and latest development on the incorporation of human factors in safety analysis; and concluded with analytical framework and modeling alternatives to quantify driver behavior being proposed, developed and evaluated.

Given myriads of data elements to be explored, availability of contributing factors and crash data issues, a three-pronged modeling approach was adopted to accommodate a broad spectrum of data aggregated over areas, sites and crash events. This approach was informed by the complex nature of crashes involving highway geometry, traffic exposure, contextual factors, driver characteristics, vehicle factors, as well as the interactions among them. The availability of direct or surrogate measures of crash contributing factors varies by spatial unit. To give an example, socioeconomic and demographic features of the driving population are available at census tract; roadway geometry and traffic variables are available for segments and intersections; while specific driver conditions are only collected when a crash took place. With the flexibility in spatial context and risk generating sources, the three-pronged approach provides direct

benefits to guide different safety applications such as planning, design, and operations; and informs different programs such as engineering and enforcement.

The area-based crash models were developed to incorporate human factors and driver behavior in the form of socioeconomic and demographic data. In particular, behavior-based crash prediction models for speed and alcohol-related crashes were developed, respectively. Results showed that driver behavior-related crashes were more correlated with socioeconomic and demographic variables than traffic and trip-related explanatory variables.

The site-specific crash models were exploited to address the effect of human factors and driver behaviors in two fronts: 1) developing rigorous statistical models to account for unobserved heterogeneity induced overdispersion when driver behavior information is not available, 2) treating behavior variables as a separate risk source in a prediction model. The first pursuit leads to the development of a mixed distribution random parameter model to explicitly account for unobserved heterogeneity. The second pursuit results in the development of a multivariate multiple risk source regression model to simultaneously predict crash count and severity. Modeling results show better model performance and valid model inferences for quantifying the effect of driver factors on crash occurrence can be achieved with proposed multiple risk source models.

The event-oriented models were utilized to evaluate the interaction between human factors and engineering variables in a crash event. Driver errors were categorized by the driver's action during a crash on a roadway segment. The modeling results identified many highway geometric features, traffic conditions, and driver characteristics as statistically correlated to different types of driver mistakes. An exploratory analysis was followed to evaluate the effect of driver mistakes on the crash injury outcomes.

The dissertation demonstrates the strength of using diverse methods and models under various circumstances to incorporate human factors and driver behavior in crash prediction. The safety professionals can choose appropriate models based on their own data availability, unit of analysis, and design effective treatments or training programs. This research shares new insights to reinforce informed decision support for cumulative safety improvement of roadway network, recognizes the opportunities to address high priority safety issue areas, and determines the appropriate countermeasures.

Keywords: Over-dispersion; Unobserved heterogeneity; Mixed model; Random parameters model; Negative binomial-Lindley; Crash causation mechanism; Multiple risk generating process; Driver-behavioral factors; Multivariate Analysis; Driver errors

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First of all, I wish to convey my profound gratitude to the Almighty Allah for enabling me to complete this research work successfully. I also wish to thank the Almighty Allah for giving me this wonderful life to explore the world, learn and share the knowledge.

I would like to express my most sincere appreciation and thanks to my advisor, Dr. Xiao Qin for encouraging, helping, motivating and guiding me throughout my entire graduate study and of course, this research wouldn't be possible without his careful supervision. He was extremely supportive in every step of my research efforts. Through his teaching and discussions, he has greatly motivated me and contributed to my concept of Highway Safety.

I would like to thank my committee members, Dr. Robert Schneider, Dr. Jie Yu, Dr. Yin Wang, Dr. Chao Zhu and Dr. Lingqian (Ivy) Hu for being my committee members, giving their valuable time to read through my research proposal and dissertation draft, and giving me their insightful suggestions and comments.

A part of this thesis work was supported by funds from Wisconsin Department of Transportation (WisDOT). Advice and help received from all the members of the research team and technical panel of this project. Specially, I will like to thank Dr. Robert Schneider and Dr. Dominique Lord for their help in different areas of my research.

Last, but by no means least, it is a great opportunity for me to express my gratitude and immense pleasure to thank my beloved parents and family without whose blessings, love and support, this success of mine wouldn't have been possible.

PREFACE

Articles in work, submitted or published in referred journals:

1. Shaon, Mohammad Razaur Rahman, Xiao Qin, Mohammadali Shirazi, Dominique Lord, and Srinivas Reddy Geedipally. "Developing a Random Parameters Negative Binomial-Lindley Model to analyze highly over-dispersed crash count data." Analytic Methods in Accident Research 18 (2018): 33-44.
2. Shaon, Mohammad Razaur Rahman, Xiao Qin, Zhi Chen, and Jian Zhang. "Exploration of Contributing Factors Related to Driver Errors on Highway Segments." Transportation Research Record (2018): 0361198118790617.
3. Shaon, Mohammad Razaur Rahman, and Xiao Qin. "Use of Mixed Distribution Generalized Linear Models to Quantify Safety Effects of Rural Roadway Features." Transportation Research Record: Journal of the Transportation Research Board 2583 (2016): 134-141.
4. Shaon, Mohammad Razaur Rahman, Xiao Qin, Amir Pooyan Afghari, Simon Washington. “Incorporating behavioral variables into prediction of crash counts by severity: a multivariate multiple risk source approach”. Working Paper, Accident Analysis and Prevention, 2019.
5. Shaon, Mohammad Razaur Rahman and Xiao Qin. “How is Injury Severity Affected by Driver Errors: A Crash Data Based Investigation”. Working Paper.

Chapter 1 Introduction

1.1 Background

According to the National Highway Traffic Safety Administration (NHTSA), roadway crash fatality and injury rates decreased significantly from 2005 to 2011. This reduction was the result of collective efforts of roadway design, law enforcement, educational programs, and vehicle safety technologies as well as reduced personal travel and goods movement due to the Great Recession. However, since 2011, the number of annual US traffic fatalities has risen, with the steepest increase in nearly 50 years (7.2 percent) from 2014 to 2015 (NHTSA 2016). More importantly, driver behavior such as speeding, driving under influence, distracted driving related crashes are on the rise. In 2015, there is an increase of 3.2 percent in alcohol-impaired driving related crashed from 2014 on US roadways (NHTSA 2016). In 2012, speeding was a contributing factor in 30 percent of all fatal crashes with an increase of 2 percent from 2011. This recent reverse in the downward trend of traffic fatalities and increase in behavior-related crashes warrant a close and careful review of the variables contributing to traffic crashes. Special consideration needs to be taken to explore human factors for a better understanding of driver behavior related crashes.

Driving is a highly complex skill that involves dynamic interleaving and execution of multiple tasks. Human needs, limitations, and capabilities are essential in performing driving tasks. To identify critical reasons for crashes, NHTSA conducted National Motor Vehicle Crash Causation Survey (NMVCCS) from 2005 to 2007 to collect on-scene information about the crash events and associated factors leading up to crashes (NHTSA 2008). The NMVCCS classified driver-related critical reasons into recognition error, decision error, performance error, and nonperformance error (sleep, etc.) and assigned 94 percent of 2,046,000 crashes to driver related

critical reason(s). In a separate study, NHTSA listed drunk driving, distracted driving, drowsy driving, seatbelt use, speeding, drug driving as major risky driving behaviors which may compromise the safety of other drivers on the road (NHTSA). Roadway safety can also be compromised by driver limitations and competency with primary age groups being young drivers under the age of 19 years and, older drivers with age 65 years or more. In 2015, 13 percent of all fatal traffic crashes involved older drivers (age 65 years or more) (NHTSA 2016). There was 29 percent nationwide increase in old driver fatalities from 2006 to 2015 whereas the nationwide increase in total traffic fatalities was approximately 17 percent over the same time period. It is indeed crucial to understand the effect of risky driving behavior and driver limitations to develop effective safety countermeasures.

Although human factors are known to be as one of the major contributors in crash occurrence, serious attention is needed to explicitly consider human factors in project planning and development for safety improvements. There are many reasons affecting the effort devoted to incorporating human factors into safety decision-making. One of the major reasons is that every safety improvement is presumed to be made upon on human factors. For example, roadway design standard in the "Policy on Geometric Design of Highways and Streets" (AASHTO, 2000) is primarily based on human capabilities and vehicle performance. Thus, no explicit consideration for human factor is needed (AASHTO 2011). Second, compared with driver education or law enforcement programs, roadway design and traffic control features may be considered to have limited influence on driver behavior. Third, engineers and designers do not have sufficient knowledge of human factors or have not been adequately trained to adjust design standards according to the demographic, socio-economic induced changes. Finally, there are not enough driver behavioral data to support designers and engineers in making project planning and

development decision for safety, nor do they have the appropriate tools to perform relevant analysis and evaluation. Thus, a research effort is essential to integrate human factors and driver behaviors into roadway safety analysis studies.

Crashes are complex events as evident by the 110 data elements recommended in the Model Minimum Uniform Crash Criteria (MMUCC) (USDOT 2014). A major challenge in traffic safety analysis is that a myriad of factors can actively or passively influence both the number and severity of crashes. To understand the causes of a crash event, Heinrich proposed the Domino Theory in the context of industrial accidents in 1931 in which accidents are treated as a step in a sequential chain of events, each of which is dependent on the previous event (Heinrich 1941). Heinrich's Domino Theory is similar to the simple linear sequential model used today: removing one of the sequential events from the chain of events would avoid the accident. Later, Haddon used epidemiological concepts to propose the Haddon matrix as a method to capture how several factors such as human, vehicle and infrastructure variables affect a crash event and the sequential nature of crash event by proposing the pre-crash, crash and post-crash phases (Haddon Jr 1968). A series of factors can be obtained by reviewing detailed crash reports, but it is difficult to form the chain of events from the pre-designed report, making it almost impossible to analyze crash as a series of events and the outcome of a myriad of factors. As crash causation analysis is difficult to conduct, predicting the crash occurrence or outcome of a crash event (e.g., crash severity) with statistically correlated data is appealing. The predictive crash analysis is performed based on a list of variables or risk factors to model crash events using mathematical equations and statistical inferences. Ideally, if the mathematical equation in crash modeling is correctly specified, it can reveal underlying safety effects and can provide useful insight into the crash underlying process (Mitra 2006). Therefore, the proper equation is vital to identify

correlated variables with a crash event. In 2010, the Highway Safety Manual (HSM) published by AASHTO, a milestone in the application of predictive crash analysis methodologies, became the first national resource for quantitative information about crash analysis and evaluation by providing means to estimate crashes based on geometry, operating characteristics, and traffic volume, (AASHTO 2010).

Previous research has established that traffic crashes are the results of chains of causal events that arise from a multitude of contributing factors associated with roadway design, traffic operations, pavement conditions, driver behavior, human factors, and environmental factors. In temporally aggregated crash dataset, various factors leading to traffic crashes at a site do not necessarily contribute equally, though; therefore, traffic crash counts at every site should be considered as the results of various risk sources, with each risk source playing either a vital or supporting role. However, conventional crash frequency models treat the total number of crashes at a roadway site as the outcomes of a single risk source by using a predictive equation estimated with Poisson or Negative Binomial (NB) distribution. While these single equation models are statistically sound and practically useful, their results may yield biased parameter estimates due to issues related with data overdispersion (Zou et al. 2015, Rahman Shaon and Qin 2016, Shirazi et al. 2016). Furthermore, single-equation models are incapable of assuming that crashes may have various risk sources, which could result in data heterogeneity.

In addition to human factors and driver behaviors, previous crash data modeling literature established the relationship between crashes and roadway geometry, traffic exposure, and other contextual variables. FHWA noted that about 22% of PDO crashes, 19% of injury crashes and 16% of fatal crashes occur on US roadways due to adverse weather based on 10 years of crash data (Federal Highway Administration (FHWA 2013). Crashes can also be affected by travel

demand and pattern, driver demographic characteristics and, broad economic trends. Recent studies funded by NHTSA analyzed factors contributing to changes in crash fatalities over time in the United States (Longthorne et al. 2010, Bush Active Project). These studies concluded that economic conditions as well as federal behavioral and vehicle safety standards contributed to the downward trend in crash fatalities in 2008 (Longthorne et al. 2010). Furthermore, safety measures implemented by local agencies depend on local factors such as weather conditions, roadway design treatments, local law enforcement activities, fuel prices and taxes, and unemployment rates. Recent practices of FHWA sponsored Systemic Safety Evaluation methodology and tool by multiple states illustrated the use of highway geometric and traffic information for safety improvement project selection (Preston and Farrington 2011, Walden et al. 2015). The New York City Department of Transportation (DOT) developed a fixed anti-icing system for a portion of the Brooklyn Bridge to improve safety and mobility in adverse weather (Ward 2002). State and local law enforcement agencies are vital partners in helping DOTs in Driving Zero Fatalities to reality. Thus, highway safety performance should be considered as an outcome of both global trend and local influences that affect people's travel needs, decisions, and behavior.

1.2 Problem Statement

Safety studies are mostly focused on evaluating the effect of roadway geometric characteristics due to their safety implications in developing crash countermeasures. Although driver factors have been recognized as a major contributor to crash occurrence, driver behavior related variables were seldom incorporated in crash prediction models. Most safety data elements are extracted from police accident reports, state and local highway inventory, traffic count data, local

weather stations, and other databases. Despite the wealth of information, these conventional databases only cover a small fraction of a large number of elements that define human behavior, vehicle and roadway characteristics, traffic characteristics, and environmental conditions that determine the likelihood of a crash occurrence and its resulting injury severity (Mannering et al. 2016). Standard procedures for collecting driver behavior data do not exist, as highway agencies are not obligated to gather such information for safety management systems. Many other elements remain unobserved during the crash data modeling process. In Econometrics, the unavailability of relevant contributing factors that are correlated with dependent variable is called "Unobserved heterogeneity". The existence of unobserved heterogeneity is a major issue in a crash dataset. Unobserved heterogeneity is usually considered as random errors in traditional crash prediction models because the effect of each covariate is restricted to be the same across all observations, which in turn causes extra dispersion problems. Such modeling strategies can cause serious model specification problems and may result in a variation of the estimated effect of observed covariates (Mannering et al. 2016). An overview of the potential for heterogeneity in driver behavior due to a variety of highway factors was highlighted by Mannering et al. (2016). The research found that varying lane and shoulder widths may have an impact on the likelihood of a crash event, but these effects can vary among observations due to time-varying traffic, weather conditions, and the driver's reaction, all of which are not available for model development. Ignoring heterogeneous effects in explanatory variables leads to biased parameter estimates and therefore, inaccurate conclusions (Mannering et al. 2016).

The choice of modeling technique usually depends on the characteristics of the data. To obtain unbiased inference from data modeling, the corresponding model should be able to account for characteristics related to the target dataset. Crash data is often characterized by a

large sample variance compared with the sample mean[1] (Lord et al. 2005, Mitra and Washington 2007). Extensive research has been devoted to modeling and analyzing this type of crash dataset. One of the most notable accomplishments is the application of negative binomial (NB) models in crash frequency data. NB models can handle data over-dispersion by assuming a gamma distribution for the exponential function of the disturbance term in the Poisson mean. However, recent studies have pointed out that biased parameter estimates in the NB model can be found in a dataset with a long tail (Zou et al. 2015, Shirazi et al. 2016). A long tail is a statistical phenomenon that occurs when sample observations have a few very high crash counts or have preponderant zero observations which shift the overall sample mean to near zero (Shirazi et al. 2016). Failure to account for data over-dispersion leads to biased and inconsistent parameter estimates, which in turn causes erroneous inferences from models and inaccurate crash prediction information.

Based on the above discussion, the issues and problems with current safety evaluation of roadway transportation system can be summarized as follows:

- Despite the improvements in different domains of roadway system, the number of injury and fatality is still increasing. Strategies are needed to analyze new data sources and identify potential variables related to the change in crash counts and their severities.
- Given the myriad of factors contributes to crash occurrence and their availability, a single crash prediction model may be inadequate to handle such a broad spectrum of data with diverse and varying characteristics. Therefore, a new

[1] In a statistical term, the sample data is over-dispersed when the variance is greater than the mean. Data over-dispersion is often caused by unobserved data heterogeneity due to unobserved, unavailable, or unmeasurable variables that are important to explain model responses.

approach based on different levels of spatial aggregation of crash data is needed to evaluate roadway system safety.

- Most studies developed crash prediction models using traditional roadway geometry and traffic variables. Although driver factors are responsible for more than 90% of crashes, the available behavioral factors were not extensively explored to model crashes in literature. Traditional models also do not distinguish between distinct sources of crash risk into model development.
- Unavailability of rigorous methodological alternatives to quantify and distinguish distinct risk from driver behavior as well as to account for issues related with crash data when behavior information is unavailable.

It is a challenge to account for driver behavior and other human factors in a crash prediction model so that the causal relationship between crash occurrence and driver factors can be well supported, determined, and established. Based on current safety practices and limitations discussed above, this dissertation focuses on both data and technical challenges on incorporating human factors into crash prediction models and developing solutions for them. This dissertation contributes to safety analysis of a roadway network by developing the framework and procedures to incorporate the human factors into the crash prediction models along with existing engineering variables. The existing safety data items are usually available at different spatial units or on different aggregation levels. Thus, a multi-level, pronged approach based on spatial aggregation of crash data is developed to evaluate roadway system safety. This approach can allow us to handle a broad spectrum of available data items from different sources at different spatial units. This dissertation contributes in developing complex methodological alternatives using mixed-

distribution and multivariate multiple risk source modeling approaches to account for unique crash data characteristics, identify distinct sources of crash risks and unobserved variables in crash data.

1.3 Research Objectives

The objective of this dissertation is to develop crash prediction models to identify potential factors related to crash occurrence with an emphasis on driver behavior related variables. More specifically, this dissertation aims to:

1. Conduct a comprehensive literature review to identify the variety of factors that contribute to the occurrence of a crash event with a major focus of driver behaviors.
2. Explore new data sources such as behavioral, socio-economic, demographic, etc. to identify the unknown and intrinsic relationship of potential new behavioral risk factors with crash occurrences.
3. Develop modeling alternative to account for unobserved heterogeneity induced overdispersion when driver factors are not available in crash data modeling.
4. Develop modeling alternative to incorporate driver behavior as a distinct source of crash risk in crash data modeling.
5. Develop modeling framework to identify factors contributing to driver errors and its effect on crash severity.
6. Evaluate predictive power of proposed methodologies to obtain more accurate crash prediction.

7. Develop a guide for appropriate data structure and modeling techniques to model crash data at different geographic units to handle a broad spectrum of available data and facilitate decision support system for safety improvement.

The incorporation of human factors will enable the safety professionals to design effective countermeasures or training programs to increase awareness and existing safety situation. Achieving the research objectives will help to evaluate safety conditions of a roadway network by incorporating new variables and with rigorous methodological alternatives. New insights will be shed to support decision-making for cumulative safety improvement of the roadway network.

1.4 Dissertation Outline

Based on the proposed research objectives, this dissertation is organized into eight chapters. The outline of dissertation organization is illustrated in Figure 1-1. The remaining chapters of this dissertation will be organized in a journal paper format as follows:

Chapter 2 provides a comprehensive review of crash prediction studies, potential explanatory variables and their effects on crash occurrence and, relevant methodological alternatives including recently proposed mixed distribution and semi-parametric models to address data issues. This chapter focuses on summarizing the advantages and limitations of those studies and identifying critical research issues.

Chapter 3 presents the model development and analysis results of area-level crash data. Census tract has been selected as an appropriate area unit in this section. A brief discussion on crash aggregation in a spatial unit and availability of behavioral, socioeconomic, demographic,

land use and engineering variables is provided in this chapter. A variety of models including new behavioral, socioeconomic and demographic variables in crash occurrence are considered. Finally, the modeling results of area-based crash data and their applications are presented.

Chapter 4 presents the model development and analysis results of site-specific crash data when important behavioral information is missing in target crash dataset. Unavailability of important variables can cause unobserved heterogeneity induced overdispersion issue in crash dataset. Rigorous and statistically robust methodological alternative is needed to address issues related to crash dataset. A random parameter Negative Binomial Lindley regression model is developed for segments in this chapter to account for higher overdispersion issue induced by missing behavioral information and excess zero crash sites. The predictive ability of proposed model is compared with traditional modeling techniques to obtain the best model for crash prediction at site-specific level.

Chapter 5 presents the model development and analysis results of site-specific crash data when important behavioral information is available in target crash dataset. Behavioral information is usually available at larger spatial unit (e.g., county) to analyze the physical and psychological status of a community and acts as a separate risk source in crash occurrence. A multiple risk source regression model is developed for segments in this chapter to incorporate behavioral factors as a separate risk source in CPM. Moreover, a multivariate multiple risk source regression model was developed to account for the correlation between injury severity levels. The proposed model results are compared with traditional model results to understand the effect of behavioral factors on crash occurrence.

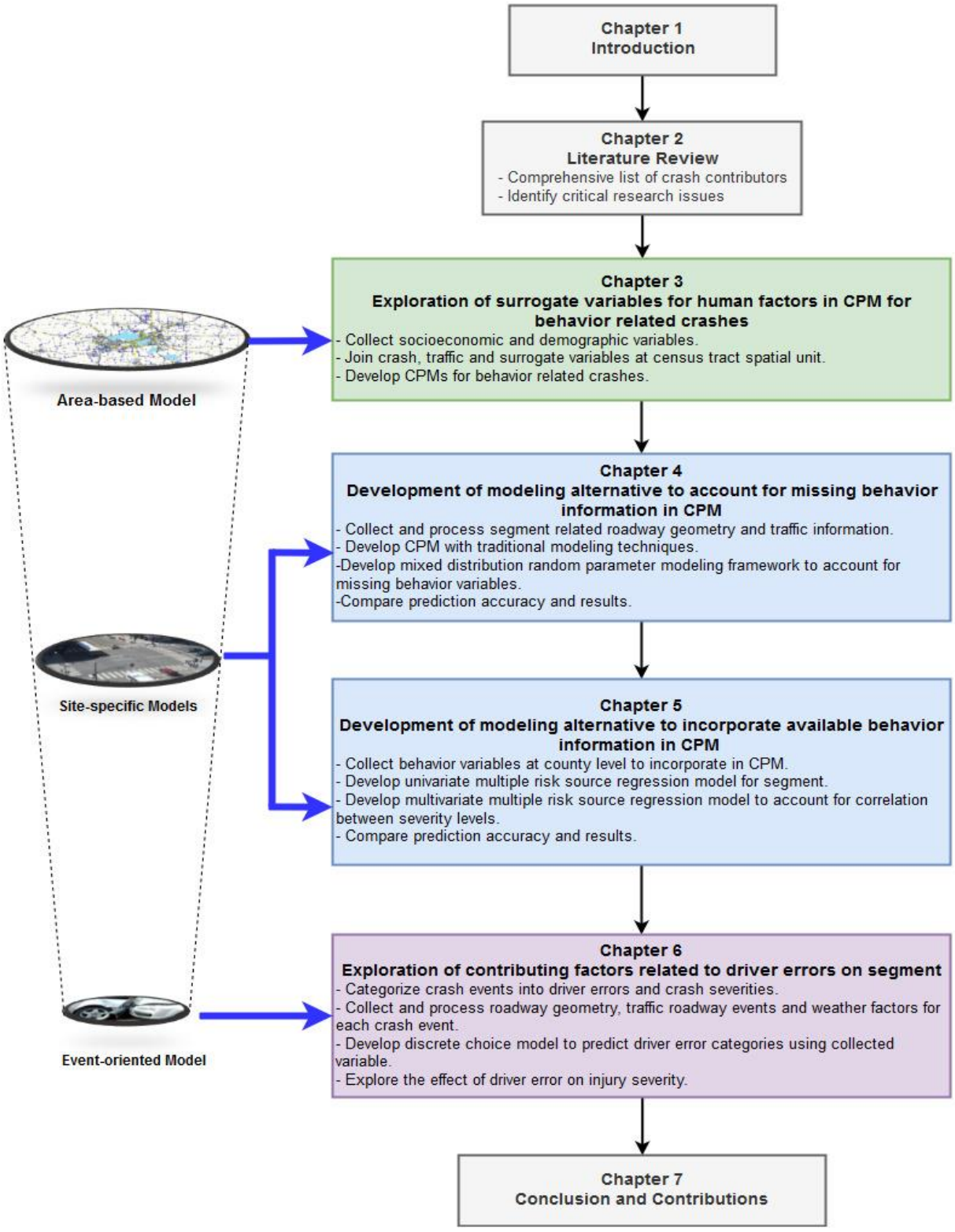

Figure 1-1 Dissertation Organization.

Chapter 6 presents the model development and analysis results of event-oriented crash data. An event-oriented model is designed from a different perspective by categorizing driver errors at segments to evaluate the interaction between human factors and engineering variables. A multinomial ordered probit model was developed to quantify the effect of roadway geometry, traffic variable, roadway event and human factors on the occurrence of driver error(s) in a crash event. Furthermore, an exploratory analysis is conducted to understand the effect of driver error(s) on resulting crash injury severity.

Chapter 7 summarizes the conclusions, contributions of the dissertation and presents future research directions.

1.5 References

AASHTO, 2010. Highway safety manual American Association of State Highway and Transportation Officials, Washington D.C.

AASHTO., 2011. A policy on geometric design of highways and streets, 2011 AASHTO.

Bush, M.S., Active Project. Identification of factors contributing to the decline of traffic fatalities in the united states. University of Michigan.

Federal Highway Administration (Fhwa), 2013. How do weather events impact roads? U.S. Department of Transportation, Washington D.C.

Haddon Jr, W., 1968. The changing approach to the epidemiology, prevention, and amelioration of trauma: The transition to approaches etiologically rather than descriptively based. American Journal of Public Health and the Nations Health 58 (8), 1431-1438.

Heinrich, H.W., 1941. Industrial accident prevention. A scientific approach. Industrial Accident Prevention. A Scientific Approach. (Second Edition).

Longthorne, A., Subramanian, R., Chen, C.-L., 2010. An analysis of the significant decline in motor vehicle traffic fatalities in 2008.

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 & Prevention 37 (1), 35-46.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic methods in accident research 11, 1-16.

Mitra, S., 2006. Significance of omitted variable bias in transportation safety studies.

Mitra, S., Washington, S., 2007. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis & Prevention 39 (3), 459-468.

NHTSA, Risky driving. US Department of Transportation.

NHTSA, 2016. 2015 motor vehicle crashes: Overview. Traffic safety facts research note 2016, 1-9.

NHTSA., 2008. National motor vehicle crash causation survey: Report to congress. National Highway Traffic Safety Administration Technical Report DOT HS 811, 059.

Preston, H., Farrington, N., 2011. Minnesota's best practices and policies for safety strategies on highways and local roads.

Shirazi, M., Lord, D., Dhavala, S.S., Geedipally, S.R., 2016. A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: Characteristics and applications to crash data. Accident Analysis & Prevention 91, 10-18.

USDOT, 2014. MMUCC guideline: Model minimum uniform crash criteria. 4th Edition ed., Washington DC.

Walden, T.D., Lord, D., Ko, M., Geedipally, S., Wu, L., 2015. Developing methodology for identifying, evaluating, and prioritizing systemic improvements.

Ward, B., Year. Evaluation of a fixed anti-icing spray technology (fast) system. In: Proceedings of the Proceedings of the 81st Annual Meeting of the Transportation Research Board, Washington, DC.

Zou, Y., Wu, L., Lord, D., 2015. Modeling over-dispersed crash data with a long tail: Examining the accuracy of the dispersion parameter in negative binomial models. Analytic Methods in Accident Research 5, 1-16.

Chapter 2 Literature Review

This chapter presents a comprehensive review of past and current crash prediction model (CPM) studies and relevant methodological options used to model crash or crash outcomes. CPM studies will be summarized to provide insight on potential variables explored in literature and their effect on crash occurrence. There are two parts in this chapter. The first part provides the overview and concept of CPM research including potential explanatory variables and their effect on crash occurrence. The second part discusses the data issues related to crash data along with the advancement of methodological options to overcome data limitations as well as to extract a large amount of information from modeling results. This section will also describe the potential methodological options considered in this dissertation and their formulation.

2.1 Concept and Overview of CPM Research

Crashes are usually caused by a combination of contributing factors, and identifying the most influential causes can be a daunting task. Factors may also have different effects at different levels of a crash event. Haddon used epidemiological concepts to propose the Haddon matrix as a method to capture how several factors such as human, vehicle and infrastructure variables affect a crash event and the sequential nature of crash event by proposing the pre-crash, crash and post-crash phases (Haddon Jr 1968). A sample of Haddon matrix is provided in Table 2-1. However, the Haddon matrix cannot explicitly incorporate exposure (e.g., traffic volume, distance traveled, time traveled) and interaction between factors.

Table 2-1 Haddon Matrix for Safety.

	Human Factor	Vehicle/ Equipment Factor	Physical Environment	Socioeconomic Factor
Pre-Crash	• Education and licensing. • Driver impairment. • Crash avoidance maneuvers (braking, turning, etc.)	• Crash avoidance equipment and technology (lights, tires, collision avoidance, etc.). • Vehicle Design. • Vehicle load.	• Adequate roadway marking. • Road hazards. • Distractions. • Weather conditions.	• Enforcement activities. • Insurance incentives. • Social norming. • Ability to use safety equipment appropriately.
Crash	• Health at time of crash. • Sitting properly in restraint • Impairment.	• Speed of travel. • Functioning of safety equipment (seat belts, air bags, child restraints). • Energy absorption of vehicle.	• Roadside features. • Guardrails. • Type and size of object struck.	• Laws concerning use of safety equipment.
Post-Crash	• Response to EMS. • Severity of injury. • Type of injury.	• Ease of extraction from vehicle. • Integrity of fuel systems and battery systems. • Automated crash notification and GPS locator.	• Distance of EMS personnel. • Notification of EMS personnel. • Accessibility to crash victims. • Rehabilitation program availability.	• Trauma system equipment, personnel, training. • Information sharing. • Resources and program for psychological recovery from trauma.

Factors related to each crash event are usually collected by a police crash report. A series of factors can be obtained by reviewing detailed crash reports, but it is difficult to form the chain of events from the pre-designed report, making it almost impossible to analyze crash as a series of events and the outcome of a myriad of factors. Most literature focuses on the pre-crash events to understand crash mechanism as it is the most influential step. NHTSA sponsored the National Motor Vehicle Crash Causation Survey (NMVCCS) from 2005 to 2007, aimed at collecting on-

scene crash information and associated factors leading up to crashes involving light vehicles (NHTSA 2008). The NMVCCS survey investigated several facets of crash occurrence such as the pre-crash movement, critical pre-crash event, critical reason, and the associated factors. The study report noted "The critical reason is the immediate reason for the critical pre-crash event and is often the last failure in the causal chain of events leading up to the crash. Although the critical reason is an important part of the description of events leading up to the crash, the critical event, the critical reason underlying the critical event, or the associated factors should not be interpreted as the cause of the crash". Several studies also investigated the precursors of a crash event. For example, Lee et al. found that variation in traffic speed and density are significant factors in crash occurrence after controlling for roadway geometry, weather and time of the day (Lee et al. 2002). Chatterjee and Davis found that drivers with a longer reaction time than following distance play a key role in crash occurrence from a stopping shockwave (Chatterjee and Davis 2016). By definition, precursor factors can be considered as critical for a crash event but cannot be explained as crash causation factors.

As crash causation analysis is difficult to conduct due to the challenges of formulating the chain of events, predicting crash occurrence or severity of a crash event with statistically correlated data is appealing. Predictive crash analysis is performed based on a list of variables or risk factors to model crash events using mathematical equations and statistical inferences. Ideally, if the mathematical equation in crash modeling is correctly specified, it can reveal underlying safety effects and can provide useful insight about the crash underlying process (Mitra 2006). Therefore, the proper equation is also vital to identify crash correlated variables. In 2010, the Highway Safety Manual (HSM) published by the American Association of State Highway and Transportation Officials (AASHTO), a milestone in the application of predictive

crash analysis methodologies, became the first national resource for quantitative information about crash analysis and evaluation by providing means to estimate crashes based on geometry, operating characteristics, and traffic volume, (AASHTO 2010).

It is also worth noting that the variable list varies between the spatial unit of crash modeling. In area-based crash modeling (where traffic analysis zones, county boundaries, etc. are used as the unit of analysis), socio-economic and, demographic factors, highway miles by speed limit, estimated total vehicle miles traveled are often used (Abdel-Aty et al. 2013, Pulugurtha et al. 2013, Wang and Kockelman 2013). In site-specific crash modeling (where specific segments and intersections are used as the unit of analysis) roadway design and traffic characteristics are often used as explanatory variables (Bauer and Harwood 1996, Greibe 2003, Qin et al. 2016); and in event-oriented modeling (where crashes and interactions are modeled in real-time), disaggregated speed, speed deviation, and volume counts are often used as predicting variables (Yu et al. 2013, Li et al. 2014, Chen et al. 2017).

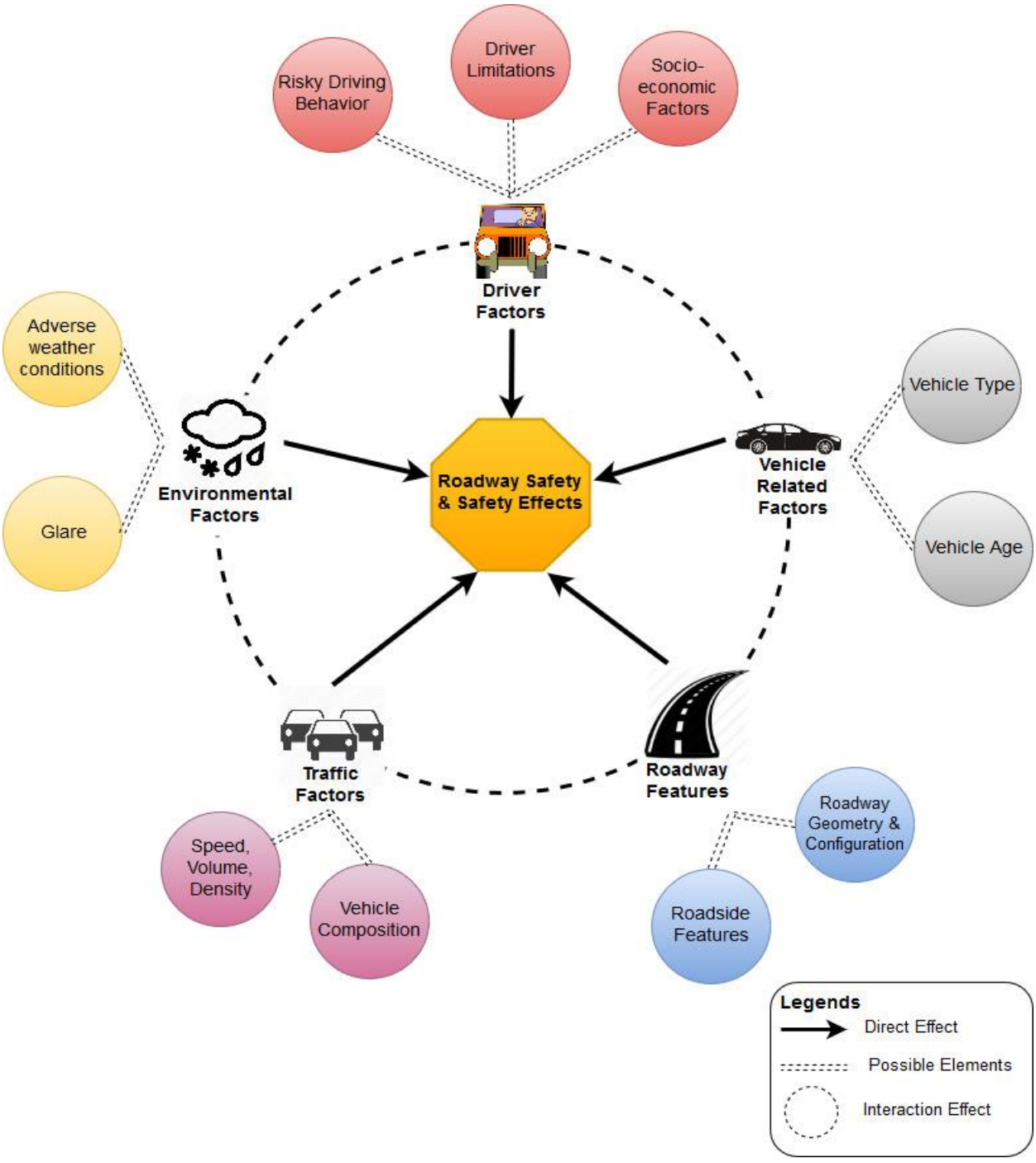

Figure 2-1 Theoretical Illustration of Crash Contributing Factors.

2.2 Safety Culture

Safety culture plays an important role in driving behavior and risk perception (Carter et al. 2014). Safety culture is the shared values and beliefs among road users that influence their decisions to drive defensively and demonstrate their commitment to safety over other competing demands and goals. Numerous studies indicate that parents play a significant role in the driving behavior of their adolescent children (Taubman et al. 2012, Taubman et al. 2013). For example, some drivers may follow the behaviors of other drivers in their community, regardless of roadway design or other site characteristics (Schneider et al. 2018). Societal expectations of acceptable transportation risk can also influence risk-taking behavior (Moeckli and Lee 2007, Rahman Shaon et al. 2018a). Safety culture is usually determined by driver's attitude (e.g., personalities), social norms (i.e., the actions of others) and perceived risk (e.g., enforcement). Safety culture is also influenced by drivers' social economic status (e.g., car dependence) and education background (e.g., respect others, risks and consequences of DUI, speeding).

Understanding what factors influence driver attitude is of paramount importance to identify potential countermeasures or design effective driver training programs to reduces crash counts and injury severity. Starting in 2008, AAA Foundation for Traffic Safety launched study to estimate traffic safety culture index in the USA using a nationally representative sample of motorists (AAA 2017). This effort is carried out annually to identify key indicators regarding the degree to which traffic safety is valued and pursued by drivers in the U.S. Result showed that albeit of strongly held concerns about risky driving attitude, many individuals admit engaging in unsafe driving practice. Taubman et al. noted that young drivers who perceived their parents to be better role models, to provide encouraging and empowering feedback for safe driving, to enable more open communication, to convey clearer messages regarding safe driving, to monitor

their driving, and to set clear limits on breaking traffic laws, tended to report taking risks less frequently, being more personally committed to safety, and driving more carefully and in a less aggressive and risky manner.

2.2.1 Demographic and Socio-Economic Factors

Collecting driver errors from the crash report or traffic citation database is a potential way to collect driver-related information, but it can only be collected after a crash or citation occurred. Human judgment, physiological and psychological behaviors of driver at the time of driving are almost impossible to capture. Therefore, it is not easy to use driver error as a predictor variable in crash modeling. A plausible solution is to use surrogate variables for driver error. Driver and passengers' age and gender are explored as human factors in literature as the driver error is not readily available (Hing et al. 2003, Mayhew et al. 2003, Sagberg and Bjørnskau 2006, NHTSA 2008, Bao and Boyle 2009, Wang et al. 2015). In psychological studies, it is noted that there is a significant statistical difference in the naturalistic decision process due to both age and gender (María L. Sanz de Acedo Lizárraga 2007). A list of studies also found a significant difference in crash risk related to driver age and gender (Massie et al. 1995, Åkerstedt and Kecklund 2001, Chang and Yeh 2007, McAndrews et al. 2013, McAndrews et al. 2017). Age cohorts and proportion of gender are usually used as an explanatory variable in predicting crash count or severity in area-level crash modeling (Hadayeghi et al. 2007, Lee et al. 2017). McAndrews et al. investigated the relationship between age, gender, race/ethnicity and travel mode with crash injury severity in Wisconsin (McAndrews et al. 2013, McAndrews et al. 2017). The authors found that adolescents and older travelers face a higher risk of fatal and injury crashes compared to adults aged 25 to 64 years (McAndrews et al. 2013). They also noted that white people are

equally safe as pedestrians and motor vehicle occupants, whereas other racial and ethnic groups are less safe in all modes of transportation (McAndrews et al. 2013, McAndrews et al. 2017). One of the major sources of demographic data used in literature is US census data. Other demographic variables such as percent of the population with a driver license, transit pass, population density, etc. were also used in literature to predict crash count (Hadayeghi et al. 2010, Jiang et al. 2016).

Similar to the demographic information, socio-economic attributes were also used in literature to predict crashes (Hadayeghi et al. 2003, Aarts and Van Schagen 2006, Hadayeghi et al. 2010, Jiang et al. 2016, Wang and Huang 2016). It is well documented in the travel demand prediction that traveler's economic conditions (e.g., median household income) is one of the key determinants to their trip making decisions, i.e., number of trips, trip lengths, and destinations. Several studies shared a common conclusion that unemployment rate is negatively associated with fatality rates (Longthorne et al. 2010), veiled from the observations that the period of low fatality rates often coincides with recessions. Another economic metric, gasoline price has been extensively studied for its impact on fatality rate. Most studies concluded that gasoline prices negatively affected fatal crashes, albeit indirectly (Grabowski and Morrisey 2004, Leigh and Geraghty 2008). According to a review of the trend in 2015 crash fatalities conducted by Wisconsin Bureau of Transportation Safety (BOTS), there is a 28.2 percent decrease in real gas price in 2015 compared to the previous year (Wisconsin Department of Transportation (WisDOT) 2017). Using "the proportionate change in the given motor vehicle fatality count" including the "real gasoline price, 1-year lag" (Grabowski and Morrisey 2004) estimated that about a 23% increase in fatalities can be explained by the decrease in gas prices and an increase in gasoline taxes substantially reduced traffic fatalities. The underlying theory is that the

increase in gasoline price deters the gasoline consumption, which resulted in the decrease in crash fatalities. Leigh et al. found that a 10% increase in the gasoline tax led to a 1.8%-2.0% decrease in fatalities after assuming base year and scenario year were equally likely (Leigh and Geraghty 2008). However, the change of income and fluctuation of gasoline prices may affect trip decisions and driver behavior (a better way to save gasoline) to an average traveler, they may not have a substantial impact on certain behavior such as alcohol consumption and DUI. Ye et al. explored the relationship between monthly DUI fatal crashes and the monthly gasoline prices and VMT in Texas (Ye et al. 2011). Although they found VMT does have a positive effect on DUI fatal crashes, they failed to show a statistically significant relationship between the monthly number of DUI fatal crashes and gasoline prices. The authors attributed this to the lack of public transportation in Texas cities because travelers must use their personal automobile to commute regardless of the gasoline prices. The findings suggest that the gasoline price does not affect the Texas drivers that drink and drive.

