DISSERTATION

REFINED-SCALE CRASH DATA ANALYSIS USING MULTI-LEVEL REGRESSION MODELS

Submitted by

Xiaoxiang Ma

Department of Civil and Environmental Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2016

Doctoral Committee:

Advisor: Suren Chen

Mazdak Arabi
Neil S Grigg
Haonan Wang

ABSTRACT

REFINED-SCALE CRASH DATA ANALYSIS USING

MULTI-LEVEL REGRESSION MODELS

Road traffic safety has long been a major public health problem for the general public and government agencies. Nevertheless, road traffic crashes continue to bring immeasurable pain and suffering to the society, as well as high financial expenses associated with medical bills and lost productivity. After identifying some key research gaps related to the existing crash modeling such as lack of insightful modeling of crash rates, time-varying explanatory variables, serial correlation, unobserved heterogeneity and multiple dependent variables, the objective of this dissertation is to narrow these gaps by systematically developing advanced multilevel models for traffic safety modeling. It is expected that series of new crash models developed in this dissertation not only contribute to the state-of-the-art crash modeling, but also add to the knowledge toward developing proactive traffic management strategy. The dissertation has eight chapters: Chapter one provides some background information and literature review. Chapter two presents crash rate analysis with data in refined scales to quantify the relation between crash rate and time-varying variables along with other contributing factors. In Chapter three, the unobserved heterogeneity issue on mountainous highways crash rates is examined by developing an advanced random parameter tobit model with panel data in refined temporal scale. Chapter four proposes a correlated random parameter marginalized two-part model as an alternative to study the relationship between crash rate and its contributing factors. Chapter five examines the differences of contributing factors towards injury severity on mountainous (MN) and non-mountainous (NM) highway crashes using mixed logit models. Chapter six studies the effects of weather and traffic characteristics on single-vehicle and multi-vehicle crashes jointly by proposing a multivariate count data model which addresses unobserved heterogeneity across multiple dependent variables. In Chapter seven, a framework of Bayesian multivariate space-time model that can address spatial correlation/heterogeneity, temporal

correlation/heterogeneity, and the correlation between different injury severities is introduced. Chapter eight concludes this dissertation by summarizing major findings and sharing some observations in terms of future research.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Suren Chen, for his continuous guidance and support through my study at Colorado State University. This dissertation would not have been possible without his encouragement and dedication. Not only has he provided insightful advice for my doctoral research, but he also served as an excellent life mentor to me. I consider myself lucky and privileged for having the opportunity to work under his supervision.

I am grateful to Dr. Mazdak Arabi, Dr. Neil S Grigg and Dr. Haonan Wang for serving on my dissertation committee and for their valuable suggestions. I also want to thank Dr. Feng Chen. His expertise in traffic safety enlightened me throughout my work and I enjoyed the cooperation with him.

To my fellow graduate students at Colorado State University and my friends in China and the United States, thank you for believing in me and encouraging me especially during difficult times. The joyful conversations we had and the fun moments we shared together will be forever treasured.

Last but not least, I would like to thank my family, my parents, my siblings, my uncle and my aunts for their unconditional love and support. They have been and always will be the source of my inspiration and motivation. I love them dearly.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 Background

Road traffic safety has long been a major public health problem for the general public and government agencies. As one of the infrastructure systems that people have to deal with every day, a safe transportation system is critical to the overall welfare of the whole society. Nevertheless, road traffic crashes continue to bring immeasurable pain and suffering to the society, as well as billions dollars of cost regarding medical expenses and lost productivity. Recently, the World Health Organization estimated that road traffic crashes claimed 1.24 million death tolls every year, and seriously injured another 20 to 50 million people (World Health Organization, 2013). In the U.S., there were 32,719 people died and an 2,313,000 people injured in the 5,687,000 police-reported traffic crashes in 2013 (NHTSA, 2014). According to Centers for Disease Control and Prevention (2015), the leading cause of death for age group from 5 to 24 in the U.S. is road traffic crashes. With decades of efforts that researchers and practitioners have been putting together to prevent and mitigate motor vehicle crashes, the overall crash-related fatalities rates are decreasing in the U.S. in recent years. However, given the sheer high number of crashes and enormous life and financial losses resulted from traffic crashes every day, there is still a long way to go to mitigate the number of crashes and the resulting injury severity.

A motor vehicle crash is a complicated process, which is usually determined by four main elements: human factors, vehicle factors, infrastructures and environmental conditions. In fact, Treat (1980) showed that human factors had contributed to around 95% of all motor vehicle crashes by playing a sole role or combined with other factors. However, this does not exonerate the infrastructure and environmental conditions from being attributable to crashes. Poch and Mannering (1996) pointed out that by identifying geometric conditions that lead to accidents, these conditions can be corrected thus accidents likelihood be reduced. Similar inference can be readily extended to other factors that contribute to accidents, such as traffic and environmental conditions. In addition, from the standpoint of a traffic safety researcher or engineer, it is of paramount importance to provide safer road infrastructure and driving environment that could help reduce the possibility of human error and mitigate the consequences should an accident occur.

Numerous accident studies have been attempted to quantify the impacts of a variety of casual factors on crash likelihood and the outcome of motor vehicle accidents. Over the years, highway safety researchers and engineers have putting an increasing emphasis on the particular role of road safety in the decision-making process from transportation planning, design to maintenance and operation. Such tasks are generally achieved by building statistical models that link various contributing factors to highway crash outcomes (i.e. crash rates, frequencies and injury severity levels). Statistical models can provide quantitative information about the impact of contributing factors on highway safety, and are vital to most highway safety studies. Two recent comprehensive reviews (Lord and Mannering, 2010; Savolainen et al., 2011) summarized the limitations and strengths of these statistical models that are widely used to study crash frequencies and injury severity outcomes. These two reviews showed highway safety researchers and practitioners rely heavily on statistical tools to draw inference on motor vehicle crashes. In Highway Safety Manual (HSM), the whole Part C is devoted to predictive methods on crashes that are based on a series of statistical models (AASHTO, 2010), highlighting the importance of statistical models in highway safety studies.

Given the importance of statistical models, it is critical to develop sound and reliable ones for investigating highway crashes. Although methodological approaches have made promising progress in recent years to deal with various issues associated with crash dataset and methodologies (Mannering and Bhat, 2014), there are still some significant limitations in existing studies that require further investigation. One notable issue is about time-varying explanatory variables. In the absence of detailed data, highway safety studies generally adopt data that is highly aggregated in time domain, usually over a year or even several years. This leads to potential loss of important information associated with time-varying explanatory variables that are critical to crashes, and can also introduce errors in model estimation (Lord and Mannering, 2010; Washington et al., 2011). Advances in studies using refined-scale crash data are clearly needed to investigate the relationship between highway safety and these time-varying variables, which can provide insightful guidance to local traffic agencies at the operational level.

1.2 Literature Review

1.2.1 Crash Occurrence Modeling

Traffic accident modeling has long been an important research area for safety researchers. Two measures are commonly adopted in crash occurrence studies: crash frequency and crash rate. Crash frequency measures the number of motor vehicle crashes happening on some transportation facility (e.g. roadway segment, intersection) over a pre-specified time period, while crash rate is an exposure-based measure that represents the number of motor vehicle crashes happening per (100) million vehicle miles traveled (VMT). Crash frequency is a direct and easy-to-perceive measure of safety, and crash rate is a relative measure of safety which is widely adopted in crash reporting systems. Each measure has its own merits and shortcomings. Most of the research efforts have been focused on crash (accident) frequency modeling so far.

1.2.1.1 Crash Frequency Models

Since crash frequency analyses require handling of the dependent variable that is non-negative integer, Poisson regression model serves as a good starting point because it deals with heteroscedasticity and preserves the count data nature of crash frequency data. One major constraint of Poisson model is that it requires the mean of the crash counts to be equal to the variance, which is dubbed equal-dispersion. If the equal dispersion assumption is violated, parameter estimation will be biased which further leads to incorrect inferences (Shankar et al., 1995). Moreover, crash frequency data are usually characterized by over-dispersion (i.e. the variance exceeds the mean). To overcome the equal dispersion restriction of Poisson model, negative binomial models (a.k.a. Poisson-gamma model) were proposed and widely adopted to analyze crash frequencies (Abdel-Aty and Radwan, 2000; Carson and Mannering, 2001; Malyshkina and Mannering, 2010a; Noland, 2003; Poch and Mannering, 1996; Shankar et al., 1995). Many other models were developed as alternatives or extensions of Poisson and negative binomial models. The Poisson-lognormal model was developed as an alternative to negative binomial model in dealing with over-dispersion, and showed more flexibility than the negative binomial model (Aguero-Valverde, 2013a; Miaou and Lord, 2003). In light of excess zero observations in crash data, Miaou (1994) proposed zero-inflated Poisson model. Zero-inflated Poisson and negative binomial models assume a dual state (crash-free and crash-prone state) and show to provide a better fit than their Poisson and negative binomial counterparts (Huang and Chin, 2010; Lord et al., 2005; Shankar et al., 1997, 2003). Lord et al. (2005), however, argued that zero-inflated models cannot reflect underlying data generating process because the safe state has zero long-term mean. Malyshkina and Mannering (2010b) proposed a zero-state Markov switching model to overcome the shortcoming of the zero-inflated model. The finite mixture model is a more recent development that assumes that observations come from several unknown subgroups. Park and Lord (2009) examined the application of finite mixture count data model and argued that it can capture group-level unobserved heterogeneity as opposed to observation-level unobserved heterogeneity. Although the development of these sophisticated statistical models has contributed greatly to extracting more information out of existing crash data sources, improvements in data quality are believed to benefit the development of safety models and further policy implications.

1.2.1.2 Crash Rate Models

In contrast to a large number of literatures about different count-data models for crash frequency analyses (Lord and Mannering, 2010), studies that focused on crash rate are very limited. From the methodological perspective, crash rates modeling involves continuous data, which are usually left-censored at zero. This is because roadway segments without any accident reported over a specified time period will simply generate zeros on accident records, setting up a data clumping at zeros. Anastasopoulos et al. (2008) dubbed this effect as censoring effect. In light of the censoring effects in crash rate data, Anastasopoulos et al. (2008) applied tobit model instead of linear regression to study 5-year highway accident rates in the state of Indiana. They have not only demonstrated tobit model as a proper approach, but also identified some important contributing factors to crash rates, including pavement characteristics, geometric features and traffic characteristics. However, their statistical models were limited by assuming fixed parameter across observations. To handle unobserved heterogeneity across observations, Anastasopoulos et al., (2012a) further developed a random-parameters tobit model, which showed superiority over its fixed parameter counterpart in terms of goodness of fit. Their results indicated that some parameters have mixed effects on accident rates. Xu et al. (2013) investigated influencing variables of crash rates on arterial streets in Las Vegas using tobit model. Other studies also extended tobit regression to a multivariate setting, and analyzed accident rates on different injury levels (Anastasopoulos et al., 2012b). All these studies have contributed to the understanding of highway safety. One common limitation of these previous studies, however, is that they all exploited highly aggregated data. Weather and traffic-related variables, which have been well established to have significant impact on highway safety, were either not adequately addressed or not even considered in the model. In contrast to various models to study crash frequencies, only tobit model was used to examine crash rates so far. Development of more appropriate models to study crash rates could potentially provide more predictive and explanatory power over tobit model.

1.2.1.3 Multilevel Models

With refined data in both spatial and temporal domains, serious correlation problems may exist among the records, setting up a structure in the data. A Multilevel model is an ideal way to deal with different structures in the data. In this dissertation, the terms "multilevel model" and "hierarchical model" will be used interchangeably. A term closely associated with multilevel models is panel data (also known as cross-sectional time-series data) models. these models have been widely used in econometric, social and behavioral analyses, and have gained popularity among safety specialists due to their capacity to address both time-series and cross-sectional variations. In crash modeling arena, panel data models can be mainly generalized into three categories: fixed effects (FE) models, random effects (RE) models, and random parameters (RP) models. A FE model has some appeals in panel data modeling because of not requiring the assumption about unobserved heterogeneity (Greene, 2008). Nonetheless, FE models usually require estimations of a vast number of parameters (e.g. site-specific or time-specific indicator variables), which may dominate the contributing factors. As compared to FE model, RE model is generally more popular in modeling crash data with repeated observations (Aguero-Valverde, 2013a;

Chin and Quddus, 2003; Qi et al., 2007; Shankar et al., 1998). Recently, random-parameter (RP) model was proposed as a general extension of RE model to allow not only the constant term but also the coefficients to vary across observations (Anastasopoulos and Mannering, 2009; Lord and Mannering, 2010). In addition to these three most frequently used types of panel data models, there were also some other research attempts to account for temporal and spatial correlations in accident occurrence, such as negative multinomial model (Caliendo et al., 2013; Ulfarsson and Shankar, 2003), generalized estimating equation (GEE) model (Lord and Persaud, 2000; Wang and Abdel-Aty, 2006), and time-series model (Mohammed A Quddus, 2008). However, most of these studies focused on crash frequency prediction by applying panel data with repeated yearly observations, rather than data in more refined scales. The policy implications inferred from these studies are then limited, in a way that they merely focused on improving geometric designs.

1.2.2 Injury Severity Models

Apart from crash occurrence, injury severity is another significant aspect of crash risk. Injury severity data from police reports are usually measured on KABCO scale: fatal injury or killed (K), incapacitating injury (A), non-incapacitating injury (B), possible injury (C), and property damage only (O). Being capable of accounting for the ordinal nature, traditional ordered probability models (ordered probit and ordered logit) have been widely applied to study injury severity outcomes (Chimba and Sando, 2009; Christoforou et al., 2010; Xie et al., 2009). However, Savolainen et al. (2011) pointed out that a traditional ordered model has two major problems when applied to injury severity data. First, it is very vulnerable to underreporting of crashes which is a well-known problem with police-reported crash data. The second limitation is associated with the restriction that it imposes on how the variables affect outcome probabilities (see Fig. 1.1). Multinomial logit models, on the contrary, do not place the same restriction on variables as traditional ordered probability models do. They can produce correct parameter estimates even in the presence of underreporting problem. On the downside, multinomial logit models ignore the ordinal nature of injury severity data and require independence of irrelevant alternatives (IIA) property. Mixed logit models (McFadden and Train, 2000) relax the IIA assumption of multinomial logit models by introducing correlation in unobserved factors. Given the flexibility and methodological superiority, mixed logit models have been the state-of-the-art technique since their inception (Chen and Chen, 2011; Milton et al., 2008). Besides random parameter models (e.g. mixed logit model), latent class models have also gained recognition recently (Cerwick et al., 2014; Eluru et al., 2012). Differing from mixed logit models, latent class models can accommodate group specific unobserved heterogeneity and relax continuous distributional assumptions for random parameters. However, latent class logit models do not deal with individual unobserved heterogeneity as mixed logit models do. No consensus has yet been made regarding which approach is superior given the fact that both approaches have their respective strengths and limitations (Xiong and Mannering, 2013).

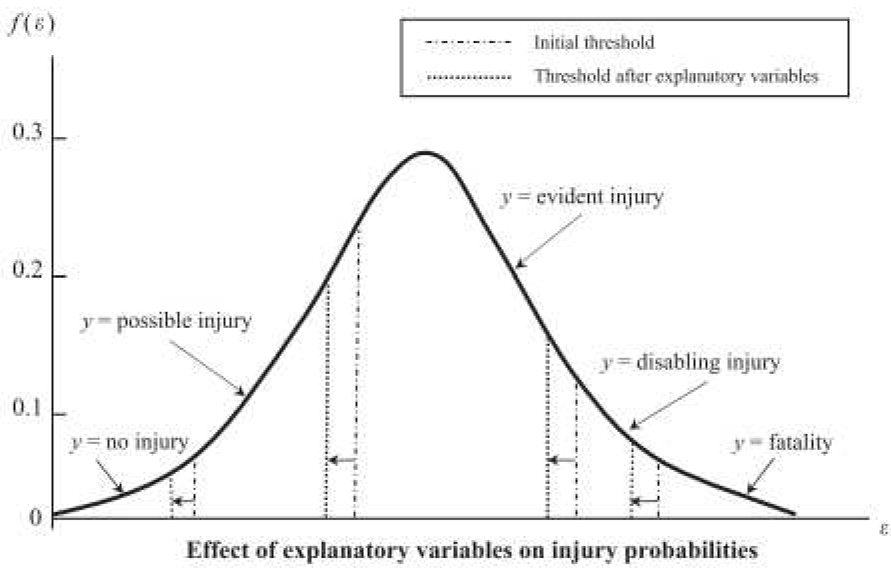

Figure 1.1 The limitations of the standard ordered probability model (Washington et al., 2011)

1.2.3 Multivariate Crash Frequency

As discussed above, traffic crashes are investigated using univariate models without further distinguishing crashes by different types (Chen et al., 2014; Chin and Quddus, 2003; Johansson, 1996; Ma et al., 2015a; Malyshkina and Mannering, 2010b; Shankar et al., 1995). Recognizing the need of accounting for unobserved factors across different types of crashes, Ma and Kockelman (2006) adopted a multivariate Poisson (MVP) model to analyze accidents by injury levels. Their results indicated that a

MVP regression is superior to its univariate counterparts. Given that MVP cannot accommodate over-dispersion, Ladrón de Guevara et al. (2004) applied a multivariate negative binomial (MVNB) model to investigate fatal, injury and property-damage crashes simultaneously. However, MVNB assumes a gamma distributed error which is motivated merely by mathematical convenience, and it does not allow negative correlation structure. In order to overcome the drawbacks of MVP and MVNB, Chib and Winkelmann (2001) proposed a multivariate Poisson lognormal (MVPL) model. It is not only capable of addressing over-dispersion but also allows a full general correlation structure. Given its strengths, MVPL has been widely applied in traffic safety studies (Aguero-Valverde and Jovanis, 2009; El-Basyouny et al., 2014; El-Basyouny and Sayed, 2009a; Ma et al., 2008; Park and Lord, 2007). It is worth noting that most of these past research endeavors have mainly focused on multivariate cross-sectional count data. Panel data, which can capture unobserved factors, has become popular in traffic safety studies recently. A variety of panel data models, such as random effects and random parameter models, have been widely applied in univariate models (Chen et al., 2014; Chin and Quddus, 2003; Ma et al., 2015a; Qi et al., 2007; Shankar et al., 1998). However, so far multi-level model has rarely been explored to study crash frequencies and injury severity levels simultaneously.

1.2.4 Factors Affecting Crash Occurrence

A large number of factors contribute to the likelihood of a motor vehicle accident. These contributory factors can be categorized into human factors, traffic flow characteristics (e.g. annual average daily traffic (AADT) and traffic speed), roadway characteristics (e.g. geometric designs and pavement conditions), and environmental conditions (e.g. weather and surface conditions). Ideally, all categories of factors should be incorporated to develop Safety Performance Functions (SPF). Nevertheless, not all these types of factors have been given equal attention in the past, possibly due to the limitation of crash datasets available to the analysts. A brief review of these contributing factors is presented in this section.

The relationship between crash occurrence and roadway geometry has been extensively documented in the literature (Anastasopoulos and Mannering, 2009; Ma and Kockelman, 2006; Milton and Mannering, 1998; Shankar et al., 1998; Venkataraman et al., 2014). For example, Milton and Mannering (1998) examined the association between various geometric features and crash frequency while controlling for traffic exposure. They found that some geometrics have significant impacts on crash frequency. Ma et al. (2008) developed a multivariate Poisson-lognormal model for crash frequencies by severity level and found that geometrics such as horizontal curve length, the degree of curvature, and vertical curve length contributed to crash likelihood.

For weather characteristics, most of the past research emphasized the impact of precipitation, including rainfall, rainfall intensity and snowfall. A recent review (Theofilatos and Yannis, 2014) pointed out that most existing studies show that precipitation tends to increase crash risks. However, the adoption of precipitation is clearly inferior to more direct variables, such as road surface condition and visibility. These variables can better reflect true effects of precipitation on crash risks. Again, these studies were conducted on highly aggregated data, which is an obvious limitation. For traffic characteristics, the relationship between speed and crash frequency has always been under debate (Wang et al., 2013a). Although some researchers showed that accidents risk increases with the mean traffic speed (Aarts and Van Schagen, 2006), some empirical evidence showed otherwise (Pei et al., 2012). Pei et al. (2012) argued that it is essential to incorporate disaggregated speed data to study the relationship between speed and road safety. These studies showed the significance of time varying weather and traffic variables, where their effects were usually not comprehensively addressed.

1.3 Current Research Gaps

Based on the above literature review, some research gaps in existing literature are then identified including the lack of study in crash rates, time-varying variables, serial correlation and unobserved heterogeneity. Detailed discussion on these research gaps is given in the following section.

1.3.1 Lack of Study in Crash Rates

For studies on the occurrence of crashes, most of the focus has been placed on crash frequencies, which measure the number of accidents on some road transport facilities over some time period. As an important alternative to crash frequencies, crash rates (measured as the number of accidents per million vehicle miles traveled (VMT)) modeling can be promising because of its popularity in traffic safety performance assessments by different stakeholders. Compared to abundant literature on crash frequencies, studies on crashes rates are very limited (Anastasopoulos et al., 2012a, 2012c, 2008; X. Xu et al., 2013). More studies on crash rates are clearly desired to fill such a gap and achieve better understanding towards motor vehicle crashes.

1.3.2 Time-varying Explanatory Variables

Past studies on crash frequencies and crash rates were usually conducted over some extended time period, often, a long one ranging from a year to several years. These studies were geared towards providing guidance for policies and countermeasures to reduce the number of crashes. However, they are unable to capture the impact of explanatory variables (e.g. weather conditions and traffic characteristics) that vary significantly over the specified time period. This limitation undermines the models' explanatory power, especially for time-varying variables (weather and traffic variables) as their variations are neglected due to aggregation. Omitting these time-varying explanatory variables not only leads to the loss of valuable information over the specified time period but also introduces error and bias in model estimation as a result of the unobservable (Lord and Mannering, 2010; Washington et al., 2011).