Social class or socio-economic status represents a person's rank in society in terms of wealth, occupational prestige and education. Piff et al. noted that higher social class individuals are more prone to unethical behavior in both naturalistic and laboratory settings which include violating traffic rules (Piff et al. 2012). Socio-economic variables can also be used as a surrogate variable for exposure. Jiang et al. investigated the importance of various features on macro-level hotzone identification for crash risks by injury levels, collision types, and non-motorized crashes separately using roadway and traffic-related features, demographic features, socio-economic and land use features (Jiang et al. 2016). Authors noted that school enrollment density and percentage of households having 2 or more automobiles are consistently important for all types of crash risks. Authors found that hotzones are associated with higher school enrollment density and a

lower percentage of households having 2 or more automobiles. Demographic and socio-economic data are often used together in crash modeling as they are intertwined. Similar to the demographic data, socio-economic information is available in census-tract spatial unit, these variables are usually used in area-level crash prediction for transportation network applications. Important socio-economic data items such as vehicle ownership, median income, employment rate, education rate, etc. are used in literature as a surrogate of drivers' personality trait and behavior. Use of these variables in the area-based crash analysis may help safety professionals for zone development and identifying communities with high crash risks.

2.2.2 Human Factors in Driving

Driving a vehicle is the most popular means of transportation in the USA. According to the 2009 statistics published by Office of Highway Policy Information (OHPI), 87 percent of the US driving-age population (16 years or older) has a driver license (FHWA 2012). In Wisconsin, there are 726 drivers per 1000 residents in 2009 (FHWA 2012). Though driving is common in all age groups, it is a highly complex task that involves dynamic interleaving and execution of multiple critical subtasks. A driver license can be obtained by demonstrating certain knowledge and skills about maneuvering a vehicle, but it does not guarantee safe driving behavior. NHTSA identified drunk driving, distracted driving, drowsy driving, seatbelt use, speeding, drug driving as major risky driving behaviors which may compromise the safety of other drivers on the road (NHTSA). In 2014, 9,262 people were killed due to speeding among 32,675 total driving-related fatalities in the USA (Administration 2015). Roadway safety can also be compromised by driver limitations with primary age groups being young drivers under the age of 19 years and, older drivers with age 65 years or more. In 2015, 13 percent of all fatal traffic crashes involved older

drivers (age 65 years or more) (NHTSA 2016). Per 2014 population estimate, 14.5 percent of total population are aged 65 years or more in USA (United States Census Bureau 2016). National Highway Statistics 2010 noted that 16 percent of total licensed drivers are aged 65 years or older. Comparing fatalities with driver percentage in old age group, it can be noted that the old driver related crashes are not overrepresented. Due to a decreasing birth rate and longer life expectancy, the proportion of Americans over the age of 65 is increasing and expected to increase in coming years. There was 29 percent nationwide increase in old driver fatalities from 2006 to 2015 whereas the nationwide increase in total traffic fatalities was approximately 17 percent over the same time period. It is indeed essential to understand the effect of risky driving behavior and driver limitations to develop effective safety training and enforcement programs.

Human factors play an important role in crash events. Human Factors Guidelines (HFG) for Road Systems noted "Road users cannot be expected to solve either highway design or traffic engineering problems without making mistakes and/or compromising operational efficiency and safety" (Campbell 2012). A crash occurrence can be attributed to errors by drivers or the interaction between driver behavior and roadway design features (Hauer 1999). Per police records, driver errors can range from a traffic infraction in which the driver is not paying attention to an intentional traffic violation such as failure to yield or significantly exceeding the speed limit. The NMVCCS study classified driver related critical reasons into recognition errors, decision errors, performance errors, and nonperformance errors and assigned 94 percent of all study crashes to certain critical reason(s) (NHTSA 2008). Recognition error includes driver inattention, internal and external distraction, inadequate surveillance, etc. Recognition error was assigned as critical driver related crash reason in 40.6 percent of the crashes. Aggressive driving behavior, driving too fast, etc. are categorized as decision error, which is the reason in 34.1

percent of crashes. Overcompensation, poor directional controls are categorized as performance error. 10.3 percent of crashes occurred due to performance error. Sleep and physical impairment are considered as nonperformance error and were assigned to 7.1 percent of crashes. Driving maneuvers (tailgating, evasive maneuver, etc.) and driver condition (e.g. sleep, inattention, drunk, etc.) in pre-crash events have been analyzed in other literature to identify major contributing factors to crashes (Najm et al. 2002, Campbell et al. 2003). Inattention, tailgating, misjudged gap, excessive speed and a few other factors are found as common contributing factors in different crash types (Campbell et al. 2003).

2.2.3 Under Alcohol and Drugs Influence

Alcohol impairment of driving skill has been identified as a major traffic safety problem since early 20th century, and it continues to be a major highway safety issue (Blomberg et al. 2005). Use of alcohol can significantly affect a driver's decision-making process. Blomberg et al. conducted a case-control study to explore the relationship of Blood Alcohol Concentration (BAC) with relative crash risk (Blomberg et al. 2005). Results showed that elevated relative risk beginning at 0.05 – 0.06% BACs with an accelerating increase in risk at BACs greater than 0.10. Alcohol-impaired driving related fatalities are almost one-third of all fatalities that occur in the USA. In 2015, 10,265 people were killed in crashes involving an alcohol-impaired driver which is 29 percent of all fatalities occurred in 2015 (NHTSA 2016). Alcohol-impaired driving related fatalities increased by 3.2 percent from 2014. Alcohol-impaired driving is also one of the major crash contributing factors in Wisconsin. According to the 2015 Wisconsin Fatal Crash Trend analysis, 34 percent of total fatal crashes occurred in Wisconsin involved alcohol-impaired driving (Wisconsin Department of Transportation (WisDOT) 2017). To consider the influence

of alcohol-impaired driving in a crash prediction model, researchers used a few surrogate variables such as bar density, liquor licenses, population per liquor license per 100 sq. miles, adult/ juvenile liquor law violation arrest, etc. (Owusu-Ababio and Feng 2006, Morrison et al. 2016). Osuwu-Ababio et al. found liquor license related and liquor arrest related variables are significant in predicting alcohol-related crashes in Wisconsin (Owusu-Ababio and Feng 2006).

Besides alcohol impairment, drug-impaired driving has recently started raising government and public concerns in the USA as well as other countries (Walsh et al. 2008, Compton et al. 2009, Asbridge et al. 2012). The Governors Highway Safety Association (GHSA) sponsored a study to explore the effect of drugged driving in the United States (Hedlund 2017). The study found that fatality due to drugged driving surpasses alcohol-impaired fatalities. In 2015, 43 percent of motorists who died in the road accident had drugs in their system, whereas 37 percent of motorists who died were tested positive for alcohol. In the FARS annual report file, 57 percent of the fatally-injured drivers were tested for drugs, of them 55 percent were detected with no drugs, a particular drug was found in 34 percent, some other drugs in 7 percent, and test results were unknown for the rest 3 percent (Hedlund 2017). Figure 4 displays the trend in nationwide alcohol-impaired and police-reported drug-impaired fatalities using FARS data. Wisconsin suffered 149 fatal crashes related to drug-impaired driving which was 27 percent of total fatal crashes (Wisconsin Department of Transportation (WisDOT) 2017). Among these drug-impaired driving related crashes in Wisconsin, 46 percent of fatal crashes involved both drugs and alcohol. Although chemical blood test or a breathing test may be used to identify alcohol-impaired drivers, it is much harder to test for drugs.

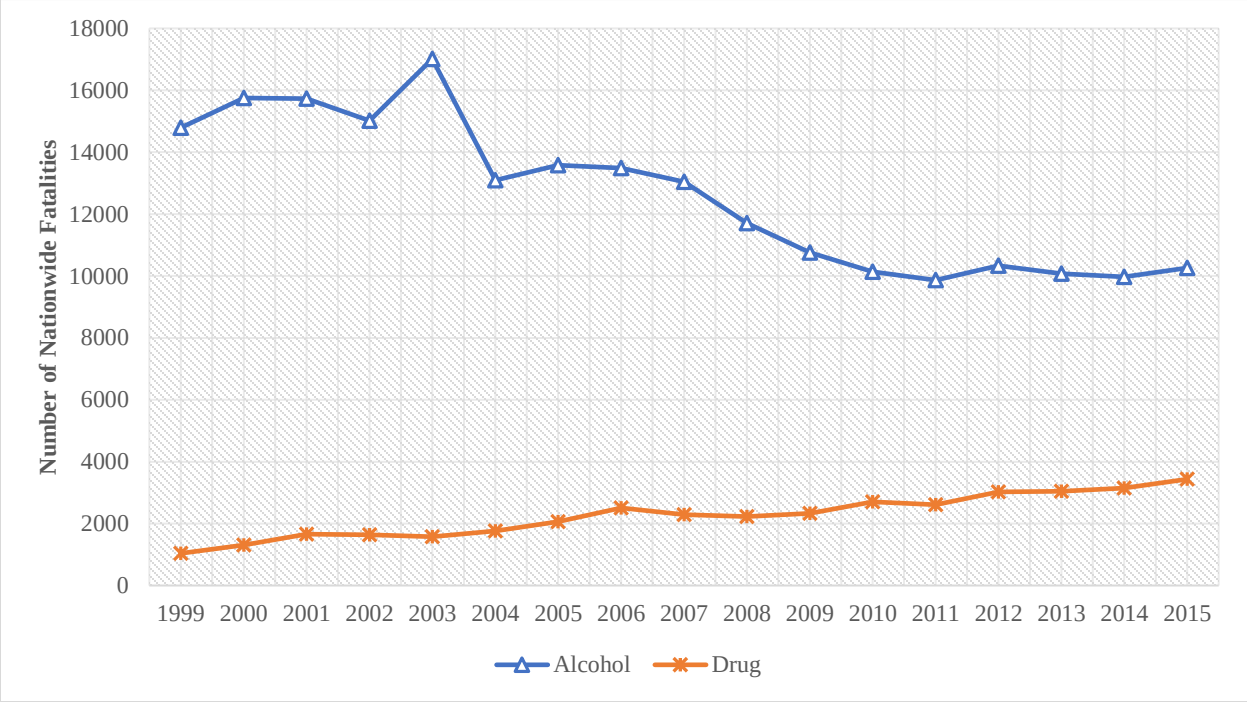

Figure 2-2 Trend in Alcohol/Drug-impaired Driving Fatalities (FARS 1999-2015).

While it is readily apparent that driving-related skills can be impaired by a wide variety of illegal substances and medications, the nature and scope of the drug-impaired driving problem have been difficult to detect and define (DuPont et al. 2012, Houwring 2013). Therefore, extensive studies are needed to explore more in this area. The recent legalization of marijuana use for medical and recreational purpose in some states has further exacerbated concern over potential risks of driving impaired by marijuana (Compton and Berning 2015).

2.2.4 Risky Driving Behaviors

Among different age cohorts, young drivers have a greater risk of crash involvement due to their risk-taking behaviors. Lack of experience and associated skills deficits, immaturity and youthfulness of young drivers can lead to intentional or unintentional risky driving behavior (Organization for Economic Co-Operation and Development 2006). Like a crash event, risky

driving behavior can be caused by a series of factors. There is strong evidence shows that the on-road risk-taking behavior does not occur in isolation of other risky lifestyle practices by young people such as alcohol or illicit substances etc. (Bingham and Shope 2004, Smart et al. 2005). For example, 41 percent of drivers killed in roadway crashes due to speeding had BAC of 0.08 G/DL or higher in their blood in the USA (NHTSA 2016). This clearly implies alcohol consumption contributes to speeding behavior. Several studies investigated the psychological behavior of young people in a driving task. Researchers noted that the developmental stage of young adulthood is characterized by immense biological and psychological contributors and can result in experimenting with a range of behavior that may result in adverse health outcomes (Steinberg and Morris 2001, Dahl 2008).

As risky driving behaviors is a broad term, few researchers investigated safety aspects of aggressive driving as a subset of risky driving. Paleti et al. noted, "A driver is characterized as acting aggressively if s/he participates in one or more of the following: speeding, tailgating, changing lanes frequently, flashing lights, obstructing the path of others, making obscene gestures, ignoring traffic control devices, accelerating rapidly from stop, and stopping suddenly" (Paleti et al. 2010). The American Automobile Association (AAA) Foundation for Traffic Safety estimated that 56% of the fatal crashes that occurred between 2003 and 2007 involved potential aggressive driving behavior, with speeding being the most common potentially aggressive action making up about 31% of total fatal crashes. Figure 4 shows the percentage of fatal crashes involving possibly-aggressive driver-level contributing factors using Fatality Accident Reporting System (FARS) 2003-2007 database. Paleti et al. found that young drivers between the age of 16 and 17 not wearing a seatbelt, under the influence of alcohol, not having a valid license, and driving a pick-up truck were found to be most likely to behave aggressively (Paleti et al. 2010).

Situational, vehicular, and roadway factors such as young drivers traveling with young passengers, young drivers driving an SUV or a pick-up truck, driving during the morning rush hour, and driving on roads with high speed limits are also found to trigger aggressive driving behavior. There is a significant gender difference in driving behavior among young drivers. Swedler found that male drivers in fatal crashes are more likely to involve BACs of 0.08 percent or more, speeding, reckless driving, night driving and felony crashes compared to female drivers (Swedler et al. 2012). Conversely, female drivers are more likely to be involved in right-angle fatal crashes.

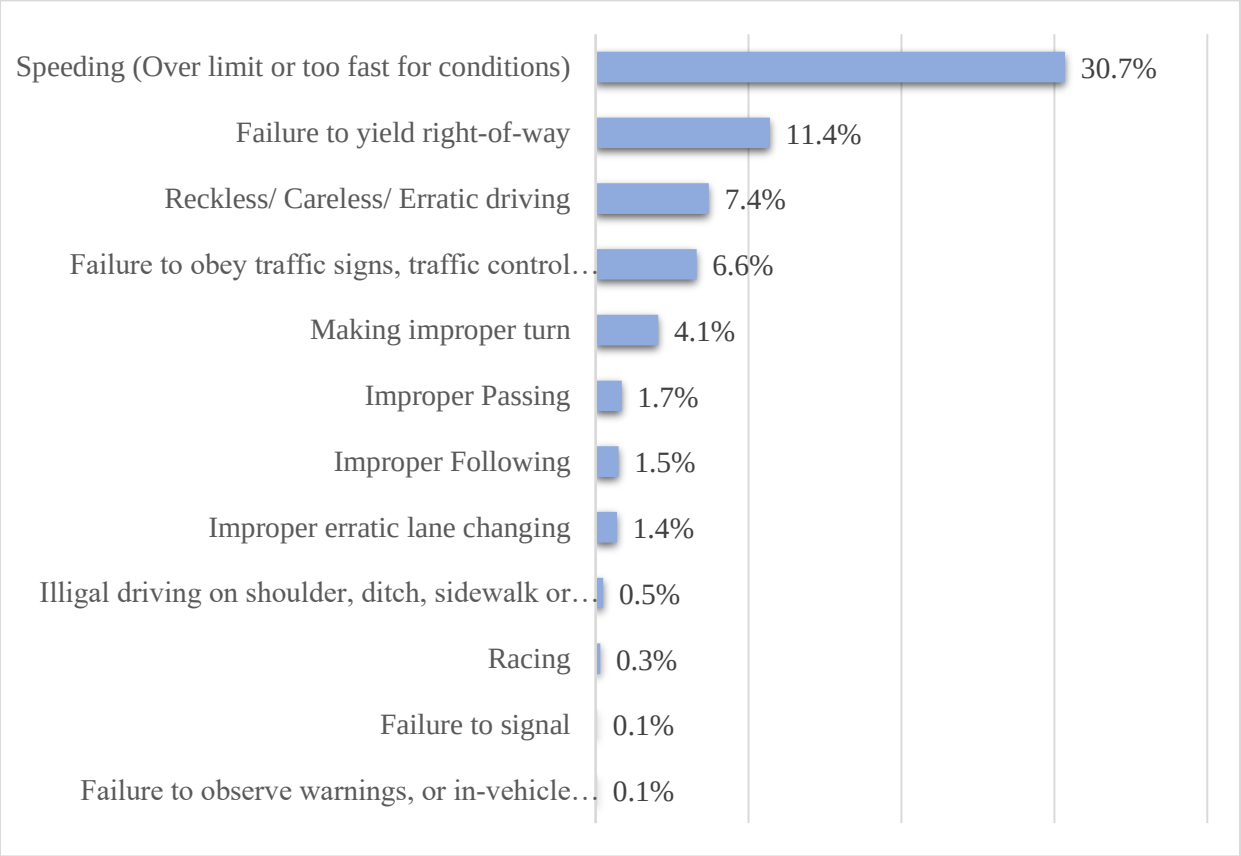

Figure 2-3 Potential Aggressive Behavior Percentage in Fatal Crashes Between 2003-2007.

Another risky driving behavior that contributes to crashes is distracted driving. Distracted driving has been defined as the "diversion of attention away from activities critical for safe driving toward a competing activity" (Redelmeier and Tibshirani 1997). For example, if a driver is traveling 55 mph and sends or reads a text that takes his or her eyes off the road for 5 seconds, it is equivalent of driving the entire length of a football field with his or her eyes closed. NHTSA statistics show that 3,477 people died due to distracted driving in 2015 on US roadways. In Wisconsin, crash fatalities saw a 46.4 percent increase in distracted driving crashes in 2015 from the prior year (Wisconsin Department of Transportation (WisDOT) 2017). Novice drivers appeared to be prone to distraction while driving (NHTSA 2010). Naturalistic driving studies showed that talking on a cell phone raises the risk of collision by more than 30% and drivers who text are at 23 times higher crash risk compared to the non-distracted drivers (Box 2009). Though driver engagement with any other activities such as eating, texting, cell-phone use, talking with passengers are all related to inattentive driving, use of electronic device garnered the public and media interest. The National Occupant Protection Use Survey (NOPUS) conducted annually by the National Center for Statistics and Analysis of NHTSA provides the only nationwide probability-based observed data on driver electronic device use in the US. NOPUS 2014 results showed that females from all age groups are more prone to use electronic devices while driving (Pickrell and Liu 2016). Electronic device use percentage was found similar in age brackets from 16 to 69 years.

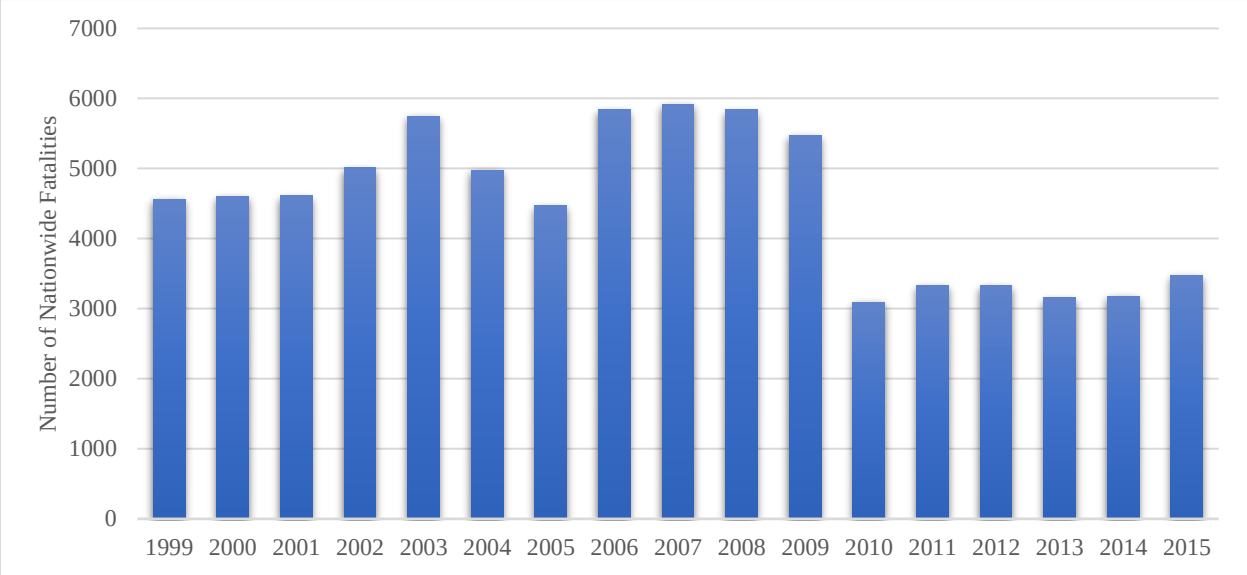

Figure 2-4 Trend in Nationwide Distracted Driving Fatalities from FARS 1999-2015.

2.2.5 Driver Limitations and Competency

Driving task requires processing information from inside and outside of the vehicle and timely execution of critical tasks. The continuous information processing and on-time execution process in driving a vehicle can be affected by driver limitation. It is known that physical strength, mental acuity, and overall health begins to deteriorate as a person ages. Mental acuity affects the visual effectiveness and information processing ability during driving task. Due to physical frailty, old drivers are more likely to be injured or die from that injury in a crash event. In Wisconsin, old drivers aged 65 years or older were involved in 18.5 percent of total fatal crashes which is higher than the national average of 13 percent (NHTSA 2017). Therefore, old people driving in an unsafe manner is a safety problem to be addressed.

To develop training programs or old-driver specific countermeasures, it is necessary to understand the kind of driving scenario or crash type over-represented by the old driver age

cohort. Reinfurt et al. analyzed FARS and GES crash data to identify crash types in which old drivers are more likely to be involved (Reinfurt et al. 2000). The authors found that with increasing age, drivers were more likely to be at-fault in left-turn crashes involving frontal and right-side impact, and when the traffic control was a stop or yield sign versus a traffic signal. Office of Behavioral Safety Research (OBSR) at NHTSA sponsored a study in 2009 to identify the behaviors and situations associated with increased crash risk for older drivers (Stutts et al. 2009). Data analysis showed that there is little evidence to suggest increased crash risk for drivers aged 60 to 69 years. The crash risk begins to demonstrate a substantial increase for drivers aged 70 to 79, with an over-representation for many crash types then increasing more sharply for drivers aged 80 years and older. A few situations such as left-turning movements, movements at stop-controlled intersections, high-speed 2-lane roadways and multi-lane roads with speed limits of 40 to 45 mi/h (e.g., suburban arterials) were associated with heightened older driver crash involvement. For fatal crashes, both "young-old" drivers with age between 60-69 years and "old-old" drivers with age 80 years or more were more likely to make errors at intersections controlled by flashing signals. Choi and Singh found that older drivers are more likely to get involved in crashes if glare from opposing headlamp at night obstructs their vision (Choi and Singh 2005).

2.2.6 Intersections between driver, roadway and environment

As noted earlier, traffic crashes occur due to interactions between roadway, traffic, environmental and driver factors. To control driver behavior up to a certain level, several traffic control elements such as traffic sign, signal, speed limit, etc. and traffic laws have been implemented in roadway systems. Drivers on a roadway section or at the intersection are

required to follow these rules to maximize efficiency and safety. Sometimes due to driver error such as recognition error, decision error, performance or non-performance error and driver limitation, drivers do not follow the rules.

Though categorization of driver error (e.g. recognition error, performance error, etc.) is straightforward, it does not provide necessary information to develop countermeasures. The interaction between driver error and roadway geometric and contextual features is still an interesting area for researchers to explore. Hauer noted that the speed at which people choose to travel is affected by roadway design and vehicle characteristics (Hauer 2009). Tate and Turner investigated the relationship between observed travel speed, road geometry and crashes in New Zealand (Tate and Turner 2007). Authors concluded that driver's speed choices were more strongly related to curve radius than curve design speed and that the approach speed environment also has a significant impact on the speed choice. Though many believe that the faster one drives the more likely one is to crash, Hauer and Tate et al. both noted that the relationship of travel speed with crash occurrence probability is inconclusive (Tate and Turner 2007, Hauer 2009). Hauer also noted that the probability of injury severity in a crash increases as a function of the change in travel speed (Hauer 2009). Kai and Qin investigated the factors contributing to driver error at uncontrolled, sign-controlled and signal-controlled intersections (Wang and Qin 2015). The authors used roadway characteristics (e.g. presence of curve, visibility, speed limit, etc.), driver characteristics (age, gender, DUI), environmental characteristics (weather condition, roadway condition, lighting condition) and vehicle type (passenger car, light truck, heavy truck) to predict driver errors collected from crash reports. They found that sign-controlled intersections have the highest percentage of driver error and reckless driving followed by signalized and stop-control. Drivers are also more prone to severe driver error if a driver's

vision is obscured. Adverse environmental characteristics such as snow, ice on roadway negatively affect driver error resulting in less severe driver error. Driver age, gender, and alcohol or drug use greatly influence the severity of error outcome. This information indicates that driver errors are not only the outcome of a driver's psychological behavior but also the interaction with other external factors during the driver's decision-making. More information on these contributing factors can help researchers and safety professionals to develop cost-effective countermeasures that might help to mitigate driver errors.

Roadway geometric features and traffic variables are mostly used as explanatory variables in crash prediction modeling. AASHTO published "A Policy on Geometric Design of Highways and Streets" or the Green Book to guide roadway design engineers to provide nominal safety (AASHTO 2001). This guide provides a minimum requirement for each roadway design element. The roadway network is grouped into different functional categories in the Green Book based on its function to accommodate a movement type of a trip. Though functional classes vary between rural and urban roads, the hierarchy of functional class usually consists of principle arterials (for main movement), minor arterials (distributors), collectors and local streets. Functional class of a roadway is important, as it is the first step of a design process. Following this guidebook can provide nominal safety by meeting design requirement for each roadway element by functional class but cannot prevent crashes from happening. One of the reasons could be the nominal design requirement that assumes a set of values for all road users. As a complement, the Human Factors Guidelines (HFG) for Roadway Systems recommends that highway designer and traffic engineers must consider the human factor characteristics of road user in conjunction with other design components. For example, drivers can experience difficulty at intersections in estimating gap size and speed of approaching vehicles, drivers can

experience problems in detecting a sharper curve after negotiating several longer radius curves (Glennon 1996, Staplin et al. 2001). Thus, it is safe to note that geometric standard below nominal level and/ or without considering human factor can stimulate driver error, hence increase crash risk.

Although the Green Book recommends safe and efficient practices for the design of roadways based on extensive research and study, it is difficult or almost impossible for the designer to characterize quantitatively how the facility will perform. For both new construction and reconstruction of roadway, decision makers want reasonable measures of the effect of geometric design decisions on the facility's performance for all users. In the NCHRP Report 785, Ray et al. developed a process framework to include both an approach for integrating performance-based analysis into geometric design decisions and information on the effects that different geometric elements have on project performance measures (Ray et al. 2014). Several tools and resources such as Interactive Highway Safety Design Model (IHSDM), SafetyAnalyst, and the HSM are now available to evaluate performance-based safety (Lum and Reagan 1995, AASHTO 2010, Harwood et al. 2010).

The HSM developed predictive methodologies by incorporating roadway geometric features and traffic variables (AASHTO 2010). For example, lane width, shoulder width and type, horizontal curve, superelevation, vertical grades, driveway density, centerline rumble strip, passing lane, two-way left-turn lane, roadside design, lighting and automated speed enforcement were recommended as variables in base conditions for a rural two-lane two-way segment. The value of crash modification factors (CMF) for each geometric element quantifies the changes in expected crash occurrence by changing it from the base condition. Griebe found that more than 72 percent of roadway link crash variability can be explained using appropriate explanatory

variables. Around 40 percent of the variability was explained by roadway geometric features and land use variables (Greibe 2003). For intersections, the HSM used intersection skew angle, number of approaches with left-turn and right-turn movement and lighting as geometric features in base conditions for rural two-lane two-way intersection facility. Bauer and Harwood noted that geometric variables can explain 5 to 14 percent of the variability in crash data at at-grade intersections (Bauer and Harwood 1996).

In a performance-based safety analysis, the statistical regression model assumes that the explanatory variables are independent, which is not always true. Certain variables may have interactions or combined effects as all roadway design elements can influence drivers' perception and therefore influence their driving behavior. Under such circumstances, the estimated effect of geometric variables can incorporate bias in the result. Wu and Lord investigated the dependence between variables on estimating CMFs (Wu and Lord 2016). The authors concluded that the regression model can produce biased CMFs if the variables are not independent. Presence of curve (horizontal and vertical) has a significant effect on crash occurrence. Previously, the safety effect of horizontal and vertical curves is separately quantified (Zegeer et al. 1992, AASHTO 2010). Bauer and Harwood noted that considering combined effect of horizontal and vertical curve in crash modeling can result in proportional relationship of crash with traffic variables (Bauer and Harwood 2013). Shaon and Qin discussed substantive safety benefits of lane width and shoulder width combinations underscoring increase in lane width or shoulder width may not always add safety benefits based on crash data analysis (Rahman Shaon and Qin 2016). The interaction or combined effect of roadway geometric features are also important in roadway improvement projects. In the NCHRP Synthesis 417 report, authors noted "state need to know what safety benefit can be derived from an improvement in any geometric element alone

and in combination with others. With "perfect" information on these relationships, states will be able to select safety improvements that will yield the largest safety return for the available funds for any specific 3R project or for their entire annual 3R program" (McGee 2011). In an active project entitled "NCHRP 15-50: Guidelines for Integrating Safety and Cost-Effectiveness into Resurfacing, Restoration, and Rehabilitation Projects" (Harwood 2017), the objective is to develop guidelines for safe and cost-effective practices for 3R projects based on current knowledge of geometric design elements, their impacts on safety and operations, and the trade-offs between costs and benefits. As current roadway design is governed by the Green Book or state statutes, roadway attributes within a specific functional class can be correlated with each other, which requires special attention in crash prediction models.

Traffic variables are usually treated as an exposure variable in crash modeling. Traffic variables such as AADT, truck percentage, turning movement are commonly used to model the likelihood of crash occurrence. Bauer and Harwood noted that traffic variables account for 16-38 percent variability in crash occurrence at the at-grade intersection (Bauer and Harwood 1996). Griebe noted 30 percent of the variability in roadway link crashes can be explained by traffic exposure (Greibe 2003). A major portion of literature focused on developing a defensible statistical relationship of exposure variable with crash count (Carroll 1971, Chapman 1973, Hauer 1982, 1995, Stewart 1998, Qin et al. 2004). A common way to define the safety of a system is the product of the probability of having a crash or crash risk given a unit of exposure and the observed level of exposure. Since the number of crashes is the only self-evident quantity in the equation, the resulting crash risk per unit exposure is determined by the selection of exposure measures and vice versa.

Hauer adopts the definition of "a unit of exposure" and call it a trial (Hauer 1982). The result of such a trial is the occurrence or non-occurrence of a crash (by type, severity). However, this exposure measure is oriented to the entity (driver or vehicle) involved, e.g. one truck trip or one pedestrian crossing. If it is applied in a site-specified situation, such as a road segment or intersection, the definition is still obscure, and the corresponding exposure is difficult to measure. Chapman describes the exposure as the number of opportunities for accidents of a certain type in a given time in a given area (Chapman 1973). These opportunities are occasions when cars cross each other's path when they are following each other, or even when a vehicle is traveling by itself on a winding road. This definition considers that the exposure should include characteristics of drivers and vehicles, characteristics of the roadway, and the environmental condition. Stewart defined exposure as a statistical measure providing information on the extent of a road user's exposure to the overall level of travel risk given the road conditions at any point in time (Stewart 1998). The author recommends "kilometers of travel" as a "meaningful, practical and applicable" measure of exposure.

It was noted in the safety literature that the relationship between crash and exposure variable is not linear (Qin et al. 2004). AADT of roadway link and vehicle miles traveled has been extensively used as exposure measure to predict crashes in the crash modeling literature (Stewart 1998, Wellner and Qin 2011). Qin et al. noted that the relationship between crashes and traffic variables changes from location to location as well as with crash types (Qin et al. 2004). Mannering explained that a multifaceted crash event involves an interaction between geometric, traffic and environmental variables as well as human response to external stimuli which creates heterogeneity in the crash dataset. The data heterogeneity may cause the effect of traffic variable/ exposure to change from location to location. Disaggregated level of traffic variables

are used in real-time or event-based crash prediction model. Traffic count, speed, deviation in traffic count and speed from upstream to downstream locations are used as predictor variables in crash event modeling (Abdel-Aty and Pemmanaboina 2006, Yu et al. 2013, Chen et al. 2017).

Considering the importance of roadway and corresponding traffic data in safety performance evaluation, the Fixing America's Surface Transportation Act (FAST) signed into law on December 4, 2015legislation identifies the need for improved and more robust safety data for better safety analysis to support the development of State's Strategic Highway Safety Plans (SHSPs) and their Highway Safety Improvement Programs (HSIPs) to achieve a significant reduction in fatalities and serious injuries on all public roads. More detailed roadway data are also needed by State Departments of Transportation (DOT) and local agencies as they implement their strategic highway safety plans and make safety assessments of various roadway treatments. Roadway Safety Data Program of FHWA published Model Inventory of Roadway Elements (MIRE) as a recommended listing of roadway inventory and traffic elements critical to safety management (Lefler et al. 2010). There is a total of 202 data elements that comprise MIRE Version 1.0. MIRE is intended as a guideline to help transportation agencies improve their roadway and traffic data inventories. Collecting additional data items as recommended in MIRE is a need to obtain more robust safety data and to use data for more accurate safety prediction. Weather conditions resulting in changes in roadway surface conditions have long been known to be contributing factors to the frequencies and severities of roadway crashes. According to WisDOT, 46 percent of injury and fatal crashes occurred during adverse weather condition such as rain, snow, sleet, fog or cloudy between 2014 to 2015 (Wisconsin Traffic Operations and Safety (TOPS) Laboratory 2017). FHWA noted that about 22% of PDO crashes, 19% of injury crashes and 16% of fatal crashes occur on US roadways due to adverse weather based on 10

years of crash data (Federal Highway Administration (FHWA) 2013). The most important factors influencing crashes in a weather event are those that affect the available friction between vehicle wheel and pavement, and/or driver visibility, resulting in crashes when the driver is unable to avoid collisions with moving or fixed obstacles. Without specifying the source of adverse weather event (e.g. rain, snow, fog, etc.), the potential effect of adverse weather conditions is described in Table 2-2.

Table 2-2 Potential Impact of Adverse Weather Condition.

Roadway	Vehicle	Driver
• Reduced capacity. • Reduced speed. • Effect on Level of Service. • Reduced pavement friction (longitudinal and side friction). • Lane submersion due to high water level. • Higher required time to achieve full potential of deicing chemicals. • Little melting effect from deicing chemicals below 10 degree Fahrenheit. • Corrosive effect from deicing chemical. • Adverse effect of deicing chemical on natural environment.	• Reduced vehicle control and stability. • Hydroplaning • Reduced tire traction. • Reduced rolling resistance. • Tire inflation. • Wheel spinning. • Weaken brake system. • Reduced fuel economy. • Exposure to rust and corrosion.	• Higher stopping distance. • Higher perceived risk. • Reduced visibility. • Higher car-following distance. • Higher degree of heterogeneity in driver behavior. • Heat stroke, exhaustion and cramp.

There is a significant amount of literature available discussing the relationship between a crash event and weather conditions (Shankar et al. 1995, Andrey et al. 2003, Aguero-Valverde and Jovanis 2007, Savolainen and Mannering 2007, Jung et al. 2010, Morgan and Mannering 2011, El-Basyouny et al. 2014a, El-Basyouny et al. 2014b). Andrey et al. found that almost all research has shown an increase in the frequency and severity of the crashes during adverse

surface conditions, but that the magnitude of the increase has varied widely across studies (Andrey et al. 2003). This variation could be a result of different drivers' recognition, and reactions to perceived deteriorations in roadway condition and/or the statistical model used to model crash data. For example, Andrey et al. found that precipitation increases crash risk by 70 percent (Andrey et al. 2001), Knapp noted that severe winter storm increases crash rate by approximately 1000% (Knapp 2000).

The effect of adverse weather condition can be categorized into three groups: 1) effect of roadway, 2) effect of vehicle and 3) effect on driver behavior. The Federal Motor Carrier Safety Administration (FMCSA) sponsored a study to investigate the potential impact of adverse weather and climate conditions on commercial motor vehicle operation and safety (Rossetti and Johnsen 2011). In this study, authors listed potential effects of adverse weather condition on roadway performance and influence on vehicle and driver behavior. Adverse weather actively affects the performance of a roadway by reducing capacity and speed, hence increasing travel time and delay (Hoogendoorn et al. 2010, Snelder and Calvert 2016). Several studies found that increasing water depth can affect pavement skid number and increase hydroplaning potential (Rose and Gallaway 1977, Ong and Fwa 2007). Ong and Fwa noted that skid resistance and hydroplaning potential also depend on vehicle speed, tire load, tire inflation pressure, type of tire and pavement texture along with water film thickness (Ong and Fwa 2007). During snow, deicing chemicals are usually used to reduce melting temperature of ice so that vehicle tire can reach the pavement to get traction with pavement. MnDOT conducted a study to evaluate the field effect of deicing and anti-icing chemicals (Druschel 2014). This study found that there is a little melting effect from deicing chemicals if the temperature is below 10 degree Fahrenheit. Ice melting by deicer is a time-sensitive process. Lower temperature can also affect the duration for

deicer to achieve full potential. Adverse weather conditions can also affect vehicle condition by reducing vehicle control and stability, tire inflation, wheel spinning and reduced fuel economy (Rossetti and Johnsen 2011). An unintended negative consequence posed by deicing salt is vehicle rust resulting in corrosion. Due to the construction of a vehicle with most of the underbody being wide open, most salt damage occurs underneath the car (Rodman 2016). Thus, it can be difficult to detect visually. Adverse weather can significantly influence driver behavior too. Multiple research studies noted that adverse weather such as fog, snowfall can significantly reduce driver visibility (Broughton et al. 2007, Hoogendoorn et al. 2010). Pisano et al. noted that many drivers do not realize that pavement friction is significantly reduced under these conditions, leading to greater stopping distances (Pisano et al. 2008). Hjelkrem and Ryeng investigated how precipitation, light conditions and surface conditions affect the drivers' risk perception using crash risk index (CRI) (Hjelkrem and Ryeng 2016). The authors found that both car and truck drivers perceive the highest risk when driving on snow covered roads. They also noted that adverse weather causes a higher degree of heterogeneity in driver behavior due to driver's perceived risk.

Weather elements such as rainfall, snowfall, wind speed, temperature, and visibility distance are used to access the effect of weather elements on crash occurrence in literature (Khattak and Knapp 2001, Eisenberg and Warner 2005, Abdel-Aty and Pemmanaboina 2006, Hermans et al. 2006, Andrey 2010, Usman et al. 2012, Bergel-Hayat et al. 2013, Yu et al. 2013). Researchers developed several ways to incorporate weather influence in the crash prediction model. Daily or hourly averaged weather variables are commonly used to predict crash occurrence and/ or severity in literature (Chang and Chen 2005, Malyshkina et al. 2009, Yaacob et al. 2010, El-Basyouny et al. 2014a, El-Basyouny et al. 2014b). Caliendo et al. used hourly

rainfall data and transformed it into a binary indicator for describing whether daily pavement status is dry or wet (Caliendo et al. 2007). Researchers found that the amount of precipitation has a positive effect on crash occurrence and higher precipitation has greater tendency to have higher crash rates (Chang and Chen 2005, Malyshkina et al. 2009, Yaacob et al. 2010). Eisenberg noted that lagged effects are important during precipitation events. Lagged effect for a weather event means the effect of weather events (rain, snow) on crash risk decreases with time if the weather event lasts for a longer period (e.g. crash risk on day 2 will be lower than day 1) (Eisenberg 2004). This lagged effect occurs as drivers gather information which allows them to adjust to the change in surface condition. Martchouk et al. also mentioned drivers' speed and headway change substantially during inclement weather as drivers seek to compensate for adverse conditions and maintain an acceptable level of safety (Martchouk et al. 2010).

Like an adverse weather event, weather elements on a bright sunny day may have influence on roadway safety. Driving on a bright sunny day can be discomforting to certain directional drivers due to sun glare. Glare can also occur from the headlamp of vehicles from the opposite direction at night. Glare effect usually reduces the visibility of a driver. Choi and Singh from FHWA found that glare from both sun and headlamp is a contributing factor in crashes (Choi and Singh 2005). The authors also noted that glare-related crashes showed a particular pattern with driver, vehicle and roadway related factors. Mitra compared observed and expected crash counts during glare and non-glare periods at signalized intersections of Tucson, Arizona (Mitra 2014). The author noted "both statistically and through reasoned logic, that sun glare (in general) affects intersection safety". Babizhayev specifically investigated the effect of headlamp glare to examine the type of visual impairment interceded by the increased glare sensitivity in adult drivers (Babizhayev 2003). Results showed that glare affects more adversely

to the older drivers. Though glare effect is a commonly observed fact now, it is less discussed in literature. One of the major reasons for this shortage of work on the adverse effect of glare from the sun is the fact that sun glare is temporal in nature. It is also very difficult for weather stations to capture and record sun glare, unlike other weather events.

2.3 Modeling Techniques

Crash modeling is an effective approach to exploring the relationship between crash frequency or crash severity and a set of predictors from the statistical perspective. Once the relationship is established, the mean crash count or the probability of an injury type can be estimated. It is anticipated that the explanatory variables are not only statistically correlated but are logically related to crash occurrence. Such a regression method assumes the error as random noise, and the mean can be represented as the true value around which observations fluctuate.