Traffic safety researchers have long recognized that time-varying variables (e.g. weather and traffic related variables) are critical factors in crash analyses (Theofilatos and Yannis, 2014; Wang et al., 2013a). However, most of these existing studies were still conducted on a high aggregation level possibly due to data availability problem. For weather characteristics, these studies adopted variables such as annual precipitation (Chang and Chen, 2005), monthly rainfall (Yaacob et al., 2010), weekly averaged precipitation (Malyshkina et al., 2009) and some proxy variables (Caliendo et al., 2007) to represent weather effects. In fact, without studies conducted at the more disaggregated level, it may result in ecological fallacy which indicates that relationship observed in aggregated level may not hold at disaggregate level (Davis, 2004; Freedman, 1999). Highway safety studies that adopt refined-scale data are still rare, thus are needed to disclose the actual impact of these time-varying variables on motor vehicle crashes.

Recent advances in Intelligent Transportation System have made detailed monitoring of real-time traffic and weather data available on most highways in the U.S. The incorporation of these real-time traffic and weather data into crash models could potentially bring more valuable and insightful understanding of crash mechanism, and makes it possible for real-time crash risk to be incorporated into the intelligent proactive traffic management system.

1.3.3 Serial Correlations

Given the limitation and potential benefits associated with time-varying variables, it is necessary to overcome these issues resulted from time-varying variables. However, it requires not only accessing high-quality data but also addressing some of the methodological challenges that come along with the adoption of high-quality data. To begin with, processing crash data into refined-scale generates multiple observations for the same roadway segments, and these repeated observations on the same roadway segments will be correlated over time. By the same token, if the roadway segments are in close proximity, it can set up correlation over space as well. This leads to serial correlation in the disturbances of the model, which violates a fundamental assumption in the regression models (identically independent distributed assumption). Panel data model was widely applied in the literature to address serial correlations in the data, however, its capability to consider spatial and temporal correlation based on neighboring structure is limited. More specialized models such as spatial model and spatio-temporal model are preferred in this regard.

1.3.4 Unobserved Heterogeneity

In general, motor vehicle crash is a highly complex process, which relates to various contributing factors. These contributing factors include roadway geometrics, traffic characteristics, environmental conditions and human elements. It is nearly impossible to collect all the data that contribute to a highway accident and its resulting injury severity, especially those related to human factors. As a result, the impacts of these unobserved factors on the likelihood of a highway accident cannot be adequately captured by the explanatory variables in the model, leading to unobserved heterogeneity problem (Mannering et al., 2016). If the unobserved heterogeneity is not appropriately addressed in the model, inconsistent parameters estimates would be estimated and erroneous inferences would be made consequently.

1.3.5 Multiple Dependent Variables

Crash frequency of different crash types or injury severity levels may share some unobserved factors. However, most literature on highway crash studied the total number of crashes over a specified time period for some roadway segments or intersections, without distinguishing subgroups of these crashes (crashes of different injury severity, different type, etc.) and capturing correlations between these subgroups. To examine the frequency of different crash type or injury severity, separate models were traditionally conducted, which suffers the loss of efficiency in parameter estimation. As discussed in section 1.2.3, multivariate count data models are applied to address multiple dependent variables in crash analysis. A major drawback of the current multivariate count data models is that they are designed only for cross-sectional data. Multivariate count data model that addresses panel data and data with spatial-temporal structure is very limited in crash modeling.

1.4 Objectives

Given the current research gaps identified in the section above, the objective of this dissertation is to narrow these gaps by systematically developing advanced multilevel models for traffic safety modeling. It is expected that series of new crash models developed in this dissertation not only contribute to the state-of-the-art crash modeling, but also add to the knowledge toward developing proactive traffic management strategy. At the same time, some empirical problems regarding two major highways I-25 and I-70 in Colorado are also studied. Specifically, the objectives of the current dissertation are as follows:

The <u>first objective</u> is to conduct crash rate analysis with data in refined scales to quantify the relation between crash rate and time-varying variables along with other contributing factors. The findings from this study help researchers gain a better understanding of the effects of the time-varying variable.

The <u>second objective</u> is to examine the unobserved heterogeneity issue on mountainous highways crash rates by developing an advanced random parameter tobit model with panel data in refined temporal scales. The results can reveal the random nature of some contributing variables on crash rates.

The third objective is to propose a correlated random parameter marginalized two-part model as an alternative to study the relationship between crash rates and its contributing factors. The proposed model is not only theoretically appealing but also provides a better fit to the data.

The fourth objective is to examine the differences of contributing factors towards injury severity on mountainous (MN) and non-mountainous (NM) highway crashes using mixed logit models. Mountainous highways usually exhibit complex geometry features such as steep gradients or sharp curves, which can cause considerably different driving behavior and vehicle performance as compared to non-mountainous ones. The results could provide directions on tailored injury mitigation countermeasures regarding MN and NM crashes.

The fifth objective is to study the effects of weather and traffic characteristics on single-vehicle and multi-vehicle crashes jointly by proposing a multivariate count data model that addresses unobserved heterogeneity across multiple dependent variables. The proposed model addresses multivariate count data model in a panel setting.

The sixth objective is to introduce a framework of Bayesian multivariate space-time model that can address spatial correlation/heterogeneity, temporal correlation/heterogeneity, and the correlation between different injury severities.

1.5 Outline of the Dissertation

To fulfill these objectives described in preceding section, the rest of the dissertation is divided into seven chapters structured as follows:

Chapter two examines the impacts of time-varying variables on hourly crash rates by combining real-time traffic data, real-time weather data, geometric data and crash data from highway I-25 in Colorado. This section aims at 1) quantifying the relation between crash rate and its contributing factors, and 2) examining the difference between daytime and nighttime crashes. The random effects tobit model was proposed to handle the panel data nature. The proposed models address left-censoring effects of the crash rates data while accounting for unobserved heterogeneity across roadway segments and serial correlations within the roadway segment in the meantime. To capture different characteristics, two separate models for daytime crashes and nighttime crashes are developed using refined-scaled data. Marginal effects are also calculated to examine the effects of each variable on daytime and nighttime crash rates.

Chapter three investigates the relationship between contributing factors and hourly crash rates on the mountainous highway I-70 by addressing unobserved heterogeneity across roadway segments using an advanced random parameter tobit model. Specifically, refined-scale weather and traffic data are incorporated with crash data to capture the varying nature of complex driving conditions on I-70. Random parameter tobit model, which offers improved capability to handle unobserved heterogeneity, is developed along with its fixed parameter counterpart. Model comparison is also conducted by performing likelihood ratio test and comparing the goodness of fit measures.

Chapter four attempts to propose a marginalized two-part model coupled with a correlated random parameter structure as an alternative to tobit model to study crash rates. The proposed methodology is demonstrated by investigating daily crash rates on I-25 in Colorado. The applicability and potential advantages of the marginalized two-part model as an alternative tool to study crash rates are also explored by addressing those methodological challenges associated with time-varying variables, temporal correlations, and unobserved heterogeneity. The proposed model shows superiority over other alternative models in terms of goodness of model fit.

Chapter five aims at investigating the differences in injury severity characteristics between mountainous and non-mountainous interstate highways through a comparative study. One major interstate highway with typical mountainous (MT) terrain and another one with non-mountainous (NM) terrain in Colorado have been selected for this chapter. A comparative investigation of the impact on injury severity from mountainous and non-mountainous highways is conducted. Separate mixed logit models are estimated for both highways with four-year detailed police reported crash data. Elasticity measures are also computed to identify most critical factors influencing injury severity outcomes for both types of highways. The mixed effects of some variables towards injury severity on MT and NM crashes are revealed.

Chapter six focuses on: (1) examining whether bivariate Poisson-lognormal with correlated segment level random effects (CREBPL) is appropriate for analyzing single-vehicle (SV) and multi-vehicle (MV) crashes under multivariate panel data; and (2) studying the impact of weather and traffic conditions in addition to exposure and geometric conditions on SV and MV crashes. In light of serial correlations in the data, two alternative models are proposed in addition to CREBPL: (1) bivariate Poisson-lognormal model (BPL), and (2) bivariate Poisson-lognormal model with independent segment specific random effects (UREBPL). To investigate the superiority of CREBPL, several goodness of fit measures are compared between different models. Finally, the effects of weather variables on SV and MV crashes are examined.

Chapter seven investigates the application of multivariate space-time models to jointly analyze crash frequency by injury severity levels in fine temporal scale. A framework of Bayesian multivariate space-time model is developed to address spatial correlation and/or heterogeneity, temporal correlation and/or heterogeneity, and correlations between crash frequencies of different injury severity level. A series of multivariate space-time models are proposed under the Full Bayesian framework with different assumptions on the spatial and temporal random effects. The proposed methodology is illustrated using one-year daily traffic crash data from the mountainous interstate highway I70 in Colorado, which is categorized into no injury crash and injury crash. The best model within the framework is identified through comparing the goodness of fit measures including DIC and posterior predictive checks.

Chapter eight concludes the dissertation by summarizing the major findings and sharing observations in terms of future research directions.

CHAPTER 2 REFINED SCALE PANEL DATA CRASH RATE ANALYSIS USING RANDOM-EFFECTS TOBIT MODEL[1]

2.1 Introduction

Traffic accident modeling has long been an important research area for safety researchers, and a lot of research efforts have focused on crash (accident) frequency. In addition to crash frequency, crash rate is another important highway safety indicator with some different appeals. For example, the adoption of crash rate offers standardized traffic safety measure which can more conveniently assess the relative risks among different road segments (Anastasopoulos et al., 2008). Several significant challenges about current crash frequency models have been identified in the comprehensive review study conducted by Lord and Mannering (2010). Among all those, some challenges are also shared by crash rate models, such as time-varying explanatory variables, temporal and spatial correlations.

Crash frequency or crash rate analyses were typically conducted in large scales in both temporal (e.g. yearly) and spatial (e.g. whole road) domains. It is known that the adoption of aggregated explanatory variables in larger scales ignores the within-period variation of explanatory variables, which will result in “the loss of potentially relevant explanatory information” (Lord and Mannering, 2010) and introduce error in model estimation due to unobserved heterogeneity (Washington et al., 2011). To develop crash risk models in refined scales requires not only the availability of the disaggregated data, but also overcoming some technical challenges, such as correlations by sharing unobserved effects among multiple observations generated from the same road segments and/or time period (Lord and Mannering, 2010; Lord and Persaud, 2000; Shankar et al., 1998; Sittikariya and Shankar, 2009; Ulfarsson and Shankar, 2003). More detailed review of these challenges will be made in Section 2.2.

As compared to normal driving conditions, refined-scale traffic safety modeling is more critical in adverse driving conditions, under which the explanatory variables often change considerably over different time instants and locations, such as weather, traffic and road surface conditions. As a result, both

[1] This chapter is developed from a research paper published by Chen et al. (2014).

time-varying and spatially varying (cross-sectional) data need to be considered in smaller scales, which make the temporal and spatial correlations even more complicated. Panel data with random effects has been used in recent years to deal with the temporal or spatial correlation issues (Qi et al., 2007; Shankar et al., 1998; Ulfarsson and Shankar, 2003). Despite all the progress on advanced statistical models of adopting panel data, most studies were primarily on crash frequency predictions. Crash rate modeling with refined scales, with or without panel data application, has rarely been reported.

In contrast to a large number of literature about different count-data models for crash frequency prediction (Lord and Mannering, 2010), the studies on crash rate are limited. From the statistical perspective, crash rate modeling involves continuous data, which is usually left-censored at zero. This is because roadway segments without any accident reported over a specified time period will just yield zero on accident rate data record. To handle the censoring problem of crash rate prediction, Anastasopoulos et al. (2008) successfully employed tobit model on the highway accident rate data in the state of Indiana. To handle unobserved heterogeneity across observations, Anastasopoulos et al. (2012a) further developed a random-parameters tobit model, which demonstrated superiority over its fixed parameter counterpart regarding the goodness of fit. Xu et al. (2013) used tobit regression to investigate the endogeneity problem between crash rate and travel speed. Other studies also include the multivariate tobit analysis on crash rate for each injury type (Anastasopoulos et al., 2012b).

Most of the existing models on crash frequency or crash rate were developed by combining both daytime and nighttime data as a whole, based on the assumption that traffic safety during daytime and nighttime share same contributing factors and characteristics. It is known that driving environments at daytime and nighttime are very different due to varying light, environmental and traffic conditions affecting driver behavior and, eventually traffic safety risks. In recent years, there have emerged a few studies modeling crash frequency for daytime and nighttime separately (Bullough et al., 2013; Dinu and Veeraragavan, 2011; Donnell et al., 2010), which have shown a considerable difference in contributing factors towards crash frequency between daytime and nighttime. So far, however, there is no such study reported on the crash rate, and it remains unclear whether separate modeling of the crash rate for daytime or nighttime is needed or not.

The present chapter aims at conducting crash rate study with refined-scale data in both temporal and spatial domains to identify the relation between crash rate and its contributing factors. The section integrates the strength from both the random effect tobit model and the panel data formulation of specific driving conditions such as geometry characteristics, traffic flow data, weather conditions, and road surface conditions. As a result, the outlined challenges associated with the censoring problem of crash rate, aggregated explanatory variables, temporal and spatial correlation can be appropriately addressed (Lord and Mannering, 2010). To capture different traffic characteristics, two separate models for daytime crashes and nighttime crashes are developed using refined-scale data. Some brief reviews of the topics related to the present chapter are made in the following section.

In most existing studies, aggregated accident data and associated contributing factors were often considered in large time intervals and spatial domains, partially due to the unavailability of detailed data. The contributing factors are generally aggregated into monthly or even yearly intervals and over region-wide or long roadway segments. In adverse driving conditions, detailed weather and other environment-related factors (e.g. precipitation, visibility, wind, humidity, pavement characteristics, road surface condition, etc.) are often found critical in causing a crash. As a result of adopting aggregating data, some potential explanatory information may be lost and some factors' effects towards crash occurrence can be masked during data aggregation (Lord and Mannering, 2010; Usman et al., 2011). Recently, researchers have started to incorporate more influencing factors in refined scales. Keay and Simmonds (2005) examined the relation between rainfall and daily collision data in Australia. Usman et al. (2010) are among the first to build disaggregated hourly model to study crash frequency incorporating weather and surface data. It was found hourly traffic volume, hourly road surface condition, and other hourly weather factors are important for collisions during winter storms. Abdel-Aty et al. (2004) analyzed the effects of traffic characteristics towards freeway crashes using real-time loop detector data. Abdel-Aty and Pemmanaboina (2006) combined the real-time ITS traffic data, the archived weather data, and the historical accident data to calibrate the crash prediction model. Other research attempts of building disaggregated data model by incorporating traffic flow, weather data, surface condition and other factors can be found in the studies by Hossain and Muromachi (2013), Usman et al. (2012) and Yu et al. (2013a, 2013b). In contrast to the progress on crash frequency modeling as summarized above, crash rate modeling in refined scales has rarely been reported.

With refined data in both spatial and temporal domains, serious correlation problems may exist among the records. Panel data (also known as cross-sectional time-series data) models, which have been widely used in the econometric, social and behavioral analysis, gained its popularity among safety specialists due to its capacity to address both time-series and cross-sectional variations. In crash modeling arena, panel data models can be mainly generalized into three categories: fixed effects (FE) models, random effects (RE) models, and random parameters (RP) models. An FE model has some appeals in panel data modeling because of not requiring the assumption about unobserved heterogeneity (Greene, 2008). Nonetheless, FE models require estimations of a vast number of parameters (e.g. site-specific or time-specific indicator variables), which may dominate the contributing factors. As compared to FE model, RE model is usually more prevalent in modeling crash data with repeated observations (Aguero-Valverde, 2013a; Chin and Quddus, 2003; Qi et al., 2007; Shankar et al., 1998). Recently, the random parameter (RP) model was proposed as a general extension of RE model to allow not only the constant term but also the coefficients to vary across observations (Anastasopoulos and Mannering, 2009; Lord and Mannering, 2010). In addition to these three most frequently used types of panel data models, there were also some other research attempts to account for temporal and spatial correlations in accident occurrence, such as negative multinomial model (Caliendo et al., 2013; Ulfarsson and Shankar, 2003), generalized estimating equation (GEE) model (Lord and Persaud, 2000; Wang and Abdel-Aty, 2006), and time-series model (Mohammed A Quddus, 2008). However, most of these studies focus on crash frequency prediction by applying panel data with repeated yearly observations, rather than data in more refined scales.

2.2 Methodology

In this section, the base tobit model proposed by Tobin (1958) will be adopted as the starting point to study the left-censored crash rate data. Under panel data formation, repeated observations are given for each group (e.g. roadway segments or intersections). As discussed above, correlations may exist among these repeated observations. A tobit model with random effects is therefore proposed as it is capable of accounting for both censoring effects and serial correlations (temporal or spatial correlations, depending on the panel data setting). The application of random effects tobit model is undertaken to account for such correlations across observations in addition to unobserved heterogeneity. In the present chapter, random effects tobit model will be developed based on a typical left-censored tobit model (with a lower limit of zero). A baseline structure for a left-censored tobit model with panel data can be described as follows:

$$Y_{it}^* = \boldsymbol{\beta}_{it}\boldsymbol{X}_{it} + \varepsilon_{it}, \qquad i = 1, \dots, N, t = 1, \dots, T_i \tag{2.1}$$

and

$$Y_{it} = \begin{cases} Y_{it}^* & \text{if } Y_{it}^* > 0 \\ 0 & \text{if } Y_{it}^* \leq 0 \end{cases} \tag{2.2}$$

where N is the number of groups (i.e. the number of roadway segments), T_i is the number of the repeated observations for roadway segment i, Y_{it} is the dependent variable (accident rate per million miles travelled) and Y_{it}^* is the latent variable which is observed only when being positive. $\boldsymbol{X}_{it}$ is a vector of explanatory variables (traffic condition, geometric characteristic, temporal characteristic weather condition, surface condition, etc.), $\boldsymbol{\beta}_{it}$ is a vector of estimable coefficients, and ε_{it} is the error term.

The random effects tobit model is formed by decomposing the error term ε_{it} into two parts:

$$\varepsilon_{it} = \mu_i + \upsilon_{it} \tag{2.3}$$

where μ_i is the random effects term which follows normal distribution with mean 0 and variance σ_μ^2, and υ_{it} is the remaining disturbance term which follows a normal distribution with mean 0 and variance σ_υ^2. μ_i and υ_{it} are constructed in a way such that the following equations hold:

if $i = j$ and $t = s$

$$\mathrm{Cov}\left[\upsilon_{it}, \upsilon_{js}\middle|\boldsymbol{X}\right] = \mathrm{Var}[\upsilon_{it}|\boldsymbol{X}] = \sigma_\upsilon^2 \tag{2.4a}$$

Otherwise

$$\mathrm{Cov}\left[\upsilon_{it}, \upsilon_{js}\middle|\boldsymbol{X}\right] = 0 \tag{2.4b}$$

if $i = j$

$$\mathrm{Cov}\left[\mu_i, \mu_j\middle|\boldsymbol{X}\right] = \mathrm{Var}[\mu_i|\boldsymbol{X}] = \sigma_\mu^2 \tag{2.5a}$$

Otherwise

$$\mathrm{Cov}\left[\mu_i, \mu_j\middle|\boldsymbol{X}\right] = 0 \tag{2.5b}$$

$$\mathrm{Cov}[\upsilon_{it}, \mu_i|\boldsymbol{X}] = 0 \text{ for all } i, t, j \tag{2.6}$$

Then

$$Corr[\varepsilon_{it}, \varepsilon_{is}|\boldsymbol{X}] = \sigma_\mu^2/(\sigma_\upsilon^2 + \sigma_\mu^2) \tag{2.7}$$

Possible serial correlation, which arises from the time-series nature of panel data, can be adequately accounted for through Eq. (2.7). The essential assumption for the random effects tobit model to distinguish itself from its fixed effects counterpart is that the heterogeneity (i.e. random effects) is assumed to be uncorrelated with the independent variables $\boldsymbol{X}'_{it}$. Thus, the corresponding log-likelihood function for the random effects tobit model is derived by obtaining the unconditional density through integrating the random effects μ_i out of the conditional density (Greene, 2012):

$$LL = \sum_{\forall i} ln \int f(Y_{it}|X_{it}, \mu_i; \theta)\, (1/\sigma_\mu)\phi(\mu_i/\sigma_\mu) d\mu_i \tag{2.8}$$

where

$$f(Y_{it}|\boldsymbol{X}_{it}, \mu_i; \theta) = \prod_{Y_{it}>0} \frac{1}{\sigma_\upsilon} \phi(\frac{Y_{it} - \boldsymbol{\beta X}'_{it} - \mu_i}{\sigma_\upsilon}) \prod_{Y_{it}=0} \Phi(\frac{-\boldsymbol{\beta X}'_{it} - \mu_i}{\sigma_\upsilon}) \tag{2.9}$$

$\phi(\cdot)$ is standard normal density function; $\Phi(\cdot)$ is standard normal distribution function.

Both Gauss-Hermite quadrature and simulation-based maximum likelihood estimations can be adopted to get the maximum of the log-likelihood function of random effect models (Greene, 2012). The first approach gives approximated estimation in which the estimation accuracy is partly determined by the integration points used. Random effects tobit model can be viewed as a particular case of random parameter tobit model, in which only the constant term is treated as a random parameter (Anastasopoulos and Mannering, 2009; Lord and Mannering, 2010) and simulation-based maximum likelihood estimation can be exploited to solve the log-likelihood function. Halton draw (Halton, 1960), a widely used simulation technique, is available to provide more efficient distribution of draws than purely random draws for simulation-based integration (Bhat, 2003; Train, 1999).

Due to the censoring effect of the data, it is known that the estimated parameters from the tobit model do not reflect the actual change of the dependent variable when the independent variables have a unit increase (Calzolari et al., 2001). In order to evaluate the effects of each variable in random effects tobit model, Calzolari et al. (2001) defined the marginal effects of a unit change in the i^{th} variable of **X** on the dependent variable Y as:

$$\frac{\partial E[Y]}{\partial \boldsymbol{X}_i} = \beta_i \Phi(z) \tag{2.10}$$

where

$$z = \boldsymbol{\beta X}/\sigma_{\epsilon} = \boldsymbol{\beta X}/\sqrt{\sigma_{\upsilon}^2 + \sigma_{\mu}^2} \tag{2.11}$$

and $\Phi(z)$ is the cumulative normal distribution function, β_i is the estimated coefficient corresponding to the i-th independent variable $\boldsymbol{X}_i$, σ_{μ} is the standard deviation of the random effects and σ_{υ} is the standard deviation of the disturbance term.

In addition, to assess how independent variables affect the probability of having crashes (this is equivalent to crash rate is greater than zero), marginal effects of a unit change in the i^{th} independent variable on the probability of an observation being above zero can be obtained as (Parisi and Sembenelli, 2003):

$$\frac{\partial \Phi(z)}{\partial \boldsymbol{X}_i} = \beta_i \varphi(z)/\sqrt{\sigma_{\upsilon}^2 + \sigma_{\mu}^2} \tag{2.12}$$

where $\varphi(z)$ is the standard normal density function, other variables are defined as before.

In this chapter, two separate tobit models, one for daytime and one for nighttime, will be developed. To test the statistical difference between the two separately estimated models and the all-data model (single model with daytime and nighttime data combined), a likelihood ratio test is performed using following statistics (Washington et al., 2011):

$$\mathrm{X}^2 = -2[LL(\beta_{all}) - LL(\beta_{Day}) - LL(\beta_{Night})] \tag{2.13}$$

where $LL(\beta_{all})$ is the log-likelihood at convergence of model estimated with all data; $LL(\beta_{Day})$ and $LL(\beta_{Night})$ are the log-likelihood at convergence of model estimated using daytime data and nighttime data, respectively; X^2 follows χ^2 distribution with the degree of freedom equaling the difference between the number of parameters of all-data model and the sum of parameters in daytime and nighttime models.

A likelihood ratio test can also be conducted to compare the statistical significance of the random effects tobit model over its fixed effects counterpart. The test statistic is described as:

$$\chi^2 = -2[LL(\beta_F) - LL(\beta_R)] \tag{2.14}$$

where $LL(\beta_F)$ is the log likelihood at convergence for Daytime/Nighttime fixed-effects tobit model, and $LL(\beta_R)$ is the log likelihood at convergence for Daytime/Nighttime random effects tobit model.

2.3 Data Description

A portion of Interstate I-25 in Colorado (between City of Castle Rock and City of Northglenn) is chosen to be the study section with a total length of 55.93 miles. It consists of a 28.55-mile northbound road starting from Mile Marker (MM) 188.49 to MM 221.03 and a 27.38-mile southbound road starting from MM 188.49 to MM 219.86. I-25 has a relatively flat terrain and bears a lot of similarities with many other highways across the United States. Detailed data related to collisions on the selected study section of I-25 from four sources are processed in this chapter:

(1): Traffic accident data over 1-year period (1 January 2010 to 31 December 2010) obtained from Colorado State Patrol (CSP);

(2): Highway geometry data (including pavement conditions) provided by Colorado Department of Transportation (CDOT);

(3): Real-time weather and road surface condition data recorded by Road Weather Information System (RWIS).

(4): Real-time traffic data recorded by 43 traffic flow monitoring stations along the study section on I-25.

In traditional traffic safety modeling, accident data were usually aggregated over an extended time interval and on a relatively long roadway segment span. In the present chapter, to form a refined-scaled panel data structure, the study section is firstly split into small homogeneous roadway segments (approximately 1 mile per segment) according to CDOT traffic station assignment. For example, one of the traffic stations is assigned to MM 188.49 - 192.99 by CDOT and there are totally 43 traffic stations in the study section. Then road segmentations were scrutinized according to the roadway characteristics inventory data from CDOT. If there exists distinct variance of road design within one road segment (e.g. variance of lane width, the number of lanes, speed limit, pavement condition, shoulder type, median type), the road segment will be re-segmented based on different geometric designs. A total of 57 homogeneous roadway segments is generated with 29 and 28 on the northbound and southbound roads, respectively. The accident data, real-time traffic data, real-time weather and surface condition data are processed to obtain hourly readings through integration. By assigning these data to corresponding roadway segments, two data sets with repeated hourly observations for each roadway segment, one for daytime and the other one for nighttime, are finally developed to facilitate the following study. Tables 2.1 and 2.2 summarize descriptive statistics of selected variables for daytime and nighttime datasets, respectively. More detailed information about the dataset is provided in the following.

2.3.1 Geometry Data

Roadway geometry data are collected from Roadway Characteristics Inventory (RCI) of CDOT. RCI contains detailed information about I-25 highway, including number of lanes, lane width, segment length, speed limit, deflect angle of curve, longitudinal grade, curvature, curve radius, shoulder type and width, median type and width, number of merging ramps, number of diverging ramps, pavement condition and remaining service life of rutting, etc. The variables which are significant in daytime and nighttime models are shown in Tables 2.1 and 2.2.

2.3.2 Weather and Road Surface Data

There are five weather stations installed along the interstate I-25 to provide motorists with real-time weather data. The real-time weather datasets are monitored in 20-minute intervals including visibility, humidity, precipitation, temperature, wind, and real-time road surface condition data such as road surface status and water depth. It should be noted that visibility is usually defined as the greatest distance that an object can be seen and recognized against the sky. To better account for its effect toward crashes, visibility is defined as the shortest distance an object can be perceived. Moreover, precipitation is often chosen as an important factor of weather condition in crash modeling. It is known that precipitation (rain, snow or hail) primarily contributes to accident occurrence through altering road surface condition (Yu et al., 2013b). The time lag may exist on the effects of precipitation towards crashes, especially under small time intervals. For example, precipitation on one segment over a certain hour may actually contribute to the crashes happening on this segment during the following hour. Such lag may become critical for hourly crash models, and therefore, real-time road surface condition will also be considered in this chapter along with precipitation status. Weather and road surface condition data are recorded every 20 minutes and the road surface condition definitions in the CDOT database include Dry, Wet, Trace Moisture, Chemically Wet (Moisture mixed with anti-icer), Ice Warning and Ice Watch, etc. In order to derive hourly records, the hourly road surface condition indicators are defined as the dominating condition within that particular hour. For instance, if wet road surface was recorded twice and the dry road surface was recorded once within a given hour, the hourly surface condition will be defined as a wet road surface. Based on the Mile Marker, each roadway segment is assigned to the closest weather station. The weather and road surface condition variables selected for the models of daytime and nighttime can be found in Tables 2.1 and 2.2, respectively.

2.3.3 Traffic Data

With 22 and 21 traffic monitoring stations implemented on the northbound and southbound roadways of the study section respectively, traffic information, such as traffic speed and traffic volume, is readily available. These data originally recorded every 2 minutes, is processed to obtain hourly records through aggregation. Due to possible malfunction and service disruption of the sensors, some real-time data records show unusually high values or leave some “empty” windows, which were discarded from the database. A total of 328,398 observations is then generated including 186,821 daytime observations and 141,577 nighttime observations. Unlike previous studies (e.g. Dinu and Veeraragavan, 2011), the daytime and nighttime records are strictly defined according to the exact time of sunset and sunrise every day based on Colorado Sunrise Sunset Calendar. Instead of adopting traffic speed directly, speed gap, the difference between the speed limit and corresponding mean traffic speed, is selected as the variable. Speed gap, defined here as the difference between the speed limit and average traffic speed, can be considered as a measurement of traffic congestion, and a higher speed gap indicates more congested traffic condition. The original real-time speed data from CDOT database will not exceed the local speed limit for each road segment, which means that if the actual average speed exceeds the local speed limit, the CDOT database will truncate it to the speed limit of the road segment. As a result, the "speed gap" only has non-negative values. Such a truncation of raw data from CDOT database will perhaps bring in some modeling bias issues because the influence of over-speeding driver behavior in some free flow cases cannot be considered in the modeling. Raw speed data without truncation can reveal more insights about the impact of real-time speed on traffic safety.

2.3.4 Crash Rate

Crash rate (the number of accidents per million VMT) is calculated as

$$Crash\ rate_{it} = \frac{Crash_{it}}{htv_{it} \times length_i / 1{,}000{,}000} \tag{2.15}$$

where $Crash\ rate_{it}$ is the number of accidents per million VMT on roadway segment i over t^{th} hour of the year; $Crash_{it}$ is the number of accidents happening on roadway segment i over t^{th} hour of the year;

htv_{it} is the hourly traffic volume on roadway segment i over t^{th} hour of the year; $length_i$ is the length of roadway segment i. If there was no accident happening on one segment over a given hour according to the data, a censored data record with crash rates equaling zero is generated. About 99.4% and 99.8% of the observations are censored with a cluster of zeroes in the daytime and nighttime datasets, respectively.

Table 2.1 Descriptive Statistics of Potential Variables (Daytime Dataset)

Variable	Mean	Standard deviation	Minimum	Maximum
Crash frequency	0.0058	0.0787	0	4
Accident rate (number of accidents per million VMT)	1.6156	28.8257	0	2451.12
Low Speed limit indicator (1 if speed limit is less or equal 55 miles per hour, 0 otherwise)	0.3805	0.4855	0	1
Speed gap(measured as the difference between speed limit and corresponding mean traffic speed)	3.5053	6.4581	0	69.18
Hourly traffic volume (in thousands)	4.1115	1.7734	0.03	14.7455
Truck percentage	5.9469	1.968	2.8	10.7
Visibility(in miles)	1.0735	0.1404	0	1.1
Cross wind speed	4.4687	4.1425	0	31.98
January (1 if it is in January, 0 otherwise)	0.0599	0.2372	0	1
February (1 if it is in February, 0 otherwise)	0.0567	0.2312	0	1
March (1 if it is in March, 0 otherwise)	0.0966	0.2954	0	1
April (1 if it is in April, 0 otherwise)	0.0821	0.2745	0	1
May (1 if it is in May, 0 otherwise)	0.1012	0.3016	0	1
June (1 if it is in June, 0 otherwise)	0.0922	0.2893	0	1
July (1 if it is in July, 0 otherwise)	0.0533	0.2247	0	1
August (1 if it is in August, 0 otherwise)	0.1009	0.3012	0	1
September (1 if it is in September, 0 otherwise)	0.1038	0.3051	0	1
October (1 if it is in October, 0 otherwise)	0.0787	0.2693	0	1
November (1 if it is in November, 0 otherwise)	0.0819	0.2742	0	1
December (1 if it is in December, 0 otherwise)	0.0928	0.2901	0	1
Weekend indicator (1 if it is weekend, 0 otherwise)	0.2719	0.4449	0	1
Number of enter ramp per lane per mile (measured as number of enter ramp/lanes/roadway segment length)	0.2543	0.2153	0	0.9259
Roadway segment length (in miles)	1.0186	0.7747	0.236	4.5
Curvature (degree of horizontal curve)	0.9578	0.6837	0	2.26

Curve ratio	0.4029	0.2838	0	0.9779
Inside shoulder width (in feet)	9.0051	2.6243	5	15
Long remaining service life of rutting (1 if the value of rut is higher than 99, 0 otherwise)	0.2295	0.4205	0	1
Wet road surface (1 if wet, 0 otherwise)	0.0233	0.1508	0	1
Snow occurring	0.0268	0.1615	0	1

Table 2.2 Descriptive Statistics of Potential Variables (Nighttime Dataset)

Variable	Mean	Standard deviation	Minimum	Maximum
Crash frequency	0.0018	0.0448	0	2
Accident rate (number of accidents per million VMT)	1.8029	75.3118	0	9529.25
Low Speed limit indicator (1 if speed limit is less or equal 55 miles per hour, 0 otherwise)	0.3539	0.4782	0	1
Speed gap(measured as the difference between speed limit and corresponding mean traffic speed)	1.5029	3.6996	0	51.1913
Hourly traffic volume (in thousands)	1.3391	1.3093	0.03	14.988
Truck percentage	6.5686	1.8003	2.8	10.7
Visibility(in miles)	1.0763	0.1304	0.1	1.1
Cross wind speed	3.7214	3.5243	0	28.466
January (1 if it is in January, 0 otherwise)	0.0832	0.2761	0	1
February (1 if it is in February, 0 otherwise)	0.0722	0.2588	0	1
March (1 if it is in March, 0 otherwise)	0.0985	0.2980	0	1
April (1 if it is in April, 0 otherwise)	0.0671	0.2501	0	1
May (1 if it is in May, 0 otherwise)	0.0638	0.2445	0	1
June (1 if it is in June, 0 otherwise)	0.0484	0.2146	0	1
July (1 if it is in July, 0 otherwise)	0.0339	0.1810	0	1
August (1 if it is in August, 0 otherwise)	0.0743	0.2622	0	1
September (1 if it is in September, 0 otherwise)	0.0932	0.2907	0	1
October (1 if it is in October, 0 otherwise)	0.09717	0.2962	0	1
November (1 if it is in November, 0 otherwise)	0.1129	0.3165	0	1
December (1 if it is in December, 0 otherwise)	0.1554	0.3623	0	1
Weekend indicator (1 if it is weekend, 0 otherwise)	0.2705	0.4442	0	1
Number of enter ramp per lane per mile (measured as number of enter ramp/lanes/roadway segment length)	0.2494	0.2138	0	0.9259
Roadway segment length (in miles)	1.0071	0.7603	0.236	4.5
Curvature (degree of horizontal curve)	0.9335	0.6783	0	2.2604

Curve ratio	0.4004	0.2863	0	0.9779
Inside shoulder width(in feet)	9.0078	2.5265	5	15
Long remaining service life of rutting (1 if the value of rut is higher than 99, 0 otherwise)	0.2194	0.4139	0	1
Wet road surface (1 if wet, 0 otherwise)	0.0545	0.2269	0	1
Snow occurring	0.0319	0.1757	0	1

2.4 Model Estimation Results

The likelihood ratio test with a p-value less than 0.001 of the two separately estimated models as opposed to the all-data model shows that the choice of the modeling crash rate under daytime and nighttime subsets is reasonable and necessary. The χ^2 value for Daytime model between the random effects tobit model and its fixed effects counterpart is equal to 70 with 4 degrees of freedom and that for Nighttime model is equal to 17.9 with 1 degree of freedom. These results indicate that the random effects tobit models are statistically better than their fixed counterparts with more than 99.99% confidence.

The estimation results of random effect daytime and nighttime tobit models for hourly crash rate are presented in Table 2.3. All variables significant at 90% level of confidence are reported. Random effects are found to be significant in both daytime and nighttime models, indicating that unobserved heterogeneity plays a major role in crash rate prediction. Four categories of contributing factors, such as traffic characteristics, environmental characteristics, road characteristics and temporal characteristics, are found to be significant to crash rate. The development of two separate crash rate prediction models makes it possible to reveal more detailed information about the impacts from different contributing factors on the crash rate at daytime and nighttime. For example, some contributing factors are found to be significant only to daytime crashes (e.g. low speed limit indicator, the number of entering ramps per lane per mile, etc.), while some others are only significant in the nighttime crash rate model (e.g. snow occurring, truck percentage). In the meantime, there are also some contributing factors that are significant in both models with same signs, such as speed gap, hourly traffic volume, curvature, etc. There are also some contributing factors which are significant in both models, but with different signs (e.g. visibility). It should be noted that these findings would not otherwise become available with a traditional pooled data model by mixing the data from daytime and nighttime.

Table 2.3 Random Effects Tobit Model Estimation Results

	Daytime model		Nighttime model	
	Parameter estimate	t-ratio	Parameter estimate	t-ratio
Constant	-2835.7	-23.33**	-11379	-10.90**
Low Speed limit indicator (1 if speed limit is less or equal 55 miles per hour, 0 otherwise)	108.37	2.08*		
Speed gap(measured as the difference between speed limit and corresponding mean traffic speed)	28.08	23.32**	124.50	10.27**
Hourly traffic volume(in thousands)	32.62	4.21**	395.76	6.80**
Truck percentage			-155.90	-2.12*
Visibility(in miles)	109.91	1.75	-1633.17	-3.17*
November (1 if it is on November, 0 otherwise)	63.54	1.86		
Weekend indicator (1 if it is weekend, 0 otherwise)	41.86	1.96		
Number of entering ramp per lane per mile (measured as number of entering ramp/lanes/roadway segment length)	-252.33	-2.12*		
Roadway segment length (in miles)	85.94	2.94 **		
Curvature (degree of horizontal curve)	102.06	2.61*	663.97	3.07**
Inside shoulder width(in feet)			106.65	1.98
Long remaining service life of rutting (1 if the value of rut is higher than 99, 0 otherwise)	141.25	2.50*	681.67	2.23*
Wet road surface (1 if wet, 0 otherwise)	198.16	3.99**	1233.57	4.26**
Snow occurring			-1063.98	-2.23*
Random effects (σ)	142.01	7.45**	668.75	4.83**
Number of observations	186821		141577	
Log-likelihood at convergence	-12854.5		-3592.55	
AIC(smaller is better)	25737		7209.1	
BIC(smaller is better)	25765		7233.6	

* 0.95 significance level
**0.99 significance level

The marginal effects for the crash rate (no. of accidents per 1 million VMT) and the probability of having a crash rate over zero are computed and the results are presented in Table 2.4. The specific model estimation results are discussed in the following by categories of the contributing factors.

Table 2.4 Marginal Effects of Estimated Random Effects Tobit Models

	Daytime model		Nighttime model	
	Overall sensitivity[1]	Zero sensitivity[2] (%)	Overall sensitivity[1]	Zero sensitivity[2] (%)
Low Speed limit indicator (1 if speed limit is less or equal 55 miles per hour, 0 otherwise)	0.6302	0.1939		
Speed gap (measured as the difference between speed limit and corresponding mean traffic speed)	0.1633	0.0502	0.2247	0.0171
Hourly traffic volume(in thousands)	0.1897	0.0584	0.7142	0.0545
Truck percentage			-0.2813	-0.0215
Visibility (in miles)	0.6392	0.1966	-2.9418	-0.2244
November (1 if it is in November, 0 otherwise)	0.3695	0.1137		
Weekend indicator (1 if it is weekend, 0 otherwise)	0.2434	0.0749		
Number of entering ramp per lane per mile (measured as number of entering ramp/lanes/roadway segment length)	-1.4675	-0.451		
Roadway segment length (in miles)	0.4998	0.1537		
Curvature (degree of horizontal curve)	0.5935	0.1826	1.1982	0.0914
Inside shoulder width (in feet)			0.1925	0.0147
Long remaining service life of rutting (1 if the value of rut is higher than 99, 0 otherwise)	0.8215	0.2527	1.2301	0.0938
Wet road surface (1 if wet, 0 otherwise)	1.1524	0.3545	2.2261	0.1698
Snow occurring			-1.9200	-0.1464

[1] Marginal effects of the overall expected value (see Eq. (2.10))
[2] Marginal effects of the probability of being above zero (see Eq. (2.12))

2.4.1 Traffic Characteristics

Speed limit is typically an important factor for traffic safety and it ranges from 55 miles-per-hour (mph) to 75 mph with an increment of 5 mph for the roadway segments being analyzed. Based on the results of best model fit, 55 mph is determined to be the threshold to define low speed limit indicator. Specifically, the indicator equals one when the speed limit is 55 mph and zero when the speed limit is higher than 55 mph. Table 2.4 shows that low speed limit indicator is found to be significant in the daytime model. The presence of low speed limit increases crash rate by 0.6302 and the probability of having a crash rate above zero by 0.1939%. Although it seems counterintuitive, this finding may be attributed to the fact that low speed limits are more likely to be assigned to crash-prone segments of roadway, as highlighted in several other studies (Donnell et al., 2010; Lao et al., 2014). In the nighttime model, however, speed limit indicator is found to be insignificant.

In the present chapter, it is found that speed gap is a major factor affecting crash rate in both daytime and nighttime models with a positive sign. Table 2.4 shows that a unit increase in the speed gap (in mph) contributes to an increase of crash rate by 0.1633 and 0.0502% higher probability of having crash rate above zero during daytime. During nighttime, a unit increase of the speed gap will increase crash rate by 0.2247 and the probability of having a crash rate over zero by 0.0171%. Some previous work (for example, Dias and Miska, 2009) also suggested that vehicle accidents are more likely to occur where traffic gets congested.

Hourly traffic volume is found to be statistically significant to crash rate during both daytime and nighttime periods. As shown in Table 2.4, a unit increase in hourly traffic volume (in thousands) is associated with an increase of 0.1897 and 0.7142 on crash rate during daytime and nighttime, respectively. Similarly, a unit increase in hourly traffic volume (in thousands) contributes to 0.0584% and 0.0545% higher probability of having a crash rate above zero during daytime and nighttime periods, respectively. The finding is different from a previous study (Usman et al., 2011) in which hourly traffic was found to be significant only in the aggregate analysis instead of the hourly disaggregate analysis for crash frequency data. It's worth mentioning that this disparity may be the result of difference in modeling crash frequency and crash rate, and further study may be needed in this regard to draw insightful conclusion.