Technically sound crash models should have the theoretical rigor and technical robustness to handle many types of data issues generated during data collection and reporting. As crashes are rare and random events, it is quite normal that individual locations do not have adequate data for drawing a valid and explicit conclusion. Crash data are often pooled from a wide range of geographic locations and at different times to enhance the analysis. Data collected at the same time and location may exhibit similarities, whereas data collected at different times and from different locations may exhibit markedly different characteristics and therefore be heterogeneous. Heterogeneity means the variance of the dependent variable changes from observation to observation, and may change as the independent variable changes (added or removed). The accuracy of the coefficient estimates will be compromised, and the statistics used

to test the hypothesis under the Gauss-Markov assumption will not be valid if heterogeneity is not carefully considered.

Advancement of methodologies for modeling crashes has been propelled by emerging problems with safety data; however, the field is just now experiencing rapid evolution. Development was tardy in the early fifties, and it was not until the seventies that researchers started using the Poisson distribution (which is a non-negative integer in nature) to model crash count. In the early eighties, the introduction of negative binomial or Poisson-gamma distribution to address crash data overdispersion marked a developmental monument. Since then, explosive growth in safety research has been witnessed, largely due to the substantial investment on standardization, modernization, and availability of safety data, as well as the advancement in computing power (Miaou and Lord 2003). The development in statistical methods for safety data have significantly improved modeling accuracy while overcoming data limitations. Representative work includes:

- Generalizing parametric count model through hurdle models such as zero-inflated Poisson and zero-inflated negative binomial (Miaou 1994, Shankar et al. 1997a, Son et al. 2011),
- Describing heterogeneous crash count data via mixture regression models (Park and Lord 2009, Heydari et al. 2016, Rahman Shaon and Qin 2016, Rahman Shaon et al. 2018b),
- Adopting a well-specified mean function (Mitra and Washington 2007b) and a properly constructed function for shape parameters (e.g. dispersion parameters) (Geedipally et al. 2009),
- Addressing the crash data heterogeneity by specifying a random parameter model (Anastasopoulos and Mannering 2009, Mitra and Washington 2012),

- Using quantile regression to analyzes the relationship between conditional quantiles of crash count and a set of explanatory variables (Qin et al. 2010).

Concerning data heterogeneity, these methodological innovations have significantly improved the unbiased estimation of standard errors for the coefficients as well as their statistical inferences. This sub-section is divided into two parts to provide an overview on methodological alternatives explored in literature in crash frequency modeling and event-based/ injury severity modeling, organized by major data challenges and issues they are designed to handle.

2.3.1 Crash Frequency Modeling

Crash frequency modeling focuses on establishing a quantitative relationship between crash count and contributing factors based on the statistical significance unveiled from the data. Decades of modeling crash data reveal that crash count data have a variety of issues, including over-dispersion, under-dispersion, time-varying explanatory variables, temporal and spatial correlation, low sample-mean and small sample size, injury-severity and crash-type correlation, under-reporting, endogenous variables, unobserved heterogeneity, functional form and fixed parameters (Lord and Mannering 2010, Mannering et al. 2016). Extensive research has been devoted to modeling and analyzing this type of crash dataset (Lord and Mannering 2010, Mannering and Bhat 2014, Mannering et al. 2016). Failure to account for any of these issues would lead to biased coefficient estimates and inaccurate conclusions regarding the properties of the data population.

One of the most notable accomplishments in developing noble methodologies is the application of Negative Binomial (NB) models in crash frequency data. NB models can handle

data over-dispersion by assuming a gamma distribution for the exponential function of the disturbance term in the Poisson mean. However, recent studies have pointed out that biased parameter estimates in the NB model can be found in dataset with a long tail (Zou et al. 2015, Shirazi et al. 2016). A long tail is a statistical phenomenon that occurs when sample observations have a few very high crash counts or have preponderant zero observations which shift the overall sample mean to near zero (Shirazi et al. 2016). For example, Guo and Trivedi (2002) noted that a negligible probability is usually assigned to higher crash counts in the NB model when modeling highly over-dispersed data with a long tail. Lord et al. (2005) pointed out that over-dispersion arises from the actual nature of the crash process. One limitation of the NB distribution is it assumes that only one underlying process affects the likelihood of crash frequency (Shankar et al. 1997b).

Mixture models are a very popular statistical modeling technique used to account for data over-dispersion because of its flexible and extensible model class (Shankar et al. 1997b, Aguero-Valverde and Jovanis 2008, Lord et al. 2008, Lord and Geedipally 2011, Geedipally et al. 2012, Cheng et al. 2013, Mannering and Bhat 2014, Rahman Shaon and Qin 2016, Shirazi et al. 2016). The mixture model is comprised of a convex combination of a finite number of different distributions. The advantage of the mixed model is that it provides flexibility by adding a mixed distribution to account for extra variance in the crash data which is caused by preponderant zero crash responses and a long tail. The NB-L GLM was recently introduced to model crash frequency data (Geedipally et al. 2012, Rahman Shaon and Qin 2016). As the name suggests, the NB-L distribution is a mixture of NB and Lindley distributions in which the Lindley distribution itself is a mixture of two gamma distributions (Lindley 1958). This count data mixture model works well when the dataset contains a large number of zero responses, when the dataset is

skewed, or when it is highly dispersed. Zamani and Ismail showed that the NB-L distribution provides a better fit compared to the Poisson and NB models when the probability of crash frequency at zero is large (Zamani and Ismail 2010). Lord and Geedipally (2011) applied the NB-L distribution to estimate the predicted probability and frequency of crashes using both simulated and observed data. The authors concluded that the NB-L distribution can handle crash datasets with preponderant zero crash observations. Recently, Rahman Shaon and Qin (2016) evaluated the effect of lane width and shoulder width with over-dispersed crash data using the NB-L model. The authors found that the NB-L GLM performed better than a traditional NB model with crash data characterized by preponderant zero responses while maintaining the core strength of an NB model. According to the model application with many different data sources, overwhelmingly positive results have been reported (Zamani and Ismail 2010, Lord and Geedipally 2011, Geedipally et al. 2012, Rahman Shaon and Qin 2016) Hallmark et al. (2013), Xu and Sun (2015)). Although the Lindley distribution has a closed form (Zamani and Ismail 2010), this distribution is not available in any standard statistical software (e.g. R, SAS, SPSS) to mix with the NB distribution in the context of GLM . Thus, a hierarchical structure is needed to estimate parameters of NB-L in the context of GLM. In previous work, researchers used the Bayesian interface to implement this model (Geedipally et al. 2012, Rahman Shaon and Qin 2016).

Among the myriad of potential variables which can significantly affect the likelihood of crash occurrence, the existing crash dataset contains only a fraction of them (Mannering et al. 2016). In a regression model, unobserved heterogeneity occurs if important covariates have been omitted during data collection, meaning their influence is not accounted for in the analysis. Unobserved heterogeneity is usually considered as random errors in traditional NB models

because the effect of each covariate is restricted to being the same across all observations, which in turn causes extra dispersion problems. Such modeling strategies can cause serious model specification problems and may result in variation of the estimated effect of observed covariates (Mannering et al. 2016). An overview of the potential for heterogeneity in driver behavior due to a variety of highway factors was highlighted by Mannering et al. (2016). The research found that varying lane and shoulder widths may have an impact on the likelihood of a crash event, but these effects can vary among observations due to time-varying traffic, weather conditions, and the driver's reaction, all of which are not available for model development. Ignoring heterogeneous effects in explanatory variables leads to biased parameter estimates and therefore, inaccurate conclusions (Mannering et al. 2016).

Much research has focused on obtaining unavailable but necessary information by utilizing statistical and econometric models[2] to account for unobserved heterogeneity. RP modeling is one approach documented in the literature (Mannering et al. 2016). The use of RP models in crash count data modeling has gained considerable attention. In RP modeling, data heterogeneity is addressed by allowing model parameters to vary from observation to observation. The parameter is treated as a random variable whose probability distribution usually is defined by modelers. Anastasopoulos and Mannering introduced the RP NB model to account for data heterogeneity from explanatory variables and other unobserved factors (Anastasopoulos and Mannering 2009). Other crash data studies in which the RP NB model has been applied have found a significant improvement in the statistical model fit (El-Basyouny and Sayed 2009,

[2] Refer to Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research 11, 1-16. for the list of methodological alternatives to account for unobserved heterogeneity.

Garnowski and Manner 2011, Venkataraman et al. 2011, Chen and Tarko 2014, Buddhavarapu et al. 2016).

Apart from data issue, parameter estimates depend on the specification of the link function used for model development (Mitra and Washington 2007a). Hauer noted that the choice of appropriate link function or model equation to be used for crash modeling is not transparent, is seldom discussed, and is not well documented. He also emphasized that the choice of regression equation is based on Occam's Razor, convention, habit, limitations of available software, and t-tests (Hauer 2010). The skepticism associated with the assumed linear-in parameter relationship between crash frequency and explanatory variables prompted the use of neural networks (NN) (Xie et al. 2007, Kononov et al. 2011) and support vector machines (Li et al. 2008) data mining techniques for crash modeling. SVM and NN have strong non-linear approximation capability and do not require any specific distributions or functional forms. These modeling explorations suggest that there may be a nonlinear- effect exists in some explanatory variables.

Recognizing the need for accounting strong non-linear-in-parameter relationship in model specification, some researchers have investigated the generalized additive models (GAM). The GAM can handle parametric, semi-parametric and non-parametric specification of covariate(s) in its additive terms. These additive terms can help to measure non-linear effect of covariate(s) simultaneously and keep the distributional flexibility of GLMs. GAMs are an extension of GLMs in which the response variable is not restricted to being a linear function of the covariates; it can also be the sum of smoothing functions applied to these covariates. In GAMs, non-parametric smoothing spline functions for explanatory variables can take many linear or non-linear forms. Zhang et al. utilized GAMs to explore the potential nonlinear

relationship between crash frequency and exposure on different types of urban roadway segments (Zhang et al. 2012). After examining crash rate, the authors found GAMs has a higher predictive accuracy than NBGLMs. In a crash frequency analysis study, Xie and Zhang found that GAMs with a negative binomial assumption have better goodness of fit than traditional GLMs (Xie and Zhang 2008). Ma and Yan investigated the GAM with logit link function to examine non-linear effect of driver's age on the odds of being fault in rear-end crashes (Ma and Yan 2014). Authors used the cubic regression spline instead of a step function to describe the effect of driver's age. A step function considers fixed rate of change in odds whereas the smooth function results showed that the rate varies across different age groups.

Previous research established that various factors leading to the number of traffic crashes at a site do not necessarily contribute equally; therefore, traffic crashes at every site should be considered as the results of various risk sources, with each risk source playing either a vital or supporting role. However, conventional crash frequency models treat the total number of crashes at a roadway site as the outcomes of a single risk source by using a predictive equation estimated with Poisson or Negative Binomial (NB) distribution. Furthermore, single-equation models are incapable of assuming that crashes may have various risk sources, which could result in data heterogeneity. Not until recently have researchers acknowledged the limitation of assuming a single risk source in crash modeling; and accordingly, the multiple risk source regression model has been developed to distinguish the distinct sources of crash contributing factors (Washington and Haque 2013, Afghari et al. 2016). Multiple risk source regression modeling is a reasonable alternative to single equation predictive models for predicting risk-level crashes, considering that the contribution of explanatory variables to the outcome (i.e. predicted crash count at a site) may change (Washington and Haque 2013, Afghari et al. 2016, Afghari et al. 2018). Furthermore, the

risk-level predicted crashes from multiple risk source modeling could be useful in identifying sites for safety improvements and developing targeted and effective safety countermeasures.

2.3.2 Event Data Modeling

Equally if not more important is the task of identifying the contributing factors and their impacts on crash outcomes such as injury severities. The ambitious goal of "Vision Zero" as well as tightening resources has made this an increasingly important task. The methodologies and techniques for crash outcome modeling, like its crash count modeling counterpart, are diverse. But unlike crash count which is a non-negative integer that can change from zero to a large figure, the crash outcome such as injury severity has a finite number of alternatives (e.g., a KABCO scale). In econometrics, discrete choice models describe, explain, and predict choices between two or more discrete alternatives. Moving from simple to complex, from weak to robust, the methodological evolvement of crash event outcome modeling benefits tremendously from the development of econometrics and from travel demand models where highway route choice and transportation model choice are typical applications for a discrete choice model. Several prevailing issues were found during model development, including under-reporting, fixed parameters, omitted-variables bias, small sample size, endogenous variables, temporal and spatial correlation, unobserved heterogeneity, ordinal nature of crash injury severity data, and within-crash correlation (Savolainen et al. 2011, Mannering et al. 2016). The major two issues are specific to crash severity data, and are discussed as follows:

- The ordinal scale is that the injury severity levels (i.e., fatal injury or killed, incapacitating injury, non-incapacitating injury, possible injury, and property damage

only) are ordered from the highest level to the lowest. There may be correlation among severity levels, and the correlation should be stronger between two closer levels.

- Within-crash correlation exists among crash severity levels of drivers or of most severely injured persons in all vehicles involved in a multi-vehicle crash due to the unobserved shared factors such as collision speed.

Several methodological alternatives have been proposed in literature to model crash outcome while accounting for data issues. Please refer to the studies conducted by Mannering and Bhat and, Mannering et al. for a detailed discussion of developed methodologies in crash outcome (Mannering and Bhat 2014, Mannering et al. 2016).

Starting with simple binary discrete outcome models such as binary logit and probit models, models evolved to consider multiple discrete outcomes (to consider a variety of injury-severity categories such as no injury, possible injury, evident injury, disabling injury and fatality). For the multiple discrete outcome models, multinomial models that do not account for the ordering of injury outcome have been widely applied from the simple multinomial logit model, to the nested logit model. Like crash frequency data, crash outcome dataset also suffers from unobserved data heterogeneity issue. The random parameters logit model was explored to account for the effect of unobserved factors across crash observations. The correlation between multiple discrete outcomes of crash event due to the unobserved shared factors is another major issue related to crash outcome dataset. Statistical methods that do not account for the correlation among injuries occurring in the same crash are likely to result in biased parameter estimates. The multinomial logit model (MNL) is built on the Independence of Irrelevant Alternatives (IIA) assumption, meaning adding or deleting an alternative will not change the ratio between the probabilities of

any pair of existing alternatives. In simple words, MNL does not allow for correlation between any pairs of existing alternatives. The multinomial probit (MNP) model relaxes the independence assumption built into the MNL model. Thus, MNP can be a potential alternative to model unordered discrete crash outcome data to account for correlation among outcomes.

2.4 References

Aarts, L., Van Schagen, I., 2006. Driving speed and the risk of road crashes: A review. Accident Analysis & Prevention 38 (2), 215-224.

Aashto, 2001. Policy on geometric design of highways and streets. American Association of State Highway and Transportation Officials, Washington, DC 1 (990), 158.

Aashto, 2010. Highway safety manual American Association of State Highway and Transportation Officials, Washington D.C.

Abdel-Aty, M., Lee, J., Siddiqui, C., Choi, K., 2013. Geographical unit based analysis in the context of transportation safety planning. Transportation Research Part A: Policy and Practice 49, 62-75.

Abdel-Aty, M.A., Pemmanaboina, R., 2006. Calibrating a real-time traffic crash-prediction model using archived weather and its traffic data. IEEE Transactions on Intelligent Transportation Systems 7 (2), 167-174.

Administration, N.H.T.S., 2015. 2014 motor vehicle crashes: Overview. Traffic safety facts research note 2015, 1-9.

Afghari, A.P., Haque, M.M., Washington, S., Smyth, T., 2016. Bayesian latent class safety performance function for identifying motor vehicle crash black spots. Transportation Research Record: Journal of the Transportation Research Board (2601), 90-98.

Afghari, A.P., Washington, S., Haque, M.M., Li, Z., 2018. A comprehensive joint econometric model of motor vehicle crashes arising from multiple sources of risk. Analytic Methods in Accident Research 18, 1-14.

Aguero-Valverde, J., Jovanis, P., 2008. Analysis of road crash frequency with spatial models. Transportation Research Record: Journal of the Transportation Research Board (2061), 55-63.

Aguero-Valverde, J., Jovanis, P.P., Year. Identifying road segments with high risk of weather-related crashes using full bayesian hierarchical models. In: Proceedings of the 86th Annual Meeting of the Transportation Research Board, Washington, DC.

Åkerstedt, T., Kecklund, G., 2001. Age, gender and early morning highway accidents. Journal of sleep research 10 (2), 105-110.

Alver, Y., Demirel, M., Mutlu, M., 2014. Interaction between socio-demographic characteristics: Traffic rule violations and traffic crash history for young drivers. Accident Analysis & Prevention 72, 95-104.

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

Andrey, J., 2010. Long-term trends in weather-related crash risks. Journal of Transport Geography 18 (2), 247-258.

Andrey, J., Mills, B., Leahy, M., Suggett, J., 2003. Weather as a chronic hazard for road transportation in canadian cities. Natural Hazards 28 (2), 319-343.

Andrey, J., Mills, B., Vandermolen, J., 2001. Weather information and road safety. Institute for Catastrophic Loss Reduction, Toronto, Ontario, Canada.

Asbridge, M., Hayden, J.A., Cartwright, J.L., 2012. Acute cannabis consumption and motor vehicle collision risk: Systematic review of observational studies and meta-analysis. Bmj 344, e536.

Babizhayev, M.A., 2003. Glare disability and driving safety. Ophthalmic research 35 (1), 19-25.

Bao, S., Boyle, L.N., 2009. Age-related differences in visual scanning at median-divided highway intersections in rural areas. Accident Analysis & Prevention 41 (1), 146-152.

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

Bauer, K., Harwood, D.W., 1996. Statistical models of at-grade intersection accidents.

Begg, D.J., Gulliver, P., 2008. A longitudinal examination of the relationship between adolescent problem behaviors and traffic crash involvement during young adulthood. Traffic injury prevention 9 (6), 508-514.

Bergel-Hayat, R., Debbarh, M., Antoniou, C., Yannis, G., 2013. Explaining the road accident risk: Weather effects. Accident Analysis & Prevention 60, 456-465.

Bingham, C.R., Shope, J.T., 2004. Adolescent problem behavior and problem driving in young adulthood. Journal of Adolescent Research 19 (2), 205-223.

Blomberg, R.D., Peck, R.C., Moskowitz, H., Burns, M., Fiorentino, D., 2005. Crash risk of alcohol involved driving: A case-control study.

Box, S., 2009. New data from vtti provides insight into cell phone use and driving distraction. Virginia Tech Transportation Institute 27.

Broughton, K.L., Switzer, F., Scott, D., 2007. Car following decisions under three visibility conditions and two speeds tested with a driving simulator. Accident Analysis & Prevention 39 (1), 106-116.

Buddhavarapu, P., Scott, J.G., Prozzi, J.A., 2016. Modeling unobserved heterogeneity using finite mixture random parameters for spatially correlated discrete count data. Transportation Research Part B: Methodological 91, 492-510.

Caliendo, C., Guida, M., Parisi, A., 2007. A crash-prediction model for multilane roads. Accident Analysis & Prevention 39 (4), 657-670.

Campbell, B.N., Smith, J.D., Najm, W.G., 2003. Examination of crash contributing factors using national crash databases.

Campbell, J.L., 2012. Human factors guidelines for road systems Transportation Research Board.

Carroll, P., 1971. Techniques for the use of driving exposure information in highway safety research. HSRI, University of Michigan.

Chang, H.-L., Yeh, T.-H., 2007. Motorcyclist accident involvement by age, gender, and risky behaviors in taipei, taiwan. Transportation Research Part F: Traffic Psychology and Behaviour 10 (2), 109-122.

Chang, L.-Y., Chen, W.-C., 2005. Data mining of tree-based models to analyze freeway accident frequency. Journal of safety research 36 (4), 365-375.

Chapman, R., 1973. The concept of exposure. Accident Analysis & Prevention 5 (2), 95-110.

Chatterjee, I., Davis, G.A., 2016. Analysis of rear-end events on congested freeways by using video-recorded shock waves. Transportation Research Record: Journal of the Transportation Research Board (2583), 110-118.

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, 86-95.

Chen, Z., Qin, X., Shaon, M.R.R., 2017. Modeling lane-change related crashes with lane-specific real-time traffic and weather data. Journal of Intelligent Transportation Systems (just-accepted).

Cheng, L., Geedipally, S.R., Lord, D., 2013. The poisson–weibull generalized linear model for analyzing motor vehicle crash data. Safety science 54, 38-42.

Choi, E.-H., Singh, S., 2005. Statistical assessment of the glare issue-human and natural elements. National Center for Statistics and Analysis.

Compton, R., Vegega, M., Smither, D., 2009. Drug-impaired driving: Understanding the problem and ways to reduce it: A report to congress.

Compton, R.P., Berning, A., 2015. Drug and alcohol crash risk. Journal of Drug Addiction, Education, and Eradication 11 (1), 29.

Dahl, R.E., 2008. Biological, developmental, and neurobehavioral factors relevant to adolescent driving risks. American journal of preventive medicine 35 (3), S278-S284.

Druschel, S., 2014. Field effects on deicing and anti-icing performance. Minnesota Department of Transportation.

Dupont, R.L., Voas, R.B., Walsh, J.M., Shea, C., Talpins, S.K., Neil, M.M., 2012. The need for drugged driving per se laws: A commentary. Traffic injury prevention 13 (1), 31-42.

Eisenberg, D., 2004. The mixed effects of precipitation on traffic crashes. Accident analysis & prevention 36 (4), 637-647.

Eisenberg, D., Warner, K.E., 2005. Effects of snowfalls on motor vehicle collisions, injuries, and fatalities. American journal of public health 95 (1), 120-124.

El-Basyouny, K., Barua, S., Islam, M.T., 2014a. Investigation of time and weather effects on crash types using full bayesian multivariate poisson lognormal models. Accident Analysis & Prevention 73, 91-99.

El-Basyouny, K., Barua, S., Islam, T., Year. Full bayesian multivariate models to assess time and weather effects on crash types. In: Proceedings of the Transportation Research Board 93rd Annual Meeting.

El-Basyouny, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. Accident Analysis & Prevention 41 (5), 1118-1123.

Federal Highway Administration (Fhwa), 2013. How do weather events impact roads? U.S. Department of Transportation, Washington D.C.

Fhwa, 2012. Our nation's highway: 2011. US Department of Transportation, Washington D.C.

Garnowski, M., Manner, H., 2011. On factors related to car accidents on german autobahn connectors. Accident Analysis & Prevention 43 (5), 1864-1871.

Geedipally, S., Lord, D., Park, B.-J., 2009. Analyzing different parameterizations of the varying dispersion parameter as a function of segment length. Transportation Research Record: Journal of the Transportation Research Board (2103), 108-118.

Geedipally, S.R., Lord, D., Dhavala, S.S., 2012. The negative binomial-lindley generalized linear model: Characteristics and application using crash data. Accident Analysis & Prevention 45, 258-265.

Glennon, J.C., 1996. Roadway defects and tort liability.

Grabowski, D.C., Morrisey, M.A., 2004. Gasoline prices and motor vehicle fatalities. Journal of Policy Analysis and Management 23 (3), 575-593.

Greibe, P., 2003. Accident prediction models for urban roads. Accident Analysis & Prevention 35 (2), 273-285.

Guo, J.Q., Trivedi, P.K., 2002. Flexible parametric models for long-tailed patent count distributions.

Hadayeghi, A., Shalaby, A., Persaud, B., 2003. Macrolevel accident prediction models for evaluating safety of urban transportation systems. Transportation Research Record: Journal of the Transportation Research Board (1840), 87-95.

Hadayeghi, A., Shalaby, A., Persaud, B., 2007. Safety prediction models: Proactive tool for safety evaluation in urban transportation planning applications. Transportation Research Record: Journal of the Transportation Research Board (2019), 225-236.

Hadayeghi, A., Shalaby, A.S., Persaud, B.N., 2010. Development of planning level transportation safety tools using geographically weighted poisson regression. Accident Analysis & Prevention 42 (2), 676-688.

Haddon Jr, W., 1968. The changing approach to the epidemiology, prevention, and amelioration of trauma: The transition to approaches etiologically rather than descriptively based. American Journal of Public Health and the Nations Health 58 (8), 1431-1438.

Hallmark, S.L., Qiu, Y., Pawlovitch, M., Mcdonald, T.J., 2013. Assessing the safety impacts of paved shoulders. Journal of Transportation Safety & Security 5 (2), 131-147.

Harwood, D., 2017. Guidelines for integrating safety and cost-effectiveness into resurfacing, restoration, and rehabilitation projects

Harwood, D.W., Torbic, D.J., Richard, K.R., Meyer, M.M., 2010. Safetyanalyst: Software tools for safety management of specific highway sites.

Hauer, E., 1982. Traffic conflicts and exposure. Accident Analysis & Prevention 14 (5), 359-364.

Hauer, E., 1995. On exposure and accident rate. Traffic engineering & control 36 (3), 134-138.

Hauer, E., 1999. A primer on traffic safety. Institute of Transportation Engineers.

Hauer, E., 2009. Speed and safety. Transportation Research Record: Journal of the Transportation Research Board (2103), 10-17.

Hauer, E., 2010. Cause, effect and regression in road safety: A case study. Accident Analysis & Prevention 42 (4), 1128-1135.

Hedlund, J., 2017. Drug impaired driving: A guide for states. Governors Highway Safety Association (GHSA).

Hermans, E., Brijs, T., Stiers, T., Offermans, C., Year. The impact of weather conditions on road safety investigated on an hourly basis.

Heydari, S., Fu, L., Lord, D., Mallick, B.K., 2016. Multilevel dirichlet process mixture analysis of railway grade crossing crash data. Analytic methods in accident research 9, 27-43.

Hing, J.Y.C., Stamatiadis, N., Aultman-Hall, L., 2003. Evaluating the impact of passengers on the safety of older drivers. Journal of safety research 34 (4), 343-351.

Hjelkrem, O.A., Ryeng, E.O., 2016. Chosen risk level during car-following in adverse weather conditions. Accident Analysis & Prevention 95, 227-235.

Hoogendoorn, R., Tamminga, G., Hoogendoorn, S., Daamen, W., Year. Longitudinal driving behavior under adverse weather conditions: Adaptation effects, model performance and freeway capacity in case of fog. In: Proceedings of the Intelligent transportation systems (itsc), 2010 13th international ieee conference on, pp. 450-455.

Houwring, S., 2013. Estimating the risk of driving under the influence of psychoactive substances Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV.

Jiang, X., Abdel-Aty, M., Hu, J., Lee, J., 2016. Investigating macro-level hotzone identification and variable importance using big data: A random forest models approach. Neurocomputing 181, 53-63.

Jung, S., Qin, X., Noyce, D.A., 2010. Rainfall effect on single-vehicle crash severities using polychotomous response models. Accident Analysis & Prevention 42 (1), 213-224.

Khattak, A.J., Knapp, K.K., 2001. Snow event effects on interstate highway crashes. Journal of cold regions engineering 15 (4), 219-229.

Knapp, K.K., Year. An investigation of volume, safety, and vehicle speeds during winter storm events. In: Proceedings of the Ninth AASHTO/TRB Maintenance Management Conference.

Kononov, J., Lyon, C., Allery, B.K., 2011. Relation of flow, speed, and density of urban freeways to functional form of a safety performance function. Transportation Research Record: Journal of the Transportation Research Board 2236 (1), 11-19.

Lee, C., Saccomanno, F., Hellinga, B., 2002. Analysis of crash precursors on instrumented freeways. Transportation Research Record: Journal of the Transportation Research Board (1784), 1-8.

Lee, J., Abdel-Aty, M., Cai, Q., 2017. Intersection crash prediction modeling with macro-level data from various geographic units. Accident Analysis & Prevention 102, 213-226.

Lefler, N., Council, F., Harkey, D., Carter, D., Mcgee, H., Daul, M., 2010. Model inventory of roadway elements-mire, version 1.0.

Leigh, J.P., Geraghty, E.M., 2008. High gasoline prices and mortality from motor vehicle crashes and air pollution. Journal of Occupational and Environmental Medicine 50 (3), 249-254.

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.

Li, Z., Ahn, S., Chung, K., Ragland, D.R., Wang, W., Yu, J.W., 2014. Surrogate safety measure for evaluating rear-end collision risk related to kinematic waves near freeway recurrent bottlenecks. Accident Analysis & Prevention 64, 52-61.

Lindley, D.V., 1958. Fiducial distributions and bayes' theorem. Journal of the Royal Statistical Society. Series B (Methodological), 102-107.

Longthorne, A., Subramanian, R., Chen, C.-L., 2010. An analysis of the significant decline in motor vehicle traffic fatalities in 2008.

Lord, D., Geedipally, S.R., 2011. The negative binomial–lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. Accident Analysis & Prevention 43 (5), 1738-1742.

Lord, D., Guikema, S.D., Geedipally, S.R., 2008. Application of the conway–maxwell–poisson generalized linear model for analyzing motor vehicle crashes. Accident Analysis & Prevention 40 (3), 1123-1134.

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 (5), 291-305.

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 & Prevention 37 (1), 35-46.

Lum, H., Reagan, J.A., 1995. Interactive highway safety design model: Accident predictive module. Public Roads 58 (3).

Ma, L., Yan, X., 2014. Examining the nonparametric effect of drivers' age in rear-end accidents through an additive logistic regression model. Accident Analysis & Prevention 67, 129-136.

Malyshkina, N.V., Mannering, F.L., Tarko, A.P., 2009. Markov switching negative binomial models: An application to vehicle accident frequencies. Accident Analysis & Prevention 41 (2), 217-226.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Analytic methods in accident research 1, 1-22.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research 11, 1-16.

María L. Sanz De Acedo Lizárraga, M.T.S.D.a.B., María Cardelle-Elawar, 2007. Factors that affect decision making: Gender and age differences. International Journal of Psychology and Psychological Therapy.

Martchouk, M., Mannering, F., Bullock, D., 2010. Analysis of freeway travel time variability using bluetooth detection. Journal of Transportation Engineering 137 (10), 697-704.

Massie, D.L., Campbell, K.L., Williams, A.F., 1995. Traffic accident involvement rates by driver age and gender. Accident Analysis & Prevention 27 (1), 73-87.

Mayhew, D.R., Simpson, H.M., Pak, A., 2003. Changes in collision rates among novice drivers during the first months of driving. Accident Analysis & Prevention 35 (5), 683-691.

Mcandrews, C., Beyer, K., Guse, C., Layde, P., 2017. Linking transportation and population health to reduce racial and ethnic disparities in transportation injury: Implications for practice and policy. International Journal of Sustainable Transportation 11 (3), 197-205.

Mcandrews, C., Beyer, K., Guse, C.E., Layde, P., 2013. Revisiting exposure: Fatal and non-fatal traffic injury risk across different populations of travelers in wisconsin, 2001–2009. Accident Analysis & Prevention 60, 103-112.

Mccartt, A.T., Shabanova, V.I., Leaf, W.A., 2003. Driving experience, crashes and traffic citations of teenage beginning drivers. Accident Analysis & Prevention 35 (3), 311-320.

Mcgee, H.W., 2011. Geometric design practices for resurfacing, restoration, and rehabilitation Transportation Research Board.

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), 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. Statistical methods and modeling and safety data, analysis, and evaluation: Safety and human performance. pp. 31-40.

Mitra, S., 2006. Significance of omitted variable bias in transportation safety studies.

Mitra, S., 2014. Sun glare and road safety: An empirical investigation of intersection crashes. Safety science 70, 246-254.

Mitra, S., Washington, S., 2007a. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis & Prevention 39 (3), 459-468.

Mitra, S., Washington, S., 2007b. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis and Prevention 39 (3), 459-468.

Mitra, S., Washington, S., 2012. On the significance of omitted variables in intersection crash modeling. Accident Analysis and Prevention 49, 439-448.

Morgan, A., Mannering, F.L., 2011. The effects of road-surface conditions, age, and gender on driver-injury severities. Accident Analysis & Prevention 43 (5), 1852-1863.

Morrison, C., Ponicki, W.R., Gruenewald, P.J., Wiebe, D.J., Smith, K., 2016. Spatial relationships between alcohol-related road crashes and retail alcohol availability. Drug and alcohol dependence 162, 241-244.

Najm, W., Koopmann, J., Boyle, L., Smith, D., 2002. Development of test scenarios for off-roadway crash countermeasures based on crash statistics.

Nhtsa, Risky driving. US Department of Transportation.

Nhtsa, 2010. Traffic safety facts: Driver electronic device use observation protocol. DOT HS 811, 361.

Nhtsa, 2016. 2015 motor vehicle crashes: Overview. Traffic safety facts research note 2016, 1-9.

Nhtsa, 2017. Traffic safety facts: Older population. National Center for Statistics and Analysis, Washington D.C.

Nhtsa, N.H.T.S.A., 2008. National motor vehicle crash causation survey: Report to congress. National Highway Traffic Safety Administration Technical Report DOT HS 811, 059.

Ong, G., Fwa, T., 2007. Prediction of wet-pavement skid resistance and hydroplaning potential. Transportation Research Record: Journal of the Transportation Research Board (2005), 160-171.

Organization for Economic Co-Operation and Development, 2006. Young drivers: The road to safety., Paris, France.

Owusu-Ababio, S., Feng, G., 2006. Predicting alcohol-related crash rate in wisconsin using neural networks. Intelligent engineering systems through artificial neural networks, volume 16. ASME Press.

Paleti, R., Eluru, N., Bhat, C.R., 2010. Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. Accident Analysis & Prevention 42 (6), 1839-1854.

Park, B.J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accident Analysis and Prevention 41 (4), 683-691.

Pickrell, T.M., Liu, C., 2016. Occupant restraint use in 2014: Results from the nopus controlled intersection study.

Piff, P.K., Stancato, D.M., Côté, S., Mendoza-Denton, R., Keltner, D., 2012. Higher social class predicts increased unethical behavior. Proceedings of the National Academy of Sciences 109 (11), 4086-4091.

Pisano, P.A., Goodwin, L.C., Rossetti, M.A., Year. Us highway crashes in adverse road weather conditions. In: Proceedings of the 24th Conference on International Interactive Information and Processing Systems for Meteorology, Oceanography and Hydrology, New Orleans, LA.

Pulugurtha, S.S., Duddu, V.R., Kotagiri, Y., 2013. Traffic analysis zone level crash estimation models based on land use characteristics. Accident Analysis & Prevention 50, 678-687.

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

Qin, X., Ng, M., Reyes, P.E., 2010. Identifying crash-prone locations with quantile regression. Accident Analysis & Prevention 42 (6), 1531-1537.

Qin, X., Rahman Shaon, M.R., Chen, Z., 2016. Developing analytical procedures for calibrating the highway safety manual predictive methods. Transportation Research Record: Journal of the Transportation Research Board (2583), 91-98.

Rahman Shaon, M.R., Qin, X., 2016. Use of mixed distribution generalized linear models to quantify safety effects of rural roadway features. Transportation Research Record: Journal of the Transportation Research Board (2583), 134-141.

Rahman Shaon, M.R., Schneider, R.J., Qin, X., He, Z., Sanatizadeh, A. and Flanagan, M.D., 2018a. Exploration of pedestrian assertiveness and its association with driver yielding behavior at uncontrolled crosswalks. Transportation research record, 2672(35), pp.69-78.

Rahman Shaon, M.R., Qin, X., Shirazi, M., Lord, D. and Geedipally, S.R., 2018b. Developing a Random Parameters Negative Binomial-Lindley Model to analyze highly over-dispersed crash count data. Analytic methods in accident research, 18, pp.33-44.

Ray, B.L., Ferguson, E.M., Knudsen, J.K., Porter, R.J., Mason, J., 2014. Performance-based analysis of geometric design of highways and streets.

Redelmeier, D.A., Tibshirani, R.J., 1997. Association between cellular-telephone calls and motor vehicle collisions. New England Journal of Medicine 336 (7), 453-458.

Reinfurt, D.W., Stewart, J., Stutts, J.C., Rodgman, E.A., 2000. Investigations of crashes and casualties associated with older drivers. Project G. 8. Chapel Hill, NC: University of North Carolina Highway Safety Research Center.

Rodman, K., 2016. Hidden hazards of road salt: Car corrosion can take a toll. AccuWeather.com.

Rose, J.G., Gallaway, B.M., 1977. Water depth influence on pavement friction. Journal of Transportation Engineering 103 (4).

Rossetti, M.A., Johnsen, M., 2011. Weather and climate impacts on commercial motor vehicle safety.

Sagberg, F., Bjørnskau, T., 2006. Hazard perception and driving experience among novice drivers. Accident Analysis & Prevention 38 (2), 407-414.

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

Savolainen, P.T., Mannering, F., Lord, D., Quddus, M.A., 2011. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. Accident Analysis and Prevention 43 (5), 1666-1676.

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

Shankar, V., Milton, J., Mannering, F., 1997a. Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. Accident Analysis and Prevention 29 (6), 829-837.

Shankar, V., Milton, J., Mannering, F., 1997b. Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. Accident Analysis & Prevention 29 (6), 829-837.

Shirazi, M., Lord, D., Dhavala, S.S., Geedipally, S.R., 2016. A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: Characteristics and applications to crash data. Accident Analysis & Prevention 91, 10-18.

Smart, D., Vassallo, S., Sanson, A., Cockfield, S., Harris, A., Harrison, W., 2005. In the driver's seat: Understanding young adults' driving behaviour.

Snelder, M., Calvert, S., 2016. Quantifying the impact of adverse weather conditions on road network performance. EJTIR 16 (1), 128-149.

Son, H., Kweon, Y.J., Park, B., 2011. Development of crash prediction models with individual vehicular data. Transportation Research Part C-Emerging Technologies 19 (6), 1353-1363.

Staplin, L., Lococo, K., Byington, S., Harkey, D., 2001. Guidelines and recommendations to accommodate older drivers and pedestrians.

Steinberg, L., Morris, A.S., 2001. Adolescent development. Annual review of psychology 52 (1), 83-110.

Stewart, D.E., 1998. Methodological approaches for the estimation, evaluation, interpretation and accuracy assessment of road travel" basic risk"," relative risk", and" relative risk odds-ratio" performance measure indicators: A" risk analysis and evaluation system model" for measuring, monitoring, comparing and evaluating the level (s) of safety on canada's roads and highways Transport Canada, Road Safety.

Stutts, J., Martell, C., Staplin, L., 2009. Identifying behaviors and situations associated with increased crash risk for older drivers.

Swedler, D.I., Bowman, S.M., Baker, S.P., Year. Gender and age differences among teen drivers in fatal crashes. In: Proceedings of the Annals of Advances in Automotive Medicine/Annual Scientific Conference, pp. 97.

Tate, F., Turner, S., 2007. Road geometry and drivers' speed choice. Road & Transport Research: A Journal of Australian and New Zealand Research and Practice 16 (4), 53.

United States Census Bureau, 2016. Fff: Older americans month: May 2016.

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 & Prevention 48, 368-378.

Venkataraman, N., Ulfarsson, G., Shankar, V., Oh, J., Park, M., 2011. Model of relationship between interstate crash occurrence and geometrics: Exploratory insights from random parameter negative binomial approach. Transportation research record: journal of the transportation research board (2236), 41-48.

Walsh, J.M., Verstraete, A.G., Huestis, M.A., Mørland, J., 2008. Guidelines for research on drugged driving. Addiction 103 (8), 1258-1268.

Wang, J., Huang, H., 2016. Road network safety evaluation using bayesian hierarchical joint model. Accident Analysis & Prevention 90, 152-158.

Wang, J., Zheng, Y., Li, X., Yu, C., Kodaka, K., Li, K., 2015. Driving risk assessment using near-crash database through data mining of tree-based model. Accident Analysis & Prevention 84, 54-64.

Wang, K., Qin, X., 2015. Exploring driver error at intersections: Key contributors and solutions. Transportation Research Record: Journal of the Transportation Research Board (2514), 1-9.

Wang, Y., Kockelman, K.M., 2013. A poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. Accident Analysis & Prevention 60, 71-84.

Washington, S., Haque, M., 2013. On the commonly accepted assumptions regarding observed motor vehicle crash counts at transport system locations.

Wellner, A., Qin, X., 2011. Highway safety metrics implementation and evaluation using a geographic information system-based screening tool. Transportation Research Record: Journal of the Transportation Research Board (2241), 1-9.

Williams, A.F., Kyrychenko, S.Y., Retting, R.A., 2006. Characteristics of speeders. Journal of Safety Research 37 (3), 227-232.

Wisconsin Department of Transportation (Wisdot), 2017. Preliminary review of 2015 crash fatality trends. Wisconsin Department of Transportation (WisDOT), Wisconsin.

Wisconsin Traffic Operations and Safety (Tops) Laboratory, 2017. The wistransportal system. University of wisconsin-madison and wisconsin department of transportation (wisdot) bureau of traffic operations (bto).

Wisdot, 2016. Citations and convictions: Facts and figures. Department of Motor Vehicle, Wisconsin.

Wu, L., Lord, D., Year. Investigating the influence of dependence between variables on crash modification factors developed using regression models. In: Proceedings of the Transportation Research Board 95th Annual Meeting, TRB, Washington DC.

Xie, Y., Lord, D., Zhang, Y., 2007. Predicting motor vehicle collisions using bayesian neural network models: An empirical analysis. Accident Analysis & Prevention 39 (5), 922-933.

Xie, Y., Zhang, Y., 2008. Crash frequency analysis with generalized additive models. Transportation Research Record: Journal of the Transportation Research Board 2061 (1), 39-45.

Xu, J., Sun, L., 2015. Modeling of excess zeros issue in crash count andysis. Journal of Jilin University (Engineering and Technology Edition) 45 (3), 769-775.

Yaacob, W.F.W., Lazim, M.A., Wah, Y.B., Year. Evaluating spatial and temporal effects of accidents likelihood using random effects panel count model. In: Proceedings of the Science and Social Research (CSSR), 2010 International Conference on, pp. 960-964.