In addition to above mentioned traffic-related factors, the truck percentage is also found to have a significant effect on crash rate during nighttime. The negative sign of this factor shows that high truck percentage is associated with lowering crash rate during night period. It should be noted that such a phenomenon is not found in the daytime model. Possible reasons behind such difference include more alerting driving behavior of surrounding drivers by the presence of more trucks on some highway segments during nighttime. Similar phenomena on reducing the crash rate with the presence of trucks were also observed in some existing studies with pooled (all-data) models (Anastasopoulos et al., 2008; Shankar et al., 1997). But different phenomena between nighttime and daytime models as demonstrated above have not been reported in any existing study.

2.4.2 Environmental Characteristics

Visibility, defined as the shortest distance that an object can be perceived against sky, is found to be statistically significant for both nighttime and daytime models. Nevertheless, the effects of visibility towards crash rate during daytime and nighttime are very different. At daytime, a unit increase in visibility will increase crash rate by 0.6392 and the probability of having a crash rate above zero by 0.1966%. At nighttime, it will decrease crash rate by 2.9418 and lower the likelihood of having a crash rate above zero by 0.2244%. By developing two separate models for daytime and nighttime with more refined temporal and spatial scales, some new findings become possible. It seems obvious that better visibility can considerably contribute to lower crash rate at night. However, it shows better visibility actually slightly increases the crash rate at daytime. Such a finding in daytime is different from several previous works (Usman et al., 2012, 2011, 2010; C. Xu et al., 2013; Yu et al., 2013b), in which visibility was found to be negatively related to crash frequency. Possible reason for this is that lower visibility (for instance, fog and snow) usually increases driver alertness leading to more cautious driving behavior, overweighting the influence stemming from extended reaction time caused by the reduced visibility. On the other hand, better visibility at daytime may potentially encourage more aggressive driving behavior. While at nighttime, the benefits from better visibility become more significant when people usually drive more cautiously at night than at daytime. This is the first time to discover the opposite effects from visibility on daytime and nighttime crash rates, thanks to developing two separate models in the present chapter. In fact, with opposite effects of visibility towards crash rate on daytime and nighttime models, it is not surprising to find out that visibility becomes insignificant in the estimated pooled random effect tobit model with all (mixed) data.

As an important environmental factor, road surface condition has long been believed to be closely related to traffic safety. Road surface conditions originally defined in CDOT database include Dry, Wet, Chemically Wet and so on. Among all these road surface indicators, wet road surface is found to be a significant factor for increasing crash rate during both daytime and nighttime periods. In fact, it has larger influence on crash rate as compared to most of the other contributing factors in the model. As shown in Table 2.4, wet road surface results in an increase of crash rate by 1.1524 and higher probability of having a crash rate above zero by 0.3345% during daytime. For nighttime, it becomes an increase of crash rate by 2.2261 and chance of having a crash rate above zero by 0.1698%. These findings are consistent with common sense and also similar to some of the existing studies (Bertness, 1980; Eisenberg, 2004; Keay and Simmonds, 2005), in which rainfall was found associated with more crashes. It is worth mentioning that precipitation status is not found significant in both daytime and nighttime models. As discussed earlier, precipitation may have lagged effect on traffic safety when time scale gets smaller. The results of the present chapter somehow support the assumption and also suggest that wet road surface may be preferred over rainfall as a contributing factor in crash rate prediction models with refined-scale data.

Snow occurring is found to be significantly associated with lower crash rate during nighttime period. According to the results of marginal effects, the occurrence of snow causes crash rate to decrease by 1.92 and lowers the probability of having a crash rate above zero by 0.1464% during nighttime. Similar to other adverse environmental factors, such as reduced visibility, snow can make a driver harder to safely operate the vehicle, but at the same time, also alert people to drive more cautiously. Like some others factors discussed above, the final outcome from the opposite effects on both vehicle operation and driving behavior is not straightforward, usually requiring specific analysis.

2.4.3 Temporal Characteristics

Among the temporal characteristics, both November and weekend are found to affect the daytime crash rate. November tends to sustain 0.3695 more on the crash rate and 0.1137% higher probability of having a crash rate above zero than the rest months of the year. This may be associated with unobserved effects caused by sudden temperature drop and early storm in Colorado during November of 2010. Other month indicators have also been tested and were found not significant. Weekends typically experience 0.2434 more daytime crash rate and 0.0749% higher probability of having a daytime crash rate above zero, as compared to weekdays. No significant change in crash rate is discovered for November and weekend during nighttime period.

2.4.4 Roadway Characteristics

In terms of roadway characteristics, the number of entering ramp per lane per mile is found to be a significant factor (with a negative sign) for vehicle crash rate in the daytime model. Table 2.4 shows that one unit increase in the number of entering ramp per lane per mile results in a decrease of crash rate by 1.4675 and a 0.451% less probability of having a crash rate greater than zero. This finding is different from some studies (e. g. Anastasopoulos and Mannering, 2009; Anastasopoulos et al., 2008), which showed positive correlation between the number of ramps per lane per mile and crash rate. Nevertheless, the finding from this chapter is consistent with the findings by Pei et al. (2012). It is possible that there are some unobserved effects or site-specific characteristics leading to different findings. It is believed that more studies in this regard are needed and hopefully more insightful findings can be made with improved statistical tools and/or datasets. During nighttime, the present model suggests that the number of entering ramp per lane per mile plays an insignificant role in crash rate.

It is shown in Table 2.3 that roadway segment length is important to traffic crash rate in the daytime model. Table 2.4 shows that every mile increment in the roadway segment length increases the number of crashes per million VMT by 0.4998 and the probability of having a crash rate above zero by 0.1537%. This finding is in line with some of the previous ones (Anastasopoulos and Mannering, 2009), which have indicated that longer the roadway segment is, a higher crash rate for the roadway segment will occur. While in some other studies (e.g. Anastasopoulos et al., 2008), roadway segment length was found to have no effect on the crash rate. It is possible that roadway segment length has some unobserved complex effects toward crash rate, which require further analysis to uncover fully.

The curvature of the road is found to have significant potential for increasing crash rate in both models. A unit increase in curvature is associated with 0.5935 more on crash rate during daytime and 1.1982 more during nighttime. A similar pattern is also found in existing studies (Abdel-Aty and Radwan, 2000; Hosseinpour et al., 2014) and it may be attributable to limited sight distance and increased vehicle maneuver difficulty on horizontal curves (Hosseinpour et al., 2014).

Inside shoulder width also plays a role in the nighttime crash rate model. The positive signs in Table 2.3, and Table 2.4 indicate that wider inside shoulder is associated with higher crash rate during night period, while Abdel-Aty and Radwan (2000) found that the increase in shoulder width reduces crash frequency. Shoulder width is known to be another contributing factor, which may cause opposite effects on both driving conditions and also driving behavior, leading to the composite outcome. For example, Venkataraman et al., (2014, 2013) found that larger shoulder width initially improves traffic safety performance, but exhibits opposite effect after exceeding a certain value.

As for pavement condition, different indicators are tried in the model, including International Roughness Index, pavement condition indicator, and ruti index. The remaining service life of rut is gauged by the ruti index in CDOT database, with 100 ruti value for .15 inches or less rut depth and the value of 50 indicates no more remaining service life (corresponds to .55 inches or higher average rut depth). The long remaining service life of rutting indicator is an instrumental variable defined as 1 when the value of ruti is greater than 99 or 0 otherwise. The threshold of 99 is determined based on best model fit. Amongst these measures of pavement condition, the long remaining service life of rutting indicator is found to be statistically significant. Table 2.4 shows that long remaining service life of rutting results in an increment of 0.8215 and 1.2301 for crash rate during daytime and nighttime, respectively. This phenomenon may be attributabl to driver-environment interaction, which implies that drivers are inclined to drive more attentively and slowly on roadway segments with deeper rut (i.e. shorter remaining service life) to maintain a certain level of riding quality.

Some other geometric variables have also been tested in the models and were found not significant including international roughness index, pavement condition, median type, surface type, outside shoulder width, the number of lanes, the number of leaving ramp per lane per mile, grade, curve and so on. One possible reason why these variables were not significant is that this segment of I-25 on the flat plateau has a small variance of road design and has no sharp curve or steep grade.

2.6 Summary

This chapter reports so far the first research attempt on developing random effects tobit models for both daytime and nighttime crash rates based on disaggregate modeling approach with panel data in refined temporal and spatial scales. 1-year accident data from I-25 in Colorado and detailed data of traffic condition, environmental condition, road geometry and road surface conditions were processed into hourly basis on 1-mile roadway segments on average to demonstrate the proposed approach. Some interesting observations were made, and the significance of the present chapter is summarized as follows:

(1) By adopting tobit model with random effects, not only the censoring effects of crash rates data can be accounted for, but the unobserved heterogeneity across observations can also be potentially captured;

(2) Comprehensive road geometry, real-time traffic, weather and road surface data in refined temporal and spatial scales (hourly record and 1-mile roadway segments on average) was integrated into the crash rate model development with panel data structure.

(3) The utilization of panel data in both refined temporal and spatial scales has great potential for capturing both spatially varying and time-varying nature of variables (e.g. hourly traffic volume, visibility, wet road surface, etc.), which was usually ignored in traditional traffic crash modeling through data aggregation.

(4) In addition to the refined scales, for the first time, crash rate was studied with two separate models developed for nighttime and daytime, which led to some new findings and more refined information than traditional pooled data model. The results showed that there was a major difference in contributing factors towards crash rate between daytime and nighttime, implying the considerable needs to consider daytime and nighttime crashes separately when refined-scale data (e.g. hourly) is studied.

(5) Although it was only demonstrated on one portion of the I-25 highway, the proposed approach can be easily applied to other highways in the United States and the rest of the world. After further studies on many similar highways with refined scales are conducted, improved understanding of contributing factors for traffic crashes can be reached. As a result, more efficient and adaptive mitigation efforts may become possible to save more people's lives from crashes. These efforts include improvements in vehicle design, highway design, traffic management or law enforcement based on these new findings. Along with this line, some future studies can be carried out such as risk-based optimal route selection for adverse driving conditions, active law enforcement/traffic control intervention and advanced resource allocation and planning for the trucking industry, etc.

CHAPTER 3 MODELING CRASH RATES FOR A MOUNTAINOUS HIGHWAY USING REFINED SCALE PANEL DATA[2]

3.1 Introduction

The modeling of crash frequency and crash rate has been the research focus of traffic safety analysis over past several decades. Various models have been developed to predict the crash frequency and identify hazardous factors that affect auto-vehicle safety (see (Lord and Mannering, 2010; Mannering and Bhat, 2014) for complete reviews on these methodological approaches). However, most of these existing literature focus on studying traffic accident counts that are usually highly aggregated over a long time period (for example a year, a month). As pointed out by Lord and Mannering (2010), the adoption of aggregated data ignores time-varying nature of some critical factors thus lead to important information loss and introduce error in model estimation. Researchers have long recognized that weather and traffic conditions are critical factors in crash occurrence analyses which can vary considerably over time and space. For example, surrogate binary variables were developed to study the safety effect of pavement surface conditions in several studies (Caliendo et al., 2007; Miaou and Lord, 2003). Abdel-Aty and Pemmanaboina (2006) developed a crash prediction model with real-time traffic flow data and archived weather data. Yu et al. (2013b) studied crash occurrence using real-time data and concluded that weather condition variables play a vital role in causing an accident. Usman et al. (2012, 2011, 2010) created a road surface index to measure road surface condition and incorporated a variety of other weather and traffic-related variables to investigate accident frequency. For mountainous highways, weather and other environment-related variables (e.g. visibility, road surface condition, etc.) become even more critical due to complex temporal- and spatial-varying nature and interaction with mountainous terrain (long and steep slopes, sharp curves). Therefore, to appropriately model crash safety risks on mountainous highways and disclose the inherent crash mechanism, refined-scale models are often desired. Although large bodies of

---

[2] This chapter is developed based on a published research paper by Ma et al. (2015a) with the permission from the publisher Transportation Research Board.

literature were contributed to study relative crash risk using real-time data (Abdel-Aty et al., 2004; Abdel-Aty and Pemmanaboina, 2006; Yu et al., 2013b), existing studies on traditional traffic safety predictions with refined scales, however, remain very limited.

Refined-scale modeling on crash counts often encounters some technical challenges, such as strong correlation and excessive zeroes. As repeated observations are generated from the same roadway section across time, temporal serial correlations are presumably to occur, which can get even stronger with more refined time scales. If left unaccounted for, the presence of temporal serial-correlations can lead to a violation of independence assumption on error terms, causing biased model estimations and erroneous inferences. To deal with the correlation challenge, some existing research efforts on handling serial correlations of panel data (time series cross-sectional data) offer some helpful experiences. For example, Ulfarsson and Shankar (2003) adopted the negative multinomial model to study median crossover accidents in a panel format. Lord and Persaud (2000) applied general estimating equation model to handle temporal correlation that present over year-to-year accident data. Chin and Quddus (2003) developed a random effect negative binomial model for yearly panel accident data at signalized intersections. However, these studies were still highly aggregated in time domain (data are aggregated over a year). Recently, Qi et al. (2007) developed random effects ordered probit model using panel accident data by dividing data into the different time period (weekday peak hour period, weekday off-peak hour period and weekend period).

As an important alternative to crash frequency, crash rate (measured as the number of accidents per million vehicle miles traveled) modeling with refined scales can be promising due to its popularity on traffic safety performance assessment by different stakeholders. By considering crash rate modeling, the excessive-zero problem of crash count modeling converts to left-censoring effects of dependent variables. Anastasopoulos et al. (2008) showed that tobit model is a good choice to model accident rates because of its capability of handling left-censoring data. Anastasopoulos et al. (2012a) further developed random parameter tobit model on aggregated crash data, and demonstrated the strength of random parameter tobit model to account for unobserved heterogeneity. Chen et al. (2014) adopted a random effects tobit model approach to study daytime and nighttime crash rate using panel data. Other studies using tobit model include those investigating endogeneity problem (X. Xu et al., 2013) or multivariate crash data (Anastasopoulos et al., 2012b). Similar to crash frequency studies, crash rate modeling in refined temporal or spatial domains has rarely been reported.

Interstate highway I-70 in Colorado is well known for its typical mountains terrain, critical role to local and national traffic, and inclement weather. The objective of this chapter is to examine crash rates for a portion of I-70 by developing an advanced random parameter tobit model with panel data in refined temporal scale. Specifically, refined-scale weather and traffic data in a panel formation are incorporated to accommodate varying nature of complex driving conditions on I-70. Demonstrated with crash data from a typical mountainous highway, this chapter considers comprehensive weather, road surface, and traffic data along with complex geometric features. Unlike traditional traffic safety modeling, this chapter forms a longitudinal panel data format by dividing data into repeated hourly records for each roadway segment. In the meantime, the study takes advantage of the strength offered by random parameter tobit model to handle serial-correlations (roadway segment-specific correlations) across time, unobserved heterogeneity and censoring effects of the crash rate data. There are two major contributions of this chapter. First, this is so far the first reported effort on integrating random parameter tobit model and refined-scale panel data to develop crash rate data models. In addition to the strength of handling unobserved heterogeneity explicitly like some recent studies (Anastasopoulos et al., 2012a; Chen et al., 2011; Dinu and Veeraragavan, 2011; El-Basyouny and Sayed, 2009b), random parameter model is adopted in the present chapter to account for serial correlations across observations within panel data for the first time. The other contribution of this chapter is the adoption of refined scale hourly-based weather and traffic data in the crash rate modeling on mountainous highways. With these time-varying factors (visibility, road surface condition, hourly traffic volume, etc.) being incorporated in refined scales, more reliable and insightful observations can be made than existing aggregated-scale studies. As a result, time-varying characteristics of some factors can be retained in the model, and their complex effects on crash rate can be further disclosed.

3.2 Data Description

In this chapter, we focus on a selected portion of a mountainous highway I-70 in Colorado. The selected portion has a total length of 56.06 miles, starting from Mile Marker 195.26 to Mile Marker 251.32. Homogeneous roadway segments are defined based on the principle that no distinct road design variance (e.g. change of lane width, shoulder type, median type, speed limit, pavement condition) exists within any single segment. The study section was divided into 100 homogeneous roadway segments with an average length around 1.08 miles, including 52 eastbound and also 48 westbound segments. The 56.06-mile study section features typical mountainous terrain with steep slopes and sharp curves. Moreover, this part of I-70 is well known for its susceptibility to inclement and fast changing weather conditions, such as snow, rain, and wind etc. The typical mountainous terrain makes detailed weather information even more critical factors in influencing road safety than many other time-invariant ones. Four types of accident-related data from this 56.06-mile section are included in this chapter: (1) traffic accident data; (2) highway geometry data (including pavement conditions); (3) real-time weather and road surface condition data; (4) real-time traffic data. By combining data from different sources in refined scale, we can perform a more insightful and comprehensive study on hazardous factors.

The traffic accident data, which was provided by Colorado State Patrol (CSP), ranges from January 2010 to December 2010. To get panel data in refined temporal scale, the authors processed accident data into hourly records for each segment according to the occurrence time of each accident. The roadway geometry data were collected from Roadway Characteristics Inventory (RCI) of Colorado Department of Transportation (CDOT). Detailed roadway design features and pavement characteristics are available in this dataset, including speed limit, segment length, number of lanes, lane width, deflection of horizontal curve, horizontal curvature, vertical grade, shoulder width, shoulder types, median types, median width, international roughness index, rutting depth, etc.

In addition to these traditional data which are often used to predict accident occurrence, real-time traffic and environmental related data provided by Road Weather Information System (RWIS) are also incorporated into this chapter. There are 24 traffic stations along the I-70 corridor being studied, providing real-time monitoring data of traffic speed and volume. The real-time traffic data were initially recorded every 2 minutes and were further processed into the hourly record to facilitate the following study. A total of 7 weather stations is installed along the study area, providing motorists with real-time weather condition such as visibility, precipitation and road surface status. The real-time weather/surface condition data were recorded in 20-minute intervals. To generate hourly record for road surface condition, road surface condition variables are determined by the governing surface status within each hour. For example, an hourly surface condition is defined as wet road surface if two or more wet road surface conditions were recorded within that hour. Roadway segments are assigned with real-time data from the closest weather and traffic stations.

The hourly crash rate (also referred to as accident rate) was determined as:

$$Crash\ rate_{it} = \frac{Crash_{it}}{htv_{it} \times length_i / 1{,}000{,}000} \tag{3.1}$$

where $Crash\ rate_{it}$ is the number of accidents per million vehicle miles travelled (VMT) on segment i during t^{th} hour of the year, $Crash_{it}$ is the number of accidents happening on segment i during t^{th} hour of the year, htv_{it} is the hourly traffic volume of roadway segment i during t^{th} hour of the year, and $length_i$ is the length of segment i.

A censored accident rate data (crash rate equals to zero) is generated when there is no accident on a given roadway segment during a given time period (in this case an hour). As a result of possible sensor malfunctions, sometimes real-time data records leave “empty” windows (i.e. no data for one or several sensors at some time). After discarding those missing values, a total of 643,322 records were generated with 480 uncensored ones (non-zero) and 642,842 censored ones (zero). A descriptive statistics are shown in Table 3.1.

Table 3.1 Descriptive Statistics of Explanatory Variables

Variable	Mean	Standard deviation	Minimum	Maximum
Crash frequency	0.0008	0.0321	0	10
Accident rate (number of accidents per million VMT)	2.0477	150.24	0	47954
Roadway characteristics				

High speed limit indicator (1 if speed limit is greater than or equal to 60 miles per hour, 0 otherwise)	0.8619	0.3450	0	1
Vertical grade	-0.0632	3.5193	-6.24	6.24
Deflection angle of curve	26.87	18.830	0	84.488
Inside shoulder width indicator (1 if inside shoulder width is larger than 5 feet, 0 otherwise)	0.1535	0.3604	0	1
Pavement characteristics				
The indexed value of the international roughness index(lower values equal rougher roads)	93.892	5.2736	80	100
Traffic characteristics				
Speed gap (measured as the difference between speed limit and corresponding average traffic speed)	3.5865	6.3055	0[a]	57.751
Temporal characteristics				
nighttime indicator (1 if it is during nighttime, 0 otherwise)	0.4293	0.4950	0	1
Weather/surface characteristics				
Visibility (in miles)	1.0154	0.2279	0	1.1
Wet road surface (1 if road surface is wet, 0 otherwise)	0.1128	0.3164	0	1
Chemical wet road surface (1 if road surface is chemically wet, 0 otherwise)	0.0573	0.2325	0	1
Icy warning road surface (1 if road surface is in freeze condition, 0 otherwise)	0.0954	0.2938	0	1

[a] If the actual average speed exceeds the local speed limit, the CDOT database will truncate it to the speed limit of the road segment. So the minimum value of "speed gap" is 0 here instead of negative values.

3.3 Methodology

Repeated hourly observations are generated by each unit (roadway segment or intersection, in this case, roadway segment) to form panel data. This type of data differs substantially from traditional cross-sectional data due to the presence of serial correlations, thus requiring different model settings. In order to develop a random parameter tobit model for panel data that can account for serial correlations, the study firstly starts with a base tobit model. The tobit model (also referred to as censored regression model) was first proposed by Tobin (1958) to handle data with a left censored or right censored dependent variable. We treat the panel accident rate data simply as pooled ones first. Thus, a typical left-censored (with a lower threshold at zero) tobit model is developed as a baseline model in the present chapter by adding the time dimension:

$$Y_{it}^{*} = \boldsymbol{\beta_i X}_{it} + \varepsilon_{it},\ \ i = 1, \dots, N, t = 1, \dots, T_i \tag{3.2}$$

and

$$Y_{it} = Y_{it}^{*} \ \text{ if } \ Y_{it}^{*} > 0 \tag{3.3a}$$

$$Y_{it} = 0 \ \text{ if } \ Y_{it}^{*} \leq 0 \tag{3.3b}$$

where N is the number of units (in this case, the number of roadway segments), and T_i is the number of the repeated observations for roadway segment i. Y_{it}^{*} is the underlying latent variable which is observed only when being positive, and Y_{it} is the dependent variable (number of accidents per million miles travelled per hour). $\boldsymbol{X}_{it}$ is a vector of explanatory variables (e.g. traffic condition, geometric characteristic, temporal characteristic weather condition, surface condition, etc.), $\boldsymbol{\beta_i}$ is a vector of estimable parameters, and ε_{it} is the error term which is normally and independently distributed with mean zero and variance σ^2. The resulting likelihood function for the tobit model is given as (Anastasopoulos et al., 2008):

$$L = \prod\nolimits_0 \left[1 - \Phi\left(\frac{\beta X}{\sigma}\right)\right] \prod\nolimits_1 \frac{1}{\sigma} \Phi[(Y_{it} - \beta X)/\sigma] \tag{3.4}$$

We account for unit-specific effects (roadway segment-specific effects) by allowing parameter sets associated with a given roadway segment to vary randomly. Such a practice renders a proper assumption that explanatory variables have different effects on different roadway segments, which offers ample room to capture heterogeneity across roadway segments and address serial-correlations within a certain roadway segment. For a given roadway segment i, the random parameter tobit specification is defined by allowing β_i to vary across units:

$$\beta_i = \beta + \varphi_i\ ,\ \ i = 1,2, \dots, N \tag{3.5}$$

where φ_i is randomly distributed (such as normally distributed with mean zero and variance σ_{φ}^2). One key assumption at this point is that the T_i observations for a given segment i are independent conditioned on φ_i. Therefore, under above construction, these T_i observations in segment i are actually correlated and jointly distributed through φ_i, which is capable of accounting for the temporal correlations across observations within segment i. The joint conditional density function for roadway segment i is shown as follows:

$$f\left(y_{i1}, y_{i2}, \dots, y_{iT_i}\middle|\varphi_i\right) = \prod_{t=1}^{T_i} f(y_{it}|\varphi_i) \tag{3.6}$$

where $f(y_{it}|\varphi_i) = f(y_{it}|\varphi_i, Y_{it}^* > 0) \times f(y_{it}|\varphi_i, Y_{it}^* \leq 0)$. The corresponding log likelihood function is then formed by integrating φ_i out of the conditional density function (Greene, 2012):

$$LL = \sum_{\forall i} ln \int_{\varphi_i} f\left(y_{i1}, y_{i2}, \dots, y_{iT_i}\middle|\varphi_i\right) g\,(\varphi_i) d\varphi_i \tag{3.7}$$

where $g(\varphi_i)$ is the probability density function of φ_i. Both approximation by Hermite quadrature and simulation-based maximum likelihood approach can be adopted to handle computational difficulties that brought by random parameter tobit model. In the present chapter, we adopt adaptive Gaussian-Hermite quadrature method to estimate random parameter tobit model.

3.4 Model Results

To compare random parameter tobit model with its fixed parameter counterpart, a likelihood ratio test is conducted to test the null hypothesis that the random parameter tobit model is statistically equivalent to the corresponding fixed parameter tobit model. The test statistic is given as (Washington et al., 2011):

$$\chi^2 = -2[LL_f\left(\beta^f\right) - LL_{rp}(\beta^{rp})] \tag{3.8}$$

where $LL_f\left(\beta^f\right)$ is the log-likelihood when the fixed parameter tobit model is converged, and $LL_{rp}(\beta^{rp})$ is the log-likelihood when the random parameter tobit is converged. The test statistic follows a χ^2 distribution, with degrees of freedom equal to the difference in the numbers of parameters between the two tested models. As shown in table 3.2, the resulting χ^2 value of the likelihood ratio test is 47 with 1 degrees of freedom. This means that we are more than 99.9% confident that the random parameter tobit model statistically outperforms its fixed counterpart.

Table 3.2 Likelihood Ratio Test

	Random parameter model	Fixed parameter model
-2Log likelihood at convergence	15911	15958
Number of parameters	13	12
Degrees of freedom	1	
$\chi^2 = -2[LL_f\left(\beta^f\right) - LL_{rp}(\beta^{rp})]$	47	
Critical χ^2 (0.999 level of confidence)	10.83	

Table 3.3 shows the model estimation results for random parameter tobit model. Only variables that are significant at a 90% level of confidence are reported. A total of 11 explanatory variables is found to significantly affect accident rate, among which the traffic related and weather/surface related variables play significant roles. As for random parameters, one parameter is determined to be randomly distributed when both the mean and the variance of the parameter density are found to yield significant estimations. In the present study, only one variable, that is speed gap, is found to be statistically significant random parameter. More details about the findings listed in Table 3.3 will be discussed in the following by categories of variables.

Table 3.3 Random Parameter Tobit Model Estimation Results

Variable	coefficient estimation	t-statistic
Roadway characteristics		
High speed limit indicator (1 if speed limit is greater than or equal to 60 miles per hour, 0 otherwise)	2304.41	4.22**
Vertical grade	-89.8549	-1.78*
Deflection angle of curve	24.2409	2.46**
Inside shoulder width indicator (1 if inside shoulder width is larger than 5 feet, 0 otherwise)	1100.99	2.39**
Pavement characteristics		
The indexed value of the international roughness index(lower values equal rougher roads)	-403.5	-21.26**
Traffic characteristics		
Speed gap (measured as the difference between speed limit and corresponding average traffic speed) [a]	184.61	7.43**
standard deviation of parameter distribution-normal	132.79	6.45**
Temporal characteristics		
nighttime indicator (1 if it is in nighttime, 0 otherwise)	-3461.74	-9.08**
Weather/surface characteristics		
Visibility (in miles)	-1789.52	-3.31**
Wet road surface (1 if road surface is wet, 0 otherwise)	2743.81	6.10**
Chemical wet road surface (1 if road surface is chemically wet, 0 otherwise)	3872.48	7.15**
Icy warning road surface (1 if road surface is in freeze condition, 0 otherwise)	3350.97	6.74**
Model statistics		
Number of observations	643322	
$LL(0)$ (log-likelihood with nothing)	-450399.5	

$LL(C)$ (log-likelihood with constant)	-8267.5
$LL(\beta)$ (log-likelihood at convergence)	-7955.5
AIC (smaller is better)	15937
BIC (smaller is better)	15971

*: significant at 90% level of confidence
**: significant at 95% level of confidence
[a] The original real-time speed data from CDOT database will do not exceed the local speed limit for each road segment, So the "speed gap" in this chapter only has non-negative values, and it didn't reflect over-speeding behaviors.

3.4.1 Roadway Characteristics

The speed limit is a major policy-related variable to regulate driving speeds. As shown in Table 3.3, high speed limit indicator is positively related to the accident rate. It indicates that if the speed limit for a roadway segment is equal to or greater than 60 miles per hour, the accident rate is higher than those with lower speed limits. Similar to the present finding, Lee and Mannering (2002) found that higher speed limit (above 85km/h) increases crash frequency. Nonetheless, some other studies (Donnell et al., 2010; Lao et al., 2014) found the opposite, i.e. higher speed limit is associated with fewer accidents. Lee and Mannering (2002) argued that the speed limit is endogenous related to accident frequency. The disparities in findings of safety effects of speed limit underscore the need for conducting more research on this complex variable in the future.

Vertical grade, which ranges from -6.24 to 6.24 (negative grade means downgrade, and positive grade means upgrade), is found to be negatively associated with accident rate. This result indicates that for downgrade slope, steeper vertical grade causes higher accident rate. While for upgrade slopes, steeper vertical grade causes lower accident rate. These findings are consistent with the common knowledge about braking distances on different slopes.

In order to assess the effect that horizontal curve has towards crash occurrence, different measurements were used in previous literatures, such as the presence of curve (Anastasopoulos et al., 2012a), degree of curvature (Ahmed et al., 2011; Ma and Kockelman, 2006), curve length (Carson and Mannering, 2001; Ma and Kockelman, 2006) , or deflection of curve (Noland and Oh, 2004) etc. In the present chapter, all of these measurements were tried separately and also in different combinations.

However, only the deflection of curve angle is found to be significant. Specifically, deflection of curve angle is positively associated with accident rate, indicating the larger deflection of curve angle the higher accident rate will be. This finding is somewhat different from earlier studies (Noland and Oh, 2004), where larger deflection of curve angle was found to be associated with fewer accidents. It is known that road curvature is likely to have mixed overall safety influence (Wang et al., 2013a), and future analyses are clearly needed to study the safety effect of road curvature further.

In addition to abovementioned roadway characteristics, the inside shoulder width indicator (1 if inside shoulder is larger than 5 feet, 0 otherwise) also plays a significant role in the accident rate model. The positive sign of estimated coefficient for inside shoulder width indicator implies that roadway segments with wider inside shoulder result in higher accident rates. Note that Anastasopoulos and Mannering (2009) used the same inside shoulder width indicator. However, they found that this indicator has a mostly negative effect on crash frequency (in fact, they found it to be a random parameter). It is typically believed that modeling accident counts and modeling accident rates are inherently different (accident rate is exposure based measure while accident count is not). So such a difference in observations is plausible, and it necessities more specific investigations on accident rate studies.

3.4.2 Pavement Characteristics

As for pavement characteristics, considerations are given to important measures such as the international roughness index (IRI) and rutting depth. Nevertheless, only the IRI-related variable is found to produce a statistically significant result. A surrogate measure of IRI, the indexed value of the international roughness index (IRII) (lower values equal rougher roads), is used. The result is seemingly counterintuitive that higher IRII results in lower accident rate. However, it can be partly explained by risk compensation (Anastasopoulos et al., 2012a; Assum et al., 1999; Winston et al., 2006), which implies drivers may become more careful when they perceive hazardous conditions such as rough road surface.

3.4.3 Traffic Characteristics

Speed is known to be a major factor that affects accident rates. In the present chapter, we use speed gap instead of absolute speed value. Speed gap herein is calculated as the difference between posted speed limit and mean traffic speed. The original real-time speed data from CDOT database do not exceed the local speed limit for each road segment, which means that if the actual average speed exceeds the local speed limit, the CDOT database will truncate it to the speed limit of the road segment. As a result, the "speed gap" in this chapter only has non-negative values, and it can reflect traffic congestions but not over-speeding behaviors. As presented in Table 3.3, speed gap was found to generate a random parameter which is normally distributed with a mean 184.61 and standard deviation 132.79 as shown in Fig. 3.1. This result indicates that an increase in speed gap leads to an accident rate increase on 92.3% of the roadway segments, and an accident rate decreases on the other 7.7% of the roadway segments. It is known that larger speed gap often occurs when traffic gets congested. Therefore, this observation may partly reflect that more accidents are likely to happen on more congested roadway segments, which is overall consistent with several previous studies (Dias and Miska, 2009; Kononov et al., 2008).

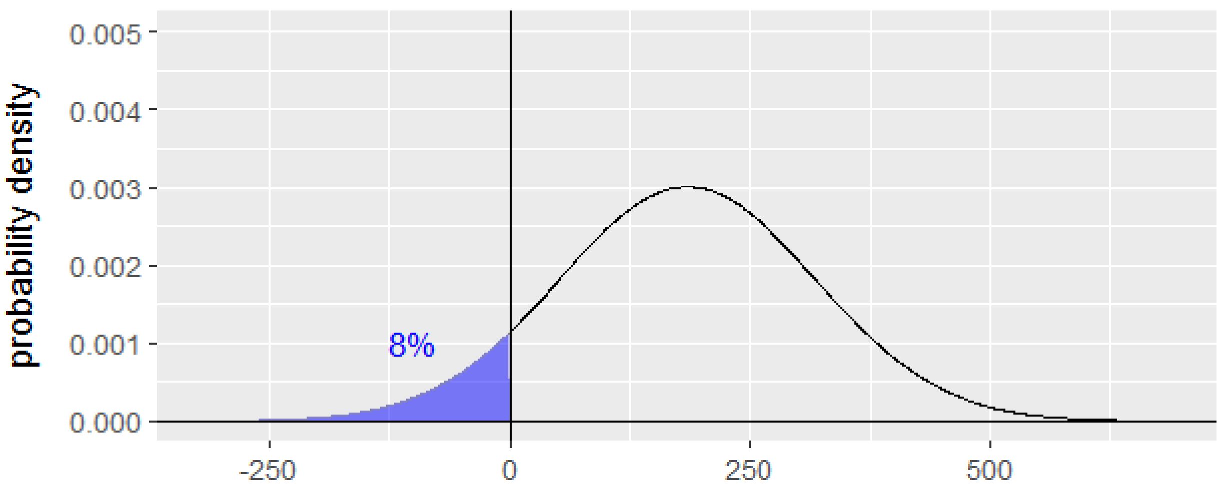

Figure 3.1 Distribution of Parameter Estimation for Speed Gap

3.4.4 Temporal Characteristics

When it comes to temporal characteristics, a variety of variables are tried during the model estimation process, including a weekend indicator, Monday indicator, Friday indicator and different month indicators. Only the nighttime indicator (1 if it is in the nighttime, 0 otherwise) is found to produce a significant result. Note that nighttime indicator is defined exactly by the 2010 Colorado Sunrise Sunset Calendar instead of using the same time period (e.g. 6 pm to 6 am) every day. Thus it can better capture nighttime driving condition. Table 3.3 shows that the estimated parameter of nighttime indicator has a negative sign, indicating that nighttime period is associated with less accident rate in contrast to daytime period.

3.4.5 Weather/Surface Characteristics

Visibility is found to be significant in the accident rate model with a negative sign, which means better visibility decreases accident rate. This observation is in line with some previous findings (Usman et al., 2012; C. Xu et al., 2013; Yu et al., 2013b), which also showed that poor visibility condition has the potential to cause more accidents.

In the CDOT database, road surface conditions were defined as dry, wet, chemically wet (moisture mixed with anti-icer), slush or ice warning (1 if the road surface is in freeze condition, 0 otherwise), etc. A variety of these indicators were tried in the model. However, only three of them are found to be significant in the estimated accident rate model. Specifically, wet road surface, chemically wet road surface and ice warning road surface are all positively associated with accident rates. The wet road surface is found to be associated with higher accident rate, which is in line with previous results. For example, Caliendo et al. (Caliendo et al., 2007) also found that wet pavement increases accidents in general. Also, chemically wet road surface and ice warning road surface both increase accident rates in contrast to the normal dry condition as expected. These phenomena, for one more time, highlight the increased risk of driving on these adverse road conditions on mountainous highways and the importance of providing timely maintenance to improve road surface conditions.

3.5 Summary

This chapter presented analyzed accident rates on mountainous highways using random parameter tobit model. By facilitating data in refined temporal scale (hourly basis), this chapter differs from previous accident rates studies in adopting disaggregated data. In addition to handling left-censoring effects and explicitly capturing unobserved heterogeneity, the random parameter tobit model can also properly deal with serial correlations that usually present in panel data. Moreover, by incorporating weather and traffic data in refined temporal scale (hourly basis) in our study, detailed phenomena were observed including those related to time-varying factors often being masked in traditional studies with aggregated data. The proposed model was developed based on real-time weather and traffic data from RWIS, a common data source around the country. Therefore, the technique developed in this chapter is easily transferrable to other highway safety studies with refined scales, exhibiting great engineering application potentials.

A typical mountainous interstate highway I-70 was selected to demonstrate the proposed methodology. A random parameter tobit model was estimated by combining different data into one comprehensive data set. The likelihood ratio test result showed the superiority of random parameter tobit model over its fixed parameter counterpart. Model results demonstrated that various factors related to traffic and weather/surface conditions, roadway geometry and pavement play significant roles in crash rates. Poor visibility is found to increase accident rate. Similarly, adverse road surface conditions, including wet road surface, chemically wet road surface and ice warning, were all found to increase accident rate. Traffic-related factor, speed gap, was also found to affect accident rate significantly. Some factors were also found to possibly have a mixed influence on accident rate, such as speed limit, and inside shoulder width indicator. In addition, speed gap was found to produce a random impact on accident rates in the model.

CHAPTER 4 CORRELATED RANDOM PARAMETER MARGINALIZED TWO-PART MODEL: AN APPLICATION TO REFINED-SCALE LONGITUDINAL CRASH RATES[3]

4.1 Introduction

With enormous economic costs and human casualties that motor-vehicle crashes continue to claim, researchers have been putting together a lot of efforts to investigate the relationships between crashes and its contributing factors using various statistical models (Mannering and Bhat, 2014). Crash frequency and rate are two primary safety measures associated with traffic crashes. As the major alternative to crash frequency, crash rate is an appealing standardized relative safety measure widely used in traffic safety studies. For both crash frequency and rate studies, most of the current literature adopt highly aggregated or averaged data over a particular time period, usually a year or even several years. By doing so, explanatory variables that vary significantly over time, such as traffic characteristic and weather information, have to be aggregated or averaged over a long time duration, leading to loss of valuable information during the predefined time period (Lord and Mannering, 2010; Washington et al., 2011). In an attempt to overcome this issue, some studies with more refined temporal and/or spatial scales have emerged to provide more insightful inferences about crash frequencies and rates (Chen et al., 2014; Ma et al., 2015; Usman et al., 2012, 2011, 2010; Qi et al., 2007). By incorporating real-time weather and traffic data obtained from Road Weather Information System (RWIS), the present chapter conducts a crash rates analysis in refined temporal scale, which can reveal the impacts of time-varying variables and provides more valuable practical guidance. As a result of adopting refined-scale data, however, multiple observations for the same roadway segment over different time periods may be generated which are correlated due to unobserved effects (Mannering and Bhat, 2014). In other words, longitudinal data will be generated instead of traditional cross-sectional one, which poses extra challenges in the modeling process.

---

[3] This chapter is developed based on a research paper by Xiaoxiang Ma, Suren Chen and Feng Chen, which was presented (TRB 16-3707) at the 95th Annual Meeting of Transportation Research Board in Washington, D.C., in January 2016.

To account for temporal and/or spatial correlations that are present within longitudinal data, researchers have developed various models such as negative multinomial models (Ulfarsson and Shankar, 2003), generalized estimating equation (GEE) models (Lord and Persaud, 2000; Mohammadi et al., 2014; Wang and Abdel-Aty, 2006), random effects models and random parameter models (Chen et al., 2014; Ma et al., 2015a; Qi et al., 2007). Among these models, random parameter models are attractive because they could not only address serial correlations, but also allow the parameters to vary across observations, avoiding potential erroneous inferences if parameters are not actually fixed (Lord and Mannering, 2010). However, most existing studies that utilized random parameter models were based on an independent assumption for random parameters. Recognizing the limitation of such an assumption, some recent studies allowed random parameters to be correlated by assuming a multivariate distribution with unrestricted variance-covariance matrix (Xiong and Mannering, 2013; Yu et al., 2015). Another challenge is that crash rates are non-negative continuous values with a point mass at zero, and modeling such data with standard linear regression leads to inconsistent and biased estimates (Washington et al., 2011). To tackle such a problem, tobit models with fixed parameters were proposed to deal with crash rates data from interstate highways in Indiana (Anastasopoulos et al., 2008). To avoid biased estimates and erroneous inferences in the presence of unobserved heterogeneity (Washington et al., 2011) for fixed parameter tobit model, Anastasopoulos et al. (2012) further proposed a random parameter tobit model. Random parameter tobit model allows parameters to vary across observations, exhibiting improved performance than its fixed parameter counterpart. Recently, Chen et al. (2014) and Ma et al. (2015) applied random effects/random parameter tobit models to address longitudinal crash rates data in refined temporal scales. Yu et al. (2015) demonstrated that correlated random parameter tobit model outperforms its uncorrelated random parameter counterpart.

Despite successful applications of tobit models on traffic safety research, some potential issues may arise when tobit model is applied to investigate crash rates. The first problem pertains to whether any censoring has actually occurred in the data. By definition crash rates apparently cannot take negative values. Therefore, the zeros in crash rate data are de facto self-representing data values (either due to no crashes occurring or underreporting), originated from underlying data generating process instead of data censoring. The tobit model was designed to address biases introduced by data censoring, thus it is theoretically applicable only to the situation when the dependent variable can be negative but is somehow censored at zero (Maddala and Lahiri, 1992; Sigelman and Zeng, 1999). Existing studies suggested that when tobit is applied to data with zeros originated from data generating process rather than censoring, it may result in biased parameter estimates (Belasco and Ghosh, 2012; Sigelman and Zeng, 1999). Another issue is the normality assumption of the latent response in the tobit model (Anastasopoulos et al., 2008). In practice, such a normality assumption barely holds, especially when data is highly disaggregated. It not only makes tobit model inflexible, but also may leads to possible model biases when normality assumption is violated for the data with a high proportion of zeros (Arabmazar and Schmidt, 1982; Belasco and Ghosh, 2012; Bera et al., 1984).

In the present chapter, thanks to the Road Weather Information System (RWIS), real-time traffic and weather data can be adopted, which will be further processed into refined time intervals (daily observations). Refined-scale data adopted in the present chapter is highly disaggregated in temporal scale as opposed to traditional safety studies, exhibiting two distinct features: (1) substantial proportion of zero observations (over 91%); and (2) right skewed positive outcomes. These two features of the present data make the tobit model, if developed, even more susceptible to the issues as discussed above. Therefore, more appropriate modeling methodology is desired to address the data with unique characteristics in the present chapter. As an alternative to address the clumping of zero values, two-part model has been widely applied in the field of econometrics, medical studies, ecological studies, etc. (Cragg, 1971; Duan et al., 1983; Liu et al., 2012; Sigelman and Zeng, 1999; Smith et al., 2014; Su et al., 2009). Basically, two-part model uses one equation to determine whether the outcome is positive, and a second one to determine the level of the outcome when it’s positive. One key feature that distinguishes two-part model from tobit model is that two-part model treats zero values as self-representing observations, instead of proximity of zeros, missing or negative values caused by left censoring. With possible potentials to tackle the aforementioned challenges associated with existing modeling of refined data, random parameters and crash rates, two-part approach seems to be a very promising and natural choice. As compared to the popular crash rates modeling methodologies, two-part model exhibits its theoretical appeal in a way that it has the potential to avoid possible biased estimates resulting from inappropriate use of tobit model under some situations, as discussed previously. However, one limitation that prevent traditional two-part model from being directly applied to study crash rates is that it is usually difficult to get a straightforward interpretation of covariates' impact on crash rates. In order to facilitate crash rates study by providing more interpretable covariate effects, marginalized two-part model proposed by Smith et al. (2014), rather than traditional two-part model, is adopted as the baseline model. To the best of authors′ knowledge, study on crash rates using any type of two-part model has not been reported so far.