Ye, F., Garcia, T.P., Pourahmadi, M., Lord, D., Year. Extension of a negative binomial garch model: Analyzing the effects of gasoline price and vmt on dui fatal crashes in texas. In: Proceedings of the 91st Annual Meeting of the Transportation Research Board.

Yu, R., Abdel-Aty, M., Ahmed, M., 2013. Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. Accident Analysis & Prevention 50, 371-376.

Zamani, H., Ismail, N., 2010. Negative binomial-lindley distribution and its application. Journal of Mathematics and Statistics 6 (1), 4-9.

Zegeer, C.V., Stewart, J.R., Council, F.M., Reinfurt, D.W., Hamilton, E., 1992. Safety effects of geometric improvements on horizontal curves. Transportation Research Record (1356).

Zhang, Y., Xie, Y., Li, L., 2012. Crash frequency analysis of different types of urban roadway segments using generalized additive model. Journal of safety research 43 (2), 107-114.

Zou, Y., Wu, L., Lord, D., 2015. Modeling over-dispersed crash data with a long tail: Examining the accuracy of the dispersion parameter in negative binomial models. Analytic Methods in Accident Research 5, 1-16.

Chapter 3 Using Proxy Variables for Driver Behavior in Area-based Crash Prediction Models

This chapter presents the exploration of proxy variables for driver behavior in area-level CPM. This area-level modeling has been developed in the 1st tier of the proposed 3-tier spatial unit approach for crash data modeling. In this regard, a series of socioeconomic and demographic variables collected at census tract by US census has been explored as surrogate measures for driver behavior and modeled to evaluate their effect on area-level crash occurrences. To better understand the effect of behavioral factors, behavior-related CPMs are developed along with total crash CPM. It is hypothesized that the exploration of surrogate measures for driver behavior may yield potential and informative variables that are highly correlated with crash occurrence. Exploration of these variables in behavior-related CPMs may provide information on the intrinsic relation between proxy variables and behavior-related crash occurrences.

3.1 Introduction

Research efforts have shifted recently to a higher level of aggregated crash analyses in which crash prediction models (CPMs) are used to relate traffic crashes aggregated by a specific spatial

scale to area-level factors such as socioeconomic status, demographic characteristics, land use, and traffic patterns. CPMs may help agencies to be more proactive in incorporating safety considerations in the long-term transportation planning process (Washington 2006).

The selection of a spatial unit is an important element of developing a macro-level CPM. A wide array of spatial units has been employed, such as regions (Washington et al. 1999), counties (Miaou et al. 2003), zip codes (Girasek and Taylor 2010), census tracts (Wang and Kockelman 2013), block groups (Levine et al. 1995), and traffic analysis zones (Xu et al. 2014). Studies related to macro-level CPMs most commonly involve aggregate CPMs that have been developed to relate roadway crashes to a variety of explanatory factors, including road network composition, traffic patterns, and area-level demographic and socioeconomic characteristics. Among spatial units explored in the literature, variables related to socioeconomic, demographic, and traffic patterns are readily available at the census tract level from the U.S. Census. For this macro-level model development, census tract was chosen as the spatial unit for the development of macro-level crash prediction model.

The highway safety literature shows that driver error is one major type of factor contributing to crashes (NHTSA 2008). Speeding has been identified as one of the main driver-related factors that contribute to crashes. In 2014, 9,262 of the total 32,675 driving-related fatalities in the United States were due to speeding (NHTSA 2015). Alcohol-impaired driving is another driver-related error that causes many crashes. According to the 2015 Wisconsin Fatal Crash Trend analysis, 34 percent of all fatal crashes that occurred in Wisconsin involved alcohol-impaired driving (Wisconsin Department of Transportation (WisDOT) 2017). In macro-level CPMs, the crash data and covariates are aggregated to a spatial unit, which in turn makes it nearly impossible to incorporate specific driver factors into CPMs. The covariates may have

different effects on different driver behaviors. The development of separate CPMs for only driver behavior-related crashes can be a potential way of exploring the effects of driver behavior on area-level CPMs. The development of a behavior-based CPM along with a CPM for all crashes can test the hypothesis that covariates may have different effects on different crash types based on driver behavior.

The objective of this section is to investigate the key contributors and their effects on various driver behavior-related crashes as well as all crashes. The negative binomial modeling approach was used to develop CPMs for all crashes, for speed-related crashes, and for alcohol-related crashes. The parameter estimates of the developed models will shed light on the effects of covariates on different crash types based on driver behavior. The model outputs can be used to identify communities with higher crash risk and help agencies develop more informative and cost-effective countermeasures.

3.2 Methodology

Negative Binomial (NB) model is one of the most notable models for crash frequency data. The model is suitable for a dependent variable that is a non-negative integer. It accounts for data over-dispersion, handles traffic exposure and offset variables, and has model parameters that are easy to estimate in any of the statistical software applications. The probability mass function (pmf) of the NB distribution can be written as:

$$\mathbf{P(Y = y; \phi, p)} = \frac{\mathbf{\Gamma(\phi+y)}}{\mathbf{\Gamma(\phi)\times y!}}\mathbf{(1-p)^{\phi}(p)^{y}};\quad \mathbf{\phi > 0, 0 < p < 1} \tag{1}$$

Where,

p = probability of success in each trial;

Φ= Inverse dispersion parameter α (i.e. Φ=1/α);

The dispersion parameter measures the dispersion of the response variable. If the dispersion parameter equals zero, the NB model becomes the Poisson model, suggesting the Poisson distribution is a limiting case of the binomial distribution. If the dispersion parameter is greater than zero, it means that the response variable is over-dispersed. Using a log-link function, the mean response can be written as:

$$\ln(\boldsymbol{\mu}) = \boldsymbol{\beta_0} + \sum_{i=1}^{q} \boldsymbol{\beta_i X} \quad (2)$$

Where,

X= Covariates used to model mean response;

β_i= Regression coefficient for covariates (q= Number of total covariates used in the model).

3.3 Data Processing and Exploratory Analysis

Census tract has been used as a spatial unit for developing area-based crash frequency prediction models to explore the effect of surrogate measures for driver behavior in this section. The census tract information from Wisconsin has been collected by the U.S. Census. The analysis in this study used 2015 TIGER/Line Shapefiles data from the U.S. Census. The 2015 TIGER/Line dataset for census tract contains 28 separate data tables with 17,812 attributes total. All data tables can be integrated using a unique census tract identification number. Based on the literature, a series of attributes were selected to explore as covariates in a crash frequency prediction model for census tract. ArcMap was used to join selected attributes to each census tract. Data were processed further based on variable definitions in order to normalize the covariates. The final dataset was represented in a percentage format.

The roadway network-related attributes in each census tract were obtained from WISLR since they were not available in the TIGER/Line database. The WISLR database contains roadway and traffic-related information, including the geographical location of the roadway, for all roadway networks in Wisconsin. The WISLR dataset was spatially joined with census tract using ArcMap in order to obtain the total roadway length, AADT, and the total number of intersections within each census tract.

Police-reported crashes that occurred on Wisconsin roadways from 2011 to 2015 were collected from the MV4000 dataset and were processed to develop area-level CPMs. The effect of human behavior was explored by extracting two subsets of crash data – speed-related and alcohol-related – from the all crash dataset based on the human factor related to each crash occurrence. The MV4000 dataset contains flags for each crash type. A speed flag and alcohol flag was used to extract the subset of each crash type. Once crash data are collected, the crashes need to be linked with the census tract based on the location information; however, crashes may not always occur within the defined census tract boundary. When existing roadways are used for defining government boundaries, a portion of crashes occurred on these census tract boundaries. A major challenge is joining crashes that occur on the census tract boundary, or the "boundary collision issue". Researchers have developed several methods for properly distributing boundary crashes among corresponding census tracts. A list of available methods are provided below:

- Equal proportion: Proportioning the crash based on the number of adjacent spatial unit
- Geo-processing methods: Data attributes aggregated for each spatial unit as they were geo-coded in ArcGIS (Wei, 2010)

- One-to-one method: Each spatial unit forming the boundary is assigned one whole collision) (Wei, 2010)
- Vehicle Kilometer Traveled (VKT) Proportion: Proportioning the boundary crashes based on the value of VKT of corresponding spatial units
- Total Lane Kilometers (TLK) Proportion: Same as above, but with measured "total lane kilometers"
- Density Probability: Aggregation of boundary collisions by density probability ratio (Cui et al. 2015)

The equal proportion method was used in this study for joining boundary crashes with the census tract. The following steps were completed in ArcMap to filter boundary crashes and join them with the census tract using equal proportion method:

1. Convert polygon shapefile of census tract to line features.
2. Create point shapefile of all crashes using location attributes (longitude and latitude) available in the MV4000 database.
3. Use "Select by location" tool to select crashes occurred within a specified distance from census tract line shapefile. In this study, a distance of 30 meters was used.
4. Based on selected crashes from "Select by location" tool, create two separate shapefiles for crashes by splitting them: "Crashes on Boundary" and "Crashes within Boundary".
5. Use "Spatial Join" tool with join option as "One-to-One" to count the number of crashes occurred with each census tract from "Crashes within Boundary" file.

6. Use “Spatial Join” tool with join option as “One-to-Many” with “Crashes on Boundary” file to count the number of zones related to each crash occurred on census boundary.
7. If a crash occurred on the boundary of “n” census tracts, split the crash value to “1/n” in each census tract.
8. Sum joining results from step 5 and step 7 to obtain total crashes in a census tract.

Table 3-1 Summary Statistics of Wisconsin Census Tract Data.

Variable	Unit	Mean	Standard Deviation	Minimum	Maximum
Total Crash	Count	351.505	259.586	4.500	2193.167
Speed-related Crash	Count	63.01	53.24	0	451.33
Alcohol-related Crash	Count	16.88	10.97	0	102.5
Area	Sq. Mile	41.1	76.471	0.068	799.8
Roadway Length	Miles	166.4	218.288	0.112	2242.551
VMT	Veh-mile	80972.3	61319	5.600	553114.024
Number of Intersections	Count	248.3	189.04	0.000	1664
Population Density	Count/Sq. Mile	2919.1	4657.4	0.000	50428.739
Male	%	0.495	0.042	0.000	1.00
White	%	0.841	0.228	0.000	1.00
Proportion w/Age <18	%	0.226	0.064	0	0.49
Proportion w/Age >64	%	0.150	0.063	0	0.53
Median Age	Years	39.52	7.646	0.000	67.5
Enrolled in school	%	0.264	0.091	0.000	0.970
Primary work commute mode (Car)	%	0.878	0.107	0.000	1.000
Primary work commute mode (Public Transit)	%	0.027	0.056	0.000	0.653
Primary work commute mode (Bicycle)	%	0.008	0.018	0.000	0.212

Primary work commute mode (Walk)	%	0.035	0.057	0.000	0.613
Median Income	USD	53744.2	19695	0.000	156250
Below Poverty	%	0.146	0.126	0.000	0.864
Less_High School	%	0.083	0.077	0.000	0.540
High School Degree	%	0.297	0.103	0.000	0.571
College degree	%	0.335	0.070	0.000	0.558
Bachelor degree	%	0.284	0.168	0.000	0.928
Unemployment	%	0.337	0.086	0.105	1.000
Number of vehicles	Count	4501.393	2403.197	0.000	19880
Number of Bars	Count	2.187	2.572	0.000	23.000

Table 3-1 shows the summary statistics of the processed dataset used to develop area-level CPMs. The explanatory variables presented as percentages were calculated from information provided by the TIGER/Line dataset. Vehicle Miles Travelled (VMT) and number of intersections are considered to be roadway information in the dataset. All other variables are extracted and calculated from the TIGER/Line dataset. Please note that all explanatory variables in the final dataset are continuous variables. Categorical variables were not generated for this pilot run analysis.

Three NB models were developed with the processed dataset to quantify the effects of explanatory variables on total crashes, speed-related crashes, and alcohol-related crashes in a census tract. The Variance Inflation Factor (VIF) was estimated for each model to check multi-collinearity. Any covariate with a VIF value greater than 5 were excluded from final models. Table 3-2 provides the summary of model coefficient estimates of area-level crash prediction models. Table 3-3 provides detailed model estimates and performance measures.

Table 3-2 Parameter Estimate Summary for Area-level Crash Prediction Model.

Variable Category	All Crashes	Speed-related Crashes	Alcohol-related Crashes
Intercept	Intercept (0.179)	Intercept (-3.044)	Intercept (-3.001)
Traffic and Trip	Log(VMT) (0.544) No. of intersections (1.285E-03) Car Trips (-0.454)	Log(VMT) (0.601) No. of intersections (1.41E-03)	Log(VMT) (0.434) No. of intersections (9.692E-04) Car Trips (-0.385)
Demographic Variables	Area (-2.815E-03) Percent Male (0.978) Male w/age <18 (-0.849) Percent White (-0.449) Median Age (-8.427E-03)	Area (-2.065E-03) Percent Land (0.427) Population Density (-3.081E-05) Percent Male (1.039) Per. Male <18 yrs (-1.458) Percent White (-0.563) Median Age (-8.409E-03)	Area (0.001496) Percent Land (-0.316) Population Density (-9.108E-06) Percent Male (1.556) Per. Male <18 yrs (-7.723E-03) Percent White (0.408) Median Age (-7.718E-03)
Socioeconomic Variables	Median Income (-5.207E-06) In Labor Force (0.846)	Less_High_School (0.724) In labor force (0.56)	Median Income (-5.998E-06) Less_High_School (1.915) In labor force (1.234) Housing (7.154E-05)

Table 3-3 Detailed Model Parameter Estimate for Area-Level CPM.

Parameters	Total Crash			Speed-related Crash			Alcohol-related Crash		
	Estimate	Std. Error	VIF	Estimate	Std. Error	VIF	Estimate	Std. Error	VIF
Intercept	1.521	0.269		-2.027	0.397		-1.607	0.279	
Log(VMT)	0.508	2.266E-02	1.739	0.612	2.516E-02	1.651	0.386	2.335E-02	1.879
No. of intersections	9.542E-04	1.474E-04	4.634	1.427E-03	1.315E-04	3.158	7.480E-04	1.333E-04	4.605
Percent car trips	-0.477	0.161	1.644	-0.380	0.190	1.783			
Area	-2.448E-03	2.828E-04	2.783	-2.221E-03	2.738E-04	2.239	-9.431E-04	2.590E-04	2.776
Population				-2.979E-	4.961E-	2.426	-8.321E-	3.998E-	2.169

Density				05	06		06	06	
Percent Male				1.006	0.420	1.233			
Percent White	-0.566	9.355E-02	2.697	-0.595	0.101	2.595	0.350	8.756E-02	2.364
Age <18	-1.199	0.300	2.214	-1.562	0.328	2.151	-1.045	0.266	1.869
Age >64							-1.564	0.276	1.938
Median HH Income	-4.096E-06	9.669E-07	2.141				-5.350E-06	9.499E-07	2.192
Per_less_high	0.702	0.265	2.478	1.129	0.286	2.350	1.584	0.256	2.532
Percent unemployed	-0.877	0.205	1.747	-1.035	0.217	1.596			
Total number of vehicles	2.974E-05	9.527E-06	3.128				3.749E-05	8.850E-06	3.212
Bar count	3.356E-02	5.568E-03	1.226				4.769E-02	5.004E-03	1.231
Theta	4.344	0.162		3.937	0.158		7.336	0.418	
Null Deviance	2614.8			3201.3			2873.8		
Residual Deviance	1451.8			1463.4			1445.4		
AIC	18025			13010			9235.1		

Table 3-4 illustrates the effects and comparisons of crash contributing factors between different crash types. The sign provided in the parenthesis indicates a positive or negative effect of the contributing factors on different types of crash occurrences.

Table 3-4 Potential Contributing Factors in Area-level Crash Occurrences.

Category	Total Crashes	Speed-related	Alcohol-related
Traffic Variable and Trip Pattern	• VMT (+) • No. of Intersections (+)	• VMT (+) • No. of Intersections (+)	• VMT (+) • No. of Intersections (+)
Travel Pattern	• Car Trip (-)	• Car Trip (-)	
Demographic Variable	• Area (-) • Percent White (-) • Age<18 (-) • Number of Vehicles (+) • Bar Count (+)	• Area (-) • Population Density (-) • Percent Male (+) • Percent White (-) • Age<18 (-)	• Area (-) • Population Density (-) • Percent White (+) • Age<18 (-) • Age>64 (-) • Number of Vehicles (+) • Bar Count (+)

Socioeconomic Variable	• Median Income (-) • Education less than High School Percentage (+) • Unemployment rate (-)	• Education less than High School (+) • Unemployment rate (-)	• Median Income (-) • Education less than High School (+)

3.4 Findings

The CPM results show that roadway, travel pattern, socioeconomic, and demographic variables were statistically significant in predicting total crashes and behavior-related crashes in a census tract. VMT and intersection density were both statistically significant in predicting total crash and behavior-related crashes. The parameter estimates of the roadway network-related variables are positive, meaning that an increase in any variable will increase the total number of crashes in a census tract overall, as well as the total number of speed-related crashes and alcohol-related crashes, specifically. The percentage of car trips is statistically significant in predicting all crashes and speed-related crashes, but is not significant in predicting alcohol-related crashes; this indicates that alcohol-related crashes are not dependent on the number of car trips made within a census tract. A higher percentage of car trips also indicates a more uniform traffic mix within a census tract. The coefficient estimate of car trips is negative with regard to total crashes and speed-related crashes; therefore, a more uniform traffic mix will decrease these types of crashes.

Among demographic features, total area, population density, number of cars, gender, and race-related variables, bar counts are associated statistically with crash frequency. The negative sign of the Area variable in the models for total crash, speed- and alcohol-related crashes means number of crashes decreases as area size increases. This relationship indicates that rural areas may have lower crash density because the size of rural census tract is usually larger. Population density can also be considered as a surrogate measure for area type, as more densely populated

areas represent urban areas. Population density is statistically significant in predicting both types of behavior-related crashes (alcohol-related and speed-related). The negative sign of the population density variable means fewer crashes may occur in more populated areas. This relationship indicates that both speed- and alcohol-related crashes occur more frequently in rural areas compared with urban areas. The positive coefficient of the percent male variable indicates that male drivers are statistically more prone to speed-related crashes than female drivers. The population composition of a census tract is represented by exploring the coefficient estimates of people less than 18 years of age and people more than 64 years of age. The negative coefficient estimate of the percentage of people less than 18 years of age means that fewer crashes occurred in areas with more people who are younger than 18. This relationship is reasonable because the percentage of licensed drivers or vehicle owners is the lowest among young people (<18) compared with other age groups. The percentage of people older than 64 years of age is only significant in predicting alcohol-related crashes. A census tract with older people usually has less alcohol-related crashes.

Among socioeconomic variables, median income, education status, and employment status were found to be statistically significant. Higher income indicates the community is more educated. Model parameter estimates also show that both total crashes and alcohol-related crashes in a census tract decrease with an increase in median income. Interestingly, median income was not statistically significant in predicting speed-related crashes, implying that a person's income or socioeconomic status does not predict speeding behaviors. A similar conclusion can be made with regard to education status. The coefficient estimates for the "less than high school degree" variable implies that all crashes, speed-related crashes, and alcohol-related crashes increase when the percentage of uneducated people within a census tract

increases. The estimated coefficient sign is consistent for all modeled crash types. Similarly, employment status (percent unemployed) reflects the status of household median income and education levels. The total number of vehicles in a census tract can be used to present trips generated from a census tract. The estimated coefficients for the total number of vehicles is positive for both total crashes and alcohol-related crashes, indicating that a census tract's crash count increases with an increase in traffic. Speed-related crashes, however, do not depend on the number of vehicles in a census tract.

3.5 Summary and Recommendations

The area-level crash frequency modeling results can help transportation agencies monitor area-level safety, identify major crash determinants, and evaluate safety programs and investment decisions. These results can be used to identify communities with a high risk of crashes and develop effective countermeasures to increase safety.

The area-level CPM analysis provides an opportunity to collect new data items for more rigorous crash analysis. The "bar count" variable collected from Business Analyst was available only for southern Wisconsin, but even with this limitation, the bar count within a census tract was found to be statistically significant in predicting total crashes and alcohol-related crashes. The CPM results also indicate that socioeconomic status and demographic variables are related to all types of crashes. Exploration and incorporation of these variables could provide a better understanding of safety issues within a census tract and help to develop effective safety countermeasures.

3.6 References

Girasek, D.C., Taylor, B., 2010. An exploratory study of the relationship between socioeconomic status and motor vehicle safety features. Traffic injury prevention 11 (2), 151-155.

Levine, N., Kim, K.E., Nitz, L.H., 1995. Spatial analysis of honolulu motor vehicle crashes: Ii. Zonal generators. Accident Analysis & Prevention 27 (5), 675-685.

Miaou, S.-P., Song, J.J., Mallick, B.K., 2003. Roadway traffic crash mapping: A space-time modeling approach. Journal of Transportation and Statistics 6, 33-58.

NHTSA., 2015. 2014 motor vehicle crashes: Overview. Traffic safety facts research note 2015, 1-9.

NHTSA, 2008. National motor vehicle crash causation survey: Report to congress. National Highway Traffic Safety Administration Technical Report DOT HS 811, 059.

Wang, Y., Kockelman, K.M., 2013. A poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. Accident Analysis & Prevention 60, 71-84.

Washington, S., 2006. Incorporating safety into long-range transportation planning Transportation Research Board.

Washington, S., Metarko, J., Fomunung, I., Ross, R., Julian, F., Moran, E., 1999. An inter-regional comparison: Fatal crashes in the southeastern and non-southeastern united states: Preliminary findings. Accident Analysis & Prevention 31 (1), 135-146.

Wisconsin Department of Transportation (WisDOT), 2017. Preliminary review of 2015 crash fatality trends. Wisconsin Department of Transportation (WisDOT), Wisconsin.

Xu, P., Huang, H., Dong, N., Abdel-Aty, M., 2014. Sensitivity analysis in the context of regional safety modeling: Identifying and assessing the modifiable areal unit problem. Accident Analysis & Prevention 70, 110-120.

Chapter 4 Estimating the Effect of Unobserved Behavior Variables: A Random Parameter Mixed Distribution Approach

This chapter discusses segment-level CPM development in the 2nd tier of the proposed 3-tier approach. The 2nd tier modeling approaches are developed to identify variables available at segment level that are correlated with crash occurrence with an emphasis on driver behavior. This chapter specifically discusses the development of a modeling technique to account for unobserved driver behaviors in CPM. Unavailability of driver behavior information in crash dataset can cause unobserved heterogeneity induced overdispersion issue in crash dataset. A mixed distribution random parameter modeling technique is developed to address unobserved heterogeneity in crash dataset. The modeling results are compared with traditional models to identify superior model based on prediction results and model inferences.

4.1 Introduction

A roadway crash is a multifaceted event involving circumstances such as highway geometry, traffic exposure, contextual factors, driver characteristics, vehicle factors, as well as the interactions among them. Identifying key crash risk factors and understanding their effects is critical to finding cost-effective strategies for the prevention and reduction of traffic crashes and their severities. Typically, a quantitative safety analysis is performed through descriptive

statistics to identify patterns and regression models are used to identify factors associated with crashes. Once the association is properly established, additional insights about the crash can be revealed and evaluated. Lastly, the mean crash count can be estimated by mathematical formulation (Mitra and Washington 2012).

Crash data are often characterized by the existence of a large sample variance compared with the sample mean[3] (Lord et al. 2005, Mitra and Washington 2007). Extensive research has been devoted to modeling and analyzing this type of crash dataset (Lord and Mannering 2010, Mannering and Bhat 2014, Mannering et al. 2016). A notable accomplishment resulting from this research is the application of the negative binomial (NB) model in analyzing crash frequency data. The NB model can handle data over-dispersion by assuming a gamma distribution for the exponential function of the disturbance term in the Poisson mean. However, recent studies have pointed out that with a heavy-tailed crash dataset, the NB model can produce biased parameter estimates (Zou et al. 2015, Shirazi et al. 2016). A heavy-tailed distribution is a statistical phenomenon that occurs when sample observations have a few very high crash counts with preponderant zero observations; this shifts the overall sample mean to near zero (Shirazi et al. 2016). Failure to account for data over-dispersion could lead to biased and inconsistent parameter estimates, which in turn causes researchers to make erroneous inferences from models and can also lead to inaccurate crash prediction values.

The mixed model is a well-known methodology used to incorporate heterogeneity into statistical analysis. Safety literature shows that mixed distribution NB models expanded the linear mixed model for continuous responses to discrete responses (e.g., crash count) by

[3] In a statistical term, the sample data is over-dispersed when the variance is greater than the mean. Data over-dispersion is often caused by unobserved data heterogeneity due to unobserved, unavailable, or unmeasurable variables that are important to explain model responses.

incorporating correlated non-normally distributed outcomes. Several mixed NB models have been proposed, including the NB-Lindley (NB-L), NB-Generalized Exponential (NB-GE), and NB-Dirichlet process (NB-DP) generalized linear models (GLMs) (Geedipally et al. 2012, Vangala et al. 2015, Rahman Shaon and Qin 2016, Shirazi et al. 2016). The advantage of using a mixed model is that it adds a mixed distribution to account for extra variance in the crash data which is caused by preponderant zero crash responses and/or a heavy-tail of crash counts (Shirazi et al., 2016). The underlying hypothesis is that the crash datasets are comprised of distinct subpopulations which have different probabilistic distributions. On the other hand, accessing all data items associated with the likelihood of crash occurrence and/or injury severity is nearly impossible. Omitting important variables causes data heterogeneity which adds extra variation in the effect of explanatory variables. Random parameters (RP) models can account for unobserved heterogeneity by allowing the parameter of variables to vary from one observation to the next and by estimating the unbiased mean effect of explanatory variables (Mannering et al. 2016). Therefore, incorporating both random parameters and mixed probabilistic distributions within a single model can be a viable alternative for handling crash data with high over-dispersion and unobserved heterogeneity.

The objective of this study was to develop and document an RPNB model with Lindley mixed effect for heterogeneous count data that features an excess number of zero responses and/or a heavy-tail. The proposed RPNB-L model was developed in a Bayesian hierarchical framework that is expanded from fixed-coefficients NB-L GLM (Geedipally et al. 2012, Rahman Shaon and Qin 2016). The study utilized two crash datasets, one from Indiana and one from South Dakota, to calibrate the parameters in RPNB-L GLM. The datasets were characterized by

over-dispersion with a very high percentage of zero responses and a heavy-tail. The model fitting and the modeling results were compared with the traditional NB, RPNB and NB-L models.

4.2 Literature Review

The existence of preponderant zero crash sites with a heavy tail can create highly over-dispersed data. The NB distribution has been used to model crash frequencies for decades because it can handle data over-dispersion, a unique attribute of crash frequency data. However, some studies have noted that the NB distribution cannot handle over-dispersion caused by a heavy tail in the crash data (Guo and Trivedi 2002, Park et al. 2010, Zou et al. 2015, Shirazi et al. 2016). Guo and Trivedi (2002) noted that a negligible probability is usually assigned to higher crash counts in the NB model during the modeling of highly over-dispersed data with a heavy tail. Lord et al. (2005) pointed out that over-dispersion arises from the actual nature of the crash process. One limitation of the NB distribution is that it assumes that only one underlying process affects the likelihood of crash frequency (Shankar et al. 1997).

A mixture model is a very popular statistical modeling technique that is often used to account for data over-dispersion because it is flexible and extensible (Shankar et al., 1997; Aguero-Valverde and Jovanis, 2008; Lord et al., 2008; Lord and Geedipally ,2011; Geedipally et al., 2012; Cheng et al., 2013; Mannering and Bhat, 2014; Rahman Shaon and Qin, 2016; Shirazi et al., 2016). The mixture model is comprised of a convex combination of a finite number of different distributions. The NB-L GLM is a mixture of the NB and Lindley distribution in which the Lindley distribution itself is a mixture of two gamma distributions (Lindley, 1958). The NB-L GLM was recently introduced to model crash frequency data (Geedipally et al., 2012; Rahman Shaon and Qin, 2016). The count data mixture model works well when the dataset contains a

large number of zero responses, is skewed, or is highly dispersed. Zamani and Ismail showed that the NB-L distribution provides a better fit compared to the Poisson and NB models when there is a large probability of crash frequency at zero (Zamani and Ismail, 2010). Lord and Geedipally (2011) applied the NB-L distribution to estimate the predicted probability and frequency of crashes using both simulated and observed crash data. The authors concluded that the NB-L distribution can handle crash datasets with preponderant zero crash observations. Recently, Rahman Shaon and Qin (2016) evaluated the effect of lane and shoulder width on over-dispersed crash data using the NB-L model. The authors found that the NB-L GLM performed better than a traditional NB model when working with crash data characterized by preponderant zero responses, and that the core strength of the NB model was maintained. Overwhelmingly positive results have been reported from applying the NB-L model with many different data sources (Zamani and Ismail, 2010; Lord and Geedipally, 2011; Geedipally et al., 2012; Hallmark et al., 2013; Xu and Sun, 2015; Rahman Shaon and Qin, 2016). Although the Lindley distribution has a closed form (Zamani and Ismail, 2010), the Lindley distribution cannot be mixed with the NB distribution in the context of GLM because it is not available in any standard statistical software (e.g. R, SAS, SPSS). Researchers have used the Bayesian method to create the hierarchical structure that is needed to estimate the parameters of NB-L in the context of GLM (Geedipally et al., 2012; Rahman Shaon and Qin, 2016).

A mixture model is a very popular statistical modeling technique that is often used to account for data over-dispersion because it is flexible and extensible (Shankar et al. 1997, Aguero-Valverde and Jovanis 2008, Lord et al. 2008, Lord and Geedipally 2011, Geedipally et al. 2012, Cheng et al. 2013, Mannering and Bhat 2014, Rahman Shaon and Qin 2016, Shirazi et al. 2016). The mixture model is comprised of a convex combination of a finite number of

different distributions. The NB-L GLM is a mixture of the NB and Lindley distribution in which the Lindley distribution itself is a mixture of two gamma distributions (Lindley 1958). The NB-L GLM was recently introduced to model crash frequency data (Geedipally et al. 2012, Rahman Shaon and Qin 2016). This count data mixture model works well when the dataset contains a large number of zero responses, is skewed, or is highly dispersed. Zamani and Ismail showed that the NB-L distribution provides a better fit compared to the Poisson and NB models when there is a large probability of crash frequency at zero (Zamani and Ismail 2010). Lord and Geedipally (2011) applied the NB-L distribution to estimate the predicted probability and frequency of crashes using both simulated and observed crash data. The authors concluded that the NB-L distribution can handle crash datasets with preponderant zero crash observations. Recently, Rahman Shaon and Qin (2016) evaluated the effect of lane and shoulder width on over-dispersed crash data using the NB-L model. The authors found that the NB-L GLM performed better than a traditional NB model when working with crash data characterized by preponderant zero responses, and that the core strength of an NB model was maintained. Overwhelmingly positive results have been reported from applying the NB-L model with many different data sources (Zamani and Ismail 2010, Lord and Geedipally 2011, Geedipally et al. 2012, Hallmark et al. 2013, Xu and Sun 2015, Rahman Shaon and Qin 2016). Although the Lindley distribution has a closed form (Zamani and Ismail 2010), the Lindley distribution cannot be mixed with the NB distribution in the context of GLM because it is not available in any standard statistical software (e.g. R, SAS, SPSS). Thus, a hierarchical structure is needed to estimate the parameters of NB-L in the context of GLM. In previous work, researchers used the Bayesian interface to implement this model (Geedipally et al. 2012, Rahman Shaon and Qin 2016).

The existing crash dataset contains only a fraction of the potential variables that can significantly affect the likelihood of crash occurrence (Mannering et al. 2016). Unobserved heterogeneity in a regression model occurs when important covariates have been omitted during the data collection process. The influence of these variables is therefore not accounted for in the analysis. Unobserved heterogeneity in traditional NB models is usually considered to be random errors because the effect of each covariate is restricted to be the same across all observations; this causes even more dispersion problems. Such modeling strategies can cause serious model specification problems and may result in a variation of the estimated effect of observed covariates (Mannering et al. 2016). An overview of the potential for heterogeneity in driver behavior was highlighted by Mannering et al. (2016). For example, the research found that varying lane and shoulder widths may have an impact on the likelihood of a crash event, but that these effects can vary among observations due to time-varying traffic, weather conditions, and/or the driver's reaction, all of which are not available for model development. Ignoring heterogeneous effects in explanatory variables leads to biased parameter estimates and therefore inaccurate conclusions (Mannering et al. 2016).

Research studies have been devoted to the task of obtaining unavailable but necessary data by utilizing statistical and econometric models[4] to account for unobserved heterogeneity. The RP modeling approach (Mannering et al. 2016)[5] has gained considerable attention for its use with crash count data. RP modeling addresses data heterogeneity by allowing the model

[4] Refer to Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research 11, 1-16. for the list of methodological alternatives to account for unobserved heterogeneity.

[5] Finite mixture models, which are a type of latent variable models or latent class models was also explored as another alternative to account for unobserved heterogeneity in literature (Peng and Lord 2011, Shirazi et al. 2016). This approach expresses the overall distribution of one or more variables as a mixture of a finite number of component distributions which prescribes the observations from different groups, subpopulations or latent classes, each can be represented by a probability distribution function. Together, a finite mixture model can handle various distributions for different sub-populations in the target dataset.

parameters to vary from observation to observation. The parameter is treated as a random variable whose probability distribution is usually is defined by the modelers. Anastasopoulos and Mannering (2009) introduced the RPNB model to account for data heterogeneity caused by explanatory variables and other unobserved factors. Crash data studies that have applied the RPNB model have found a significant improvement in the statistical model fit (El-Basyouny and Sayed 2009, Garnowski and Manner 2011, Venkataraman et al. 2011, Chen and Tarko 2014, Buddhavarapu et al. 2016).

In summary, the RP model incorporates the effect of unobserved variables by allowing model parameters to vary from observation to observation, but this method is susceptible to observations generated from different data sources. The mixed model also did not resolve the issue of omitted variables that could affect the likelihood of crashes. However, joint mixture distributions and random parameters can both identify groups of observations with homogeneous variable effects within each group and can allow for the consideration of varying parameters so that the effects of unobserved variables are included (Peng and Lord 2011). Buddhavarapu et al. (2016) developed a spatial finite-mixture RPNB model that relaxed the distributional assumptions of RP. The study outlined in this paper pursued the same goal by utilizing the strengths and flexibility of both methods. Although the NB-L does not literally generate multiple homogeneous groups, it offers flexibility to account for skewness in crash observation which occurs when preponderant zero crash sites with a heavy tail are present. The unobserved heterogeneity in explanatory variables is assumed to be addressed when estimated parameters are allowed to vary across observations in NB-L.

4.3 NB-Lindley GLM

The NB-L distribution re-parameterized in a GLM context can be formulated in Eq. (1) (Geedipally et al. 2012, Rahman Shaon and Qin 2016):

$$P(Y = y|\, \mu, \phi, \theta) = \int NB(y; \phi, \varepsilon\mu) Lindley(\varepsilon; \theta)\, d\varepsilon \quad (1)$$

In Equation 1, *f (u; a, b)* is the distribution of the variable u, with parameters *a* and *b*. Following this explanation, given ε, the variable *Y* follows a NB distribution with a mean and inverse-dispersion parameter of $\varepsilon\mu$ and ϕ (ϕ =1/α), respectively. The variable ε follows a Lindley distribution with parameter θ.

If we assume that the crash count follows the *NB-L (y; μ, ϕ, θ)* distribution, the mean response function can be structured as follows (Geedipally, Lord & Dhavala, 2012; Rahman Shaon & Qin, 2016):

$$E(Y = y) = \mu \times E(\varepsilon) \quad (2)$$

where, $\mu = e^{\beta_0 + \sum_{i=1}^{q} \beta_i X}$ and $E(\varepsilon) = \frac{\theta+2}{\theta(\theta+1)}$

By replacing the value of μ and E(ε), the mean response function can be written as follows:

$$E(Y) = \left(e^{\beta_0 + \sum_{i=1}^{q} \beta_i X}\right) \times \frac{\theta+2}{\theta(\theta+1)} = e^{\left\{\beta_0 + log\left[\frac{\theta+2}{\theta(\theta+1)}\right]\right\} + \sum_{i=1}^{q} \beta_i X} = e^{\beta_0' + \sum_{i=1}^{q} \beta_i X} \quad (3)$$

where, $\beta_0^{'} = \beta_0 + log\left[\frac{\theta+2}{\theta(\theta+1)}\right]$

The Lindley distribution is a mixture of two gamma distributions. Therefore, the Lindley distribution can be rewritten as (Geedipally, Lord & Dhavala, 2012; Rahman Shaon & Qin, 2016):

$$\varepsilon \sim \frac{1}{1+\theta} Gamma(2,\theta) + \left(1 - \frac{1}{1+\theta}\right) Gamma(1,\theta) \quad (4)$$

which can be restructured as:

$$\varepsilon \sim \sum Gamma(1+z,\theta)\ Bernoulli\left(z; \frac{1}{1+\theta}\right) \quad (5)$$

The NB-L GLM can be written as the following multi-level hierarchical structure using Eqs. (1)-(5):

$$P(Y = y; \phi, \mu|, \varepsilon) = NB(y; \phi, \varepsilon\mu)$$

$$\mu = e^{\beta_0 + \sum_{i=1}^{q} \beta_i X}$$

$$\varepsilon \sim Gamma(\varepsilon; 1+Z, \theta)$$

$$Z \sim Bernoulli\left(z; \frac{1}{1+\theta}\right) \quad (6)$$

The above formulation is similar to a Generalized Linear Mixed Model (GLMM) (Booth, Casella, Friedl, & Hobert, 2003), where the mixed effects follow the Lindley distribution. In this modeling structure, the crash count follows an NB distribution which is conditional on a site-specific frailty term. The site-specific frailty term ε was assumed in order to accommodate extra variance in the crash data. The Lindley mixed effect, in hierarchical terms, is formulated by

adding a site-specific offset (constant) term in the log-transformed domain of the mean response of the NB distribution.

The specification of prior distributions for the parameters is necessary for obtaining the Bayesian estimate. Prior distributions are meant to describe a prior knowledge about the parameters of interest. The site-specific frailty term follows a non-informative prior of the gamma distribution. The shape parameter in the gamma distribution follows a Bernoulli distribution with a probability parameter of $1/(1 + \theta)$. A weakly informative prior may yield a model output in which the parameter estimate for the Lindley distribution may contribute more than the NB distribution. The Markov chain Monte Carlo (MCMC) can suffer from poor mixing due to the correlation between the intercept and the site-specific frailty term. According to the literature, prior knowledge should be used to formulate the informative priors (if known) (Bedrick, Christensen, & Johnson, 1996; Schlüter, Deely, & Nicholson, 1997). A prior should be used to ensure $E(\varepsilon) = 1$ in order to limit the contribution of the mixed effect from the Lindley distribution. Geedipally, Lord, and Dhavala (2012) suggested using a prior for $1/(1 + \theta)$ that follows a beta distribution. The reasonable choice for prior distribution is Beta (n/3, n/2), where n is the total observations (Geedipally et al., 2012).

4.4 Random Parameters NB-Lindley GLM

Let x_{ij} denote the j-th covariate associated with i-th site. In a RP model, the coefficient β_{ij} is assumed to be random, and is written as:

$$\beta_{ij} = b_j + w_{ij} \tag{7}$$

where b_j denotes the fixed term (the mean parameter estimate), and w_{ij} denotes the random term. The random term is assumed to follow a predefined distribution such as a normal distribution with a mean equal to zero and a variance of σ^2. The random parameter β_{ij} should be used if the standard deviation of the random term w_{ij} is significantly different from 0 (under the frequentist approach; more discussion on that is provided below); otherwise, a fixed parameter or coefficient should be applied over all the individual observations (Anastasopoulos and Mannering 2009, El-Basyouny and Sayed 2009). Considering the above parameterization, the probability mass function (pmf) for RPNB model can be written as:

$$p(y_i) = \frac{\Gamma(\phi+y_i)}{\Gamma(\phi)\Gamma(y_i+1)}(1-p_i)^{y_i}p_i^{\phi}; \quad \phi > 0, 0 < p_i < 1 \tag{8}$$

where, $p_i = \frac{\phi}{\mu_i+\phi}$

Technically, the NB-L GLM itself can also be considered as a random parameters model because the intercept (or the mixed effect) that follows the Lindley distribution varies from observation to observation. The coefficients of explanatory variables are considered as random variables when developing a full RPNB-L GLM. In this paper, the NB-L model can be referred to as RPNB-L if the coefficient of any covariates can be considered a random variable. Recalling the hierarchy developed for NB-L GLM, the RPNB-L GLM can be written as the following multi-level structure:

$$P(y_{i;}\phi, \mu_i|\varepsilon_i) = NB(y_i; \phi, \varepsilon_i\mu_i)$$

$$\log(\mu_i) = \beta_0 + \sum_{j=1}^{q}\beta_{ij}x_{ij}$$

$$\varepsilon_i \sim Gamma(\varepsilon; 1 + z_i, \theta)$$

$$z_i \sim Bernoulli(z; \frac{1}{1+\theta})$$

$$\beta_{ij} = \beta_j + w_{ij}$$

$$w_{ij} \sim Normal(0, {\sigma_j}^2) \quad (9)$$

The MCMC chains in RPNB-L may suffer from poor mixing due to potential correlations between the intercept and regression coefficients, especially since both vary across observations. One simple way to overcome this difficulty is to center or standardize covariates before using them in the model. The traditional way of standardizing a covariate can be written as follows:

$$x_{ij}^* = \frac{x_{ij} - m_j}{s_j} \quad (10)$$

where,

i = 1,2…., n denotes the number of observations;

j = 1,2…., q denotes the number of covariates; and

m_j and s_j are the mean and standard deviation of j-th covariate.