This chapter attempts to demonstrate applicability of marginalized two-part model as an alternative to tobit model to study crash rates, while investigating the impact of time-varying variables on crash occurrence at the same time. Random parameter model is developed to address longitudinal nature of the data due to adopting refined temporal scale (daily). Moreover, to handle the inappropriate independent assumption of random parameters, a correlated random parameter model is developed to capture possible correlations between random parameters. As a summary, a correlated random parameter marginalized two-part model (CRPMTP) is designed to study crash rates with a refined-scale panel structure. The proposed methodology is demonstrated by investigating crash rates on an urban freeway in Colorado. The applicability and potential advantages of marginalized two-part model as an alternative tool to study crash rates by addressing those outlined methodological challenges associated with time-varying variables, temporal correlations and fixed parameters are also explored.

4.2 Data Preparation

In order to more comprehensively investigate the relationships between crash rates and its contributing factors, a representative portion of one major freeway I-25 in Colorado with detailed traffic and weather data is selected for this chapter. The selected portion of I-25 is located between City of Northglenn and City of Castle Rock. Both the southbound and the northbound are included in the current study: the 28.55-mile northbound ranges from Mile Marker (MM) 188.49 to MM 221.03, and the 27.38-

mile southbound ranges from MM 188.49 to MM 219.86. To capture the time-varying nature of some explanatory variables, real-time weather and traffic data are acquired through Road Weather Information System (RWIS) and traffic flow monitoring stations respectively.

Four types of datasets are utilized in the present chapter, (1) one year crash data (Jan. 2010- Dec. 2010) provided by Colorado State Patrol (CSP), (2) roadway geometry (including pavement conditions) from the Roadway Characteristics Inventory (RCI), (3) real-time weather and road surface condition data provided by RWIS, (4) real-time traffic data recorded by traffic flow monitoring stations along the selected highways. The latter two datasets are the key components that make this disaggregate analysis possible. The real-time weather and road surface condition data were recorded at a twenty-minute interval by the weather stations installed along the highway. Road surface condition types originally defined in the CDOT database include dry, wet, chemically wet (moisture mixed with anti-icer), ice warning, ice watch, etc. Precipitation status initially defined in the CDOT database include no precipitation, rain occurring, snow occurring and others. Because road surface condition and precipitation status often vary within a day especially under the adverse weather, percentages of road surface condition and precipitation status are defined instead of indicator variables. The real-time traffic data were recorded at a two-minute interval, which is aggregated into daily records for the present chapter. By combining data from different sources and processing into refined-scale, a more insightful and thorough study of potential contributing factors that may lead to crashes can be conducted.

The chosen highway section is split into homogeneous roadway segments through two-stage data segmentation regime. First, the roadway segments are defined according to CDOT traffic station assignment. Then the segments are further divided into homogeneous roadway segments based on the variation of geometric features, such as longitudinal grade, deflect angle of curve, curve radius, the number of lanes, median type, shoulder type, pavement condition, speed limit, etc. A total of 57 homogeneous roadway segments is obtained with 29 segments from the northbound and 28 ones from the southbound. The segments have an average length of 1 mile. Since traffic and weather stations do not always work as expected, there may be “empty” windows in the original data due to sensor malfunction or service disruption. After removing these incomplete data records, a total of 17,776 observations is acquired in the final dataset, with 91.44% of which has zero accidents.

Table 4.1 Summary of Descriptive Statistics

Variable	Mean	Std Dev.	Minimum	Maximum
Dependent variable				
Crash rate	1.741	8.347	0	231.216
Crash frequency	0.099	0.348	0	6
Independent variable				
Speed gap (measured as the difference between speed limit and corresponding mean traffic speed)[a]	2.585	2.652	0	31.796
Maximum crosswind speed (largest cross wind speed of the day) [b]	7.341	5.739	0	34.502
Weekend indicator (1 if during weekend, 0 otherwise)	0.275	0.447	0	1
Curvature (degree of horizontal curve)	0.971	0.690	0	2.260
November indicator (1 if during November, 0 otherwise)	0.090	0.286	0	1
Percentage of snow occurring (ratio of snow occurring regarding precipitation status)	0.0276	0.103	0	0.775
Number of enter ramp per lane per mile (measured as number of enter ramp/lanes/roadway segment length)	0.259	0.216	0	0.926
Outside shoulder length (in feet)	10.340	2.190	6	15
Poor pavement indicator (1 if the pavement condition for the primary direction is good, 0 otherwise)	0.356	0.479	0	1

[a]:If the actual average speed exceeds the local speed limit, the CDOT database will truncate it to the speed limit. So the minimum value of "speed gap" is 0 here instead of negative values.
[b] Crosswind is the perpendicular component of wind to the direction of travel.

The daily crash rates are calculated with the following equation:

$$Crash\ rate_{it} = \frac{Crash_{it}}{DTV_{it} \times len_i/1{,}000{,}000} \tag{4.1}$$

where Crash rate_{it} is the number of crashes per million vehicle miles traveled (VMT); Crash_{it} is the number of crashes happened on segment i during day t; DTV_{it} is the daily traffic volume; and len_i is the length of roadway segment. The subscript i denotes i-th roadway segment, and subscript t denotes t-th day of the year. During the study period, a total of 1,761 crashes are recorded on the selected portion of Interstate I25.

Multi-collinearity issue is investigated to circumvent the inclusion of highly correlated explanatory variables. Groups of variables that are possibly correlated are identified at first. For example, poor pavement indicator, the indexed value of international roughness index and the life for rutting are all variables that measure pavement conditions. Special attentions are given to these variables with possible collinearity regarding variable selections. Different variables within these groups are tested separately in the model, and the inclusion of one variable is based on the log-likelihood value. The largest variance inflation factor (VIF) value for the variables included in the final model is 1.75, indicating no inclusion of highly correlated variables. Table 4.1 presents the summary statistics for the variables used in the dataset.

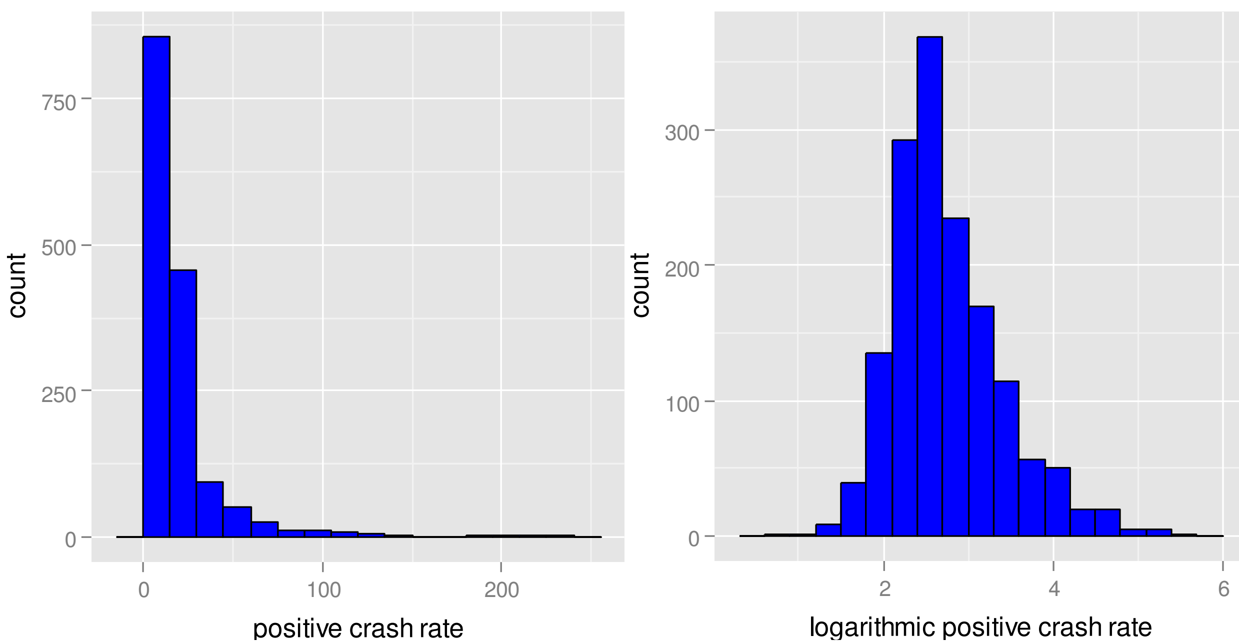

Figure 4.1 Histogram of Positive Crash Rates and Logarithmic Positive Crash Rates

Figure 4.1 provides histograms of positive crash rates and logarithmic of positive crash rates respectively. The left sub-graph shows that the distribution of positive crash rates is right-skewed with a heavy tail. Such phenomenon is echoed by the right sub-graph in Figure 4.1 where a log transformed positive crash rates data shows normality, which indicates that the normal assumption of positive crash rates may be inappropriate.

4.3 Methodology

Based on the work by Tooze et al. (2002) on correlated random effects model and the work by Smith et al. (2014) on the cross-sectional data model, the authors extend the marginalized two-part model to accommodate longitudinal data with correlated random parameters in this section. The proposed model relates the two components of the model by assuming a multivariate normal distribution for the random parameters. Firstly, the binary part of the model is introduced. For a random variable Y_{it}, which represents crash rates with observed value y_{it} for a roadway segment i at time t, R_{it} is defined as:

$$R_{it} = \begin{cases} 0, & if\ Y_{it} = 0 \\ 1, & if\ Y_{it} > 0 \end{cases} \tag{4.2}$$

with the probabilities:

$$\Pr(R_{it} = r_{it}) = \begin{cases} 1 - p_{it}, & if\ r_{it} = 0 \\ p_{it}, & if\ r_{it} > 0 \end{cases} \tag{4.3}$$

Same as conventional two-part models, the binary part of the proposed model is usually modeled using a logistic regression:

$$logit(p_{it}) = \alpha_1 X_{1it} + \beta_1 Z_{1it} \tag{4.4}$$

where X_{1it} is the vector of explanatory variables with fixed parameter vector α_1, and Z_{1it} is the vector of explanatory variables with random parameter vector β_1.

Secondly, for the continuous part of the model, a conventional two-part model links the explanatory variables as:

$$E(ln\ Y_{it} |\ Y_{it} > 0) = X\beta \tag{4.5}$$

In this case, the estimated coefficient β can be interpreted as the effect of a unit increase in the corresponding covariate on the conditional mean of $ln\ Y_{it}$ given that Y_{it} is greater than zero. With regard to crash rates study, such an interpretation from Eq. (4.5) refers to the impact of the variable on crash rates for those segments with crashes. However, traffic safety practitioners would be more interested in the impact of the variable on crash rates for all segments (both with and without crashes). In order to obtain a better understanding of the overall population-level effects (segments with and without crashes) of variables, the following link proposed by Smith et al. (2014) for marginalized two part model is employed:

$$E(Y_{it}) = \exp(X\beta) \tag{4.6}$$

In the case where the log-normal distribution is the density function of Y_{it} with a mean of μ_{it} and a variance of σ_2 on the log scale:

$$E(Y_{it}) = \exp(\alpha_2 X_{2it} + \beta_2 Z_{2it}) = p_{it}\exp(\mu_{it} + \sigma^2/2) \tag{4.7}$$

where X_{2it} is the vector of explanatory variables with fixed parameter vector α_2, and Z_{2it} is the vector of explanatory variables with random parameter vector β_2. The random parameters are allowed to be correlated by assuming multivariate normal distribution for vector $\boldsymbol{\beta}$ with a mean of $\boldsymbol{\mu}$ and a variance-covariance matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \end{bmatrix} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{4.8a}$$

with

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{bmatrix}, \text{ and } \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{\mathbf{12}} & \cdots & \sigma_{\mathbf{1},\boldsymbol{m-1}} & \sigma_{\mathbf{1},\boldsymbol{m}} \\ \sigma_{\mathbf{21}} & \sigma_2^2 & \cdots & \sigma_{\mathbf{2},\boldsymbol{m-1}} & \sigma_{\mathbf{2},\boldsymbol{m}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_{\boldsymbol{m-1},\mathbf{1}} & \sigma_{\boldsymbol{m-1},\mathbf{2}} & \cdots & \sigma_{m-1}^2 & \sigma_{\boldsymbol{m-1},\boldsymbol{m}} \\ \sigma_{\boldsymbol{m},\mathbf{1}} & \sigma_{\boldsymbol{m},\mathbf{2}} & \cdots & \sigma_{\boldsymbol{m},\boldsymbol{m-1}} & \sigma_m^2 \end{bmatrix} \tag{4.8b}$$

For correlated random parameter model, the variance-covariance matrix $\boldsymbol{\Sigma}$ is not structured, which means the off-diagonal elements of $\boldsymbol{\Sigma}$ are not restricted to zero. Under this formulation, both the within part (within the binary part or the continuous part) and cross part (between the binary part and the continuous part) correlations can be accounted for. When all the off-diagonal elements of $\boldsymbol{\Sigma}$ are restricted to zeros, the model will be reduced to uncorrelated random parameter model.

Based on the above model parameterization, the resulting likelihood function is shown in the following:

$$L(\alpha_1, \beta_1, \alpha_2, \beta_2, \sigma, \boldsymbol{\mu}, \boldsymbol{\Sigma} | y, X, Z)$$
$$= \prod_{i=1}^{N} \iint \prod_{t=1}^{T_i} (1 - p_{it})^{1(y_{it}=0)} \left\{ \frac{p_{it}}{y_{it}\sqrt{2\pi\sigma}} exp\left[-\frac{1}{2\sigma^2} (lny_{it} - \mu_{it})^2 \right] \right\}^{1(y_{it}>0)} f(\beta_1, \beta_2 | \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\boldsymbol{\beta} \tag{4.9}$$

where

$$p_{it} = \frac{\exp(\alpha_1 X_{1it} + \beta_1 Z_{1it})}{1 + \exp(\alpha_1 X_{1it} + \beta_1 Z_{1it})} \tag{4.10}$$

and

$$\mu_{it} = \alpha_2 X_{2it} + \beta_2 Z_{2it} - lnp_{it} - \sigma^2/2 \tag{4.11}$$

N is the number of roadway segments, and T_i is the number of hourly observations on roadway segment i. $f(\beta_1, \beta_2 | \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the probability density function of vector $\boldsymbol{\beta}$, which follows a multivariate normal distribution. Such a likelihood function is analytically intractable. However, maximization of such a likelihood function can be achieved by using quasi-Newton optimization which approximates a likelihood function by adaptive Gaussian quadrature (Pinheiro and Bates, 1995). The lognormal distribution is adopted because it is an appropriate choice for the data in the present chapter. Note that other distributions, such as gamma, Weibull and log-skewed-normal distributions, can also be used to model skewed, nonnegative, and continuous data when appropriate. This feature makes two-part model much more flexible than tobit model.

4.4 Model Comparisons

4.4.1 Comparing Nested Models

Three marginalized two-part models are estimated for the present chapter, including fixed parameter marginalized two-part model (FPMTP), uncorrelated random parameter marginalized two-part model (UCRPMTP) where all the off-diagonal elements in the variance-covariance matrix of random parameters are zeros, and correlated random parameter marginalized two-part model CRPMTP model where all the off-diagonal elements in the variance-covariance matrix of random parameters are not restricted to zeros. Given that these models are nested with each other, we conduct likelihood ratio tests to determine if there are any statistical differences between FPMTP, UCRPMTP and CRPMTP models using the test statistic (Washington et al., 2011):

$$\mathrm{X}^2 = -2[LL(\beta_{Full}) - LL(\beta_{Reduced})] \tag{4.12}$$

where $LL(\beta_{Full})$ is the log-likelihood at convergence of the full model (it is UCRPMTP between FPMTP and UCRPMTP, and CRPMTP between UCRPMTP and CRPMTP), $LL(\beta_{Reduced})$ is the log-likelihood at convergence of the reduced model (it is FPMTP between FPMTP and UCRPMTP, and UCRPMTP between UCRPMTP and CRPMTP). The test statistic is again χ^2 distributed with the degree of freedom equivalent to the difference of the numbers of parameters between the full model and reduced model.

In comparing UCRPMTP and FPMTP models, the χ^2 value is 1095 with four degrees (four important random parameters) of freedom, and the corresponding P-value is less than 0.001. When comparing between CRPMTP and UCRPMTP models, the χ^2 value is 76 with three degrees (three significant covariance terms) of freedom, and the corresponding P-value is less than 0.001. Consequently, it can be concluded that UCRPMTP model is statistically superior to FPMTP, and CRPMTP model is statistically superior to UCRPMTP. Therefore, CRPMTP is concluded to be the best amongst the three candidate models concerning the goodness of model fit, which reveals that CRPMTP can not only account for unobserved heterogeneity among observations but also address possible correlations between those unobserved heterogeneities.

4.4.2 Comparing Non-nested Models

A corresponding tobit model is also developed, and the estimation result is not reported for the sake of space. To compare the performance between tobit model and FPMTP model, log-likelihood value, Akaike information criterion (AIC) and mean absolute deviance (MAD) are studied with the results shown in Table 4.2.

Table 4.2 Measures of Model Goodness of Fit

	Log-likelihood	AIC	MAD
FPMTP	-10287.5	20598	3.00
tobit	-11040	22101	3.07

Although a direct comparison of log-likelihood value and AIC suggests FPMTP is to be favored according to Kullback-Leibler divergence, MAD does not show substantial evidence that FPMTP should be preferred. To further test whether FPMTP is better than tobit model, a hypothesis test proposed by Vuong (1989) for non-nested models is conducted. To test equivalence of two parametric classes of models F_{θ_*} and G_{γ_*}, a null hypothesis is constructed as:

$$H_0: E^0 \log \frac{F_{\theta_*}}{G_{\gamma_*}} = 0 \tag{4.13}$$

where E^0 is the expectation taken at the true data generating process, θ_* and γ_* are the corresponding pseudo true values. According to Vuong (1989), under H_0:

$$n^{-1/2} \frac{LR_n(\hat{\theta}, \hat{\gamma})}{\widehat{\omega}_n} \longrightarrow N(0,1) \tag{4.14}$$

where $LR_n(\hat{\theta}, \hat{\gamma})$ is the difference in maximum log-likelihood values, n is the number of observations, and

$$\widehat{\omega}_n = \frac{1}{n} \sum (\log \frac{F_{\hat{\theta}}}{G_{\hat{\gamma}}})^2 - (\frac{1}{n} \sum \log \frac{F_{\hat{\theta}}}{G_{\hat{\gamma}}})^2 \tag{4.15}$$

When testing against the hypothesis that FPMTP is superior to tobit model, the test statistic is calculated to be 80.65 (critical value at 0.999 level of confidence is 3.09). Therefore, it can be concluded that FPMTP performs better than its corresponding tobit model.

4.5 Model Estimation Results

The CRPMTP models developed in the present chapter can be termed as a logistic-lognormal-normal model, where ‘logistic’ refers to binary part (part I) of the model, ‘lognormal’ to the continuous part (part II) of the model, and ‘normal’ to the multivariate normal distribution of random parameters. Other types of CRPMTP models can also be developed. For instance, logistic-‘log-skew-normal’-normal models are also estimated for the same dataset. However, the estimate of the skewness parameter in the logistic-‘log-skew-normal’-normal FPMTP model is found insignificant, and the likelihood ratio test indicates there is no difference between logistic-lognormal-normal FPMTP and logistic-‘log-skew-normal'-normal FPMTP. The estimation results for logistic-lognormal-normal FPMTP, UCRPMTP, and CRPMTP models are presented in Table 4.3, Table 4.4 and Table 4.5 respectively.

4.5.1 Fixed-Parameter Model Results

Thanks to the real-time data acquired from RWIS and the adoption of refined temporal scale in this chapter, time-varying variables are obtained and tested in addition to geometric features and temporal characteristics. Amongst all these traffic/weather/road surface related variables, speed gap, maximum crosswind speed and percentage of snow occurring are found to have a significant impact on crash occurrence.