The standardized estimated coefficients need to be transformed back to the original scale after convergence, for ease of interpretation and inference. The following formulas describe the transformation (Gelfand et al. 1995):

$$\beta_1 = \frac{\beta_1^*}{s_1}$$

•••

$$\beta_q = \frac{\beta_q^*}{s_q}$$

$$\beta_0 = \beta_0^* - \sum_{i=1}^{q} \frac{\beta_q m_q}{s_q} \quad (11)$$

Where, β_q^* is the standardized coefficient and β_q is the transformed coefficient in the original scale of the covariate.

The current formulation of the random part w_{ij} is defined with a prior that follows a normal distribution with a zero mean value. However, even though the prior is considered to have a mean value of zero for w_i, the posterior mean of the parameter will not necessarily be zero. Hence, this causes a conflict with the fixed effect parameter estimate of β which results in poor mixing in MCMC chains, and identifiability issues in parameter estimates will therefore occur. A simple but effective method of centering the fixed effect parameter in the mean of the defined random coefficient can help to overcome this issue. The random coefficient definition in the model can be structured as:

$$\beta_{ij} \sim Normal\left(\beta_j, \sigma_j^2\right)$$

$$1/\sigma_j^2 \sim Gamma(0.01, 0.01) \quad (12)$$

Previous literature explored several distributions such as normal, lognormal, uniform, triangular, gamma etc. in Equation 15. The normal distribution was found to provide the best statistical fit (Li et al. 2008, Anastasopoulos and Mannering 2009). Thus, normal distribution

was adopted for this study. A good mixing in the MCMC chains was achieved by using the above formulation.

4.5 Model Estimation

The RPNB-L model was formulated and estimated in a Bayesian framework using WinBUGS (Lunn et al. 2000). The traditional fixed-coefficients NB, the random parameters NB, the fixed-coefficients NB-L models were also implemented in a Bayesian framework for comparison purposes. A total of three (3) Markov chains were used in the model estimation process with 80,000 iterations per chain for each model. In order to reduce autocorrelation, a thinning factor of three (3) was used in WinBUGS. The first 25,000 iterations were discarded as burn-in samples. The remaining iterations were used for estimating the model coefficients. The Gelman-Rubin (G-R) convergence statistic and Monte Carlo (MC) error were used to verify that the simulation runs converged properly. In the analysis, the research team ensured that the G-R statistic was less than 1.1. Mitra and Washington (2007) suggested that convergence was achieved when the G-R statistic was less than 1.2. The MC error of each parameter estimate was tested to ensure it was less than 3 percent of the estimated posterior standard deviation.

It is important to note that the estimation of RP models in a Bayesian framework is somewhat different compared to the frequentist or Maximum Likelihood Estimate (MLE) approach. In a Bayesian framework, the RP approach provides additional modeling flexibility by adding another level of hierarchy in the model parameterization. The variance in the model parameters is assumed to come from unobserved data heterogeneity and is estimated by adding another level of hierarchy for the variance. Thus, unlike the MLE estimates, any parameter defined as random in a Bayesian framework will have a positive variance. In short, although the

parameters may have the same mean estimates, the identification of which variables are random will be completely different. The parameters will always be random in Bayesian models if the Bayesian hierarchical model is defined as such, but the variables in MLE are considered random only if they meet a specific statistical criterion (i.e., $\sigma^2 > 0$ at a 5% significance level for example). The goodness-of-fit (GOF) of the models under investigation is also influenced by this difference in parameters.

Marginal effects are used to determine the impact of each covariate on the expected mean value of the dependent variable [6]. The marginal effect represents the effect of a unit change in the independent variable on the expected mean of the dependent variable. The marginal effect can be estimated as $\frac{\delta \mu_i}{\delta x_{ik}} \times \frac{x_{ik}}{\mu_i} = \beta_{ik} x_{ik}$, where μ_i is the expected mean outcome in each modeling approach (Washington et al. 2010). In the case of RP models, it is important to note that the marginal effects were estimated considering variation in estimated model parameters. The parameter means for each site were estimated after the MCMC chains converged in WinBUGS, and then were used to estimate the marginal effect of each observation.

4.6 Data Description

The characteristics of the two datasets used in this study are described in this section, which is divided into two subsections. The first subsection summarizes the characteristics of the data collected at 338 rural interstate roadway segments in Indiana. The second subsection describes the characteristics and summary statistics of the data collected at rural two-lane two-way highways in South Dakota. Both datasets are highly dispersed and characterized by a heavy tail,

[6] The marginal effects were estimated for each observation and the mean value of all marginal effects are represented in Table 4 and Table 6. It is important to note that, the marginal effect for each covariate significantly varies from site-to-site.

and both contain several variables which were used in model development to minimize the omitted-variable bias problem that can plague the development of crash prediction models (Lord and Mannering 2010).

4.6.1 Indiana Data

The Indiana dataset contains crash, roadway geometry, and traffic data collected over a five-year period (from 1995 to 1999) on 338 rural interstate roadway segments in the state of Indiana. The Indiana dataset has been used in several previous research studies, such as Washington et al. (2010), Geedipally et al. (2012). and Shirazi et al. (2016). In this dataset, 120 out of the 338 highway segments did not have any reported crashes over the five-year period (~36% are 0s). Table 4-1 presents the summary statistics of the variables used for developing the models in this study.

Table 4-1 Summary Statistics for the Indiana Dataset.

Variables	Description	Mean	Standard Deviation	Minimum	Maximum
Crash	Number of Crashes in 5 years	16.973	36.297	0	329
Log(ADT)	Logarithm of Average daily traffic over the 5 years	10.036	0.681	9.153	11.874
Friction	Minimum friction reading in the road segment over the 5-year period	30.514	6.674	15.900	48.2
Pavement	Pavement surface type (1 if asphalt, 0 if concrete)	0.769	0.422	0	1
Median Width	Median width in feet	66.984	34.169	16	194.7
Barrier	Presence of median barrier (1 if present, 0 if absent)	0.160	0.367	0	1
Rumble	Interior rumble strips	0.725	0.447	0	1

Length	Segment length in miles	0.710	1.225	0.009	4.054

4.6.2 South Dakota Data

The South Dakota dataset is characterized by a preponderant number of zero responses and a heavy tail. In this dataset, the roadway geometric characteristics and traffic data elements were collected from the South Dakota Department of Transportation (SDDOT). Multiple event tables from the SDDOT Roadway Inventory System (RIS) were combined to generate homogeneous segments. Crash data between 2008 and 2012 were spatially joined with the roadway data according to their spatial distance. The original dataset for rural two-lane two-way highway segments in South Dakota contains 16,827 segments. A sample of 10,000 observations from the total segments was used to evaluate the performance of the RPNB-L model in this study. The rural two-lane two-way segment database was previously used by Rahman Shaon and Qin (2016) to evaluate the performance of the NB-L model. The summary statistics of the sample data from South Dakota data are provided in Table 4-2.

Table 4-2 Summary Statistics for the South Dakota Dataset.

Variable	Definition	Mean	Standard Deviation	Min	Max
Crash	Count of Crashes	0.614	2.493	0	88
AADT	Annual Average Daily Traffic	917.933	913.790	45.00	21396.00
Segment Length	Segment Length in Miles	0.383	1.035	0.010	16.494
Speed Limit	Posted Speed Limit	57.273	10.712	20.00	65.00
Radius	Radius of curvature in miles	0.081	0.184	0.00	1.084
Lane Width	Lane width in feet	12.955	2.098	9.00	24.00
Shoulder Width	Shoulder width in feet	3.046	2.553	0.00	15.00
Vertical Grade	Yes		21.58%		
	No		78.42%		

In the South Dakota dataset, 78 percent of the 10,000 sample segments did not experience any crashes during the study period. The mean and standard deviation of the crash count for the 10,000 sample observations are equal to 0.614 and 2.493, respectively. Due to preponderant zero crash sites, the estimated skewness of the crash count was equal to 11.624, which shows that the crash count is highly skewed to the right. Annual average daily traffic (AADT), segment length, lane width, shoulder width, speed limit and radius of curvature of the horizontal curve were used as continuous explanatory variables to model crash data. Vertical grade is the only binary variable (1 if Yes, 0 if No) included in the model.

4.7 Results and Discussions

Detailed modeling results from the application of the RPNB-L GLM to both Indiana and South Dakota datasets are presented in this section. The first subsection that follows documents the modeling results for the Indiana dataset. The second subsection provides the modeling results for the South Dakota dataset. The performance of the RPNB-L model was compared to the NB, RPNB, and the NB-L GLMs for both datasets.

4.7.1 Indiana Data Results

Table 4-3 and Table 4-4, respectively, summarize the modeling results and the estimated marginal effects for the Indiana dataset. The segment length variable was considered as an offset variable in all modeling approaches, as developed in previous studies that utilized this dataset (listed above). Therefore, it is assumed that the number of crashes will increase linearly as the segment length increases. In Table 4-3, the results of the RPNB-L model were compared to the fixed and random parameters NB and the fixed parameters NB-L model. In all models, the

estimated 95 percent marginal posterior credible intervals for all coefficients did not include zero. Hence, it can be concluded that all coefficients are statistically significant at a 5 percent significance level. In this section, only the modeling results for the application of the RPNB-L GLM are discussed. Anastasopoulos and Mannering (2009) and Geedipally et al. (2012) provide further discussions on the parameter estimates for the random parameters NB and the fixed parameters NB-L, respectively.

Table 4-3 Modeling Results for the Indiana Dataset.

Parameters	NB		RPNB		NB-L		RPNB-L	
	Value	Std. Dev.	Value	Std. Dev.	Value	Std. Dev.	Value	Std. Dev.
Parameter Mean								
Intercept	-4.449	0.067	-5.486	0.035	-3.947	0.162	-4.443	0.206
Log(ADT)	0.689	0.133	0.816	31.750	0.651	0.145	0.717	0.231
Friction	-0.027	0.011	-0.029	0.133	-0.027	0.012	-0.032	0.015
Pavement	0.422	0.189	0.588	0.012	0.445	0.210	0.605	0.281
Median Width	-0.005	0.002	-0.012	0.240	-0.006	0.002	-0.012	0.004
Barrier	-3.031	0.308	-6.614	0.003	-3.282	0.338	-6.152	0.898
Rumble	-0.405	0.186	-0.288	0.437	-0.404	0.207	-0.329	0.260
$\alpha = 1/\phi$	0.950	0.122	0.137	0.035	0.239	0.083	0.128	0.028
θ					1.464	0.180	1.414	0.173
Std. Deviation of Random Parameters								
Log(ADT)			0.302	0.172			0.232	0.137
Friction			0.057	0.011			0.056	0.011
Pavement			0.326	0.216			0.291	0.200
Median Width			0.028	0.003			0.028	0.003
Barrier			2.390	0.399			1.925	0.709
Rumble			0.379	0.242			0.310	0.183
Model Performance								
Dbar	1891.93		1481.09		1585.93		1422.70	
Dhat	1883.01		1296.86		1469.51		1276.00	
pD	8.92		184.22		116.41		146.30	
DIC	1900.84		1665.31[†]		1702.34		1569.00	

MAD[7]	6.92	6.90	6.88	6.71

Note: † With the MLE RPNB, only three variables (logarithm of ADT, presence of median barrier and interior rumble strips) were found to be random. This increased the Deviance Information Criterion or DIC to 1736.

The parameter mean for the traffic flow variable was estimated using the RPNB-L model to be less than one, indicating that the crash risk increases at a decreasing rate as the value of the traffic flow variable increases. A similar or consistent trend was observed for all other modeling approaches. The estimated marginal effect of the traffic flow variable also indicates that this variable has a positive influence on crash occurrence. Although the magnitude of coefficient can vary from site-to-site using the RPNB-L GLM, all estimated coefficients for the traffic flow variable have a value that is greater than zero.

The sign of the parameter mean estimates for both the roadway geometry and pavement-related variables are consistent with those found in Geedipally et al. (2012) using the same dataset. In this study, the RPNB-L helps to provide more details about the parameter estimates by combining the RP structure with the NB-L framework. The friction variable, which represents the minimum friction reading on the road segment over the five-year period, shows that the majority of sites (71.6 percent of normal density function) have estimated model coefficients with a value of less than zero while the rest of the sites have a coefficient that is greater than zero; this indicates that the friction variable has a mixed (both positive and negative) effect on crash risk. The marginal effect illustrates that the overall impact of the friction variable has a decreasing effect on crash risk. A similar pattern can also be observed with the median width variable. In this case, 66.6 percent of the estimated coefficients have a negative value while the

[7] Mean Absolute Deviance (MAD) provides a measure of the average miss-prediction of the model which can be estimated as $\frac{1}{n}\sum_{i=1}^{n}|Predicted\ value - Observed\ value|$. A value close to 0 suggests that, on average, the model predicts the observed data well.

rest are positive. More than 98 percent of the normal density function for pavement type has a value greater than zero with an estimated parameter mean of 0.422, meaning a change in pavement type from concrete to asphalt almost always increases the probability of a crash. A similar observation can also be obtained for the median barrier variable, which supports the effect of the median barrier variable as observed by Anastasopoulos and Mannering (2009).

Table 4-4 Average marginal effects for the Indiana Dataset.

Variables	Model			
	NB	RPNB	NB-L	RPNB-L
Log(ADT)	6.915	8.189	6.533	7.537
Friction	-0.812	-0.897	-0.824	-0.896
Pavement	0.325	0.452	0.343	0.578
Median Width	-0.351	-0.771	-0.412	-0.785
Barrier	-0.484	-1.057	-0.524	-1.181
Rumble	-0.293	-0.209	-0.293	-0.378

4.7.2 South Dakota Data Results

The model parameter estimates and marginal effects of the covariates for the South Dakota data are provided in Table 4-5 and Table 4-6, respectively. The first part of Table 5 provides the estimates of the parameter means, and the second part of the table provides the estimated standard deviation of the random parameters. Unlike the model development for the Indiana data, the segment length variable was defined as a random parameter rather than as an offset. All covariates were also defined as random parameters in both RPNB and RPNB-L models. The estimated standard deviation of all random parameters was found to be statistically significant at a 5 percent significance level.

The parameters mean for the lane width variable is not statistically significant at a 5 percent significance level when the NB distribution is used, as indicated in Table 4-5. Yet, for

the purpose of comparison between different models used in the analysis, this variable was kept in the model. The parameters mean for lane width did become significant when more advanced modeling alternatives (i.e.: RPNB, NB-L, and RPNB-L) were applied to this dataset. In addition, the results in Table 4-5 indicate that the parameters mean for the shoulder mean variable is not significant for all modeling approaches; however, since the standard deviation of the parameters is significant, this variable was kept in the model. The location of the mean of the coefficient distribution is not necessarily critical as long as the likelihood function improves with the significant standard deviation of the parameters (Anastasopoulos and Mannering 2009). While the parameters mean for all explanatory variables have a similar sign in all applied models, the magnitude of the estimates is different. Interestingly, the standard deviations of parameters for lane width and shoulder width are both statistically significant at a 5 percent significance level. The parameters mean is significant at a 5 percent confidence level for all other variables.

Table 4-5 Modeling Results for the South Dakota Dataset.

Parameters	NB		RPNB		NB-L		RPNB-L	
	Value	Std. Dev.	Value	Std. Dev.	Value	Std. Dev.	Value	Std. Dev.
Parameter Mean								
Intercept	-7.609	0.027	-7.879	0.052	-7.546	0.038	-7.676	0.043
log(AADT)	0.751	0.031	0.744	0.032	0.754	0.031	0.738	0.032
Segment Length	0.674	0.020	0.745	0.026	0.658	0.018	0.740	0.026
Speed Limit	0.025	0.003	0.033	0.003	0.026	0.003	0.030	0.003
Lane Width	***-0.006***	***0.011***	-0.037	0.013	-0.010	0.009	-0.026	0.013
Shoulder Width	***-0.001***	***0.010***	***-0.009***	***0.011***	***-0.002***	***0.010***	***-0.002***	***0.012***
Radius	-0.501	0.129	-0.564	0.124	-0.516	0.131	-0.506	0.121
Vertical Grade	-0.992	0.073	-1.389	0.133	-1.013	0.073	-1.066	0.089
$\alpha = 1/\phi$	1.228	0.063	0.406	0.083	0.260	0.049	0.114	0.014
θ					1.501	0.033	1.495	0.034
Std. Deviation of Random Parameters								
log(AADT)			0.317	0.057			0.121	0.049
Segment			0.235	0.022			0.195	0.021

Length				
Speed Limit	0.038	0.004	0.033	0.003
Lane Width	0.117	0.021	0.101	0.020
Shoulder Width	0.092	0.019	0.069	0.013
Radius	0.437	0.301	0.384	0.136
Vertical Grade	1.274	0.180	0.550	0.176
Model Performance				
Dbar	14321	13450	12238.6	11550
Dhat	14310.3	12780	11166.8	10150
pD	8.981	669.9	1071.81	1393
DIC	14330	14120†	13310.4	12940
MAD	6.92	6.88	6.72	6.64

Note: Parameter estimates not significant under 5 percent significance level are shown in italic and bold fonts.
† With the MLE RPNB, all the variables except speed limit were found to be random. This increased the DIC to 14132.

The RPNB-L model has smaller standard deviation estimates for all model coefficients (random parameters). The smaller standard deviation for the random parameter estimates means that the normal distribution of a covariate parameter is more centered around the mean value when using the RPNB-L model; this may be a result of the site-specific frailty term used in the NB-L formulation that accounts for a portion of data variation.

The segment length and the AADT variables in the RPNB-L model showed a positive relationship with crash count for almost all segments, but with varying magnitude. The estimated marginal effect for AADT also emphasizes the positive effect on crash occurrence of AADT. A similar trend is also observed for the segment length variable. More than 81.8 percent of the sites have parameter estimates that are greater than zero for the speed limit variable. The estimated marginal effect for the speed limit variable indicates that there is an overall increase in crash occurrence with a unit increase in the speed limit variable.

Table 4-6 Average marginal effects for the South Dakota Dataset.

Variables	Model			
	NB	RPNB	NB-L	RPNB-L
log(AADT)	4.83	4.769	4.85	4.69
Segment Length	0.258	0.269	0.252	0.262
Speed Limit	1.419	2.027	1.473	1.672
Lane Width	-0.078	-0.566	-0.131	-0.361
Shoulder Width	-0.003	-0.025	-0.006	-0.005
Radius	-0.041	-0.05	-0.042	-0.04
Vertical Grade	-1.698	-3.189	-1.754	-2.127

The distribution of Radius of curvature has a crash count that decreases with the increase in the radius of curvature, but the magnitude varies among sites, as expected. The standard deviation of parameter estimate for the radius of curvature indicates that more than 90 percent of sites have negative coefficients. Similar observations can be made for the lane width variable, where 60.2 percent of the random parameter estimates have a value of less than 0. This trend also applicable to the shoulder width variable, where more than 51 percent of the parameter estimates have a value of less than zero. Rahman Shaon and Qin (2016) used the same dataset and made similar observations. The authors noted that lane width may have mixed safety effects, and an increasing lane width or shoulder width or combination of both may not always bring additional safety benefits. Further research should look into whether or not an increase in lane width leads to an increase in safety. One interesting finding is that the estimated marginal effect of the shoulder width variable is quite similar between NB, NB-L, and RPNB-L (between -0.003 to -0.006), whereas it is quite different for the RPNB model. One possible explanation for this variation could be that the mean estimate of the shoulder width itself is not statistically significant in all models. The model parameters estimate and marginal effect of the grade variable indicates that the presence of vertical grade reduces crash occurrence for almost all sites (97.4 percent of the

distribution has value less than zero). The estimate of the dispersion parameter is also the smallest for the RPNB-L model. The Poisson regression is a limiting case of the NB regression because the dispersion parameter approaches zero. The mean estimates in the RPNB-L model are less affected by the data dispersion, which means it captures more variation in the data than the other three models.

4.7.3 Model Performance

The last section of Table 4-3 and Table 4-5 provides the model performance estimates based on the Deviance Information Criterion (DIC) for the Indiana and South Dakota datasets, respectively[8]. The DIC is a widely used GOF statistic for comparing models in a Bayesian framework (Spiegelhalter et al. 2002). It is worth pointing out that the model parameterization can influence the estimation of the DIC value, and the comparisons with DIC should be made only between models that have similar parameterizations (Geedipally et al. 2014). All developed models can be adequately compared using the DIC measure because both the NB-L and RPNB-L models are developed based on the NB model parameterization. The DIC consists of two components: (a) measures of how well the model fits the data, Dbar ($\overline{(D(\theta))}$) and (b) a measure of model complexity (pD). Thus, DIC can provide a better comparison between models that are characterized by different complexities.

A comparison of the DIC values between models illustrated that the RPNB-L model performed better than the NB-L and RPNB. Table 4-3 and Table 4-5 show that the DIC value is highest in the traditional NB model. The small pD value illustrates that the NB model is less complex than other model alternatives used in this study. According to the estimated pD value, the RPNB-L

[8] DIC is a hierarchical modeling generalization of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), defined as $DIC = \overline{D(\theta)} + pD \; and \; pD = \overline{D(\theta)} - D(\bar{\theta})$, where θ represents the collection of parameters.

model is the most complex of all the models due to its mixed distribution and random components in the explanatory variables. The point estimate of deviance illustrated by Dhat shows that the RPNB-L model has the smallest deviance in both datasets. Dbar represents almost the same information as Dhat except that it represents the posterior mean of deviance rather than a point estimate. The RPNB-L model, despite having the highest penalty value of pD, has a 5.8 percent and 7.8 percent improvement in DIC values for the Indiana dataset when compared with the RPNB and fixed parameters NB-L model, respectively. The MAD estimates indicate that the fixed-coefficient NB-L model has better predictive ability than RPNB even though the estimated DIC value is smaller with RPNB compared to the fixed parameters NB-L model. The improvement in DIC with the RPNB-L model compared to the RPNB and NB-L models for the South Dakota dataset are 8.4 percent and 2.8 percent, respectively. The MAD estimates illustrate that RPNB-L has the lower mean absolute error compared to other models in both datasets. Due to the frailty terms that explain additional data heterogeneity along with random parameters, RPNB-L compensates for increased model complexity by improving the predictive modeling ability, which is reflected in the MAD that considers both bias and variance.

4.8 Summary and Conclusions

Researchers can experience challenges when it comes to understanding the underlying crash generating process, producing reliable model coefficients, and making statistical inferences from crash data. This study proposed the application of a RPNB-L GLM for analyzing crash data by implementing an NB-L model with coefficients that varied from site to site. The model was applied to two observed datasets, one collected in Indiana and the other in South Dakota. The model results were compared to the traditional NB, RPNB, and fixed parameters NB-L models.

Results showed that both the fixed coefficient NB-L (especially compared to the MLE RPNB) and newly developed RPNB-L GLMs performed better than a fixed and random parameters NB GLM. The estimated effects of covariates using RPNB-L were less dispersed compared to the RPNB model, according to the standard deviation of random parameters. The RPNB-L model's proficiency in accounting for highly dispersed data led to its ability to achieve around 6 percent and more than 8 percent improvement in DIC, respectively, for the Indiana and South Dakota data. The estimated skewness of the crash count was 11.624 for the South Dakota data. Shirazi et al. (2017) recommended that the NB-L (and RPNB-L) should be used over the NB when the skewness value exceeds 1.92. In conclusion, both the fixed and random parameters of NB-L GLMs offer a viable alternative to the traditionally both fixed and random parameters NB GLMs when analyzing over-dispersed crash datasets.

The random parameters defined in this study were independent and characterized by a single normal distribution to account for unobserved heterogeneity in crash occurrences. The independence assumption restricts the interaction between random parameters. It is possible that the sources of heterogeneity are correlated due to the interactions between explanatory variables (Mannering et al. 2016). Mannering et al. (2016) suggested developing a random parameters model with correlated parameters to account for correlation among random parameters; however, using a simple distribution to characterize the random parameter mean and variance may not fully capture the underlying nature of unobserved heterogeneity in the dataset which could result in erroneous model inferences. Unobserved heterogeneity can be tracked in a more sophisticated manner when heterogeneity is included in the mean and variance, by providing additional flexibility in the heterogeneity capturing process (Behnood and Mannering 2017b, a, Seraneeprakarn et al. 2017). The proposed model should be developed further, and more reliable

parameter estimates should be obtained by applying an RPNB-L with correlated random parameters and an RPNB-L with heterogeneity in the mean and variance. Additionally, more work should be performed to examine the "identification" of random parameters under the Bayesian framework in order to match those identified under the frequentist approach.

4.9 References

Aguero-Valverde, J., Jovanis, P., 2008. Analysis of road crash frequency with spatial models. Transportation Research Record: Journal of the Transportation Research Board (2061), 55-63.

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.

Bedrick, E.J., Christensen, R., Johnson, W., 1996. A new perspective on priors for generalized linear models. Journal of the American Statistical Association 91 (436), 1450-1460.

Behnood, A., Mannering, F., 2017a. Determinants of bicyclist injury severities in bicycle-vehicle crashes: A random parameters approach with heterogeneity in means and variances. Analytic Methods in Accident Research 16, 35-47.

Behnood, A., Mannering, F., 2017b. The effect of passengers on driver-injury severities in single-vehicle crashes: A random parameters heterogeneity-in-means approach. Analytic Methods in Accident Research 14, 41-53.

Booth, J.G., Casella, G., Friedl, H., Hobert, J.P., 2003. Negative binomial loglinear mixed models. Statistical Modelling 3 (3), 179-191.

Buddhavarapu, P., Scott, J.G., Prozzi, J.A., 2016. Modeling unobserved heterogeneity using finite mixture random parameters for spatially correlated discrete count data. Transportation Research Part B: Methodological 91, 492-510.

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, 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, 38-42.

El-Basyouny, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. Accident Analysis and Prevention 41 (5), 1118-1123.

Garnowski, M., Manner, H., 2011. On factors related to car accidents on german autobahn connectors. Accident Analysis and Prevention 43 (5), 1864-1871.

Geedipally, S.R., Lord, D., Dhavala, S.S., 2012. The negative binomial-lindley generalized linear model: Characteristics and application using crash data. Accident Analysis and Prevention 45, 258-265.

Geedipally, S.R., Lord, D., Dhavala, S.S., 2014. A caution about using deviance information criterion while modeling traffic crashes. Safety science 62, 495-498.

Gelfand, A.E., Sahu, S.K., Carlin, B.P., 1995. Efficient parametrisations for normal linear mixed models. Biometrika 82 (3), 479-488.

Guo, J.Q., Trivedi, P.K., 2002. Flexible parametric models for long-tailed patent count distributions.

Hallmark, S.L., Qiu, Y., Pawlovitch, M., Mcdonald, T.J., 2013. Assessing the safety impacts of paved shoulders. Journal of Transportation Safety & Security 5 (2), 131-147.

Li, W., Carriquiry, A., Pawlovich, M., Welch, T., 2008. The choice of statistical models in road safety countermeasure effectiveness studies in iowa. Accident Analysis and Prevention 40 (4), 1531-1542.

Lindley, D.V., 1958. Fiducial distributions and bayes' theorem. Journal of the Royal Statistical Society. Series B (Methodological), 102-107.

Lord, D., Geedipally, S.R., 2011. The negative binomial–lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. Accident Analysis and Prevention 43 (5), 1738-1742.

Lord, D., Guikema, S.D., Geedipally, S.R., 2008. Application of the conway–maxwell–poisson generalized linear model for analyzing motor vehicle crashes. Accident Analysis and Prevention 40 (3), 1123-1134.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transportation Research Part A 44 (5), 291-305.

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.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. Winbugs-a bayesian modelling framework: Concepts, structure, and extensibility. Statistics and computing 10 (4), 325-337.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Analytic methods in accident research 1, 1-22.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research 11, 1-16.

Mitra, S., Washington, S., 2007. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis and Prevention 39 (3), 459-468.

Mitra, S., Washington, S., 2012. On the significance of omitted variables in intersection crash modeling. Accident Analysis and Prevention 49, 439-448.

Park, B.-J., Lord, D., Hart, J.D., 2010. Bias properties of bayesian statistics in finite mixture of negative binomial regression models in crash data analysis. Accident Analysis and Prevention 42 (2), 741-749.

Peng, Y., Lord, D., 2011. Application of latent class growth model to longitudinal analysis of traffic crashes. Transportation Research Record: Journal of the Transportation Research Board (2236), 102-109.

Rahman Shaon, M.R., Qin, X., 2016. Use of mixed distribution generalized linear models to quantify safety effects of rural roadway features. Transportation Research Record: Journal of the Transportation Research Board (2583), 134-141.

Schlüter, P., Deely, J., Nicholson, A., 1997. Ranking and selecting motor vehicle accident sites by using a hierarchical bayesian model. Journal of the Royal Statistical Society: Series D (The Statistician) 46 (3), 293-316.

Seraneeprakarn, P., Huang, S., Shankar, V., Mannering, F., Venkataraman, N., Milton, J., 2017. Occupant injury severities in hybrid-vehicle involved crashes: A random parameters approach with heterogeneity in means and variances. Analytic Methods in Accident Research 15, 41-55.

Shankar, V., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. Accident Analysis and Prevention 29 (6), 829-837.

Shirazi, M., Dhavala, S.S., Lord, D., Geedipally, S.R., 2017. A methodology to design heuristics for model selection based on characteristics of data: Application to investigate when the negative binomial lindley (nb-l) is preferred over the negative binomial (nb). Accident Analysis and Prevention Forthcoming.

Shirazi, M., Lord, D., Dhavala, S.S., Geedipally, S.R., 2016. A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: Characteristics and applications to crash data. Accident Analysis and Prevention 91, 10-18.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64 (4), 583-639.

Vangala, P., Lord, D., Geedipally, S.R., 2015. Exploring the application of the negative binomial–generalized exponential model for analyzing traffic crash data with excess zeros. Analytic methods in accident research 7, 29-36.

Venkataraman, N., Ulfarsson, G., Shankar, V., Oh, J., Park, M., 2011. Model of relationship between interstate crash occurrence and geometrics: Exploratory insights from random parameter negative binomial approach. Transportation research record: journal of the transportation research board (2236), 41-48.

Washington, S.P., Karlaftis, M.G., Mannering, F., 2010. Statistical and econometric methods for transportation data analysis CRC press.

Xu, J., Sun, L., 2015. Modeling of excess zeros issue in crash count andysis. Journal of Jilin University (Engineering and Technology Edition) 45 (3), 769-775.

Zamani, H., Ismail, N., 2010. Negative binomial-lindley distribution and its application. Journal of Mathematics and Statistics 6 (1), 4-9.

Zou, Y., Wu, L., Lord, D., 2015. Modeling over-dispersed crash data with a long tail: Examining the accuracy of the dispersion parameter in negative binomial models. Analytic Methods in Accident Research 5, 1-16.

Chapter 5 Incorporating Behavior Variables into Crash Prediction: a multivariate multiple risk generating process approach

This chapter continues the discussion on segment-level CPM development in the 2nd tier of the proposed 3-tier spatial unit approach. This chapter specifically discusses the development of a modeling technique to incorporate available information related to driver behavior into CPM. In traditional CPMs, temporally aggregated crashes are modeled as a single chain of events which assumes all covariates used in model development contributes equally in each crash occurred on a segment or intersection. Crash risks generate from different risk sources with each risk source playing either a vital or supporting role. This chapter presents the development of multiple risk source regression modeling approach to incorporate driver behavior related information as a separate risk source from roadway geometry and traffic variables into CPMs. In addition, development of a multivariate multiple risk source model was also presented to account for the correlated between injury severity levels into CPM and predict crash count and injury severity simultaneously. The modeling results are then compared with traditional single risk source NB model, and their marginal effects are also estimated.

5.1 Introduction

The frequency and severity of traffic crashes have been largely used in transportation safety as two indicators of crash risk (Washington et al. 2018). These two indicators form the overall risk at transport network locations and thus mitigating one without paying attention to the other one is incomplete and can be wrong. While road agencies and departments of transportation aim to

reduce the frequency and severity of traffic crashes, highway safety improvement programs are primarily focused on preventing severe and fatal crashes, as the cost per person of a fatal crash is almost 250 times higher than a non-injury crash (Harmon et al. 2018). As a result, considering crash severity in conjunction with crash frequency is paramount in crash causal analysis and modeling.

Crash prediction models have been widely used to study crash frequency and investigate crash contributing factors at transport network locations. These models have been traditionally applied to the crash frequencies aggregated over different crash severity levels. To incorporate crash severity into crash prediction models, numerous studies have modeled crash frequency of a particular severity level at a specific intersection or segment (Hauer et al. 1988, Shankar et al. 1995, Poch and Mannering 1996, Abdel-Aty and Radwan 2000, Lord and Persaud 2000, Lyon et al. 2003, Qin et al. 2004, Abdel-Aty et al. 2005, Tarko et al. 2008, Geedipally et al. 2010, Geedipally and Lord 2010). However, the independent (i.e., univariate) modeling of crash frequency with various injury severities may not be accurate because crash frequencies may be correlated across different severities due to the presence of shared effects from engineering, spatial, and unobserved factors, (Wang et al. 2017). Neglecting such correlations may lead to biased parameter estimates and inaccurate inferences about crash contributing factors (Ma et al. 2008, Mannering and Bhat 2014, Serhiyenko et al. 2016). Empirical evidence has shown that multivariate crash frequency models (e.g., multivariate Poisson lognormal model) can provide better predictive accuracy than its univariate counterparts (Ma and Kockelman 2006, Wang et al. 2017). Hence, multivariate models have gained popularity, as they can model crash counts of different severities simultaneously and explore the effects of covariates in a more accurate fashion.

Research has established that traffic crashes are the results of chains of causal events that arise from a multitude of contributing factors associated with roadway design, traffic operations, pavement conditions, driver behavior, human factors, and environmental factors. These factors do not necessarily contribute equally to crashes at a site, though; therefore, it may be more plausible to consider traffic crashes at every site as the results of multiple risk sources, with each risk source playing either a primary or supporting role. However, conventional crash frequency models treat the crash count at a roadway site as the outcomes of a single risk source by using a single predictive equation estimated with Poisson or Negative Binomial (NB) distribution. While these single equation models are statistically sound and practically useful, their results may yield biased parameter estimates due to issues related with data overdispersion[9] (Zou et al. 2015, Rahman Shaon and Qin 2016, Shirazi et al. 2016, Rahman Shaon et al. 2018). Furthermore, single-equation models are incapable of assuming that crashes may have various risk sources, which could result in data heterogeneity. Not until recently have researchers developed the multiple risk source regression model to distinguish the distinct sources of crash contributing factors (Washington and Haque 2013, Afghari et al. 2016).

Multiple risk source regression modeling is a reasonable alternative to single equation predictive models for predicting risk-level crashes, considering that the contribution of explanatory variables originated from distinct risk sources to the outcome (i.e., predicted crash count at a site) may change. The feasibility of using a generalized structure for modeling crashes by multiple sources of risk has been investigated, and the models have been developed for univariate crash prediction (e.g., total crashes) (Washington and Haque 2013, Afghari et al. 2016,

[9] Crash data are often characterized by the existence of a large sample variance compared with the sample mean. In a statistical term, the sample data is over-dispersed when the variance is greater than the mean. Data over-dispersion is often caused by unobserved data heterogeneity due to unobserved, unavailable, or unmeasurable variables that are important to explain model responses.

Afghari et al. 2018). However, these models ignored the possible correlation across crash counts of different crash severity levels, and thus the parameter estimates are prone to bias. Therefore, there is a need to incorporate crash severity into multiple risk source modeling of crashes.

Significant amount of research has been devoted to identifying and quantifying the effect of contributing factors on crash occurrence. Crash risk originated from driver behavior has been recognized as major crash contributors in highway safety literature (Sabey and Staughton 1975, Rumar 1985, NHTSA 2008, Shaon et al. 2018a). Albeit of universal acceptance, incorporation of contributing factors originated from behavioral risk source into crash frequency modeling is limited due to data unavailability. There is no established method available to collect driver behavior related variables at crash sites. Alcohol-impaired, drug-impaired driving, distraction and speeding behaviors are frequently identified as contributing factors to crash occurrence (Sabey and Staughton 1975, Rumar 1985). The absence of these important pieces of behavioral information in crash data can cause unobserved heterogeneity and modeling result can yield biased parameter estimates (Mannering et al. 2016).

This study extends the idea of using a multiple risk source structure to develop a multivariate multiple risk source methodological approach to estimate both crash counts and severity, simultaneously. Similar to multivariate crash prediction models, it is hypothesized that the multivariate multiple risk source modeling approach will provide improved accuracy than a univariate model because it considers the correlation between crash counts of different severities and accommodates for unobserved heterogeneity which could result from the omission of multiple risk sources in modeling equations. In regard to multiple risk sources, the two risk sources - engineering and behavioral risk source related crash contributing factors were explored in the proposed model. Considering data limitation related to driver behaviors, a few behavioral

variables that are uncorrelated with engineering factors and solely originated from a different source (e.g., physical and psychological characteristics) were incorporated as behavioral variables in this study. Furthermore, the risk-level predicted crashes from multiple risk source modeling could be useful in identifying sites for safety improvements and developing targeted and effective safety countermeasures.

5.2 Literature Review

A large number of studies in the road safety literature estimated crash frequency by crash severity to evaluate the safety implications of contributing factors (Hauer et al. 1988, Shankar et al. 1995, Poch and Mannering 1996, Abdel-Aty and Radwan 2000, Lord and Persaud 2000, Lyon et al. 2003, Qin et al. 2004, Abdel-Aty et al. 2005, Tarko et al. 2008, Geedipally et al. 2010, Geedipally and Lord 2010). In this context, several previous studies noted that crash counts across different injury severity are likely to be correlated (Ma et al. 2008, Wang et al. 2017). Therefore, incorporating the correlation between crash counts of different injury severities is an important practice when estimating crash counts and severities, simultaneously. Such correlation can be effectively handled by multivariate regression models (Ye et al. 2009, Pei et al. 2011, Wang et al. 2011, Chiou and Fu 2015, Zeng et al. 2016) (Please refer to Mannering and Bhat (2014) and Mannering et al. (2016) for comprehensive list of literature on crash data modeling). Both multivariate Poisson and multivariate Poisson lognormal models are popular choices, but the latter is more effective for overdispersed data (Chib and Winkelmann 2001, Ma and Kockelman 2006, Park and Lord 2007, Ma et al. 2008, Ye et al. 2008, Ye et al. 2009). The covariance structure used in the multivariate Poisson lognormal model allows for estimating model parameters with smaller standard errors while maintaining the core strength of the Poisson

distribution. Studies have shown that this model of crashes outperforms the univariate models in terms of statistical fit (Chib and Winkelmann 2001, Park and Lord 2007, Ma et al. 2008).

Substantial effort has been devoted to identify primary risk factors contributing to crashes at a site and quantify their effects on crash occurrences (Miaou et al. 1992, Milton and Mannering 1998, Garber and Ehrhart 2000, Persaud 2001, Lee and Mannering 2002, Bahar et al. 2004, Tarko and Kanodia 2004). Roadway design factors and traffic operational characteristics dominate this list of variables in the crash data modeling related literature (Shankar et al. 1995, Abdel-Aty and Radwan 2000, Quddus et al. 2001, Chin and Quddus 2003, Oh et al. 2004, Qin et al. 2004, Anastasopoulos and Mannering 2009, El-Basyouny and Sayed 2009, Fitzpatrick et al. 2010, Geedipally et al. 2012, Mitra and Washington 2012, Islam et al. 2014a, Montella and Imbriani 2015, Qin et al. 2016, Qin et al. 2018, Rahman Shaon and Qin 2016, Shaon et al. 2018b). The findings show that roadway geometric features such as lane width, shoulder width, and horizontal and vertical alignments are statistically significant in their correlation with crash occurrence. In addition, traffic operational variables such as Average Annual Daily Traffic (AADT), truck traffic and posted speed limit have been shown to have a significant influence on safety. Since these variables represent the engineering principles and practices in highway design and capacity analysis, they are often referred to as engineering variables. Understanding the safety performance of engineering variables is instrumental in identifying effective engineering solutions. Most proven safety countermeasures involve the modification and improvement of roadway and roadside design features as well as controlling traffic features on specific roadway sites (FHWA 2017). The prevalence of studying these variables is also due to the availability and quality of data, as transportation agencies are required to collect and maintain them for highway

performance monitoring, planning and program development, design, and operations, as well as maintenance activities.

Driver behavior variables, however, are not readily available even though they are considered universally as a major contributor to crashes (Sabey and Staughton 1975, Rumar 1985, NHTSA 2008, Washington and Haque 2013, Afghari et al. 2018, Shaon et al. 2018a). Standard procedures for collecting driver behavior data do not exist, as highway agencies are not obligated to collect such information for safety management systems. The behavior data collected from crash data represent a very small portion of driver activities in traffic events. The most relevant source for obtaining this information is perhaps the crash report where police officers may record information regarding driver's condition and his or her opinion of the possible contributing factors. This type of information, albeit extremely valuable, is often incomplete, underreported, and inconsistent. Reports show that risky driving behaviors such as distracted driving, impaired driving, speeding are often identified as major contributors to crash occurrences (Redelmeier and Tibshirani 1997, Box 2009, NHTSA 2010). Such information, however, is usually available only for severe crashes in which thorough investigations are performed. One of the most exhaustive studies conducted so far is the National Motor Vehicle Crash Causation Survey administered by National Highway Traffic Safety Administration for which a group of experts reviewed a nationally representative sample of 5,471 crashes during a 2.5-year period. Commonly used roadway or environmental conditions were found as the primary reason for only 135 crashes from this study – a mere 2.5 percent, which shows the necessity of incorporating driver behaviors into crash prediction models.