Table 4.3 Fixed-Parameter Estimation Results

Parameter	Estimate	Standard Error	P-value
Part I: Pr(Y>0)			
Intercept	-3.6430	0.0820	<0.0001
Speed gap (measured as the difference between speed limit and corresponding mean traffic speed)	0.1637	0.0081	<0.0001
Maximum cross wind speed (largest cross wind speed of the day)	-0.0193	0.0048	<0.0001
Weekend indicator (1 if during weekend, 0 otherwise)	-0.1361	0.0369	0.0002
November indicator (1 if during November, 0 otherwise)	0.1095	0.0473	0.0206
Curvature	0.7837	0.0431	<0.0001
Poor pavement indicator (1 if the pavement condition for the primary direction is good, 0 otherwise)	0.1510	0.0394	<0.0001
Part II: E(Y)			
Intercept	1.0360	0.1031	<0.0001
Speed gap (measured as the difference between speed limit and corresponding mean traffic speed)	0.1111	0.0072	<0.0001
Maximum crosswind speed (largest cross wind speed of the day)	-0.0119	0.0049	0.0147
Percentage of snow occurring (ratio of snow occurring regarding precipitation status)	0.7097	0.1161	<0.0001
Number of enter ramp per lane per mile (measured as number of enter ramp/lane/roadway segment length)	-0.3745	0.0818	<0.0001
Curvature	0.3809	0.0427	<0.0001
Outside shoulder width (in feet)	-0.1117	0.0066	<0.0001

Sigma (σ)	0.5538	0.0101	<0.0001
Model statistics			
Number of observations		17776	
-2Log-likelihood		20575	
AIC		20605	
BIC		20722	

Moving to the specific estimation results for FPMTP model, there are six variables found significant in the binary part of the model. Speed gap is significant with a positive sign, indicating higher speed gap leads to higher probability of crashes. Maximum crosswind speed is found to be significant with a negative sign, which means a larger maximum cross wind speed of the day will decrease the occurrence of crashes. This result seems counter-intuitive. However, per risk compensation hypothesis, drivers tend to alter their driving behavior when experiencing perceived changes in riskiness (Dulisse, 1997). Given that wind within the study area is relatively moderate, it is possible that the presence of moderate cross wind triggers drivers' alertness and leads to overall fewer crashes. As for temporal characteristics, it is found that weekends are less likely to have accidents as opposed to weekdays. This result is consistent with crash statistics from National Highway Traffic Safety Administration's crash report (NHTSA, 2010), where weekends show fewer accidents than weekdays. Another temporal characteristic, November, is found to be associated with higher likelihood of crashes. This is likely because of a sudden temperature drop and early winter storms in Colorado during November 2010, which coincides with the hypothesis that the onset of a major snow event (especially the first of the season) has a significant impact on the occurrence of crashes (El-Basyouny et al., 2014). For segments with curvature, it is found that those segments are more likely to incur crashes. Some previous studies also indicated (Chen et al., 2014; Ma et al., 2008; Ma and Kockelman, 2006) that degree of curvature was inclined to increase crash frequency for crashes with different severity levels. Concerning pavement condition, poor pavement indicator is found to increase the crash probability.

As for the continuous part of FPMTP model, speed gap, maximum crosswind speed, the percentage of snow occurring, the number of entering ramp per lane per mile, curvature, and outside shoulder width are found to be significant. The results in Table 4.3 indicates that larger speed gap is associated with higher crash rates, which is consistent with our previous studies (Chen et al., 2014; Ma et al., 2015a). On the contrary, higher maximum cross wind speed is associated with lower crash rates. As discuss above, such a counter-intuitive result can also be attributed to risk compensation theory. With regard to other weather-related variables, serious multicollinearity is detected between visibility, the percentage of wet road surface, the percentage of chemically wet road surface, the percentage of rain occurring and percentage of snow occurring. Among these variables, the percentage of snow occurring is included in the final model as it produces best model fit. It is found that an increase in the percentage of snow occurring leads to an increase in crash rates as expected. Regarding roadway characteristics, the number of entering ramp per lane per mile is found to decrease crash rates. Similar trends were also found in several studies (Chen et al., 2014; Pei et al., 2012). In addition, it is estimated that higher curvature leads to higher crash rates. Moreover, Table 4.3 also shows that outside shoulder width plays an important role in traffic crash rates. It is found that increase outside shoulder width leads to decrease in crash rates. Previous studies (Anastasopoulos et al., 2008) also indicated that outside shoulder width is negatively related to crash rates.

4.5.2 Uncorrelated Random Parameter Model Results

Given the longitudinal nature (repeated observations on the same roadway segment) of the dataset, it is reasonable to adopt random parameter model that accommodates panel setting to capture possible temporal correlations and unobserved heterogeneities. For that matter, UCRPMTP model is estimated where random parameters are assumed to be independently and normally distributed, and the results are shown in Table 4.4. Regarding random parameter formulation, a random parameter is determined only when both the mean and the standard deviation of the parameter density are statistically significant (Anastasopoulos et al., 2012a). When a parameter is determined to be a random parameter, its effect varies across roadway segments. Four parameters are found to be random parameters: Speed gap and curvature in the binary part, speed gap, and percentage of snow occurring in the continuous part. Interpretation of these random parameters will be given in the next section. Besides likelihood ratio test, log-likelihood at convergence, AIC, and BIC values also show that UCRPMTP outperforms FPMTP model.

Table 4.4 Uncorrelated Random Parameter Estimation Results

Parameter	Estimate	Standard Error	P-value
Part I : Pr(Y>0)			
Intercept	-3.6018	0.1071	<0.0001
Speed gap (measured as the difference between speed limit and corresponding mean traffic speed)	0.1874	0.0153	<0.0001
Standard deviation	0.0701	0.0124	<0.0001
Maximum cross wind speed (largest cross wind speed of the day)	-0.0112	0.0052	<0.0355
Weekend indicator (1 if during weekend, 0 otherwise)	-0.1736	0.0295	<0.0001
November indicator (1 if during November, 0 otherwise)	0.0615	0.0359	0.0923
Curvature	0.4505	0.1365	0.0017
Standard deviation	0.7403	0.0961	<0.0001
Poor pavement indicator (1 if the pavement condition for the primary direction is good, 0 otherwise)	0.2577	0.0941	0.0084
Part II: E(Y)			
Intercept	0.8730	0.1913	<0.0001
Speed gap (measured as the difference between speed limit and corresponding mean traffic speed)	0.1607	0.0148	<0.0001
Standard deviation	0.0730	0.0123	<0.0001
Maximum cross wind speed (largest cross wind speed of the day)	-0.0132	0.0048	0.0082
Percentage of snow occurring (ratio of snow occurring in terms of precipitation status)	0.6825	0.1146	<0.0001
Standard deviation	0.4160	0.1312	0.0025
Number of enter ramp per lane per mile (measured as number of enter ramp/lane/roadway segment length)	-0.4563	0.1727	0.0109
Curvature	0.3688	0.0561	<0.0001
Outside shoulder width (in feet)	-0.1074	0.0149	<0.0001
Sigma (σ)	0.3719	0.0072	<0.0001
Model statistics			
Number of observations		17776	
-2Log-likelihood		19480	
AIC		19518	
BIC		19556	

4.5.3 Correlated Random Parameter Model Results

Although UCRPMTP has the potential to capture temporal correlation and unobserved heterogeneities in the data, it imposes an independence assumption on random parameters' distributions. Such an assumption may be dubious as there may be correlations between random parameters, and it may result in biased estimations if possible correlations between random parameters are not properly accounted for (Conway and Kniesner, 1991). The author relaxes the independence assumption on random parameters by employing a multivariate normal distribution for the random parameters, leading to a CRPMTP model. A covariance term in the variance-covariance matrix is determined only when it produces a statistically significant estimate. Otherwise, it is set to zero. Table 4.5 provides the estimation results for CRPMTP model.

Table 4.5 Correlated Random Parameter Estimation Results

Parameter	Estimate	Standard Error	P-value
Part I : Pr(Y>0)			
Intercept	-3.6640	0.1089	<0.0001
Speed gap (measured as the difference between speed limit and corresponding mean traffic speed)	0.2026	0.0221	<0.0001
Standard deviation	0.1310	0.0184	<0.0001
Maximum cross wind speed (largest cross wind speed of the day)	-0.0109	0.0053	<0.0440
Weekend indicator (1 if during weekend, 0 otherwise)	-0.1852	0.0293	<0.0001
November indicator (1 if during November, 0 otherwise)	0.0603	0.0356	0.0958
Curvature	0.5078	0.1359	0.0005
Standard deviation	0.7578	0.0971	<0.0001
Poor pavement indicator (1 if the pavement condition for the primary direction is good, 0 otherwise)	0.2967	0.0833	0.0008
Part II: E(Y)			
Intercept	0.9485	0.2072	<0.0001
Speed gap (measured as the difference between speed limit and corresponding mean traffic speed)	0.1670	0.0207	<0.0001
Standard deviation	0.1253	0.0170	<0.0001
Maximum crosswind speed (largest cross wind speed of the day)	-0.0128	0.0049	0.0116
Percentage of snow occurring (ratio of	0.6891	0.1169	<0.0001

snow occurring regarding precipitation status)			
Standard deviation	0.4400	0.1338	0.0018
Number of entering ramp per lane per mile (measured as number of entering ramp/lane/roadway segment length)	-0.4714	0.1823	0.0126
Curvature	0.3567	0.0634	<0.0001
Outside shoulder width (in feet)	-0.1147	0.0165	<0.0001
Sigma (σ)	0.3708	0.0072	<0.0001
Model statistics			
Number of observations		17776	
-2Log-likelihood		19404	
AIC		19448	
BIC		19492	

Variance covariance matrix of the random parameters

	Speed gap (part I)	Curvature (Part I)	Speed gap (Part II)	Percentage of snow occurring (Part II)
Speed gap (part I)	0.0172			
Curvature (part I)	-0.0405	0.5743		
Speed gap (part II)	0.0160	-0.0397	0.0157	
Percentage of snow occurring (part II)	0	0	0	0.1936

Model estimation results from CRPMTP are quite consistent with those from UCRPMTP. One notable difference is that the estimated standard deviations of the random parameters are noticeably larger in CRPMTP, indicating CRPMTP captures more unobserved heterogeneities as compared to UCRPMTP. With regard to random parameters in the model, speed gap in the binary part is normally distributed with a mean of 0.2026 and a standard deviation of 0.131. This indicates that an increase in speed gap increases crash probability on 94% roadway segments while decreases crash probability on 6% roadway segments. Curvature in the binary part is normally distributed with a mean of 0.5078 and a standard deviation of 0.7578, which means large curvature increase crash probability on 75% roadway segments while decrease crash probability on 25% roadway segments. Speed gap in the continuous part is also normally distributed with a mean of 0.167 and a standard deviation of 0.1253, indicating an increase in speed gap increase crash rates on 91% roadway segments and decrease crash rates on 9% roadway segments. Percentage of snow occurring in the continuous part is normally distributed with a mean of 0.6891 and a standard deviation of 0.44, suggesting increasing snow precipitation increases crash rates on 94% roadway segments while decreases crash rates on 6% roadway segments.

Table 4.6 Correlation Matrix of Random Parameters

	Speed gap (part I)	Curvature (Part I)	Speed gap (Part II)	Percentage of snow occurring (Part II)
Speed gap (part I)	1			
Curvature (part I)	-0.407	1		
Speed gap (part II)	0.974	-0.418	1	
Percentage of snow occurring (part II)	0	0	0	1

Apart from random parameters, correlations between those random parameters are also revealed by developing CRPMTP model. The correlation matrix for the random parameters is shown in Table 4.6. Speed gap parameter in the binary part is positively correlated with speed gap parameter in the continuous part. Curvature parameter in the binary part is negatively correlated with speed gap parameters in both the binary part and the continuous part. As a result of adopting CRPMTP model, these findings can be finally revealed and improve the understanding of crash occurrence. In addition to the likelihood ratio test discussed in section 4.4.1, AIC and BIC values also provide evidence that CRPMTP should be favored over the other two competing models.

It is worth mentioning that the real-time traffic and real-time weather-related variables are found to play important roles in crash occurrence. If it were not for the adoptions of refined scale analysis with real-time data, some important findings as summarized above would not otherwise be made possible.

4.6 Summary

Traffic crashes are greatly affected by time-varying variables, such as weather and traffic conditions. It is, therefore, desirable to develop crash occurrence model with those time-varying variables. This chapter studied crash rates using refined-scale longitudinal data with excess zeros by developing the correlated random parameter two-part (CRPMTP) model. One-year crash data with detailed real-time traffic and weather related data on one major highway in Colorado were investigated in this chapter. The novelty of the proposed methodology is reflected from following aspects:

1) The marginalized two-part model was adopted for the first time to study crash rates. By comparing model performance between marginalized two-part model and tobit model, it shows that marginalized two-part model outperforms tobit model;

2) A correlated random parameter model, as opposed to the uncorrelated one employed in most existing literature, was developed to avoid the inappropriate independence assumption on random parameters. Likelihood ratio test along with AIC and BIC measures indicated correlated random parameter model is superior to corresponding uncorrelated one. This finding supports our hypothesis that correlated random parameter model could account for both the unobserved heterogeneities across roadway segments and the correlations between those unobserved heterogeneities;

3) Through employing real-time data from RWIS, a refined temporal scale analysis with daily data was conducted. By doing do, the impacts of time-varying variables were revealed, which would otherwise be unavailable if yearly crash data were used like most existing studies.

In order to adequately address the associated challenges, FPMTP model was first developed as a baseline model. Then, UCRPMTP model and CRPMTP model were developed to account for temporal correlations and unobserved heterogeneities. Likelihood ratio tests showed that the CRPMTP model was the best among the three models regarding goodness of fit. It was also found that tobit model was not the preferred model choice in this chapter. By facilitating a multivariate normal distribution of the random parameters, the CRPMTP model not only accounts for unobserved heterogeneity but also captures the correlations between random parameters. This chapter demonstrates that two-part model may be a better alternative to tobit model in analyzing crash rates when the data is right-skewed with a large proportion of zero values.

Moreover, speed gap in the binary part, curvature in the binary part, speed gap in the continuous part and percentage of snow occurring in the continuous part were found to have mixed effects on crash occurrence. Correlations between those random parameters were also revealed by adopting CRPMTP model. These finding can improve the understanding of the relationship between crash occurrence and contributing factors. Developing crash models that incorporate time-varying variables on a daily basis not only contributes to the improved understanding of the crash occurrence but also bears the potential to provide road users and policy makers with more detailed and relevant crash risk information. The present chapter also has some limitations: the data used only covers one highway over one year period, to provide more general insights of crash risks with refined temporal scale on the main highways, studies with data in longer time durations and also on different highways are desired in the future.

CHAPTER 5 EMPIRICAL ANALYSIS OF CRASH INJURY SEVERITY ON MOUNTAINOUS AND NON-MOUNTAINOUS INTERSTATE HIGHWAYS[4]

5.1 Introduction

Traffic safety on highways is a major concern to both transportation agencies and researchers (Christoforou et al., 2010; Yasmin and Eluru, 2013). In order to implement more effective and customized injury mitigation strategy, it is crucial to investigate injury severity and associated risk factors of crashes on a specific highway. Mountainous highways, where steep gradients and sharp curves are usually present, can cause considerably different driver behavior and vehicle performance as compared to non-mountainous counterparts. In addition to geometric complexness, mountainous highways are usually more susceptible to harsh weather conditions. Despite sharing a lot of similarities, such as traffic volume, driver population, vehicle composition, different highways in the same area may exhibit varying traffic injury risks with different contributing factors. This is especially so for those regions with both mountainous and non-mountainous highways where a uniformed traffic safety performance function across the region may not be sufficient.

Considerable research efforts have been made to analyze injury severity on typical highways during the past decades in terms of different categories of crashes, such as different vehicle types and crash types (Chang and Mannering, 1999; Mouyid Islam and Hernandez, 2013), different numbers of vehicles involved in crashes (Chen and Chen, 2011; Savolainen and Mannering, 2007; Xie et al., 2012), and different driver demographic and road surface conditions (Morgan and Mannering, 2011; Ulfarsson and Mannering, 2004). Based on these studies, transportation practitioners and researchers have gained good knowledge about crash severity on common non-mountainous highways. There are limited studies focusing on crash severity on mountainous highways. Yu and Abdel-Aty (2014a) examined crash injury severity for two roads, a mountainous freeway, and an urban expressway, using real-time traffic and weather data. In their subsequent study, Yu and Abdel-Aty (2014b) examined crash severity using three

[4] This chapter is modified from a published paper by Ma et al. (2015b).

different models for a mountainous freeway. These studies offered some good insights about severity based on separate investigations on non-mountainous and mountainous highways in different regions. However, little study has been reported about comparative research focusing on the unique contribution from mountainous nature on crash injury risk.

This chapter aims at investigating the injury severity characteristics on mountainous interstate highways through a comparative investigation. Two major interstate highways in Colorado, one being non-mountainous and the other one being mountainous are selected. Studying both mountainous and non-mountainous highways in the same region can offer some unique advantages on investigating the impacts specifically contributed by mountainous nature through excluding influences from many other factors including driver population and so on. In addition, different from some existing studies on mountainous highways (e.g. Yu and Abdel-Aty 2014a, 2014b), detailed police reported data is used in this chapter to consider as many critical factors as possible (Christoforou et al., 2010). Although being criticized of possibly suffering from underreporting (Savolainen et al., 2011), the police reported crash data is believed to provide more insights than non-crash-specific data, avoid small sample problems, and maintain statistical inferences from detailed crash data (Anastasopoulos and Mannering, 2011). With some new findings for the first time, the present chapter can provide better insights about contributing factors and associated mechanisms related to mountainous nature, which can add to the state-of-the-art of understanding injury severity risks and potential mitigation efforts. The findings of this chapter will provide scientific guidance to improve the current highway design and traffic management policy, and propose next-generation safety initiatives for mountainous highways in order to reduce the injury severity, life and financial losses caused by crashes.

5.2 Data and Empirical Setting

Colorado State Patrol (CSP) has detailed traffic crash data of Colorado highways, which contains crash, driver, vehicle, roadway design and environmental information. To study different characteristics of crash injury severity on mountainous and non-mountainous interstate highways, two major interstate highways which both cross Colorado are selected: I70 and I25. The Interstate I70 Mountain Corridor, ranging from Denver to Grand Junction through the Rocky Mountains, is a typical mountainous highway segment. Along the corridor, steep grades and sharp curves, accompanied by fast changing weather conditions, pose considerable safety threats on passing vehicles. I25 in Colorado goes through the Great Plains and shares a lot of similarities with other non-mountainous highways in the United States. Therefore, in the following comparative study, mountainous-highway (MT) crashes refer to those happening on I70 Mountain Corridor, and non-mountainous-highway (NM) crashes refer to those happening on I25 with plain terrain.

Four-year detailed crash data (2007-2010) on I70 Mountain Corridor and I25 are utilized in this chapter. After removing the crash records without crash location information, there are totally 16,057 crash data records during the 4-year time period in the CSP database, with 7,467 records on MT highway (I70 Mountain Corridor), and 8,590 records on NM highway (I25). The selected mountainous highway has a total length of 259.94 miles with an average 27859 AADT, while the selected non-mountainous highway has a total length of 298.879 miles with an average 69664 AADT. In the CSP database, the variable "Highest Inj Level" means the highest level of injury in a crash on a scale from 0 to 4, representing no injury, possible injury, non-incapacitating injury, incapacitating injury and fatal, respectively. To ensure that each category has a decent number of observations, they are regrouped into three categories (1) no injury (NI), (2) possible injury/non-incapacitating injury (PI/NII), and (3) incapacitating injury/fatal (II/F). To simplify the following presentation, incapacitating injury/fatal is referred to as severe injury, and possible injury/non-incapacitating injury is referred to as moderate injury. Among 7,467 crashes happening on MT highway, 5,739 (76.9%) crashes had no injury, 1,465 (19.6%) crashes had a moderate injury, and 263 (3.5%) crashes had a severe injury. Among 8,590 crashes happening on NM highway, 6,089 (70.9%) crashes had no injury, 2,194 (25.5%) crashes had a moderate injury, and 307 (3.6%) crashes had a severe injury.

Detailed crash characteristics in the CSP data are categorized into five groups: (1) roadway characteristics, (2) temporal and environmental characteristics, (3) driver characteristics, (4) crash characteristics, and (5) vehicle characteristics. In the remainder of this chapter the driver characteristics, vehicle characteristics and accident characteristics refer to the characteristics of at-fault driver or at fault vehicle. In order to limit potential estimation biases, all explanatory variables are carefully screened, and some are redefined. For example, snowy and icy road surface indicators are combined together, as it is possible that an crash reported to occur on snowy surface was actually caused by icy road surface (Morgan and Mannering, 2011). Additionally, the driver had no insurance indicator, and the driver had no proof of insurance indicator are combined. Tables 5.1 and 5.2 give the number of observations and the percentage distribution across the injury severity for MT and NM data sets, respectively.