Although site-specific driver behavior variables may not be readily available, behavioral variables are sometimes collected at a larger geographic scale (e.g., county) to analyze the

physical and psychological status of a community. Driver behavior is determined by drivers' commitment to the values and beliefs in safety, which is influenced by attitudes, social norms, and perceived risk. Social norms play an important role in driving behavior and risk perception (Carter et al. 2014). For example, some drivers may follow the behaviors of others in their community, regardless of roadway design or site characteristics (Schneider et al. 2018). Societal expectations of acceptable transportation risk can also influence risk-taking behavior (Moeckli and Lee 2007). Proxy variables can be used to substitute driver risk factors in crash count modeling for measuring the effect of behavioral risk on crash occurrence, including total number of speeding offenses (Afghari et al, 2018), operating while intoxicated citation count (Smith 2000, Nagle 2012), drug arrest count (Walsh et al. 2008, Compton et al. 2009, Asbridge et al. 2012), violent crime rate (Weiss 2013, Carter and Piza 2017, Ando et al. 2018), and liquor license rate (LaScala et al. 2000).

Understanding the effects of crash data generating mechanisms provides useful information about the sources of variance in crash data. Peng et al. (2014) used a generalized waring model to differentiate between different distinct sources of crash heterogeneity using different variance terms. The authors separated the observed variability into random errors; the proneness, which refers to the internal differences between observations, and the liability, which refers to the variance caused by unobserved exogenous variables. This new modeling structure has a better performance compared to the NB model and showed that a crash may originate from different sources through different processes, which contributes to additional variances. Explicitly incorporating the heterogeneous sources into crash modeling can be challenging, but one logical approach is to group contributing factors by risk source (e.g., environmental factors, roadway geometric design features, driver behavior) and assume that variables within the same

risk source affect crash occurrence in a similar manner, but that sources contribute to crashes to different extents. This assumption resonates with what Lord, Washington, and Ivan have noted in their seminal work that concludes that over-dispersion arises from the actual nature of the crash process (Lord et al. 2005). The NB distribution is therefore limited in that it assumes that only one underlying process affects the likelihood of crash frequency (Shankar et al. 1997).

Recently, researchers introduced a multiple risk generating process regression model in which crashes at a given site are assumed to have originated from distinct sources of risk, and their relationships are represented by multiple equations (Washington and Haque 2013, Afghari et al. 2016, Afghari et al. 2018). The authors argued that the single risk source assumption in traditional crash prediction models is statistically sound but cannot sufficiently address unobserved heterogeneity. The application of single risk source traditional models in blackspot identification centers on the assumption that operational causal factors such as roadway geometry or traffic factors operate in a single chain to form the total crash count. The result is that other risk sources such as driver behavioral factors, which are the cause of more than 50% of crashes, are neglected (NHTSA 2008, Washington and Haque 2013, Shaon et al. 2018a). A single risk source may attribute behavioral factors to operational factors, resulting in biased parameter estimates and erroneous model prediction. A multiple risk source regression model can add flexibility to estimate crashes based on their originating risk sources and provide meaningful parameter estimates. The empirical evidence shows that assumption of multiple risk sources in modeling equation provide improved model fit and can account for unobserved heterogeneity that results from ignoring risk sources (Afghari et al. 2016, Afghari et al. 2018).

5.3 Research Hypothesis

To estimate crash counts and injury severity, simultaneously, a multivariate framework in regression modeling is needed to accommodate the correlation between crash counts of different injury severities. Equally important part in crash data modeling is to distinguish between the sources of crash risk. In spite of the importance of behavioral factors, a limited amount of research has directly incorporated these risks into crash prediction modeling because of the lack of site-specific driver behavior-related factors. Alternatively, the effect of site-specific variables, in combination with the influence of a broader safety culture represented by driver behavior, would provide many helpful insights.

The hypothesis of the methodological approach in this study is described below, and includes the theoretical support for this type of crash modeling:

- This study hypothesizes that a single risk source model (e.g., Poisson, NB) cannot sufficiently account for unobserved heterogeneity in crash data. Considering multiple underlying risk sources in crash data modeling may allow researchers to account for unobserved heterogeneity at each risk-generating source.
- This study hypothesizes that risk sources can be categorized based on distinct sources of data and their physical meaning. Two distinct risk sources, engineering, and behavioral risk sources are considered which simultaneously contribute to the crash occurrence on a roadway segment.
- Crash counts of different severities are correlated. Considering the correlation of crash severities in the modeling structure allows for the simultaneous estimation of crash frequency and severity and thus, reduces bias in the estimated model

parameters. The multivariate structure is considered for two injury severities in this study: injury crashes and non-injury crashes.

Based on the above-mentioned hypothesis, the unstructured covariance matrix is used to define the correlation between injury severity levels, which contributes to the estimation of more precise model parameters. Multiple risk sources are considered to have varying contributions to crashes of all severities at each site and across sites. Site-specific risk-level weights (also vary between injury severity) are used to generate multiple proportions of total crashes. A bivariate (e.g., two risk sources) random error term at each risk level is used to account for unobserved heterogeneity and define the correlation between risk sources.

5.4 Methodology

Assuming observed crash count Y_i at location i, summed across underlying risk sources j, it can be hypothesized that each risk source is responsible for contributing to a proportion of the total observed crashes which are unobserved or latent at the crash location. To determine the latent probabilities of unobserved crash counts from different risk sources, let's assume the total observed crash count follows a Poisson distribution with a total predicted mean μ_i:

$$Y_i \sim Poisson(\mu_i) \quad (1)$$

A latent mixture modeling approach can be used to link multiple risk-generating sources with the mean of Poisson distribution. The latent mixture approach requires the decomposition of the mean function of the Poisson distribution (μ_i) into multiple mixture components (Afghari et al. 2016, Afghari et al. 2018):

$$\mu_i = \sum_{j=1}^{J} \mu_{ji} \quad \text{and} \quad \mu_{ji} = w_{ji}\mu_i \tag{2}$$

Where, $\sum_{j=1}^{J} w_{ji} = 1$ and w_{ji} is the proportion (or weight) of the predicted crash count at site i attributed from latent risk source j and J is the total number of underlying risk sources. Assuming exponential functions for the decomposed means of the Poisson distribution, each of the above-mentioned predicted means is a function of a variety of contributing factors associated with unique risk sources:

$$\mu_{ji} = \alpha_{j0} F_i^{\alpha_1} \exp(\sum \beta_j X_{ji} + \varepsilon_{ji}) \tag{3}$$

Where,

F_i = measure of exposure (shared between risk sources),

X_{ji} = explanatory variables for risk source j at site i,

α_j, β_j = estimated regression parameters,

ε_{ji} = model errors independent of all explanatory variables.

To account for unobserved heterogeneities arising from overdispersion, error terms (ε_{ji}) are allowed to vary across observations. In addition, to account for the correlation between the underlying risk sources, the error terms are defined to follow a Multivariate Normal distribution which can be constructed as follows:

$$\varepsilon_i \sim Multivariate\ Normal\ (0, \Sigma_R)$$

Where,

$\varepsilon_i = [\varepsilon_{1i}\ \varepsilon_{2i} \dots \varepsilon_{Ji}]$

and $\Sigma_R = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \dots\dots & \sigma_{1J}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \dots\dots & \sigma_{2J}^2 \\ \dots & \dots & \dots\dots & \dots \\ \sigma_{J1}^2 & \sigma_{K2}^2 & \dots\dots & \sigma_{JJ}^2 \end{bmatrix}$

Please note that these error terms at the risk source level can also account for unobserved and/or unavailable factors that may have contributed to crash occurrence. The univariate risk source regression model, however, does not distinguish injury crashes from non-injury crashes. Thus, a multivariate modeling approach is needed for such a distinction where crashes by injury severity are modeled simultaneously. In a set of crash data at n roadway segments, let assume crashes are classified into K categories which represent K crash severities. Let $Y_{ki} = (y_{1i,}\ y_{2i,} \dots\dots\dots, y_{Ki})'$ be a K-dimensional vector that denotes the total crash count at i-th ($i = 1, 2, \dots, n$) roadway segment that belongs to k-th (k=1, 2,….,K) injury severity. Assuming crash counts by crash severity follows the Poisson distribution with mean μ_{ki} for k=1, 2, …., K and following a similar crash generating mechanism, crashes in each severity category are generated from multiple risk sources which are summed to obtain the total crash count at a location. The regression equation can be constructed as follows:

$$\ln(\mu_{ki}) = \ln(\mu_{ki}^*) + \varepsilon_{ki}$$

$$\mu_{ki}^* = \sum_{j=1}^{J} \mu_{jki}$$

$$\mu_{jki} = w_{jki}\mu_{ki} \qquad (5)$$

Where, w_{jki} is the proportion (or weight) of the total predicted crash counts for crash severity k at site i attributed from latent risk source j, and $\sum_{j=1}^{J} w_{jki} = 1$. The new error term $\boldsymbol{\varepsilon_{ki}}$ denotes the random effect which is uncorrelated with the explanatory variables and accounts for the unobserved heterogeneity arising from different crash severity levels. This new error term is also assumed to be multivariate normally distributed across crash counts of different severity levels. Let's assume $\varepsilon_i = (\varepsilon_{1i}, \varepsilon_{2i}, ., \varepsilon_{Ki})'$ represents a vector of random effects at each location i and it follows a K-dimensional normal distribution:

$$\varepsilon_i \sim Multivariate\ Normal\ (0, \Sigma_M)$$

Where, **0** is a K-dimensional zero vector and Σ_M is a J×J variance-covariance matrix. Following the above specification of the error term, it is equivalent to $\exp(\varepsilon_i) \sim Lognormal\ (0, \Sigma)$. The variance-covariance matrix Σ accounts for unstructured error and unobserved heterogeneous effects and can be formulated as follows:

$$\Sigma_M = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \dots\dots & \sigma_{1K}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \dots\dots & \sigma_{21}^2 \\ \dots & \dots & \dots\dots & \dots \\ \sigma_{K1}^2 & \sigma_{K2}^2 & \dots\dots & \sigma_{KK}^2 \end{bmatrix} \quad (6)$$

The diagonal elements σ_{kk}^2 of the variance-covariance matrix represents the heterogeneous variance of ε_{ki}, and the off-diagonal elements σ_{rs}^2 represents the heterogeneous covariance between ε_{ri} and ε_{si} where $r \neq s$.

Following a similar concept of univariate modeling, the mean response from each risk source λ_{jti} can be expressed as follows:

$$\mu_{jki} = \alpha_{jk0} F_i^{\alpha_{t1}} \exp(\sum \beta_{jk} X_{ji} + \varepsilon_{jki}) \quad (7)$$

Where, ε_{jki} is another error term used to account for unobserved heterogeneity and correlated between underlying risk-generating sources within each crash severity. This indicates the incorporation of additional K number of errors into the modeling structure. For example, let's assume we have crash data from 2 crash severity types (e.g., injury and non-injury) and there are 2 underlying risk sources (e.g., engineering and behavioral). Under multivariate multiple risk source modeling, $\varepsilon_{1i} = [\, \varepsilon_{11i}, \varepsilon_{12i}]$ and $\varepsilon_{2i} = [\, \varepsilon_{21i}, \varepsilon_{22i}]$ will account for unobserved heterogeneity and correlations between underlying risk-generating sources for crash severity 1 and crash severity 2, respectively. Because of multivariate modeling, there will be another error term $\varepsilon_i = [\, \varepsilon_{1i}, \varepsilon_{2i}]$ to account for unstructured errors and unobserved heterogeneous effects for each crash severity. The expected mean, variance, and covariance for multivariate multiple risk source regression model can be expressed as follows:

$$E[Y_{ki}] = \mu_{ki}^* \times \exp(\frac{\sigma_{jj}}{2}) \quad (8)$$

$$Var[Y_{ki}] = E[Y_{ki}] + (E[Y_{ti}])^2 \times [\exp(\sigma_{ii}) - 1] \quad (9)$$

$$Cov[Y_{ri}, Y_{si}] = E[Y_{ri}] \times [\exp(\sigma_{rs}) - 1] \times E[Y_{si}] \quad (10)$$

The multivariate crash data modeling using Poisson-lognormal mixture can accommodate overdispersion in the data. From Eq. (8) and (9), it can be noted that $Var[Y_{ki}] > E[Y_{ki}]$ since the diagonal elements of $\Sigma, e.g.\ \sigma_{kk}^2 > 0$. Additionally, the multivariate structure can incorporate the correlation among the components in a response vector as described in Eq. (10).

The proposed multivariate and univariate multiple risk source regression models are formulated and estimated in the Bayesian framework using OpenBUGS (Spiegelhalter et al. 2007). In the Bayesian framework, the regression parameters are estimated by maximizing the posterior which is a combination of the likelihood function and the defined prior information. It is necessary to specify a prior distribution for the parameters to obtain the Bayesian estimate. Prior distributions are meant to reflect prior knowledge about the parameters of interest. In the absence of solid prior information, uninformative priors can be assumed to estimate both univariate and multivariate multiple risk source models in the Bayesian framework:

$\beta_{jk} \sim Normal(0,100)$

$w_{jki} \sim Uniform(a, b)$

$\Sigma_R^{-1} \sim Wishart(I_J, J)$

$\Sigma_M^{-1} \sim Wishart(I_K, K)$

Where, I_J and I_K represents J×J and K×K dimensional identity matrix, respectively. Defining the value for lower limit *a* and upper limit *b* for the prior information of a risk-level weight should be approached with caution. Prior knowledge, if available, can be used to define these values (Afghari et al. 2016). The Markov chain Monte Carlo (MCMC) can also suffer from poor

mixing, and the effective number of parameters of estimated models can be negative (it is guaranteed to be positive for properly defined and converged models) if there is a conflict between data and prior information. Several preliminary models were developed with different prior values for a and b to obtain good mixing between Monte Carlo chains and model convergence. The model performance of preliminary models was also compared to choose the optimal value. The final model is estimated with a=0.45 and b=0.95.

In addition, marginal effects are used to determine the impact of each covariate on the expected mean value of the dependent variable. The marginal effect represents the effect of a unit change in the independent variable on the expected mean of the dependent variable. For a multiple risk source model, the marginal effect for each explanatory variable can be estimated using the following equation (Washington et al. 2010):

$$E_{x_{jik}}^{\mu_i} = \frac{\delta \mu_i}{\mu_i} \times \frac{x_{jik}}{\delta x_{jik}} = w_j \beta_{jk} x_{jik} \quad (11)$$

In Eq. (11), the risk-level weights (w_j) will influence the estimated marginal effect of an explanatory variable in j-th risk source.

5.5 Data Description

To empirically test the proposed methodology, the multivariate multiple risk source model is applied to traffic crashes along rural two-lane highway segments in Wisconsin, United States. Crash data for this network is available in "KABCO" scale containing a total count of crashes that occurred on Wisconsin state highways between 2011 and 2015. The KABCO injury codes

presented in the dataset were consolidated into two levels in this study – injury crashes (K, A, B, and C) and non-injury crashes (O) to ensure that a sufficient number of observations was available in each crash severity level. A similar approach has been used by other researchers to ensure sufficient sample size for model estimation (Milton et al. 2008, Islam et al. 2014b, Uddin and Huynh 2018).

Crash contributing factors were collected for two distinct risk sources (J=2) including engineering and behavioral factors. The factors within engineering risk source include typical roadway geometric factors and traffic features for rural two-lane highways which were collected from MetaManager, a data management system developed and maintained by the Wisconsin Department of Transportation (St Clair 2001). Typical roadway geometry related variables such as segment length, lane width, and shoulder width variables were collected from roadway inventory data table. The percent passing and posted speed limit variables were collected from the mobility data table. Please note that the segmentation of Wisconsin two-lane highways does not involve horizontal and vertical curves. This means the highway is not segmented at the starting or ending of a horizontal or vertical curve. To provide information on curves, no passing zone is generated from MARKINGview, an asset management tool used to capture and maintain traffic marking information and location of no passing zones. No passing zones are usually marked with a solid yellow line placed on hills or curves where you cannot see far enough ahead to pass safely. In the mobility data table, no passing zones are expressed in percentage of segment length. AADT and percent of truck on each segment were collected as traffic-related features for model development in this study.

To conform to the study hypothesis of multiple risk sources, the behavioral factors used in this study are solely generated from a different risk source than engineering risk factors. The

behavioral factors used in this study were collected from the Uniform Crime Reporting (UCR) program in Wisconsin. The UCR program provides crime and arrest data from local law enforcement agencies to the Federal Bureau of Investigation and Bureau of Justice Information and Analysis (FBI). The collected behavioral factors include the operating a motor vehicle while intoxicated (OWI) rate, drug-related arrest rate, and violent crime rates for each county between 2013 and 2014. In this context, the rate is defined as the total count per 1 million people within a defined geographic area (e.g., county). Collected variables were averaged for each county over a two-year period to represent the behavior risk source. Community-level factors that heighten the risk of experiencing problems with alcohol include the per capita number of alcohol outlets in a community. The liquor license rate, which can be defined as the number of liquor outlet licenses per 500 people, was collected from the Wisconsin Department of Revenue for 2013 and 2014, and the average value was used (Wisconsin Department of Revenue).

Alcohol impairment of driving skill has been identified as a major traffic safety problem since early 20th century (Blomberg et al. 2005). National Survey on Drug Use and Health's (NSDUH) "State Estimates of Drunk and Drugged Driving" report released in 2012 indicates the prevalence of alcohol-impaired driving in Wisconsin are among the highest in the nation (Substance Abuse and Mental Health Services Administration 2012). Besides alcohol impairment, drug-impaired driving has recently started raising government and public concerns in the USA as well as other countries (Walsh et al. 2008, Asbridge et al. 2012). In Wisconsin, the drug law violations are defined as the violation of laws prohibiting the production, distribution, and/or use of certain controlled substances and the equipment or devices utilized in their preparation and/or use. This includes the unlawful cultivation, manufacture, distribution, sale, purchase, use, possession, transportation, or importation of any controlled drug or narcotic

substance. Based on above definition and observed consequences of impaired driving, both OWI and drug arrest rate can be considered as direct measures of crash risks the transportation network is exposed to in a community (e.g., county). Violent crimes used in this study includes murder and nonnegligent manslaughter, rape, robbery, and aggravated assault.

The collected behavior variables were then linked to the roadway database using the spatial join tool in ArcMap, a geographic information system software package. The complete rural two-lane highway dataset contains 9,605 segments extended over 8,669 miles after cleaning for missed observations for target variables and very short segments. A sample of 6,000 segments extended over 5,400 miles from the complete dataset was used to evaluate the proposed methodology in this study. Table 1 provides the summary statistics of the variables used for this study.

Table 5-1 Summary Statistics of Wisconsin Dataset.

Variables	Description	Mean	Standard Deviation	Minimum	Maximum
Crash Data					
"K+A" Crashes	Serious Injury Crashes	0.353	0.685	0	8
"B+C" Crashes	Minor Injury Crashes	1.511	2.786	0	67
"K+A+B+C" Crashes	Total Injury Crashes	1.864	3.067	0	70
"O" Crashes	Non-Injury Crashes	3.414	5.422	0	116
Exposure					
AADT	Annual Average Daily	3916.855	3106.853	80	66712

	Traffic				
Engineering Risk Source					
Length	Segment Length in miles	0.900	0.423	0.050	2.600
LW	Lane Width in feet	12.020	0.875	9	20
SW	Shoulder Width in feet	6.771	2.797	0	15.500
Truck	Percentage of Heavy Truck	10.797	4.334	0	34.700
Speed	Posted Speed Limit in miles per hour	52.667	6.398	30	70
Passing	Percent Passing	0.464	0.270	0	1
Behavioral Risk Source					
OWI	Operating While Intoxicated citation rate per Million population	46.864	24.653	6.100	335.850
Drug Arrest	Drug Arrest rate per Million population	37.052	20.745	3.750	195.750
Violent Crime	Violent Crime rate per Million population	13.715	11.922	0.700	95.100
Liquor License	Liquor license rate per 500 population	2.220	1.256	0.900	8.300

5.6 Results and Discussion

This section of the paper explains the application of the multivariate multiple risk source regression model to estimate crash count and injury severity simultaneously. The models developed for this study were designed to estimate injury and non-injury crashes (K=2). The performance of the multivariate multiple risk source model was compared with the NB model, a

univariate multiple risk source model, for both severity levels to determine whether the multivariate approach of multiple risk source model was theoretically sound and offered improved model performance. A total of two (2) MCMC chains were used to implement all models in the Bayesian framework. Model convergence was obtained through 130,000 iterations, and 30,000 samples were used as burn-in period. The Gelman–Rubin convergence statistics (G-R statistics) were reviewed to verify the model convergence (i.e., when the G-R statistic is less than 1.2) (Mitra and Washington 2007).

Table 2 summarizes the modeling results for both injury and non-injury crashes in the study dataset. Table 2 shows that for the NB model, the estimated 95 percent posterior credible intervals for all coefficients in both injury and non-injury crashes did not include zero; hence, all coefficients are statistically significant at a 5 percent significance level. Though the drug arrest rate variable was statistically significant in predicting no injury crashes for the NB model, the variable was not statistically significant at 5 percent with both univariate and multivariate multiple risk source models. Afghari et al. (2016) found similar results when comparing the NB and multiple risk source model. It is found that six out of nine explanatory variables were not statistically significant in the multiple risk source model, but they were statistically significant in the single source NB model. The drug arrest rate was statistically significant in predicting injury crashes in all models, indicating driving under the influence of drug results in more injury crashes. All other posterior mean estimates of explanatory variables were statistically significant at a 5 percent significance level in both univariate and multivariate multiple risk source models.

Table 5-2 Non-Injury and Injury Crash Modeling Results.

Parameters	**Non-Injury**	**Injury**

	NB Model	Univariate Multiple Risk Source Model	Multivariate Multiple Risk Source Model	NB Model	Univariate Multiple Risk Source Model	Multivariate Multiple Risk Source Model
Exposure: AADT	0.814	0.813	0.817	0.780	0.778	0.780
	(0.021)	(0.022)	(0.021)	(0.025)	(0.025)	(0.024)
Inverse-Dispersion	0.460			0.490		
	(0.016)			(0.022)		
Engineering Variables						
Constant	-4.162	-3.680	-3.578	-4.408	-3.901	-3.956
	(0.058)	(0.339)	(0.224)	(0.195)	(0.745)	(0.532)
Length	0.930	1.219	1.252	1.010	1.146	1.184
	(0.033)	(0.082)	(0.056)	(0.038)	(0.049)	(0.046)
Lane Width	-0.034	-0.045	-0.046	-0.074	-0.103	-0.095
	(0.014)	(0.022)	(0.021)	(0.017)	(0.027)	(0.024)
Shoulder Width	-0.042	-0.041	-0.047	-0.040	-0.033	-0.035
	(0.005)	(0.008)	(0.007)	(0.006)	(0.008)	(0.007)
Truck Percentage	-0.016	-0.020	-0.021	-0.015	-0.015	-0.016
	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Speed Limit	-0.023	-0.023	-0.025	-0.017	-0.010	-0.011
	(0.002)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)
Percent Passing	0.215	0.432	0.402	0.154	0.305	0.280
	(0.046)	(0.073)	(0.063)	(0.054)	(0.067)	(0.064)
Behavioral Variables						
Constant		-6.716	-6.355		-7.501	-7.010
		(1.110)	(0.690)		(1.294)	(0.906)
OWI Rate	-0.004	-0.014	-0.015	-0.004	-0.016	-0.025
	(0.001)	(0.004)	(0.004)	(0.001)	(0.005)	(0.005)
Drug Arrest Rate	0.002	***0.006***	***0.005***	0.003	0.017	0.022
	(0.001)	***(0.003)***	***(0.004)***	(0.001)	(0.007)	(0.007)
Violent Crime Rate	0.007	0.017	0.022	0.009	0.018	0.023
	(0.001)	(0.003)	(0.005)	(0.001)	(0.003)	(0.003)

Liquor License Rate	-0.136 (0.117)	-0.773 (0.206)	-0.726 (0.162)	-0.125 (0.014)	-0.985 (0.316)	-1.293 (0.29)
Risk-level Weights						
Engineering Risk		0.700 (0.116)	0.700 (0.132)		0.700 (0.121)	0.700 (0.138)
Behavioral Risk		0.300 (0.116)	0.300 (0.132)		0.300 (0.121)	0.300 (0.138)
Correlation between Risk Sources						
$\boldsymbol{\sigma_{11}}$		0.539 (0.159)	0.214 (0.101)		0.524 (0.162)	0.198 (0.091)
$\boldsymbol{\sigma_{22}}$		1.133 (0.497)	0.411 (0.240)		1.295 (0.610)	0.456 (0.304)
$\boldsymbol{\sigma_{12} = \sigma_{21}}$		0.652 (0.226)	***-0.154*** ***(0.134)***		0.697 (0.303)	***-0.066*** ***(0.135)***

Note: 1) Parameter estimates presented in bold and italic font is not significant at 5 percent significance level; 2) The estimated standard error of mean parameter estimate is presented in parenthesis.

Note that the posterior mean of the estimated parameters of explanatory variables cannot be directly compared between single risk and multiple risk source models, except for the exposure measure, because of the associated risk level weights of each variable. In non-injury crashes, the posterior mean of the estimated parameters for AADT ranges from 0.813 to 0.817 across three modeling approaches, indicating the multiple risk source modeling technique can maintain enough strength to estimate the Poisson mean while considering multiple risk sources. The posterior mean of the estimated parameter for AADT is positive for both for both severity levels implying that this variable has increasing effect on the number of injury and non-injury crashes.

In multiple risk source models, factors contributing to crashes are separated into two distinct sources: engineering and behavioral factors. The mean of posterior parameter estimate of

risk-level weights indicates that on average, 70% of both injury and no injury crashes occur due to the engineering risk source, whereas behavioral risks contribute to 30% of the injury and no injury crashes in both univariate and multivariate multiple risk source models. The statistically significant covariates are similar across all models with regard to the engineering risk source for both injury and non-injury crashes. All parameter-mean estimates for explanatory variables in the engineering risk source have similar signs, indicating similar positive or negative effects on crash risk across modeling alternatives. The parameter-mean estimates for the engineering risk variable are similar when comparing between univariate and multivariate modeling approaches; this indicates that expanding the multiple risk source methodology to a multivariate structure can provide stable parameter estimates. The estimated standard deviation of most mean posterior parameter estimates in the multivariate multiple risk source model is smaller than the estimates for the univariate model; this indicates that more accurate parameters can be estimated using the multivariate structure, as noted in the literature (Park and Lord 2007).

The posterior parameter estimates for the behavioral risk source variables paint a similar picture as the variables in the engineering risk source. The estimated parameter-mean values are mostly similar in both univariate and multivariate multiple risk source models. For instance, the mean of the posterior parameter estimates for OWI rate yielded a negative impact on crash risk, indicating that both injury and non-injury crash rates tend to decrease with an increase in OWI rate. The mean of the posterior parameters estimated for the liquor license variable for both injury and non-injury crashes suggests a negative impact on crash risk. Both estimates from the data seem counterintuitive if the OWI rate or the number of liquor licenses is regarded as the positive effect of liquor consumption to driving. OWI rate can be further regarded as a proxy of the number of drunk drivers who are more likely to be involved in a crash than sober ones.

Unfortunately, such findings do not necessarily lead to a conclusive explanation as high OWI arrests may suggest intensive enforcement activities or more effective enforcement strategies. A meta-analysis shows that drink-driving checkpoints reduce alcohol-related crashes by 17% at a minimum and all crashes by 10-15% (Erke et al. 2009). In spite of an endeavor to collect information on enforcement, the data were incomplete and inconsistent and not helpful for this study. Another caveat in the UCR dataset is that liquor licenses are not separated by bar, restaurant, and off-premise liquor outlets, as several studies noted that crash risk increases with bar and off-premise liquor outlets but decreases for restaurants with a liquor license (Treno et al. 2007, Gruenewald and Johnson 2010). Hence, the effects of these behavioral variables on crashes can be revealed and estimated via crash modeling but defining a cause-effect relationship requires additional information.

The mean posterior parameter estimates for the covariance matrix in the univariate multiple risk source model were found statistically significant at 5% significance level. This indicates that the two risk sources considered in the model are distinct and related. With the multivariate modeling approach, the posterior mean estimates of the covariance term between risk sources (mean: -0.154, std. dev.=0.134) indicate that they are no longer statistically interrelated at a 5% significance level. Based on the posterior density of Σ_M, statistically significant positive correlations ($\boldsymbol{\sigma_{12}} = \boldsymbol{\sigma_{21}} = 0.629$) exist between crash counts at different levels of severity within a segment. The univariate risk source model is a special case of the multivariate multiple risk source model, with off-diagonal elements of $\Sigma_M$equal to zero. By incorporating a statistically significant correlation in the modeling structure, the correlation in injury severity counts was incorporated into the model framework.

Based on the above discussion on modeling results, it can be noted that a significant correlation exists between crash counts for different injury severity level. As described in the methodology section, this correlation influences the estimation of model parameters (Park and Lord 2007). However, the posterior mean of the covariance matrix for the risk-level error term is no longer significant when the correlation between crash counts for different injury severity level is considered. The variance estimates indicate that the risk sources are indeed distinct for both injury and non-injury crashes; this suggests that the statistically significant correlation between engineering and behavioral risk sources can be a statistical artifact resulting from the absence of injury severity in the model. Hence, the multivariate multiple risk source regression model can provide informative parameter inferences with the existence of Σ_M and uncorrelated risk sources. A modified model was estimated with an uncorrelated error structure between risk sources. Modeling results show that parameter estimates for all covariates in both the engineering and behavioral risk sources yielded similar coefficients; thus, the results without correlation structure between risk sources were not presented here.

For the convenience of comparing the effect of individual factors from both engineering risk source and behavior risk source, marginal effects are estimated. The marginal effects of explanatory variables for both PDO and injury crashes are presented in Table 3.

Table 5-3 Average Marginal Effects for Non-Injury and Injury Crashes.

Variables	**Non-Injury**			**Injury**		
	Single Source NB Model	Univariate Multiple Risk Source Model	Multivariate Multiple Risk Source Model	Single Source NB Model	Univariate Multiple Risk Source Model	Multivariate Multiple Risk Source Model

Exposure: AADT	6.489	6.481	6.513	6.218	6.202	6.242
Engineering Risk						
Length	0.837	0.768	0.789	0.909	0.722	0.746
Lane Width	-0.408	-0.379	-0.387	-0.889	-0.866	-0.799
Shoulder Width	-0.284	-0.194	-0.223	-0.271	-0.156	-0.166
Truck Percentage	-0.173	-0.151	-0.159	-0.162	-0.113	-0.121
Speed Limit	-1.211	-0.848	-0.922	-0.895	-0.369	-0.405
Percent Passing	0.100	0.140	0.130	0.071	0.099	0.091
Behavioral Risk						
OWI Rate	-0.187	-0.197	-0.211	-0.187	-0.525	-0.722
Drug Arrest Rate	0.074	0.067	0.056	0.111	0.189	0.245
Violent Crime Rate	0.096	0.070	0.091	0.123	0.074	0.095
Liquor License Rate	-0.302	-0.515	-0.483	-0.277	-0.656	-0.861

According to Table 3, single risk models usually overestimate the contribution of a specific variable if this factor originates from a dominating source. For majority of the engineering risk factors, the estimated marginal effect is higher with a single risk source NB model than with a multiple risk source model. The estimated marginal effects of behavioral risk variables indicate that the single risk source NB model underestimates the effect of some variables related to the behavioral risk source; this may be why the effects of some variables related to the engineering risk source are overestimated. For example, the estimated marginal effect for OWI citation rate using single source NB model indicates that injury crash counts can decrease by 0.187 unit with a unit increase in OWI citation rate, whereas the multiple risk source model yielded a 0.525 and 0.722 unit increase in injury crash counts with an increase in OWI

citation rate for the univariate and multivariate modeling approach, respectively. For injury crashes, the effect of shoulder width and speed limit may be overestimated in single risk models. Moreover, the marginal effect estimates indicate that single risk source models may underestimate the effect of OWI citation rate and liquor license rate. Comparing both posterior mean of parameter estimates and marginal effects, it can be noted that all variables have similar direction (positive or negative) in both single source and multiple source regression models. But the estimated marginal effects are significantly different for variables originated from supporting risk source such as behavioral risk variables in this study. Thus, a caution should be used while interpreting the parameter estimated from single source model if variables used for model development are generated from different risk sources.

5.7 Prediction Accuracy

Table 4 provides the performance comparison for all models based on the Deviance Information Criterion (DIC). The DIC is a widely used GOF statistic for comparing models in a Bayesian framework (Spiegelhalter et al. 2002). The DIC consists of two components: (a) a measure of how well the model fits the data, Dbar ($\overline{D(\theta)}$) and (b) a measure of model complexity (pD). Thus, DIC can provide a better comparison between models that are characterized by different complexities. The likelihood of a Bayesian model can be represented by Dbar ($\overline{D(\theta)}$) and Dhat ($\widehat{D(\theta)}$). Dbar is the posterior mean of the deviance, whereas Dhat is a point estimate of the deviance. Mean Absolute Deviance (MAD) was estimated for each model to compare predictive accuracy. MAD can be calculated as follows:

$$MAD = \frac{1}{N}\sum |y_{it} - \hat{y}_{it}| \quad (13)$$

Where, N indicates the number of observations in the dataset.

Table 5-4 Comparison of Model Performance.

Methodology	Crash Type	$\bar{D}$	$\hat{D}$	DIC	pD	MAD
Single Source NB model	PDO	25040	25030	25060	14.13	2.325
	Injury	19880	19860	19890	13.08	1.413
	Total	**44920**	**44890**	**44950**	**27.21**	**1.869**
Univariate Multiple Risk Source model	PDO	20640	18900	22370	1734	2.286
	Injury	16910	15440	18370	1464	1.371
	Total	**37550**	**34340**	**40740**	**3198**	**1.829**
Multivariate Multiple Risk Source model w/ Correlated Error Structure	PDO	20360	18960	21750	1398	2.280
	Injury	16460	14910	18020	1308	1.353
	Total	**36820**	**33870**	**39770**	**2706**	**1.817**
Multivariate Multiple Risk Source model w/o Correlated Error Structure	PDO	20106	19554	20660	552.4	2.271
	Injury	16142	15057	16862	718.8	1.322
	Total	**36108**	**34766**	**37382**	**1271.2**	**1.797**

A comparison of the DIC values between models illustrates that the multivariate multiple risk source regression model with uncorrelated error structure between risk sources performed better than other models. It is evident that excluding the correlation structure will result in a smaller effective number of parameter (pD) which will influence the estimation of the DIC value. The Dbar estimate indicates that the posterior mean of deviance is the smallest for the multivariate multiple risk source model without a correlated error structure compared with all other models. There is a significant improvement in DIC value with the multivariate modeling

approach compared with the univariate multiple risk source model. The MAD estimates also indicate that multivariate risk source regression models can better predict both PDO and injury crashes compared with other models in this study.

5.8 Practical Implications

One major benefit of the multiple risk source model over a single risk model is that risk-level predicted crash counts can be obtained from the former model which is not possible with the latter model. In the literature, driver error and engineering risk factors are identified as two major sources for crash occurrences (Shaon et al. 2018). In Wisconsin, detailed crash report for each crash occurred on state trunk network is documented in Wisconsin Motor Vehicle Accident Reporting Form 4000 (MV4000) by investigating police officer(s) (WisDOT, Parker and Tao 2006). The crash report for each crash includes fourteen specific driver-related factors and thirteen specific highway factors (e.g., geometry and pavement condition) that contribute to the occurrence of each crash. Table 6 describes the list of crash contributing circumstances listed in the MV4000 database.

Table 5-5 Possible Crash Contributing Circumstances listed in the MV4000 Database.

Driver-related Factors	Highway-related Factors

• Driver condition • Physically disabled • Disregard traffic control • Following too close • Failure to yield • Failure to keep vehicle under control • In conflict • Inattentive driving	• Improper overtake • Improper turn • Left of center • Exceeding speed limit • Too fast for conditions • Unsafe braking • Others	• Snow/ Ice/ Wet • Narrow shoulder • Soft shoulder • Loose gravel • Rough pavement • Debris prior to accident	• Other debris • Sign obscured/ missed • Narrow bridge • Construction zone • Visibility obscured • Others

A crash can be linked to behavior or engineering risk related crashes using these specific contributing factors noted in the MV4000 crash report. Using crash dataset from Wisconsin, Shaon and Qin found that 79% of total crashes are related to driver error (Shaon et al. 2018). Please note that the engineering risk source variables used in this study were collected for each segment whereas the behavior variables are collected for each county and used as a proxy variable for behavioral risk in crash occurrence. Although the modeling results indicate that 30% of total crashes are generated due to behavioral risk source, this statement does not validate with the information listed by investigating police officer(s) for each crash. This may be because of the unavailability of important behavioral risk source variables that has a significant contribution to crash occurrence such as speeding behavior, fatigue or distracting driving, etc. Considering above-mentioned limitations, a comparison of predicted crashes between single and multiple risk source modeling was conducted to illustrate the strength of multiple risk source modeling. Table 6 described the predicted crash comparison for five rural two-lane sites based on observed non-injury crashes related to different risk sources.

Table 5-6 Comparison of Observed and Predicted Crash Counts between Single Source and Multiple Source Models.

Site ID	County	Total Observed Crashes	Observed Behavior Risk Related Crashes	Observed Engineering Risk Related Crashes	Predicted Crashes from Single Source NB Model	Predicted Behavior Crashes	Predicted Engineering Crashes
A	RICHLAND	6	1	5	4	1	4
B	GREEN	7	3	4	5	1	5
C	DODGE	7	2	5	2	1	2
D	JEFFERSON	7	2	5	9	3	6
E	JUNEAU	13	7	6	5	1	5

In Table 6, the comparison sites were selected where at least one behavior crash was observed. From the overall comparison between single and multiple risk source models in Table 6, it can be noted that the latter model predicted more crashes compared to single source model. Afghari et al. also found similar information while identifying crash blackspots using multiple risk source model (Afghari et al. 2016).

The single risk source NB model can only predict total crash counts for a specific site. The decomposition of observed crashes described in Table 6 indicates that crashes may come from different risk sources. For example, For Site “A”, there were a total of 6 crashes observed on that segment which includes 1 and 5 crashes occurred due to behavior and engineering risk source, respectively. The predicted value from the single source model was 4 which indicates both engineering and behavior risk variables contributes to all 4 crashes. On the other hand, multiple risk source model predicts there are 4 crashes occurred due to engineering risk source

and 1 crashes occurred due to behavioral risk source. The proposed model can also indicate limitation of important information by risk source. For Site “E”, 13 crashes were observed which include 7 and 6 crashes due to behavior and engineering risk, respectively. The predicted value from the single risk model was 5 indicating underestimation of crashes for that site. The multiple risk source model predicted 1 and 5 crashes occurred due to behavior and engineering risk, respectively. While multiple risk source model predicted engineering crashes near observed value (predicted 5 crashes out of 6 observed crashes), the behavior-related crashes were underestimated. This indicates important behavior information is needed to explain observed crashes. Considering the data limitation in this study, it can be noted that multiple risk source model is capable of predicting risk-level crashes which can help safety professionals to identify crash black-spots by risk source and design effective crash countermeasures.

5.9 Conclusions

Previous studies have explored many factors that could contribute to crash occurrence. Understanding crash-generating mechanisms, adopting appropriate hypotheses, and producing reliable parameter estimates from modeling crash data are challenging for researchers and traffic safety professionals. This study explored the influence of factors from distinct risk sources on crash occurrences while estimating crash frequency and injury severity simultaneously. While engineering risk factors were extensively utilized in crash modeling literature, use of behavioral risk factors are limited due to data unavailability. This study explored behavioral variables collected at larger geographic scale representing existing social norms that can influence driving behavior within a community. In association with engineering risk factors, these behavioral

variables are considered to compose crash risk which is originated from a distinct source. The underlying hypothesis of the proposed modeling approach is that crash counts of different injury severities are correlated, and unobserved heterogeneity cannot be sufficiently captured using a single equation crash frequency model. While a large number of studies explored multivariate models to account for the correlation between injury severities, they did not distinguish between sources of crash risk. The complicated crash generation process can be addressed by considering multiple risk sources through the proposed method. Expanding univariate multiple risk source regression modeling to a multivariate framework enabled the incorporation of both injury count correlation and distinguish between crash risk from different sources.

The proposed models were applied to a crash count dataset from Wisconsin rural two-lane highways. Two distinct risk-generating sources – engineering and behavioral – were considered. The modeling results were compared with a single equation NB model and univariate multiple risk source model. Results showed that the multivariate multiple risk source regression model has the best prediction performance among all developed models. A sample crash count comparison for five sites indicated that the proposed model can predict crashes from each risk source separately which cannot be obtained from single equation modeling. The study not only demonstrates a unique approach to explicitly incorporating behavioral factors into crash prediction models but also provides more insight into the sources of crash risk, which can be used to better inform safety practitioners and guide roadway improvement programs.

The proposed multivariate multiple risk source regression model was developed using the Bayesian framework. Despite the potential of the proposed methodology, the modeling framework may introduce computational complexity and data-specific effects. The risk-level weights used to link predicted crash risk from each risk-generating source to total crash count is

solely data dependent. Future research should explore prior knowledge of risk distribution and use it as prior information in model development. Unobserved heterogeneity can be a major issue with crash datasets. It is also important to understand the source of unobserved heterogeneity so that appropriate caution can be taken during model development. Random parameters modeling is a well-accepted methodology to address unobserved data heterogeneity in crash datasets. Though there are no theoretical limitations with regard to implementing random parameters into the multiple risk source structure, the proposed multivariate models were assumed to have fixed parameters. The random parameters structure for covariates in each risk source can be explored in future studies to improve the accuracy of the proposed model and better understand the sources of heterogeneity in crash datasets.

5.10 References

Abdel-Aty, M., Keller, J., Brady, P., 2005. Analysis of types of crashes at signalized intersections by using complete crash data and tree-based regression. Transportation Research Record: Journal of the Transportation Research Board (1908), 37-45.

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

Afghari, A.P., Haque, M.M., Washington, S., Smyth, T., 2016. Bayesian latent class safety performance function for identifying motor vehicle crash black spots. Transportation Research Record: Journal of the Transportation Research Board (2601), 90-98.

Afghari, A.P., Washington, S., Haque, M.M., Li, Z., 2018. A comprehensive joint econometric model of motor vehicle crashes arising from multiple sources of risk. Analytic Methods in Accident Research 18, 1-14.

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

Ando, R., Higuchi, K., Mimura, Y., 2018. Data analysis on traffic accident and urban crime: A case study in toyota city. International Journal of Transportation Science and Technology 7 (2), 103-113.

Asbridge, M., Hayden, J.A., Cartwright, J.L., 2012. Acute cannabis consumption and motor vehicle collision risk: Systematic review of observational studies and meta-analysis. Bmj 344, e536.

Bahar, G., Erwin, T., Mackay, M., Smiley, A., Tighe, S., Zein, S., 2004. Canadian guide to in-service road safety reviews Transportation Association of Canada, Ottawa.

Blomberg, R.D., Peck, R.C., Moskowitz, H., Burns, M., Fiorentino, D., 2005. Crash risk of alcohol involved driving: A case-control study.

Box, S., 2009. New data from vtti provides insight into cell phone use and driving distraction. Virginia Tech Transportation Institute 27.

Carter, J.G., Piza, E.L., 2017. Spatiotemporal convergence of crime and vehicle crash hotspots: Additional consideration for policing places. Crime & Delinquency, 0011128717714793.

Carter, P.M., Bingham, C.R., Zakrajsek, J.S., Shope, J.T., Sayer, T.B., 2014. Social norms and risk perception: Predictors of distracted driving behavior among novice adolescent drivers. Journal of Adolescent Health 54 (5), S32-S41.

Chib, S., Winkelmann, R., 2001. Markov chain monte carlo analysis of correlated count data. Journal of Business & Economic Statistics 19 (4), 428-435.

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

Chiou, Y.-C., Fu, C., 2015. Modeling crash frequency and severity with spatiotemporal dependence. Analytic Methods in Accident Research 5, 43-58.

Compton, R., Vegega, M., Smither, D., 2009. Drug-impaired driving: Understanding the problem and ways to reduce it: A report to congress.

El-Basyouny, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. Accident Analysis & Prevention 41 (5), 1118-1123.

Erke, A., Goldenbeld, C., Vaa, T., 2009. The effects of drink-driving checkpoints on crashes—a meta-analysis. Accident Analysis & Prevention 41 (5), 914-923.

Fbi, Uniform crime reporting (ucr) program Federal Bureau of Investigation.

Fhwa, 2017. Proven safety countermeasures. Office of Safety.

Fitzpatrick, K., Lord, D., Park, B.-J., 2010. Horizontal curve accident modification factor with consideration of driveway density on rural four-lane highways in texas. Journal of transportation engineering 136 (9), 827-835.

Garber, N., Ehrhart, A., 2000. Effect of speed, flow, and geometric characteristics on crash frequency for two-lane highways. Transportation Research Record: Journal of the Transportation Research Board (1717), 76-83.

Geedipally, S., Patil, S., Lord, D., 2010. Examination of methods to estimate crash counts by collision type. Transportation Research Record: Journal of the Transportation Research Board (2165), 12-20.

Geedipally, S.R., Lord, D., 2010. Investigating the effect of modeling single-vehicle and multi-vehicle crashes separately on confidence intervals of poisson–gamma models. Accident Analysis & Prevention 42 (4), 1273-1282.

Geedipally, S.R., Lord, D., Dhavala, S.S., 2012. The negative binomial-lindley generalized linear model: Characteristics and application using crash data. Accident Analysis & Prevention 45, 258-265.

Gruenewald, P.J., Johnson, F.W., 2010. Drinking, driving, and crashing: A traffic-flow model of alcohol-related motor vehicle accidents. Journal of studies on alcohol and drugs 71 (2), 237-248.

Harmon, T., Bahar, G., Gross, F., 2018. Crash costs for highway safety analysis.

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

Islam, M.S., Ivan, J.N., Lownes, N.E., Ammar, R.A., Rajasekaran, S., 2014a. Developing safety performance function for freeways by considering interactions between speed limit and geometric variables. Transportation Research Record 2435 (1), 72-81.

Islam, S., Jones, S.L., Dye, D., 2014b. Comprehensive analysis of single-and multi-vehicle large truck at-fault crashes on rural and urban roadways in alabama. Accident Analysis & Prevention 67, 148-158.

Lascala, E.A., Gerber, D., Gruenewald, P.J., 2000. Demographic and environmental correlates of pedestrian injury collisions: A spatial analysis. Accident Analysis & Prevention 32 (5), 651-658.

Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: An empirical analysis. Accident Analysis & Prevention 34 (2), 149-161.

Lord, D., Persaud, B., 2000. Accident prediction models with and without trend: Application of the generalized estimating equations procedure. Transportation Research Record: Journal of the Transportation Research Board (1717), 102-108.

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 & Prevention 37 (1), 35-46.

Lyon, C., Oh, J., Persaud, B., Washington, S., Bared, J., 2003. Empirical investigation of interactive highway safety design model accident prediction algorithm: Rural intersections. Transportation Research Record: Journal of the Transportation Research Board (1840), 78-86.

Ma, J., Kockelman, K.M., 2006. Bayesian multivariate poisson regression for models of injury count, by severity. Transportation Research Record 1950 (1), 24-34.

Ma, J., Kockelman, K.M., Damien, P., 2008. A multivariate poisson-lognormal regression model for prediction of crash counts by severity, using bayesian methods. Accident Analysis & Prevention 40 (3), 964-975.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Analytic methods in accident research 1, 1-22.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic methods in accident research 11, 1-16.

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 (1376).

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

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

Mitra, S., Washington, S., 2007. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis & Prevention 39 (3), 459-468.

Mitra, S., Washington, S., 2012. On the significance of omitted variables in intersection crash modeling. Accident Analysis & Prevention 49, 439-448.

Moeckli, J., Lee, J.D., 2007. The making of driving cultures. Improving Traffic Safety Culture in the United States 38 (2), 185-192.

Montella, A., Imbriani, L.L., 2015. Safety performance functions incorporating design consistency variables. Accident Analysis & Prevention 74, 133-144.

Nagle, M., 2012. Indiana traffic safety facts: Driver history and crash outcomes 2011.

Nhtsa, 2008. National motor vehicle crash causation survey: Report to congress. National Highway Traffic Safety Administration Technical Report DOT HS. National Highway Traffic Safety Administration, pp. 059.

Nhtsa, 2010. Traffic safety facts: Driver electronic device use observation protocol. DOT HS 811, 361.

Oh, J., Washington, S., Choi, K., 2004. Development of accident prediction models for rural highway intersections. Transportation Research Record: Journal of the Transportation Research Board (1897), 18-27.

Park, E., Lord, D., 2007. Multivariate poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record: Journal of the Transportation Research Board (2019), 1-6.

Parker, S.T., Tao, Y., 2006. Wistransportal: A wisconsin traffic operations data hub. Applications of advanced technology in transportation. pp. 611-616.

Pei, X., Wong, S., Sze, N.-N., 2011. A joint-probability approach to crash prediction models. Accident Analysis & Prevention 43 (3), 1160-1166.

Peng, Y., Lord, D., Zou, Y., 2014. Applying the generalized waring model for investigating sources of variance in motor vehicle crash analysis. Accident Analysis & Prevention 73, 20-26.

Persaud, B.N., 2001. Statistical methods in highway safety analysis.

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

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

Qin, X., Rahman Shaon, M.R., Chen, Z., 2016. Developing analytical procedures for calibrating the highway safety manual predictive methods. Transportation Research Record: Journal of the Transportation Research Board (2583), 91-98.

Qin, X., Chen, Z. and Rahman Shaon, R., 2018. Developing jurisdiction-specific SPFs and crash severity portion functions for rural two-lane, two-way intersections. Journal of Transportation Safety & Security, pp.1-13.

Quddus, M., Chin, H., Wang, J., 2001. Motorcycle crash prediction model for signalised intersections. WIT Transactions on the Built Environment 52.

Rahman Shaon, M.R., Qin, X., 2016. Use of mixed distribution generalized linear models to quantify safety effects of rural roadway features. Transportation Research Record: Journal of the Transportation Research Board (2583), 134-141.

Rahman Shaon, M.R., Qin, X., Shirazi, M., Lord, D. and Geedipally, S.R., 2018. Developing a Random Parameters Negative Binomial-Lindley Model to analyze highly over-dispersed crash count data. Analytic methods in accident research, 18, pp.33-44.

Redelmeier, D.A., Tibshirani, R.J., 1997. Association between cellular-telephone calls and motor vehicle collisions. New England Journal of Medicine 336 (7), 453-458.

Rumar, K., 1985. The role of perceptual and cognitive filters in observed behavior. Human behavior and traffic safety. Springer, pp. 151-170.

Sabey, B.E., Staughton, G., 1975. Interacting roles of road environment vehicle and road user in accidents. Ceste I Mostovi.

Schneider, R.J., Sanatizadeh, A., Shaon, M.R.R., He, Z., Qin, X., 2018. Exploratory analysis of driver yielding at low-speed, uncontrolled crosswalks in milwaukee, wisconsin. Transportation Research Record, 0361198118782251.

Serhiyenko, V., Mamun, S.A., Ivan, J.N., Ravishanker, N., 2016. Fast bayesian inference for modeling multivariate crash counts. Analytic methods in accident research 9, 44-53.

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

Shankar, V., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. Accident Analysis & Prevention 29 (6), 829-837.

Shaon, M.R.R., Qin, X., Chen, Z., Zhang, J., 2018a. Exploration of contributing factors related to driver errors on highway segments. Transportation Research Record, 0361198118790617.

Shaon, M.R.R., Qin, X., Shirazi, M., Lord, D., Geedipally, S.R., 2018b. Developing a random parameters negative binomial-lindley model to analyze highly over-dispersed crash count data. Analytic Methods in Accident Research 18, 33-44.

Shirazi, M., Lord, D., Dhavala, S.S., Geedipally, S.R., 2016. A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: Characteristics and applications to crash data. Accident Analysis & Prevention 91, 10-18.

Smith, D.J., Year. Human factors and traffic crashes. In: Proceedings of the Proceedings of 2000 Midwest Transportation Consortium Transportation Scholars Conference.

Spiegelhalter, D., Thomas, A., Best, N., Lunn, D., 2007. Openbugs user manual, version 3.0. 2. MRC Biostatistics Unit, Cambridge.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64 (4), 583-639.

St Clair, B., Year. Wisdot's meta-manager. In: Proceedings of the 4th National Transportation Asset Management Workshop, September.

Substance Abuse and Mental Health Services Administration, 2012. The nsduh report: State estimates of drunk and drugged driving.

Tarko, A., Inerowicz, M., Ramos, J., Li, W., 2008. Tool with road-level crash prediction for transportation safety planning. Transportation Research Record: Journal of the Transportation Research Board (2083), 16-25.

Tarko, A.P., Kanodia, M., 2004. Hazard elimination program. Manual on improving safety of indiana road intersections and sections. Joint Transp. Res. Program, 1-2.

Treno, A.J., Johnson, F.W., Remer, L.G., Gruenewald, P.J., 2007. The impact of outlet densities on alcohol-related crashes: A spatial panel approach. Accident Analysis & Prevention 39 (5), 894-901.

Uddin, M., Huynh, N., 2018. Factors influencing injury severity of crashes involving hazmat trucks. International journal of transportation science and technology 7 (1), 1-9.

Walsh, J.M., Verstraete, A.G., Huestis, M.A., Mørland, J., 2008. Guidelines for research on drugged driving. Addiction 103 (8), 1258-1268.

Wang, C., Quddus, M.A., Ison, S.G., 2011. Predicting accident frequency at their severity levels and its application in site ranking using a two-stage mixed multivariate model. Accident Analysis & Prevention 43 (6), 1979-1990.

Wang, K., Ivan, J.N., Ravishanker, N., Jackson, E., 2017. Multivariate poisson lognormal modeling of crashes by type and severity on rural two lane highways. Accident Analysis & Prevention 99, 6-19.

Washington, S., Afghari, A.P., Haque, M.M., 2018. Detecting high-risk accident locations. Safe mobility: Challenges, methodology and solutions. Emerald Publishing Limited, pp. 351-382.

Washington, S., Haque, M., 2013. On the commonly accepted assumptions regarding observed motor vehicle crash counts at transport system locations.

Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2010. Statistical and econometric methods for transportation data analysis CRC press.

Weiss, A., 2013. Data-driven approaches to crime and traffic safety (ddacts): An historical overview.

Wisconsin Department of Revenue, Alcohol license overview for wisconsin.

Wisdot, Wisconsin crash data user guide.

Ye, X., Pendyala, R.M., Shankar, V.N., Konduri, K., 2008. Simultaneous equations model of accident frequency by severity level for freeway sections.

Ye, X., Pendyala, R.M., Washington, S.P., Konduri, K., Oh, J., 2009. A simultaneous equations model of crash frequency by collision type for rural intersections. Safety science 47 (3), 443-452.

Zeng, Q., Huang, H., Pei, X., Wong, S., 2016. Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. Analytic methods in accident research 10, 12-25.

Zou, Y., Wu, L., Lord, D., 2015. Modeling over-dispersed crash data with a long tail: Examining the accuracy of the dispersion parameter in negative binomial models. Analytic Methods in Accident Research 5, 1-16.

Chapter 6 Identifying Contributors to Driver Errors and Their Impacts on Crash Severity

This chapter discusses event-oriented CPM development in the 3rd tier of the proposed 3-tier approach. Most of the driver behavior related information are available in the 3rd tier modeling approaches as the information are collected at each crash event. Although a significant amount of research has discussed event-oriented approach by modeling crash severity, exploration of what factors contribute to driver error has not been discussed in literature. To avoid a crash, a driver needs to detect a hazard, decide the safest driving maneuvers, and execute them properly. Driver errors at any of these sequential phases may lead to a crash; therefore, it is necessary to identify the contributing factors and assess their influence on driver behavior. To assist this investigation, a multinomial probit model was employed to study driver errors reported in crashes in both rural and urban state highways in Wisconsin. Furthermore, a quantitative exploration is conducted to quantify the effect of driver errors on crash severity. The broad and insightful information will help researchers and safety professionals to understand when, where, and how the driver error may lead to a crash, how they affect injury severity and to develop cost-effective preventive.

6.1 Introduction

Highway safety analysis is mostly focused on analyzing crash occurrence or severity of a crash, where highway and traffic engineering related data such as roadway geometric characteristics and traffic conditions are used as explanatory variables. It is well known that human factors probably contribute to over 90 percent of the crashes. According to the National Motor Vehicle Crash Causation Study (NMVCCS), almost 94% of the crashes are caused by driver errors (Administration 2008). Without specifically considering driver factors, crash modeling results

may be biased for the effects of explanatory variables. Thus, understanding why drivers make mistakes and how to incorporate human factors and driver behavior into crash prediction has become an increasingly important topic among safety researchers.

Crashes are complex events, evident by the 110 data elements recommended in the Model Minimum Uniform Crash Criteria (MMUCC) (USDOT 2014). Most information can be obtained directly by reviewing detailed crash reports, including police officers' judgment on driver factors contributing to the occurrence of a crash. Crash information can be augmented by socioeconomic, demographic, land use, and traffic patterns to substantiate the knowledge regarding how a driver interacts with and his/her behavior is influenced by roadway design, traffic conditions, and other contextual factors.

For any driver, there is a four-phase process of seeing and reacting to a hazard, i.e., perception, intellection, emotion, and volition, or "PIEV". An error can happen during any of the four phases. In this study, we are particularly interested in understanding when, where, and how drivers make mistakes that attribute to a crash, drawing cues from a comprehensive list of variables ranging from roadway geometry, traffic conditions, roadway, weather, and lighting conditions, events such as construction zone, debris on roadway as well as driver information such as age, gender, vehicle types, etc. Specifically, we followed the categorization method in the NMVCCS study that grouped driver errors into four categories: recognition errors, decision errors, performance errors, and non-performance errors. As each error type is specific and unique, the relating explanatory variables identified through statistical models may be different and more informative to safety professionals. The new insight of the circumstances for driver error and a better understanding of possible causes will shed light on the development of

tangible, practical, and more importantly, targeted and cost-effective enforcement strategies, driver education and training programs, engineering solutions, and vehicle technologies.

6.2 Literature Review

Crash occurrence may be attributed to errors by drivers or the interaction between driver behavior and roadway design features (Hauer 1999). Per police records, driver errors can range from a traffic infraction in which the driver is not paying attention to an intentional traffic violation such as failure to yield or significantly exceeding the speed limit. However, according to the Human Factors Guidelines (HFG) for Road Systems noted: "Road users cannot be expected to solve either highway design or traffic engineering problems without making mistakes and/or compromising operational efficiency and safety" (Campbell 2012).

Understanding the interaction between driver errors and roadway geometric and contextual features is crucial. It has been well established from crash count modeling that roadway geometry, traffic conditions and contextual factors such as weather events are related to crash occurrence, which in-turn have an effect on driver behavior. The American Automobile Association (AAA) Foundation for Traffic Safety estimated that 56% of the fatal crashes that occurred between 2003 and 2007 involved potential aggressive driving behavior, in which speeding was the most common aggressive action that makes up about 31% of total fatal crashes. Hauer noted that the speed at which people choose to travel is affected by roadway design and vehicle characteristics (Hauer 2009). Tate and Turner investigated the relationship between observed travel speed, road geometry, and crashes in New Zealand (Tate and Turner 2007) and concluded that driver's speed choices were more strongly related to curve radius than curve design speed and that the approach speed environment has a significant impact on the speed

choice. Liu and Chen documented that "driving too fast for conditions" was more likely to occur on roads with higher speed limits (50+ mph) as compared to other crashes (Liu and Chen 2009). The authors also noted that a significant proportion of speeding-related crashes occurred on adverse road surface conditions such as "Snowy/Slushy/Icy/Slippery" and "Wet" road pavement compared to other crashes.

Distracted driving is another major driver error that contributes to crashes. Novice drivers appeared to be prone to distraction while driving (NHTSA 2010). Naturalistic driving studies showed that talking on a cell phone raises the risk of collision by more than 30% and drivers that text are at 23 times higher crash risk compared to the non-distracted drivers (Box 2009). The National Occupant Protection Use Survey (NOPUS) conducted annually by the National Center for Statistics and Analysis of NHTSA results showed that females from all age groups are more prone to use electronic devices while driving (Pickrell and Liu 2016). Electronic device use percentage was found similar in age brackets from 16 to 69 years. This suggests that gender might be a more important variable than age cohorts in distracted driving error.

Work zone is a roadway event that has been reported to increase crash rate according to previous literature (Juergens 1962, Graham et al. 1977, Rouphail et al. 1988). Drivers in a construction/work zone encounter a complex array of warning signs, barrels, pylons, construction equipment, and machines, which can create hazards for drivers. The new traffic patterns and challenging roadway configurations in work zones such as stopping or slow traffic, trucks merging from the ramp, uneven pavement, narrowed lanes, absence of shoulders require drivers to operate their vehicles with extra cautions and impose considerable stress on their driving tasks.

Previously, impaired driving has been identified as a contributing factor to driver error (Blomberg et al. 2005). Use of alcohol can significantly affect a driver's decision-making process. Blomberg et al. conducted a case-control study to explore the relationship of Blood Alcohol Concentration (BAC) with relative crash risk (Blomberg et al. 2005). Results showed that elevated relative risk beginning at 0.05 – 0.06% BACs with an accelerating increase in risk at BACs greater than 0.10. In 2015, 41 percent of drivers killed in roadway crashes due to speeding had a BAC of 0.08 G/DL or higher in their blood in the USA (NHTSA 2016). Besides alcohol impairment, drug-impaired driving also has a significant effect on driving behavior (Walsh et al. 2008, Compton et al. 2009, Asbridge et al. 2012). The Governors Highway Safety Association (GHSA) sponsored a study, which found that fatality due to drugged driving surpasses alcohol-impaired fatalities in the United States (Hedlund 2017). In 2015, 43 percent of motorists that died in a road accident had drugs in their system, whereas 37 percent of motorists that died were tested positive for alcohol.

Several studies have investigated driver errors for segment and intersection-related crashes (Bao and Boyle 2009, Liu and Chen 2009, Devlin et al. 2011, Wang and Qin 2015, Dingus et al. 2016), most of which discussed factors contributing to driver error for specific type of crashes such as speed-related crashes. An overall discussion of the factors contributing to driver error is rarely discussed in literature. Wang and Qin investigated the factors contributing to driver errors at uncontrolled, sign-controlled and signal-controlled intersections (Wang and Qin 2015). The authors categorized driver error based on citation type and used roadway characteristics (e.g. presence of curve, visibility, speed limit, etc.), driver characteristics (age, gender, DUI), environmental characteristics (weather condition, roadway condition, lighting condition) and vehicle type (passenger car, light truck, heavy truck) to predict driver errors

collected from crash reports. Sign-controlled intersections are found to have the highest percentage of driver error and reckless driving, followed by signalized and stop-control. Drivers are also more prone to serious errors if their vision is obscured. Adverse environmental characteristics such as snow, ice on the roadway negatively affect the severity of driver error. Driver age, gender, and alcohol or drug use greatly influence the severity of error outcome. The findings confirmed that driver errors are not only the outcome of a driver's psychological behavior but also the interaction with other external factors during the driver's decision-making. Based on previous research, this study is an attempt to investigate the relationship between driver errors and observable factors on highway segments in both rural and urban areas. More information on these contributing factors would help researchers and safety professionals to develop cost-effective countermeasures.

6.3 Methodology

The Multinomial Probit model (MNP) is a discrete outcome model that considers a response variable with three or more levels without accounting for order between levels. Two popular choices to model multiclass categorical variables are Multinomial Logit (MNL) and MNP models. The MNL model is built on the Independence of Irrelevant Alternatives (IIA) assumption, meaning adding or deleting an alternative will not change the ratio between the probabilities of any pair of existing alternatives. In simple words, MNL does not allow for correlation between any pairs of existing alternatives. This may not be always true if the dependent variable categories are correlated. The MNP model relaxes the independence assumption built into the MNL model (Borooah 2002). The driver error categories defined in this study are not independent as they are categorized based on sequential events. From a practical

perspective, it is obvious that performance error depends on the decision of activity the driver tends to execute and decision error depends on the recognition of hazardous situation perceived by the driver. Considering dependency between driver error categories, the MNP model was assumed to be an appropriate choice for model development in this study. The utility function of the MNP model that determines the preference or possible value of attaining the outcome i (i = 1, 2, . . . , I) for observation n can be written as (Greene 2000):

$$U_{in} = \beta_i X_{in} + \varepsilon_{in} \quad (1)$$

$$[\varepsilon_{1n}, \varepsilon_{2n}, \varepsilon_{3n}, \dots\dots, \varepsilon_{in}] \sim MVN(0, \Sigma)$$

Where,

$\mathbf{X_{in}}$ = vector of independent variables for nth observation with ith outcome,

$\boldsymbol{\beta_i}$ = vector of corresponding unknown coefficients, and

$\boldsymbol{\varepsilon_{in}}$ = disturbance term that accounts for unobserved effects.

The disturbance term $\boldsymbol{\varepsilon_{in}}$ for i-th driver error type has a mean of zero and they can be correlated among different error type. Thus, the disturbance vector is defined by a multivariate normal distribution. In terms of log-likelihood that corresponds to the choice of i-th drive error, the choice of i-th driver error can be written as:

$$Prob[Choice_{in}] = Prob[U_{in} > U_{jn}, \ \ j = 1, 2, 3, \dots\dots, I; \ \ i \neq j] \quad (2)$$

Using above formulation, the probability of occurrence of i-th driver error can be written as:

$$Prob[Choice_{in}|X_n] = Prob[(\varepsilon_{in} - \varepsilon_{1n}) > X_n'(\beta_1 - \beta_i), (\varepsilon_{in} - \varepsilon_{2n}) > X_n'(\beta_2 - \beta_i), \ldots, (\varepsilon_{in} - \varepsilon_{(i-1)n}) > X_n'(\beta_{(i-1)} - \beta_i), (\varepsilon_{in} - \varepsilon_{(i+1)n}) > X_n'(\beta_{(i+1)} - \beta_i), \ldots. (\varepsilon_{in} - \varepsilon_{In}) > X_n'(\beta_1 - \beta_I)] \quad (3)$$

The estimated coefficient β_i can be interpreted as the marginal effect of X_i on the log odds-ratio of i-th alternative to the baseline alternative. A "margin" is a statistic computed from predictions from a model while manipulating the values of the covariates. The marginal effect of X_i on the probability of choosing i-th alternative can be expressed as follows:

$$\frac{\partial P_r(Y = Y_n|X_{in})}{\partial X_i} = \frac{\partial E(Y_n|X_i)}{\partial X_i} = \emptyset(\beta_i X_{in})\beta_i \quad (4)$$

Where, $\emptyset$ represents cumulative normal density function. Note that the marginal effect need not have the same sign of β_i.

6.4 Data Description

Segment-related crashes that occurred on the Wisconsin state trunk network system between 2013 to 2015 were collected, excluding deer-related crashes (Parker and Tao 2006). After cleaning all crashes without good location information, 48,441 rural crashes and 46,221 urban crashes were available. Specific driver errors were extracted from the Wisconsin Motor Vehicle Accident Reporting Form 4000 (MV4000), in which the investigating police officers documented detailed accident information (WisDOT, Parker and Tao 2006). There is a list of

fourteen driver-related factors. When a crash is associated with multiple driver factors, the most severe driver-factor is noted based on the police investigation.

Modeling fourteen choices may not be effective because of the sample size, strong correlation between some error types, and difficulties of interpretation. Based on the similarities in driver errors, the NMVCCS study classified driver related critical reasons into recognition errors, decision errors, performance errors, and nonperformance errors (Administration 2008). Recognition error includes driver inattention, internal and external distraction, inadequate surveillance, etc.; aggressive driving behavior, driving too fast, etc. are categorized as decision error; overcompensation, poor directional controls are categorized as performance error; sleep and physical impairment are considered as nonperformance error. This categorization combines driver errors with similar traits. The driver factors in Wisconsin crash data were grouped into NMVCCS' four driver error categories based on the physical meaning of each category. Table 1 shows the NMVCCS driver errors types, and corresponding Wisconsin driver factors along with summary statistics for each category.

Table 6-1 Categorization and Distribution of Driver Error.

Error Type	NMVCCS Criteria	Wisconsin Criteria	Rural	Urban
Recognition Error	• Inadequate surveillance • Internal distraction • External distraction • Inattention	• Inattentive driving	8659 (17.88%)	9044 (19.57%)
Decision Error	• Too fast for conditions • Too fast for curve • False assumption of other's action • Illegal maneuver • Misjudgment of gap or other's action • Following too closely • Aggressive driving	• Too Fast for condition • Exceed Speed Limit • Disregard traffic control • Following too close • Improper overtake • Improper turn	17139 (35.38%)	17662 (38.21%)

	behavior			
Performance Error	• Overcompensation • Poor directional control • Panic/Freezing • Other performance error	• Failure to keep vehicle under control • Left of center • Unsafe backing • Failure to yield	10288 (21.24%)	9867 (21.35%)
Non-Performance Error	• Sleep • Heart attack • Other non-perf. Error	• Disability • Driver Condition • Others	2402 (4.96%)	3030 (6.55%)
No Error			9953 (20.55%)	6620 (14.32%)

The broad categorization of driver errors follows a sequence of information processing. When driving, a driver needs to detect and identify a hazard, decide what to do, and react accordingly. Driver errors proceeding to a crash are also categorized following the above-described sequence of driving task. NMVCCS incorporated all the driver factors that may lead to lack of awareness or failure in recognition of hazardous situations. A driver's recognition efficiency can be affected by any internal or external distraction or any form of inattentive driving. In Wisconsin, 18 percent and 20 percent of total crashes that occurred between 2013 to 2015 were due to inattentive driving in rural and urban areas, respectively.

A driver's decision on what to do directly leads to the type of consequence, whether it is a decision after detecting a hazard or a decision while driving. A bad maneuver decision after recognizing a hazardous situation may cause a crash. A reckless decision like "exceeding the speed limit" may go wrong even if there are no imminent hazards. In Wisconsin, 35 percent and 38 percent of crashes occurred due to decision error in rural and urban areas, respectively.
If a maneuver is not performed properly, it may lead to a crash event. The poorly performed driving tasks are categorized as performance error, which is dependent on driver's experience and skills. Although non-performance error is not related to driver behavior, it represents driver's health conditions, fatigue, level of impairment, or other non-performance issues.

The crash dataset does not contain roadway geometric information at the crash location. Roadway geometry, pavement characteristics, mobility, safety and other roadway-related data tables stored in Meta-Manager in Wisconsin Department of Transportation (WisDOT) were linked with crash data using spatial join in ArcGIS. The joined dataset contains all information collected by crash investigating police officer, roadway geometry, and traffic information for each crash. Table 2 provides the summary statistics.

Table 6-2 Summary Statistics of Explanatory Variables.

Variable	Description	Type	Rural		Urban	
			Mean	Std. Dev.	Mean	Std. Dev.
AADT	Annual Average Daily Traffic (In thousand unit)	Continuous	21610.37	26554.21	55450.19	48566.43
Truck	Truck Percentage (%)	Continuous	11.44	4.59	7.735	2.86
Speed	Posted Speed Limit (MPH)	Continuous	57.49	11.31	46.95	14.17
Lane	Number of lanes (Count)	Continuous	2.13	0.43	2.59	0.715
LW	Lane width (feet)	Continuous	12.10	0.83	12.34	1.08
SW	Shoulder width (feet)	Continuous	8.60	3.87	5.58	5.49
Rut	Pavement rutting (inch)	Continuous	0.088	0.08	0.07	0.07
Percent Passing	Passing percentage (%)	Continuous	26	31.90	3.25	15.85
Highway Type	Interstate	Categorical with 3 levels	9840 (20.31%)		12012 (25.99%)	
	State Highway		37377 (77.16%)		30062 (65%)	
	Other state roadway		1224 (2.53%)		4147 (9%)	
Roadway Type	Undivided	Categorical with 3 levels	24177 (49.91%)		8225 (17.79%)	
	Divided		23889 (49.32%)		36324 (78.59%)	
	One Way		375 (0.77%)		1672 (3.62%)	
Presence of Median	No	Categorical with 2 levels	31530 (65.09%)		20170 (43.64%)	
	Yes		16911 (34.91%)		26051 (56.36%)	
Roadway	Dry	Categorical	27830 (57.45%)		31740 (68.67%)	

Condition	Wet	with 4 levels	5255 (10.85%)	7517 (16.26%)
	Snow		10281 (21.22%)	5307 (11.48%)
	Ice		5075 (10.48%)	1657 (3.58%)
Weather Condition	Clear	Categorical with 5 levels	20591 (42.51%)	22378 (48.42%)
	Fog/Cloudy		13619 (28.11%)	14671 (31.74%)
	Wind		1041 (2.15%)	140 (0.3%)
	Rain		3057 (6.31%)	4157 (8.99%)
	Snow/Sleet		10133 (20.92%)	4875 (10.55%)
Lighting Condition	Day	Categorical with 3 levels	33065 (68.26%)	30046 (73.66%)
	Night-Unlit		13477 (27.82%)	3026 (6.55%)
	Night-Lit		1899 (3.92%)	9149 (19.79%)
Horizontal Curve	No	Categorical with 2 levels	39390 (81.32%)	41750 (90.33%)
	Yes		9051 (18.68%)	4471 (9.67%)
Vertical Curve	No	Categorical with 2 levels	38865 (80.23%)	40542 (87.71%)
	Yes		9576 (19.77%)	5679 (12.29%)
Age group	Adolescent (<18 years)	Categorical with 5 levels	2363 (4.88%)	1789 (3.87%)
	Young Adults (18-25 years)		11206 (23.13%)	11271 (24.39%)
	Adults (26-35 years)		10309 (21.28%)	11406 (24.68%)
	Middle Age (36-65 years)		20223 (41.47%)	18294 (39.58%)
	Old (>65 years)		4340 (8.96%)	3461 (7.49%)
Gender	Male	Categorical with 2 levels	30090 (62.12%)	26989 (58.39%)
	Female		18351 (37.88%)	19232 (41.61%)
Vehicle	Passenger car	Categorical with 4 levels	35498 (73.28%)	38051 (82.32%)
	Motorcycle		888 (1.83%)	545 (1.18%)
	Light truck		8096 (16.71%)	4898 (10.6%)
	Heavy truck		3959 (8.17%)	1727 (5.9%)
Alcohol	No	Categorical with 2 levels	45725 (94.39%)	44470 (96.21%)
	Yes		2716 (5.61%)	1751 (3.79%)
Drug	No	Categorical with 2 levels	47920 (98.92%)	45881 (99.26%)
	Yes		521 (1.08%)	340 (0.74%)
Visibility Obscured	No	Categorical with 2 levels	48078 (99.25%)	46013 (99.55%)
	Yes		363 (0.75%)	208 (0.45%)

Work Zone	No	Categorical with 2 levels	47625 (98.32%)	45339 (98.09%)
	Yes		816 (1.68%)	882 (1.91%)
Debris on road	No	Categorical with 2 levels	47695 (98.46%)	45860 (99.22%)
	Yes		746 (1.54%)	361 (0.78%)

6.5 Result Analysis

The coefficient estimates of the final MNP models for rural and urban crashes are presented in Tables 3 and 4, respectively. The STATA command "mprobit" was used to estimate the coefficient of MNL model (Long and Freese 2006). In both tables, the coefficient estimates represent the log-odds ratio between the probabilities of defined driver error type and no error category with a positive sign for increase and a negative sign for decrease. "No error" category was considered as the base outcome in the MNP model.

The variables of lane width, passing percentage, and median variable were removed from the final model because they were not statistically significant in predicting any of the four driver error categories. The modeling results for "Non-performance error" were excluded as this error category does not include behavioral driver factors. For a quick summary, middle age and old age groups are more prone to non-performance error. Alcohol and drug consumption also increase the probability of non-performance error compared to no error.

Table 6-3 Coefficient Estimates for MNP Model for Driver Errors in Rural Crashes.

Variable	**Recognition Error**		**Decision Error**		**Performance Error**	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
AADT	**0.002**	**0.000**	**0.010**	**0.000**	**0.004**	**0.000**
Truck	**-0.005**	**0.003**	**-0.006**	**0.002**	**-0.011**	**0.002**
Speed	**-0.02**	**0.002**	**-0.019**	**0.002**	**-0.01**	**0.002**

Lanes		-0.046	0.032	**-0.084**	**0.03**	**-0.12**	**0.032**
Shoulder Wid		**0.012**	**0.004**	0.001	0.004	**-0.014**	**0.004**
Pavement Rutting		**-0.462**	**0.166**	**-0.515**	**0.154**	**-0.518**	**0.158**
Highway Type	Interstate	Base Condition					
	State Highway	-0.033	0.037	**-0.102**	**0.031**	**0.08**	**0.034**
	Other state roadway	-0.114	0.077	**-0.186**	**0.072**	**0.338**	**0.072**
Roadway Type	Undivided	0.07	0.047	**-0.146**	**0.043**	4E-5	0.044
	Divided	Base Condition					
	One Way	**-0.36**	**0.143**	0.036	0.123	-0.026	0.13
Horizontal Curve	No	Base Condition					
	Yes	**0.153**	**0.031**	**0.255**	**0.027**	**0.356**	**0.028**
Vertical Curve	No	Base Condition					
	Yes	-0.029	0.029	**0.065**	**0.025**	**0.054**	**0.026**
Roadway Condition	Dry	Base Condition					
	Wet	-0.026	0.048	**0.413**	**0.045**	**0.208**	**0.046**
	Snow	**-0.892**	**0.057**	**1.072**	**0.039**	**0.284**	**0.042**
	Ice	**-1.56**	**0.076**	**0.947**	**0.038**	**0.18**	**0.041**
Weather Condition	Clear	Base Condition					
	Fog/Cloudy	**0.157**	**0.026**	**0.16**	**0.025**	**0.19**	**0.026**
	Wind	**-0.895**	**0.169**	0.054	0.070	-0.068	0.077
	Rain	**-0.163**	**0.065**	**0.269**	**0.057**	0.019	0.06
	Snow/Sleet	**-0.335**	**0.063**	**0.17**	**0.041**	0.085	0.044
Lighting Condition	Day	Base Condition					
	Night-Unlit	**-0.207**	**0.026**	**-0.341**	**0.023**	**-0.116**	**0.024**
	Night-Lit	-0.020	0.059	**-0.281**	**0.056**	**-0.194**	**0.058**
Visibility	No	Base Condition					
	Yes	**-0.364**	**0.132**	-0.046	0.108	-0.105	0.114
Work Zone	No	Base Condition					
	Yes	**0.210**	**0.082**	**0.565**	**0.076**	-0.074	0.089
Debris on road	No	Base Condition					
	Yes	**-2.136**	**0.117**	**-1.863**	**0.094**	**-1.762**	**0.101**
Age group	Adolescent	Base Condition					
	Young Adults	**-0.181**	**0.056**	**-0.174**	**0.052**	**-0.178**	**0.054**
	Adults	**-0.420**	**0.056**	**-0.341**	**0.052**	**-0.278**	**0.054**
	Middle Age	**-0.561**	**0.053**	**-0.527**	**0.050**	**-0.44**	**0.051**

	Old	**-0.339**	**0.061**	**-0.582**	**0.058**	**-0.211**	**0.059**
Gender	Male	0.022	0.024	0.006	0.021	**-0.056**	**0.022**
	Female	Base Condition					
Vehicle	Passenger car	**0.194**	**0.040**	**0.332**	**0.037**	**0.216**	**0.039**
	Motorcycle	**-0.564**	**0.091**	**0.225**	**0.081**	**0.488**	**0.078**
	Light truck	**0.193**	**0.046**	**0.325**	**0.041**	**0.237**	**0.044**
	Heavy truck	Base Condition					
Alcohol	No	Base Condition					
	Yes	**1.120**	**0.067**	**1.282**	**0.066**	**1.459**	**0.065**
Drug	No	Base Condition					
	Yes	**0.754**	**0.129**	**0.740**	**0.129**	**0.875**	**0.128**
Intercept		**1.277**	**0.255**	**1.324**	**0.244**	**1.138**	**0.252**

[Note: Variables that are statistically significant at 90% confidence interval are presented in bold font]

In Table 3, it can be noted that both traffic variables: AADT and truck percentage are significantly related with all driver error categories. For a thousand-unit change in AADT results in increased probability 1.002 (e0.002) times, 1.01 (e0.010) times and, 1.004 (e0.004) times in Recognition Error, Decision Error and, Performance Error compared to No Error, respectively. The signs of estimated coefficients of truck percentage, speed, number of lanes, shoulder width, and pavement rutting represent the reduction in the probability of an error compared to no error because of a unit increase in the independent variable.

Interesting results found in roadway classification show that the highway type is significantly related to both Decision and Performance errors but not Recognition error. This suggests driver's recognition/inattentive driving error does not depend on highway type. Decision error mostly occurs on Interstate highway whereas the occurrence of performance error is least on the Interstate. The change in the probability of performance error is the highest in other highways which include rural city or town roads. One-ways reduce recognition error and undivided highway lead to less decision error. Horizontal and vertical curves significantly increase the

probability of all error categories with a maximum increase in performance error for horizontal and decision error for vertical curves.

Roadway events have a significant effect on driver errors. A comparison between roadway and weather condition variables illustrates a few important observations. For example, snowy pavement increases 4.13 times in decision error from no error whereas snow precipitation only increases by 1.24 times. Another important observation is that snowy pavement has a higher increase in probability than icy pavement. Drivers tend to be more cautious during adverse weather events because of the negative impact for recognition error. A construction zone increases the probability of decision and recognition error but is not statistically significant for performance error. The negative impact of roadway debris on all types of errors suggests drivers may be more vigilant towards unusual objects on the roadway.

Driver age, gender, vehicle type, alcohol, and drug impairment were found statistically significant in predicting all driver error categories. Adolescents are more prone to driver errors compared with all other age groups. For decision error, the probability gradually reduces with the increase in age. But for performance and recognition error, old drivers are more prone to error compared to young and middle-aged drivers. Decision and recognition error does not depend on driver's gender whereas female drivers were found to have a higher probability of performance error. Motorcycle drivers are least likely to have a recognition error, but they are most likely to commit a performance error. Alcohol or drug impairment increases the probability of all error categories with a maximum increase in performance error.

Table 4 provides the coefficient estimates for MNP model with urban crash data. Except for median variable, all explanatory variables were found statistically significant at 10% significance level to predict driver error categories in urban crashes.

Table 6-4 Coefficient Estimates for MNP Model for Driver Errors in Urban Crashes.

Variable		Recognition Error		Decision Error		Performance Error	
		Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
AADT (In thousand)		-3E-05	0.000	**0.002**	**0.000**	-0.0002	0.000
Truck		**-0.015**	**0.004**	**-0.015**	**0.004**	**-0.01**	**0.004**
Speed		**-0.006**	**0.001**	-0.001	0.001	**-0.011**	**0.001**
Lanes		**0.057**	**0.019**	**0.049**	**0.018**	**0.095**	**0.019**
Lane Wid		**0.047**	**0.011**	**0.045**	**0.011**	**0.036**	**0.011**
Shoulder Wid		0.003	0.003	**0.006**	**0.003**	**0.017**	**0.003**
Pavement Rutting		**-0.482**	**0.162**	**-0.393**	**0.152**	**0.387**	**0.159**
Percent Passing		**0.005**	**0.001**	**0.003**	**0.001**	**0.002**	**0.001**
Highway Type	Interstate	Base Condition					
	State Highway	0.001	0.033	-0.017	0.029	**-0.244**	**0.032**
	Other state roadway	**-0.12**	**0.049**	**-0.335**	**0.045**	**-0.146**	**0.047**
Roadway Type	Undivided	0.043	0.043	**-0.105**	**0.04**	**-0.069**	**0.042**
	Divided	Base Condition					
	One Way	-0.056	0.063	**-0.171**	**0.06**	**-0.248**	**0.064**
Horizontal Curve	No	Base Condition					
	Yes	**-0.196**	**0.043**	0.053	0.037	**0.283**	**0.038**
Vertical Curve	No	Base Condition					
	Yes	**-0.082**	**0.036**	-0.029	0.032	**-0.12**	**0.034**
Roadway Condition	Dry	Base Condition					
	Wet	-0.069	0.045	**0.157**	**0.042**	**0.203**	**0.044**
	Snow	**-0.802**	**0.063**	**0.323**	**0.049**	**0.104**	**0.052**
	Ice	**-1.991**	**0.123**	0.061	0.055	**-0.408**	**0.063**
Weather Condition	Clear	Base Condition					
	Fog/Cloudy	**0.145**	**0.026**	**0.168**	**0.025**	**0.198**	**0.026**
	Wind	-0.456	0.303	0.119	0.173	-0.074	0.2
	Rain	**-0.153**	**0.06**	**0.219**	**0.054**	0.077	0.057
	Snow/Sleet	**-0.296**	**0.07**	**0.181**	**0.052**	**0.12**	**0.056**
Lighting Condition	Day	Base Condition					
	Night-Unlit	**-0.19**	**0.05**	**-0.321**	**0.043**	**-0.159**	**0.047**

	Night-Lit	**-0.243**	**0.029**	**-0.37**	**0.027**	**-0.225**	**0.028**
Visibility	No	Base Condition					
	Yes	**-0.588**	**0.166**	**-0.682**	**0.153**	0.22	0.142
Construction Zone	No	Base Condition					
	Yes	-0.043	0.083	**0.289**	**0.075**	-0.052	0.084
Debris on road	No	Base Condition					
	Yes	**-2.04**	**0.149**	**-1.895**	**0.113**	**-1.925**	**0.14**
Age group	Adolescent	Base Condition					
	Young Adults	**-0.202**	**0.065**	**-0.213**	**0.061**	**-0.188**	**0.065**
	Adults	**-0.376**	**0.065**	**-0.431**	**0.061**	**-0.296**	**0.065**
	Middle Age	**-0.393**	**0.064**	**-0.5**	**0.059**	**-0.307**	**0.063**
	Old	**-0.271**	**0.073**	**-0.531**	**0.069**	**0.001**	**0.072**
Gender	Male	**-0.069**	**0.024**	-0.011	0.022	-0.035	0.023
	Female	Base Condition					
Vehicle	Passenger car	**0.413**	**0.046**	**0.512**	**0.042**	**0.44**	**0.046**
	Motorcycle	**-0.536**	**0.118**	0.02	0.099	**0.436**	**0.099**
	Light truck	**0.44**	**0.056**	**0.512**	**0.051**	**0.434**	**0.054**
	Heavy truck	Base Condition					
Alcohol	No	Base Condition					
	Yes	**0.917**	**0.086**	**0.849**	**0.083**	**1.226**	**0.083**
Drug	No	Base Condition					
	Yes	**0.667**	**0.18**	**0.527**	**0.176**	**0.783**	**0.176**
Intercept		0.035	0.195	-0.011	0.184	0.036	0.194

[Note: Variables that are statistically significant at 90% confidence interval are presented in bold font]

There are dissimilarities found in the urban crash analysis compared with rural crashes. AADT is only significant in predicting decision error. This means the probability of making performance or recognition error in an urban setting does not vary by AADT. It is counterintuitive that posted speed limit does not affect decision error because one of the major driver errors in this category is “Exceeding Speed Limit”. Plausibly, speed violation related

crashes may occur at any posted speed limit. Besides, the number of lanes, lane width, shoulder width, and passing percent have a positive effect on driver errors.

For the Highway type variable, both decision and performance errors mostly occur on interstate highways in urban areas. For recognition error, other highway types have the highest increase in probability compared to no error. The roadway type variable is not significant in predicting recognition error but it is significant for both decision and performance error at all levels. Divided highways increase the probability of both decision and performance error compared to no error. In urban areas, drivers are least likely to make performance mistakes with ice on the roadway. For other explanatory variables, similar trends as discussed for rural crashes are observed.

6.6 Discussion of Contributing Factors to Driver Errors

With numerous factors contributing to driver errors, it is challenging to surmise their individual effects to identify effective ways to reduce driver errors. Thus, a review of contributing factors by error type is necessary. These error types can be observed in Table 5 where the marginal effect of MNP is modeled for rural crashes. The marginal effect has varying definitions based upon the variable type. For a continuous variable, the marginal effect is the difference in the probability at each level following a one-unit change in the independent variables; for a categorical variable, the marginal effect is calculated as the changes in the probabilities for each level caused by a change in the value from its base level.

Table 6-5 Review of Marginal Effects for Rural Crashes.

Variable	Recognition Error	Decision Error	Performance Error
Traffic Variables	-	Truck (-0.0001)	Truck (-0.0001)
	AADT (0.0002)	AADT (0.0001)	-

Roadway Geometry	-	-	Lanes (-0.015)
	Speed (-0.002)	Speed (-0.003)	Speed (0.001)
Highway Type (base: Interstate)	Other highways (-0.028)	Other highways (-0.076)	Other highways (0.014)
	-	State highways (-0.032)	State highways (0.006)
Roadway Type (base: Divided)	Undivided (0.026)	Undivided (-0.042)	Undivided (0.015)
	One-way (-0.043)	One-way (0.041)	-
Alignment	Ver.Curve = Yes (0.010)	Hor.Curve = Yes (0.0177)	Hor.Curve = Yes (0.048)
	Hor.Curve = Yes (0.009)	Ver.Curve = Yes (0.0129)	Ver.Curve = Yes (0.010)
Pavement	-	Rutting (-0.047)	-
Roadway Condition (base: Dry)	Snow (-0.197)	Snow (0.342)	
	Wet (-0.048)	Wet (0.094)	Wet (0.015)
		Ice (0.340)	
Weather Condition (base: Clear)	Fog/Cloudy (0.007)	Fog/Cloudy (0.011)	Fog/Cloudy (0.017)
	Snow/Sleet (-0.085)	Snow/Sleet (0.061)	Snow/Sleet (0.025)
	Rain (-0.045)	Rain (0.082)	-
	Wind (0.137)	-	-
Lighting Condition (base: Day)	Night-Unlit (-0.012)	Night-Unlit (0.061)	Night-Unlit (0.026)
	Night-Lit (0.021)	Night-Lit (0.056)	Night-Lit (0.020)
Events	Debris = Yes (-0.145)	Debris = Yes (-0.226)	Debris = Yes (-0.131)
	-	Work Zone = Yes (0.145)	Work Zone = Yes (0.084)
	Visibility: Yes (-0.058)	-	-
Impairment	Alcohol (-0.028)	Alcohol (0.022)	Alcohol (0.073)
	-	Drug (0.021)	Drug (0.060)
Age (base: Adolescent)	Old (-0.022)	Old (-0.112)	Old (0.023)
	Adult (-0.040)	Adult (-0.033)	-
	Middle age (-0.046)	Middle age (-0.058)	Middle age (-0.016)
Gender (base: Female)	-	-	Male (-0.003)
Vehicle type (base: Heavy truck)	Motorcycle (-0.104)	Motorcycle (0.032)	Motorcycle (0.133)
	-	Passenger car (0.053)	Passenger car (0.014)
	-	Light truck (0.047)	-

[Note: Marginal effect presented with "-" is not significant at 90% confidence interval]

Traffic and roadway variables significantly affect the probability of decision error. For example, a unit increase in AADT increases the probability of decision error whereas an increase in truck percentage decreases the probability. In the other error types, the marginal effect of truck percentage and AADT are not statistically significant. However, posted speed limit decreases the probability of recognition error but increases the probability of performance error.

Compared with no errors, the higher probability of recognition error is likely to happen on undivided highways and/or at the places where vertical and horizontal curves are present. While on foggy/cloudy and/or windy days, drivers are more likely to make recognition mistakes. Furthermore, night time with (street) light also has a positive impact on recognition error. On the other hand, drivers are less likely to commit a recognition error on other highways and one-way streets than the Interstate and state highways. In addition, recognition error is low when the pavement is either wet or covered in snow, or weather type is snow/sleet/rain, or nighttime without light. This suggests that drivers may exercise caution when traveling in adverse weather or dark conditions. Similarly, when visibility is low or roadway debris is present, the probability of making a recognition error is low. Another source of low recognition error is people that are older than 18, with middle-aged drivers having the lowest probability of recognition error. For different vehicle types, motorcyclists have the lowest probability of recognition error.

Despite the fact that recognition error shares many similar circumstances with decision error that affect the chance a mistake, the latter is more likely to take place on Interstate highways or one-way streets, but less likely to take place on undivided highways, whereas the opposite pattern is observed for recognition error. A deterrent for decision errors seems to be poor pavement condition (i.e., large rutting value). However, the probability of decision error is higher under adverse weather (e.g., fog/cloudy, snow/sleet, rain) and/or on slippery pavement

(snow, ice, and wet) as well as the night condition irrespective of the availability of street lighting. Finally, heavy trucks have the lowest probability of decision error among all vehicle types, plausibly due to the imposed safety regulations on drivers. In contrast, work zones may see a higher probability of decision error. Decision error may also be increased by the use of alcohol and/or drugs.

Compared with the other error types, performance error is the most probable with the change in roadway geometry and traffic configuration. When making comparisons with no error, the probability of performance error is high for all highway types, horizontal or vertical alignments, adverse weather, wet pavement surfaces, and during the night. Similar to decision error, performance error is more likely to occur when a work zone is present and augmented by the use of alcohol and drugs. Finally, both motorcycles and passenger cars are associated with a higher probability of performance error than truck drivers. On the other hand, middle aged male drivers have the lowest probability of committing performance error.

Table 6 provides the estimates of marginal effects of covariates for urban crashes. Similar to rural crashes, the estimated marginal effect statistically significant at 90% confidence interval were shown in the table.

Table 6-6 Review of Marginal Effects for Urban Crashes.

Variable	Recognition Error	Decision Error	Performance Error
Traffic Variables	-	Truck (-0.002)	Truck (0.0001)
	AADT (-1.45E-7)	AADT (4.89E-7)	-
Roadway Geometry	-	Speed (0.002)	Speed (-0.002)
	-	-	Lanes (0.013)
	Lane Wid (0.004)	Lane Wid (0.006)	-
	-	-	Shoulder Wid (0.003)
	Percent Passing (0.001)	-	-

Highway Type (base: Interstate)	Other highways (0.011)	Other highways (-0.070)	-
	State highways (-0.015)	State highways (-0.019)	State highways (-0.057)
Roadway Type (base: Divided)	Undivided (0.022)	Undivided (-0.029)	-
	-	One-way (-0.026)	One-way (-0.039)
Alignment	Hor.Curve = Yes (-0.060)	-	Hor.Curve = Yes (0.075)
	Ver.Curve = Yes (-0.006)	Ver.Curve = Yes (0.014)	Ver.Curve = Yes (-0.017)
Pavement	Rutting (-0.084)	Rutting (-0.105)	Rutting (0.164)
Roadway Condition (base: Dry)	Snow (-0.159)	Snow (0.157)	Snow (0.031)
	Wet (-0.040)	Wet (0.037)	Wet (0.041)
	Ice (-0.210)	Ice (0.200)	-
Weather Condition (base: Clear)	-	Fog/Cloudy (0.011)	Fog/Cloudy (0.016)
	Snow/Sleet (-0.074)	Snow/Sleet (0.073)	Snow/Sleet (0.030)
	Rain (-0.056)	Rain (0.067)	-
	-	Wind (0.091)	-
Lighting Condition (base: Day)	Night-Unlit (-0.004)	Night-Unlit (0.061)	Night-Unlit (0.008)
	Night-Lit (-0.002)	Night-Lit (0.059)	Night-Lit (0.002)
Events	Debris = Yes (-0.142)	Debris = Yes (-0.230)	Debris = Yes (-0.142)
	Work Zone = Yes (-0.003)	Work Zone = Yes (0.096)	Work Zone = Yes (-0.040)
	Visibility: Yes (-0.075)	Visibility: Yes (-0.162)	Visibility: Yes (0.183)
Impairment	Alcohol (-0.033)	Alcohol (-0.074)	Alcohol (0.064)
		Drug (-0.051)	-
Age (base: Adolescent)	Yound Adult (-0.006)	-	-
	Adult (-0.052)	Adult (-0.053)	-
	Middle age (-0.015)	Middle age (-0.075)	Middle age (-0.011)
	-	Old (-0.134)	Old (0.075)
Gender (base: Female)	Male (-0.013)	-	-
Vehicle type (base: Heavy truck)	Passenger car (0.023)	Passenger car (0.079)	Passenger car (0.031)
	Motorcycle (-0.095)	-	Motorcycle (0.151)
	Light truck (0.028)	Light truck (0.074)	Light truck (0.027)

[Note: Marginal effect presented with "-" is not significant at 90% confidence interval]

Similar to rural crashes, traffic variables significantly affect the probability of decision error. While increase in AADT increase the probability of decision error, increase in truck percentage decrease the probability of decision error. In urban crashes, changes in roadway geometric configuration affects performance error more compared to other error categories. The effect of highway type, roadway type and existence of horizontal and vertical curve on recognition error in urban crashes are almost similar to rural crashes. The weather and roadway conditions are more likely to affect decision error whereas only wet or snowy roadway surface is responsible for performance error.

The gender variable is only statistically significant in predicting the probability of recognition error. Males are less likely to conduct recognition error in urban crashes. Having alcohol in blood while driving contributes to the probability of all driver error. But drug only affects the probability of decision error. This coincides with practical knowledge that having drug will decrease the attentiveness of the driver and eventually increase the probability of making decision error. The interpretation of marginal effect all other variables can be expressed in similar way as discussed for rural crashes.

6.7 Effect of Driver Errors on Injury Severity

Although driver errors have been recognized as a major crash contributor, it's effect on crash severity has not been explored in literature. This section discusses an exploratory analysis conducted to evaluate the effect of different combinations of driver errors on crash severity. The severity for each crash is listed in "KABCO" scale in the MV4000 crash database. The "KABCO" scale of crash injury severity can be defined as: Fatality (K), Incapacitating injury (A), Non-incapacitating injury (B), Possible injury (C) and No injury (O). It is a common

practice to consolidate KABCO into three levels—major injury (K+A), minor injury (B+C), and no injury (O) to ensure that a sufficient number of observations is available in each injury severity level. Similar approach has been used by researchers to ensure sufficient sample size for model estimation (Milton et al. 2008, Islam et al. 2014, Uddin and Huynh 2018).

Based on the driver error categorization used in this study, one crash event may involve one or multiple driver error categories. There can be 16 possible combinations of driver errors (${}^4C_0+{}^4C_1+{}^4C_2+{}^4C_3+{}^4C_4=16$, where C represents combination) out of 4 different driver error categories. These combinations include no driver error, any one of the driver error categories, any two driver errors categories, any three driver error categories and all driver error categories. For example, a driver failed to yield to another driver while driving over speed limit resulted in a crash can represent a combination of decision and performance error. Another example can be a driver failed to keep the vehicle under control on a horizontal curve because of using cell phone while driving can represent a combination of recognition and performance error. A cross-classification table was generated between all 16 combinations of driver errors and injury severities to explore the effect of driver error on injury severity. The cross-classification table for rural crashes is provided in Table 6-7.

Table 6-7 Cross Classification Table for Driver Error Combinations and Injury Severity.

Combination	Recognition Error	Decision Error	Performance Error	Non- Performance Error	No Error	Major Injury	Minor Injury	No Injury	Major to No Injury Ratio (Rank)	Minor to No Injury Ratio (Rank)
						Frequency (Percentage)	Frequency (Percentage)	Frequency (Percentage)		

EC1	✓					333 (11.31%)	2186 (15.33%)	4357 (11.43%)	0.99 (12)	1.34 (12)
EC2		✓				349 (11.85%)	2779 (19.49%)	8619 (22.61%)	0.52 (15)	0.86 (15)
EC3			✓			727 (24.69%)	2791 (19.57%)	6993 (18.35%)	1.35 (11)	1.07 (13)
EC4				✓		214 (7.27%)	818 (5.74%)	1593 (4.18%)	1.74 (10)	1.37 (11)
EC5					✓	238 (8.08%)	1655 (11.61%)	8825 (23.15%)	0.35 (16)	0.50 (16)
EC6	✓	✓				67 (2.28%)	486 (3.41%)	919 (2.41%)	0.94 (13)	1.41 (10)
EC7	✓		✓			207 (7.03%)	729 (5.11%)	1089 (2.86%)	2.46 (7)	1.79 (7)
EC8	✓			✓		51 (1.73%)	170 (1.19%)	235 (0.62%)	2.81 (6)	1.93 (6)
EC9		✓	✓			280 (9.51%)	1326 (9.30%)	3928 (10.31%)	0.92 (14)	0.90 (14)
EC10		✓		✓		38 (1.29%)	132 (0.93%)	212 (0.56%)	2.32 (8)	1.66 (9)
EC11			✓	✓		167 (5.67%)	458 (3.21%)	493 (1.29%)	4.39 (3)	2.48 (4)
EC12	✓	✓	✓			58 (1.97%)	231 (1.62%)	347 (0.91%)	2.16 (9)	1.78 (8)
EC13	✓	✓		✓		13 (0.44%)	52 (0.36%)	48 (0.13%)	3.51 (5)	2.89 (2)
EC14	✓		✓	✓		79 (2.68%)	219 (1.15%)	235 (0.62%)	4.35 (4)	2.49 (3)
EC15		✓	✓	✓		90 (3.06%)	162 (1.14%)	136 (0.36%)	8.57 (1)	3.18 (1)
EC16	✓	✓	✓	✓		33 (1.12%)	65 (0.46%)	85 (0.22%)	5.03 (2)	2.04 (5)

The first 6 columns in Table 6-7 represents the description of error combinations (EC) developed in this study. For example, EC3 represents crashes where only performance error is

involved as driver error. Similarly, EC12 represents a combination of recognition, decision and performance error involved in a crash event. The latter 3 columns represent the frequency and percentage of crashes in three crash severity levels considered in this study. The percentage provided here are estimated for each crash severity outcomes. The frequency and percentage statistics provided in Table 6-7 may not indicate which driver error combination is riskier to crash injury severity.

A "major to no injury" ratio and "minor to no injury" ratio was estimated for each driver error combination to explore the influence of driver error categories on injury severity. In case of crash severity, it is desirable to have a no injury crash compared to minor or major injury crash due to societal and economic impact of injury suffered by the crash victims. The "major to no injury" ratio represents the ratio of percentage of major injury to percentage of no injury. Similarly, "minor to no injury" ratio represents the ratio of percentage of minor injury to percentage of no injury. Ideally, a value of 1 for both ratio represents the driver error combination does not significantly affect the injury severity levels. A value more than 1 represents injury level tends to increase in that corresponding driver error combination and vice versa. Based on estimated ratio, driver error combinations were ranks for both ratios and presented in Table 6-7.

The summary statistics provided in Table 6-7 indicates that only performance error (EC3) results in maximum proportion of major injury severity in rural crashes. Although EC3 dominates the major injury severity for rural crashes, the major to no injury ratio indicates there is 35 percent higher major injury crashes compared to no injury and ranked 11th among all error combinations. The EC15: combination of decision, performance and non-performance driver errors ranked 1st in major to no injury ratio among all error combinations. The major to no injury

ratio of 8.57 for EC15 indicates that there are 757% higher major injury crashes compared to no injury crashes in that error combination. This error combination also ranked 1st in minor to no injury crash ratio. From major to no injury ratio and its corresponding rank presented in Table 2, it can be noted that occurrence of multiple driver error categories results in crashes with higher injury severity. The trend is also similar with minor to no injury ratio. The maximum minor to no injury ratio was also found in EC15 (3.18) which indicates there are 218% higher minor injury crashes occur with EC15 compared to no injury crashes. A chi-square test was also conducted to test whether the driver error combinations and injury severity levels are independent or not. The critical chi-square value with 5 percent level of significant and degrees of freedom =34 (18 driver error combinations and 3 injury severity levels) was found 48.6. The estimated chi-square value for rural crash severities in 18 driver error combination was 2876.7 which is way higher that the critical chi-square value. This indicates that the driver error combinations and crash injury severities were not statistically independent.

Based on cross-classification tables of driver error combinations by crash severity and the chi-square test results, it is evident that driver error influences the three levels of injury severities considered in this study. Therefore, the influence of driver error categories needs to be controlled while modeling crash injury severities.

6.8 Conclusion

More than 90 percent of crashes that occurred on a roadway segment involved driver error. Driver error can be categorized as recognition, decision, performance and non-performance based on the physical definition of each error category introduced in NMVCCS. The reasons

behind these errors can be complicated, including highway and traffic characteristics, environmental factors, roadway events, driver characteristics, and vehicle types.

This study established a statistical relationship between driver errors with a series of factors including roadway, traffic and crash data elements. MNP models were applied to quantify the effect of each explanatory variable. The model results suggest that many of the roadway geometry, highway classification, traffic characteristics, roadway event, and driver-related variables are statistically correlated with different driver error categories in both rural and urban areas. Dissimilarities were found by comparing results between rural and urban crashes, which suggest possible influence of safety culture.

To better understand the impact of driver error contributing factors, a review was conducted for rural crashes using marginal effects from the MNP model. The marginal effect of each explanatory variable represents the quantity of increase or decrease in the probability of a specific driver error type. Thus, each error category can be characterized by a combination of unique variables that help to differentiate future safety treatments. These findings provide evidence-based information to support safety professionals in developing cost-effective engineering countermeasures, safety enforcement or driver training programs focused on specific driver errors.

The exploration of statistical dependency between different combinations of driver error categories with ordered levels of crash severity indicated that occurrence of different combination of driver errors during a crash event can influence resulting crash severity. The chi-square test result indicated that different combinations of driver error categories and injury severity levels are not statistically independent. A major to no injury error and minor to no injury error ratio was estimated and ranked to identify risker combination of driver errors that may

result in more severe crashes. Ranks of estimated ratio indicate that more severe crashes tend to occur when drivers make multiple driver errors. These findings may help researchers and safety professionals to develop specific countermeasures and advanced vehicle features focused on driver errors to reduce the impact of driver errors on injury severity outcome.

6.9 References

NHTSA., 2008. National motor vehicle crash causation survey: Report to congress. National Highway Traffic Safety Administration Technical Report DOT HS 811, 059.

Asbridge, M., Hayden, J.A., Cartwright, J.L., 2012. Acute cannabis consumption and motor vehicle collision risk: Systematic review of observational studies and meta-analysis. Bmj 344, e536.

Bao, S., Boyle, L.N., 2009. Age-related differences in visual scanning at median-divided highway intersections in rural areas. Accident Analysis & Prevention 41 (1), 146-152.

Blomberg, R.D., Peck, R.C., Moskowitz, H., Burns, M., Fiorentino, D., 2005. Crash risk of alcohol involved driving: A case-control study.

Borooah, V.K., 2002. Logit and probit: Ordered and multinomial models Sage.

Box, S., 2009. New data from vtti provides insight into cell phone use and driving distraction. Virginia Tech Transportation Institute 27.

Campbell, J.L., 2012. Human factors guidelines for road systems Transportation Research Board.

Compton, R., Vegega, M., Smither, D., 2009. Drug-impaired driving: Understanding the problem and ways to reduce it: A report to congress.

Devlin, A., Candappa, N., Corben, B., Logan, D., 2011. Designing safer roads to accommodate driver error. Psychopharmacology 10, 193-212.

Dingus, T.A., Guo, F., Lee, S., Antin, J.F., Perez, M., Buchanan-King, M., Hankey, J., 2016. Driver crash risk factors and prevalence evaluation using naturalistic driving data. Proceedings of the National Academy of Sciences 113 (10), 2636-2641.

Graham, J.L., Paulsen, R.J., Glennon, J.C., 1977. Accident and speed studies in construction zones.

Greene, W.H., 2000. Econometric analysis Prentice Hall, Upper Saddle River, N.J.

Hauer, E., 1999. A primer on traffic safety. Institute of Transportation Engineers.

Hauer, E., 2009. Speed and safety. Transportation Research Record: Journal of the Transportation Research Board (2103), 10-17.

Hedlund, J., 2017. Drug impaired driving: A guide for states. Governors Highway Safety Association (GHSA).

Islam, S., Jones, S.L., Dye, D., 2014. Comprehensive analysis of single-and multi-vehicle large truck at-fault crashes on rural and urban roadways in alabama. Accident Analysis & Prevention 67, 148-158.

Juergens, W., 1962. Construction zone, detour and temporary connection accidents.

Liu, C., Chen, C.-L., 2009. An analysis of speeding-related crashes: Definitions and the effects of road environments.

Long, J.S., Freese, J., 2006. Regression models for categorical dependent variables using stata Stata press.

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

NHTSA, 2010. Traffic safety facts: Driver electronic device use observation protocol. DOT HS 811, 361.

NHTSA, 2016. 2015 motor vehicle crashes: Overview. Traffic safety facts research note 2016, 1-9.

Parker, S.T., Tao, Y., 2006. Wistransportal: A wisconsin traffic operations data hub. Applications of advanced technology in transportation. pp. 611-616.

Pickrell, T.M., Liu, C., 2016. Occupant restraint use in 2014: Results from the nopus controlled intersection study.

Rouphail, N.M., Yang, Z.S., Fazio, J., 1988. Comparative study of short-and long-term urban freeway work zones.

Tate, F., Turner, S., 2007. Road geometry and drivers' speed choice. Road & Transport Research: A Journal of Australian and New Zealand Research and Practice 16 (4), 53.

Uddin, M., Huynh, N., 2018. Factors influencing injury severity of crashes involving hazmat trucks. International journal of transportation science and technology 7 (1), 1-9.

USDOT, 2014. Mmucc guideline: Model minimum uniform crash criteria. 4th Edition ed., Washington DC.

Walsh, J.M., Verstraete, A.G., Huestis, M.A., Mørland, J., 2008. Guidelines for research on drugged driving. Addiction 103 (8), 1258-1268.

Wang, K., Qin, X., 2015. Exploring driver error at intersections: Key contributors and solutions. Transportation Research Record: Journal of the Transportation Research Board (2514), 1-9.

WisDOT, Wiscosnin crash data user guide.

Chapter 7 Summary and Future Work

7.1 Findings and Contributions

Researchers can experience challenges when it comes to understanding the underlying crash generating process, producing reliable model coefficients, and making statistical inferences from crash data. The driver errors have been universally accepted as a major contributor to crash occurrences. Albeit, limited research has been conducted to explicitly consider driver behaviors into CPMs. Unavailability of such important variables in crash data analysis can yield inaccurate prediction results. On the other hand, special consideration needs to be made while incorporating direct or proxy measures of driver behavior variables into CPM as they are generated from a distinct risk source. The in-depth understanding of the effect of driver behavior and evident-based findings can support safety professionals in the development of cost-effective safety countermeasures, safety enforcement, or driver training programs to improve existing safety conditions. This dissertation attempts to address following described research gaps.

Chapter 2 of this dissertation provides a comprehensive literature review on factors contribute to crash occurrence. A major challenge in traffic safety analysis is that a myriad of factors can actively or passively influence both the count and severity of crashes. This chapter provides the basic concept of safety theories and overviews of the roles and interactions of variables that contribute to crash occurrence with an emphasis on human factors and driver behaviors. The summary of literature review can advise on variables that are correlated with crash events and their availability in different spatial units. This chapter also shed lights on the existing issues related with highway safety data and methodological alternatives to account for these issues.

Chapter 3 develops area-based CPMs to identify potential proxy variables for driver behavior that are correlated with crash occurrence. Census tract has been used as a spatial unit for developing area-based CPMs using 2015 TIGER/Line data from the U.S. Census. Behavior-based CPM (e.g., speed-related, alcohol-related crashes) were explored along with total crashes to identify the intrinsic relationship between surrogate measures for driver behavior and behavior related crash frequencies. The CPM results show that roadway, travel pattern, socioeconomic, and demographic variables were statistically significant in predicting total crashes and behavior-related crashes in a census tract. The behavior-related crash modeling results provided additional insight into the effect of surrogate variables on crash occurrences. Results indicated more socioeconomic and demographic features are correlated with behavior-related crashes compared with modeling results for all crashes. The area-level crash frequency modeling results can help transportation agencies monitor area-level safety, identify major crash determinants, and evaluate safety programs and investment decisions. These results can be used to identify communities with a high risk of crashes and develop effective countermeasures to increase safety.

Chapter 4 and 5 of this dissertation discuss the development of methodological alternatives to account for driver behavior into crash data modeling framework. Chapter 4 discussed the development of methodological alternative when important driver behavior related variables are missing whereas Chapter 5 provides discusses appropriate modeling technique to use when behavior-related explanatory variables are available in segment level crash dataset. Unavailable driver behavior information introduces unobserved heterogeneity into crash dataset. On the other hand, available driver behavior information should be incorporated as separate risk source into CPM as they are distinct by generating source. An RP NBL model was developed to account for unobserved heterogeneity and tested with crash dataset from South Dakota and

Indiana. Results showed that RP NBL can outperform traditional modeling techniques based on model performance while keeping the core strength of NB distribution. To identify driver behavior as a distinct of source of risk into CPM, a multivariate multiple risk source regression model was developed using two risk sources (engineering and behavioral risk source) and for injury and no injury crashes. The proposed model was tested with crash dataset from Wisconsin rural two-lane highways. Modeling results from this newly proposed model indicated that multivariate setting of the proposed model can account for the correlation between crash severities and provide stable parameter estimates. The estimated marginal effects indicated that traditional models may underestimate the effect of covariates from behavior risk source as segment-level data availability from this source is not rich compared with engineering variables.

Chapter 6 discusses an alternative event-based crash data analysis approach focusing on driver errors. The fourteen driver-related factors listed in Wisconsin MV4000 crash database were classified into recognition errors, decision errors, performance errors, and non-performance errors based on driver's "PIEV" process. The multinomial probit model (MNP) was then employed to study driver errors reported in crashes in rural and urban areas. The modeling results identified many highway geometric features, traffic conditions, roadway events, and driver characteristics as statistically correlated to different types of driver error. To evaluate the effect of driver errors on crash injury outcomes, an exploratory analysis was conducted using possible combinations of driver error categories and its related severity outcome in each crash. The chi-square test result indicated that different combinations of driver error categories and injury severity levels are not statistically independent. A major to no injury error and minor to no injury error ratio was estimated and ranked to identify riskier combination of driver errors that

may result in more severe crashes. Ranks of estimated ratio indicate that more severe crashes tend to occur when drivers make multiple driver errors sequentially.

The contribution of this dissertation can be summarized as follows:

1. This dissertation provides a guideline on how to incorporate driver behaviors into different aspects of crash data modeling. A comprehensive literature review of safety studies indicates that driver characteristics, limitations, and errors play an important role in crash occurrence. Exclusion of such variables in quantitative safety data analysis can yield unreliable and inaccurate inference on the effect of covariates on crash occurrence.
2. This dissertation identifies surrogate measures for driver behaviors by developing behavior-based CPMs. Based on modeling results, appropriate driver training programs focusing target socioeconomic group(s) can be designed to improve safety conditions.
3. This dissertation contributes to the development of a methodological alternative to account for unobserved heterogeneity induced overdispersion issue in crash data due to unavailable behavioral factors. An RP NBL model was developed and tested with multiple crash dataset. Results showed that the proposed model can explicitly account of unobserved overdispersion while keeping the core strength on NB distribution.
4. This dissertation contributes to the development of modeling alternative to distinguish between distinct sources of crash risk and incorporate available driver behavior variables into CPM. A multivariate multiple risk source regression model was developed to predict crash frequency and injury severity simultaneously. Results showed that expanding multiple risk source modeling approach can provide consistent parameter estimate and superior model performance.

5. This dissertation proposes an alternative modeling approach to identify factors contributing to driver errors at each crash event. The modeling results identified many highway geometric features, traffic conditions, roadway events, and driver characteristics as statistically correlated to different types of driver error. Moreover, an exploratory analysis of driver errors on crash severity is conducted to identify the effect of driver error categories on crash severity. The results can help safety professionals to understand when, where, and how the driver error may lead to a crash, how they affect the crash severity outcome to develop cost-effective preventive measures.

7.2 Future Direction

This dissertation explored several study designs based on different spatial aggregation of crash data and developed methodological alternatives to account for driver behaviors in each of the proposed 3-tier approaches. Despite the potential and contributions of this dissertation, the idea of quantifying the effect of driver behaviors on crash occurrence can still be expanded in several directions. A few potential continuations of this current research that can be pursued in future are discussed as follows:

1. The area-level CPMs developed in this dissertation only focused on identifying potential surrogate measures for human factors and driver behaviors and their intrinsic relation with crash events. Although use of surrogate variables is popular among researchers, a more direct relation of crashes with driver behavior variables need to be explored. Traffic citation information (e.g., speeding citation, OWI, traffic rule violation citation, etc.) can be a potential source of direct measures for driver behavior. State Department of Motor Vehicles usually collect and maintain registered driver characteristics (e.g., age, gender,

marital status etc.) and citation information. These pieces of information can be collected and utilized in CPM in future to identify the underlying relationship between direct measures for driver behavior and crash events.

2. Spatial correlation has shown to be present in crash data, specially at larger spatial units (e.g., county, traffic analysis zone, etc.). Spatial correlation in crash data has been modeled as an error term using different neighboring structures in previous crash data modeling literature to reduce bias in estimated parameters. Although considering spatial correlation in crash modeling structure can improve model performance, it does not explain the source of spatial correlation. Use of spatial variables in explaining spatial correlation in crash data can add benefit in identifying intrinsic spatial relationships in crash data analysis.
3. This dissertation developed two modeling alternatives to account for driver behaviors at site-specific CPM. Although both modeling techniques contribute to explain heterogeneity in crash data, they are developed to solve specific issue(s) related to crash dataset. A methodological alternative needs to be developed combining both mixed distribution RP and multiple risk source modeling approaches to understand the source of unobserved heterogeneity and how to use complex modeling alternatives as a tool to analyze different datasets.
4. This dissertation discussed when, where and how drivers make error and how driver errors contribute to crash severity. A joint discrete choice model needs to be developed to explain how roadway geometry, traffic characteristics, roadway, and environmental conditions and events contribute to both driver error and its resulting injury severity.

5. This dissertation developed CPMs in a 3-tier approach to account for crash contributors that are available at different spatial units. A common platform is required to integrate data from different spatial units. To utilize data from different spatial units, a joint modeling framework is needed which can model crashes at different spatial unit with a transition of information from macro to micro level predicted crashes.

CURRICULUM VITAE
Mohammad Razaur Rahman Shaon

Education

Ph.D. Candidate, Civil & Environmental Engineering, University of Wisconsin-Milwaukee, Milwaukee, WI (Anticipated Graduation: 2018);
Major: Civil Engineering (Transportation Engineering); Current GPA: 3.85
M.S. Civil & Environmental Engineering, South Dakota State University, Brookings, SD, 2015
Major: Civil Engineering (Transportation Engineering), GPA: 3.81
B.Sc. Civil Engineering, Bangladesh University of Engineering and Technology, Dhaka, Bangladesh, 2012
Major: Civil Engineering

Research Experience

- Graduate Research Assistant, Civil & Environmental Engineering, University of Wisconsin-Milwaukee, 2016.8- present
    - Project: Identifying Highly Correlated Variables Relating to the Potential Causes of Reportable Wisconsin Traffic Crashes, 09/2016 – present
    - Project: SHRP2 IAP Round 7, Application of SHRP2 Reliability and Analysis Tool Bundle for Wisconsin Department of Transportation, 01/2017 – present
    - Project: Wisconsin Traffic Sensor Data Quality Improvement, 06/2017 – present
    - Project: Evaluation of Wisconsin High Visibility Enforcement Program, 2016
- Graduate Research Assistant, Civil & Environmental Engineering, South Dakota State University, 08/2013- 06/2015
    - Lead Researcher, Calibration of Highway Safety Manual Predictive Methods for State and Local Rural Highways, 2013 - 2015
    - Secondary Researcher, Developing A Pavement Management System for Local and Small Communities, 01/2015 – 06/2015
- Undergraduate Research Scholar, Bangladesh University of Engineering and Technology, 2011-2012
    - Surveyor, Data Analyst, Dhaka-Asulia Elevated Expressway Project.
    - Surveyor, Data Analyst, Gulistan-Sadar Ghat Elevated Expressway Project.

Teaching Experience

- Graduate Teaching Assistant, Civil & Environmental Engineering, University of Wisconsin-Milwaukee, 08/2015 – 5/2017
    - GTA for CE 202 Statics, 8/2015 – 7/2016
    - GTA for CE 303 Strength of Material, 8/2016 - 5/2017

Peer-Reviewed Publications

1. **Shaon, Mohammad Razaur Rahman**, Xiao Qin, Mohammadali Shirazi, Dominique Lord, and Srinivas Reddy Geedipally. "Developing a Random Parameters Negative Binomial-Lindley Model to analyze highly over-dispersed crash count data." Analytic Methods in Accident Research 18 (2018): 33-44.
2. **Shaon, Mohammad Razaur Rahman**, Xiao Qin, Zhi Chen, and Jian Zhang. "Exploration of Contributing Factors Related to Driver Errors on Highway Segments." Transportation Research Record (2018): 0361198118790617.

3. **Shaon, Mohammad Razaur Rahman**, Robert J. Schneider, Xiao Qin, Zhaoxiang He, Aida Sanatizadeh, and Matthew Dreis Flanagan. "Exploration of Pedestrian Assertiveness and Its Association with Driver Yielding Behavior at Uncontrolled Crosswalks." Transportation Research Record (2018): 0361198118790645.
4. Schneider, Robert J., Aida Sanatizadeh, **Mohammad Razaur Rahman Shaon**, Zhaoxiang He, and Xiao Qin. "Exploratory Analysis of Driver Yielding at Low-Speed, Uncontrolled Crosswalks in Milwaukee, Wisconsin." Transportation Research Record (2018): 0361198118782251.
5. Qin, Xiao, Zhi Chen, and **Mohammad Razaur Rahman Shaon**. "Developing jurisdiction-specific SPFs and crash severity portion functions for rural two-lane, two-way intersections." Journal of Transportation Safety & Security (2018): 1-13.
6. **Shaon, Mohammad Razaur Rahman**, and Xiao Qin. "Use of Mixed Distribution Generalized Linear Models to Quantify Safety Effects of Rural Roadway Features." Transportation Research Record: Journal of the Transportation Research Board 2583 (2016): 134-141.
7. Qin, Xiao, **Mohammad Razaur Rahman Shaon**, and Zhi Chen. "Developing Analytical Procedures for Calibrating the Highway Safety Manual Predictive Methods." Transportation Research Record: Journal of the Transportation Research Board 2583 (2016): 91-98.
8. Chen, Zhi, Xiao Qin, and **Mohammad Razaur Rahman Shaon**. "Modeling lane-change-related crashes with lane-specific real-time traffic and weather data." Journal of Intelligent Transportation Systems 22, no. 4 (2018): 291-300.
9. Qin, Xiao, Zhi Chen, Elizabeth Schneider, Yang Cheng, Steven Parker, **Mohammad Razaur Rahman Shaon**. "Designing a Comprehensive Procedure for Flagging Archived Traffic Data." Accepted for publication in Transportation Research Record, 2019.
10. Al-Mahameed, Farah, Xiao Qin, Robert Schneider, **Mohammad Razaur Rahman Shaon**. "Analyzing Pedestrian and Bicyclist Crashes at the Corridor Level: A Structural Equation Modeling Approach." Under-Review: Transportation Research Board Annual Meeting, 2019.

Conference Proceeding and Working Papers

1. **Shaon, Mohammad Razaur Rahman**, Xiao Qin, Amir Pooyan Afghari, Simon Washington. "Incorporating behavioral variables into prediction of crash counts by severity: a multivariate multiple risk source approach". Working Paper.
2. **Shaon, Mohammad Razaur Rahman** and Xiao Qin. "How is Injury Severity Affected by Driver Errors: A Crash Data Based Investigation". Working Paper.
3. **Shaon, Mohammad Razaur Rahman**, Xiao Qin, Ambily Pankaj, Elizabeth Schneider, Benjamin Rouleau. "Developing Procedures to Calibrate Travel Time Reliability Using NPMRDS Data." Accepted for presentation in Transportation Research Board Annual Meeting, 2019.
4. **Shaon, Mohammad Razaur Rahman**, and Xiao Qin. " Improving Crash Prediction Methods with a Generalized Additive Model." Presented at Transportation Research Board Annual Meeting, 2015.
5. **Shaon, Mohammad Razaur Rahman**, and Xiao Qin. " Using a Multivariate Missing Data Imputation Scheme for Missing Dual Loop Detector Data." Presented at Transportation Research Board Annual Meeting, 2017.

Skills

- Mastery of data analytics and statistics analysis (e.g., programming in R, SAS, MATLAB) and its application in traffic safety
- Experienced in Bayesian model development and application in WinBUGS and OpenBUGS
- GIS application and Spatial Analysis: Experienced in ArcMap
- Strong knowledge of traffic engineering and proficient in using traffic analysis software (e.g., HCS, SYNCHRO, CORSIM, AutoCAD)
- Capable of programming using C++
- Proficient in using Microsoft Office

- Collaboration and Team management
- Experienced in multitasking and handling work pressure
- Ability to communicate and express ideas and recommendations in both writing and verbally.

Awards and Honors

- 2018, 2017, 2016, Chancellor's Graduate Student Award, UWM
- 2017, 2016, Graduate Student Travel Award, Graduate School, UWM
- 2016, Honorable Mention for Student Paper Competition, AASHTO GIS for Transportation Symposium: Quantifying Safety Effects on Rural Highway System Using HSM Predictive Models.
- 2014, HR Green Scholarship for Excellence in Academic Achievement, South Dakota State University.
- 2007-2011, Technical Scholarship, Bangladesh University of Engineering and Technology.
- 2005-2006, Talent pool Scholarship for Excellence in Secondary School Certificate Examination, Bangladesh.