Table 5.1 Descriptive Statistics for MT model

Variable	No injury		Moderate injury		Severe injury		Total
Roadway characteristics							
Wide median (median width>=50ft)	2005	74.23%	557	21.36%	119	4.41%	2701
No rut indicator (rut index=100)	967	72.22%	303	22.63%	69	5.15%	1339
Heavy traffic (AADT/number of lanes>=7.5k)	1579	78.79%	362	18.06%	63	3.14%	2004
Highway interchange	149	73.04%	42	20.59%	13	6.37%	204
Low truck percentage (<=4%)	122	68.93%	49	27.68%	6	3.39%	177
Temporal and Environmental characteristics							
Monday	856	74.69%	234	20.42%	56	4.89%	1146
Snowy/icy road surface	2364	79.95%	546	18.46%	47	1.59%	2957
Snow/sleet/hail	1896	80.72%	420	17.88%	33	1.40%	2349
Darkness-road lighted	274	77.62%	74	20.96%	5	1.42%	353
Driver characteristics							
Young driver (age<=25)	1898	78.11%	453	18.64%	79	3.25%	2430
Old driver (age>=60)	470	72.87%	140	21.71%	35	5.43%	645
Female driver	1580	72.88%	516	23.80%	72	3.32%	2168
DUI alcohol/drug use	145	51.06%	93	32.75%	46	16.20%	284
Driver was asleep	122	52.14%	82	35.04%	30	12.82%	234
Careless/reckless driving	463	59.06%	233	29.85%	87	11.10%	784
Driver was fatigued	49	58.33%	24	28.57%	11	13.10%	84
Driver had no insurance/no proof of insurance	586	63.01%	264	28.39%	80	8.60%	930
Crash characteristics							
Only one vehicle involved	3828	78.31%	880	18.00%	180	3.68%	4888
More than two vehicles involved	189	55.92%	128	37.87%	21	6.21%	338
Animal caused	766	91.19%	69	8.21%	5	0.60%	840
Exceeded legal speed	51	58.62%	25	28.74%	11	12.64%	87
Overturn	555	59.55%	286	30.69%	91	9.76%	932
Followed too closely	459	74.76%	149	24.27%	6	0.98%	614
Front to rear collision	806	69.90%	318	27.58%	29	2.52%	1153
Side to side collision with vehicles in same direction	540	87.10%	75	12.10%	5	0.81%	620

Collision with parked motor vehicle	45	76.27%	6	10.17%	8	13.56%	59
Collision with guard rail	611	76.86%	161	20.25%	23	2.89%	795
Collision with cable rail	107	91.45%	9	7.69%	1	0.85%	117
Collision with concrete highway barrier	650	79.46%	153	18.70%	15	1.83%	818
Collision with bridge structure	59	76.62%	18	23.38%	0	0.00%	77
Collision with vehicle debris or cargo	138	93.88%	8	5.44%	1	0.68%	147
Collision with embankment	337	78.19%	71	16.47%	23	5.34%	431
Collision with delineator post	172	67.19%	65	25.39%	19	7.42%	256
Vehicle characteristics							
Truck (10001 lbs or over)	460	81.27%	83	14.66%	23	4.06%	566
Passenger car/van	2398	77.28%	623	20.08%	82	2.64%	3103
Pickup truck/utility	1237	77.65%	304	19.08%	52	3.26%	1593
Pickup truck/utility with trailer	146	84.39%	24	13.87%	3	1.73%	173
SUV	1362	74.79%	381	20.92%	78	4.28%	1821
SUV with trailer	28	87.50%	2	6.25%	2	6.25%	32

Table 5.2 Descriptive Statistics for NM model

Variable	No injury		Moderate injury		Severe injury		Total
Roadway characteristics							
Wide median (median width>=50ft)	3010	71.91%	1005	24.01%	171	4.09%	4186
Deep rut indicator (rut index<=88)	2140	72.30%	714	24.12%	106	3.58%	2960
High speed limit (speed limit=75mph)	4762	70.42%	1736	25.67%	264	3.90%	6762
Depressed median	2828	69.42%	1066	26.17%	180	4.42%	4074
Heavy traffic (AADT/number of lanes>=7.5k)	4880	71.22%	1771	25.85%	201	2.93%	6852
Temporal and Environmental characteristics							
Monday	850	66.98%	347	27.34%	72	5.67%	1269
Snowy/icy road surface	1349	73.24%	436	23.67%	57	3.09%	1842
Wet road surface	546	69.20%	228	28.90%	15	1.90%	789
Driver characteristics							
Young driver (age<=25)	2269	72.38%	770	24.56%	96	3.06%	3135
Old driver (age>=60)	532	69.27%	201	26.17%	35	4.56%	768
Female driver	1956	67.99%	828	28.78%	93	3.23%	2877
DUI alcohol/drug use	209	50.61%	151	36.56%	53	12.83%	413
Driver was asleep	139	53.88%	98	37.98%	21	8.14%	258
Careless/reckless driving	867	60.88%	470	33.01%	87	6.11%	1424
Driver was fatigued	58	57.43%	34	33.66%	9	2.93%	101
Illness/medical	47	54.65%	33	38.37%	6	6.98%	86
Distracted by passenger	50	61.73%	24	29.63%	7	8.64%	81
Driver had no insurance/no proof of insurance	734	63.22%	348	29.97%	79	6.80%	1161
Driver's license had been denied	43	56.58%	25	32.89%	8	10.53%	76
Crash characteristics							
Only one vehicle involved	3056	71.99%	1003	23.63%	186	4.38%	4245
More than two vehicles involved	329	48.53%	320	47.20%	29	4.28%	678

Animal caused	479	88.54%	57	10.54%	5	0.92%	541
Exceeded legal speed	24	54.55%	10	22.73%	10	22.73%	44
Stopped in traffic	20	44.44%	22	48.89%	3	6.67%	45
Backing	40	95.24%	1	0.33%	1	0.33%	42
Overturn	427	47.13%	371	40.95%	108	11.92%	906
Followed too closely	1047	70.13%	431	28.87%	15	1.00%	1493
Improper passing	42	68.85%	13	21.31%	6	9.84%	61
Front to front collision	34	48.57%	25	35.71%	11	15.71%	70
Front to rear collision	1700	66.69%	805	31.58%	44	1.73%	2549
Front to side collision	189	58.33%	118	36.42%	17	5.25%	324
Side to side collision with vehicles in same direction	724	84.58%	118	13.79%	14	1.64%	856
Collision with guard rail	501	70.86%	180	25.46%	26	3.68%	707
Collision with cable rail	351	86.67%	49	12.10%	5	1.23%	405
Collision with concrete highway barrier	328	65.60%	163	32.60%	9	1.80%	500
Collision with vehicle debris or cargo	247	93.92%	15	5.70%	1	0.38%	263
Collision with embankment	99	70.21%	38	26.95%	4	2.84%	141
Vehicle characteristics							
Passenger car/van	3247	71.98%	1131	25.07%	133	2.95%	4511
Pickup truck/utility	1068	70.82%	388	25.73%	52	3.45%	1508
Pickup truck/utility with trailer	216	83.08%	39	15.00%	5	1.92%	260
SUV	886	65.63%	411	30.44%	53	3.93%	1350
Defective tires	57	62.64%	33	36.26%	1	1.10%	91

5.3 Methodology

In the present chapter, we focus on the differences of crash injury severities between crashes on mountainous and non-mountainous highways. Three crash injury severity outcomes are considered: severe injury (incapacitating injury/fatal); moderate injury (possible injury/non-incapacitating injury); and no injury. Over the years, researchers have adopted a variety of discrete outcome models to analyze crash-severity data, such as ordered logit or probit models, multinomial logit models, Markov switching multinomial logit models, nested logit models, random parameter logit (mixed logit) models, and latent class models (see Mannering and Bhat (2014), and Savolainen et al. (2011) for complete reviews on those methodological approaches). Among most frequently used methodological approaches, a mixed logit model is a well-developed approach which relaxes IIA assumption and IID errors assumption and allows for unobserved heterogeneity as compared to multinomial logit model (Jones and Hensher, 2007). Besides random parameter models (e.g. mixed logit model), latent class logit models have also gained popularity recently (Cerwick et al., 2014; Eluru et al., 2012). Differing from mixed logit models, latent class models can accommodate group specific unobserved heterogeneity and relax continuous distributional assumptions for random parameters. However, latent class logit models do not deal with individual unobserved heterogeneity as mixed logit models can. No consensus has yet been made regarding which approach is superior given the fact that both approaches have both strengths and limitations (Xiong and Mannering, 2013). Given that latent class model and random parameter models may not adequately accommodate the unobserved heterogeneity in some cases, Xiong and Mannering (2013) proposed a finite mixture (latent class) random parameter model to accommodate both the group specific heterogeneity and individual heterogeneity within each group, which results in a complex model structure. Following recent work (Chen and Chen, 2011; McFadden and Train, 2000; Milton et al., 2008), the current chapter adopts a mixed logit model approach.

Let $P_n(i)$ be the probability of crash n causing injury severity category I (Ulfarsson and Mannering, 2004):

$$P_n(i) = P\left(\beta_i X_{ni} + \varepsilon_{ni} \geq \beta_j X_{nj} + \varepsilon_{nj}\right) \forall j \in I, j \neq i \tag{5.1}$$

where I is a set of all possible discrete injury outcomes, i.e. severe injury, moderate injury, and no injury in the present chapter. β_i and β_j are the vectors of estimable coefficients corresponding to different injury severity alternatives i and j respectively, while X_{ni} and X_{nj} are the vectors of explanatory variables for crash n which determine the injury severity alternatives i and j respectively. ε_{ni} and ε_{nj} are error terms which are assumed to be generalized extreme value distributed (Mansky and McFadden, 1981). The mixed logit model formation is derived when parameter β_i is allowed to vary across observations as follows (Train, 2003):

$$P_n(i|\varphi) = \int \frac{e^{\beta_i X_{ni}}}{\sum_{\forall I} e^{\beta_j X_{nj}}} f(\beta_i|\varphi) d\beta_i \tag{5.2}$$

where $P_n(i|\varphi)$ is the probability of injury severity alternative i conditioned on $f(\beta_i|\varphi)$, and $f(\beta_i|\varphi)$ is the density function of β_i with a vector of parameters φ of the density function (mean and variance).

Simulation-based likelihood method is adopted to estimate mixed logit models using Halton sequence, which has been found to be a more efficient way of drawing values than purely random draws (Bhat, 2003; Train, 2003). Methods with 200 Halton draws are used in the forthcoming model estimations (Bhat, 2003; Gkritza and Mannering, 2008; Milton et al., 2008). With a sample size of 7,467 for mountainous highway crashes and a sample size of 8,590 for non-mountainous highway crashes, both datasets are much larger than the sample size requirements suggested by Ye and Lord (2014).

Elasticity is calculated to measure the effect of explanatory variables on injury severity probability. Since all the explanatory variables in this chapter are indicator variables, direct pseudo-elasticity is calculated to assess percent effect on severity probability $P_n(i|\varphi)$ when a particular indicator changes from 0 to 1 or reverse as follows (Ulfarsson and Mannering, 2004):

$$E_{X_{nk}}^{P_n(i)} = \left[e^{\beta_{ik}} \frac{\sum_{\forall\, i' \in I} [e^{\beta_{i'} x_n}]_{x_{nk=0}}}{\sum_{\forall\, i' \in I} [e^{\beta_{i'} x_n}]_{x_{nk=1}}} - 1 \right] \times 100 \tag{5.3}$$

where $E_{X_{nk}}^{P_n(i)}$ is the direct pseudo-elasticity of the k^{th} variable from the vector X_n for observation *n*. X_{nk} is the value of the variable *k* for the outcome *n*. β_{ik} is the k^{th} component of the vector β_i of injury-severity outcome *i*. $[e^{\beta_{i'} x_n}]_{x_{nk=0}}$ is the value of $e^{\beta_{i'} x_n}$ with $X_{nk} = 0$, and $[e^{\beta_{i'} x_n}]_{x_{nk=1}}$ is the value of $e^{\beta_{i'} x_n}$ with $X_{nk} = 1$. Average pseudo-elasticity is reported by taking average of the elasticity across all observations.

5.4 Model Comparison

To determine that separate developments of MT and NM models are statistically justified, a likelihood ratio test is performed. All data model is estimated using combined MT and NM datasets. The test statistic adopted in the likelihood ratio test is (Ulfarsson and Mannering, 2004):

$$\mathrm{X}^2 = -2[LL_N(\beta) - LL_{N_{mt}}(\beta^{mt}) - LL_{N_{nm}}(\beta^{nm})] \tag{5.4}$$

where $LL_{N(\beta)}$ is the log-likelihood at convergence of the all data model, with a parameter vector β. $LL_{N_{mt}}(\beta^{mt})$ and $LL_{N_{nm}}(\beta^{nm})$ are the log-likelihood at convergence of the model estimated on the MT data subset, and the NM data subset, respectively. The X^2 test statistic follows χ^2distribution with the degrees of freedom equal to the sum of the number of the parameters estimated in the MT and NM models minus the number of the parameters estimated in all dataset models. Based on the test result with P<0.001, we can conclude that the choice of modeling MT and NM crashes separately in the present chapter is warranted.

We also conduct a likelihood ratio test to compare the differences between the random parameter models (i.e. mixed logit models) and their fixed parameter counterparts (i.e. base multinomial models), using the test statistic (Washington et al., 2011):

$$X^2 = -2[LL(\beta_{random}) - LL(\beta_{fixed})] \tag{5.5}$$

where $LL(\beta_{random})$ and $LL(\beta_{fixed})$ are the log-likelihood at convergence of mixed logit model and fixed parameter model estimated using the same dataset (e.g. MT or NM dataset), respectively. The test statistic is χ^2 distributed with the degrees of freedom equal to the difference of the numbers of estimated parameters between the two models. The χ^2 value of the test is 16.08 with three degrees of freedom for MT model. The χ^2 value is 10.82 with two degrees of freedom for NM model. Thus, the corresponding P-value is 0.0045 for MT model and 0.001 for NM model respectively. Therefore, we are more than 99.5% confident that the mixed logit models are statistically superior.

5.5 Empirical Results

The estimated model results for MT and NM crashes are given in Tables 5.3 and 5.4 respectively. The results reveal substantial differences in contributing factors towards crash injury severity between MN and NM crashes. No injury outcome is chosen to be the base alternative among the three pre-defined injury outcomes. All estimated coefficients included in the MT and NM models are statistically significant at a 95% confidence level. Tables 5.3 and 5.4 show that both severity models have an overall good fit with McFadden pseudo-ρ^2 equal to 0.4886 for the MT model and 0.4253 for the NM model, respectively.

Table 5.3 MT Crash Injury Severity Model Estimation Results

Variable	**Injury outcome**	**Estimates**	**t-Statistics**
Constant	PI/NII	-0.683	-3.00
Constant	II/F	-1.615	-4.89

Roadway characteristics			
Wide median (median width>=50ft)	PI/NII	0.179	2.25
No rut indicator (rut index=100)	PI/NII	0.199	2.18
Heavy traffic (AADT/number of lanes>=7.5k)	PI/NII	-0.334	-3.91
Highway interchange	II/F	0.882	2.64
Low truck percentage (<=4%)	PI/NII	0.721	3.21
Temporal and Environmental characteristics			
Monday	II/F	0.378	2.14
Snowy/icy road surface	PI/NII	-0.943	-3.26
(Std. dev. of parameter distribution-normal)		1.775	3.98
Snowy/icy road surface	II/F	-1.062	-4.70
Snow/sleet/hail	II/F	-0.827	-3.47
Darkness-road lighted	II/F	-1.076	-2.21
Driver characteristics			
Young driver (age<=25)	PI/NII	-0.796	-2.95
(Std. dev. of parameter distribution-uniform)		1.398	3.26
Young driver (age<=25)	II/F	-0.389	-2.31
Old driver (age>=60)	II/F	0.541	2.40
Female driver	PI/NII	0.442	5.36
DUI alcohol/drug use	PI/NII	0.620	3.61
DUI alcohol/drug use	II/F	1.316	5.47
Driver was asleep	PI/NII	0.837	4.44
Driver was asleep	II/F	0.852	3.12
Careless/reckless driving	PI/NII	0.382	3.33
Careless/reckless driving	II/F	0.824	4.51
Driver was fatigued	II/F	0.830	2.09
Driver had no insurance/no proof of insurance	PI/NII	0.596	5.42
Driver had no insurance/no proof of insurance	II/F	1.060	6.11
Crash characteristics			
Only one vehicle involved	PI/NII	-0.746	-6.45
Only one vehicle involved	II/F	-1.000	-4.24
More than two vehicles involved	PI/NII	1.156	7.01
More than two vehicles involved	II/F	1.747	5.59
Animal caused	PI/NII	-0.727	-4.49
Animal caused	II/F	-1.528	-3.14
Exceeded legal speed	PI/NII	0.934	2.98
Exceeded legal speed	II/F	1.563	3.86
Followed too closely	II/F	-1.763	-3.61
Overturn	PI/NII	1.329	8.79
Overturn	II/F	1.957	9.74
Front to rear collision	II/F	-0.676	-2.29
Side to side collision with vehicles in same direction	PI/NII	-0.912	-5.75
Side to side collision with vehicles in same direction	II/F	-2.257	-4.52

Collision with parked motor vehicle	II/F	1.026	2.14
Collision with guard rail	PI/NII	0.423	3.04
Collision with cable rail	PI/NII	-0.969	-2.29
Collision with concrete highway barrier	PI/NII	0.391	2.67
Collision with bridge structure	PI/NII	0.846	2.24
Collision with vehicle debris or cargo	PI/NII	-1.550	-3.99
Collision with vehicle debris or cargo	II/F	-2.408	-2.31
Collision with embankment	II/F	1.078	3.74
Collision with delineator post	PI/NII	0.748	3.69
Collision with delineator post	II/F	1.206	3.82
Vehicle characteristics			
Truck (10001 lbs or over)	PI/NII	-0.740	-2.91
Truck (10001 lbs or over)	II/F	-0.779	-2.23
Passenger car /van	PI/NII	-0.608	-2.73
Passenger car /van	II/F	-1.474	-5.07
Pickup truck/utility	PI/NII	-0.646	-2.84
Pickup truck/utility	II/F	-2.825	-2.57
(Std. dev. of parameter distribution-triangular)		2.229	2.58
Pickup truck/utility with trailer	PI/NII	-1.208	-3.69
Pickup truck/utility with trailer	II/F	-2.400	-3.63
SUV	PI/NII	-0.513	-2.26
SUV	II/F	-0.849	-2.88
SUV with trailer	PI/NII	-2.033	-2.55
Model statistics			
Number of observations		7467	
Log likelihood at zero		-4194.95	
Log likelihood at convergence		-8203.34	
McFadden Pseudo R-squared		0.4886	

Table 5.4 NM Crash Injury Severity Model Estimation Results

Variable	**Injury outcome**	**Coefficient**	**t-Statistic**
Constant	PI/NII	-0.940	-11.95
Constant	II/F	-1.849	-6.56
Roadway characteristics			
Wide median (median width>=50ft)	PI/NII	-0.372	-2.98
(Std. dev. of parameter distribution-normal)		0.976	3.08
Deep rut indicator (rut index<=88)	PI/NII	-0.176	-2.8
High speed limit (speed limit=75mph)	II/F	0.514	2.89
Depressed median	PI/NII	0.232	3.47
Heavy traffic (AADT/number of lanes>=7.5k)	II/F	-0.441	-3.07
Environmental and temporal characteristics			
Monday	II/F	0.483	3.22
Snowy/icy road surface	PI/NII	-0.806	-2.98

(Std. dev. of parameter distribution-uniform)		1.570	3.6
Snowy/icy road surface	II/F	-0.806	-4.56
Wet road surface	II/F	-0.871	-3.1
Driver characteristics			
Young driver (age<=25)	PI/NII	-0.162	-2.59
Old driver (age>=60)	II/F	0.436	2.18
Female driver	PI/NII	0.334	5.17
DUI alcohol/drug use	PI/NII	0.750	5.6
DUI alcohol/drug use	II/F	1.621	8.13
Driver was asleep	PI/NII	0.749	4.6
Driver was asleep	II/F	0.819	3.02
Careless/reckless driving	PI/NII	0.200	2.61
Driver was fatigued	II/F	0.762	1.99
Illness/medical	PI/NII	0.758	2.95
Distracted by passenger	II/F	1.322	3.03
Driver had no insurance/no proof of insurance	PI/NII	0.184	2.14
Driver had no insurance/no proof of insurance	II/F	0.529	3.43
Driver's license had been denied	II/F	0.939	2.17
Crash characteristics			
Only one vehicle involved	PI/NII	-0.599	-6.24
Only one vehicle involved	II/F	-1.109	-5.57
More than two vehicles involved	PI/NII	1.206	11.01
More than two vehicles involved	II/F	1.356	5.7
Animal caused	PI/NII	-0.750	-4.38
Animal caused	II/F	-1.497	-3.17
Exceeded legal speed	II/F	2.004	4.88
Stopped in traffic	PI/NII	0.911	2.53
Stopped in traffic	II/F	1.634	2.43
Backing	PI/NII	-2.641	-2.54
Overturn	PI/NII	1.572	11.54
Overturn	II/F	1.724	10.16
Followed too closely	II/F	-1.021	-3.37
Improper passing	II/F	1.468	3.11
Front to front collision	II/F	0.934	2.4
Front to rear collision	II/F	-1.392	-5.84
Front to side collision	PI/NII	0.539	3.61
Side to side collision with vehicles in same direction	PI/NII	-1.016	-8.26
Side to side collision with vehicles in same direction	II/F	-2.066	-6.46
Collision with guard rail	PI/NII	0.434	3.44
Collision with cable rail	PI/NII	-0.532	-2.72
Collision with cable rail	II/F	-0.930	-1.97
Collision with concrete highway barrier	PI/NII	0.854	6.1
Collision with vehicle debris or cargo	PI/NII	-1.753	-5.98

Collision with vehicle debris or cargo	II/F	-3.226	-3.17
Collision with embankment	PI/NII	0.614	2.57
Vehicle characteristics			
Passenger car/van	PI/NII	-0.230	-3.63
Passenger car/van	II/F	-0.836	-4.76
Pickup truck/utility	II/F	-0.645	-3.05
Pickup truck/utility with trailer	PI/NII	-0.902	-4.26
Pickup truck/utility with trailer	II/F	-1.693	-3.48
SUV	II/F	-0.444	-2.12
Defective tires	PI/NII	0.715	2.71
Model Statistics			
Number of observations		8590	
Log likelihood at zero		-9437.08	
Log likelihood at convergence		-5423.15	
McFadden Pseudo R-squared		0.4253	

With regard to the random parameter density function, four types of distributions are considered: normal, lognormal, triangular, and uniform distributions. Three variables are found to produce statistically significant random parameters in the MT model, and two random parameters are significant in the NM model. In the MT model, it is found that the snowy/icy road surface indicator variable is normally distributed for moderate injury with the mean and standard deviation being 0.943 and 1.775, respectively. This indicates that 70.2% of the MT crashes that happened on snowy/icy road surface increase the probability of moderate injury, while 29.8% of the MT crashes that occurred on snowy/icy road surface decrease the likelihood of moderate injury. Such phenomena reflect the complex tradeoff between more cautious driving behavior and the increased difficulties of operating the vehicles on snowy/icy roads. Young driver indicator is uniformly distributed for moderate injury which has the mean and the standard deviation -0.796 and 1.398, respectively. It suggests that the probability of moderate injury increases for 30.6% of MT crashes involving young driver, but decreases for the rest (the other 69.4%). This phenomenon is perhaps because of the mixed effects from relatively imprudent driving behavior and less driving experience, and yet shorter reaction time of young drivers. Pickup truck/utility indicator of having severe injury crashes is triangularly distributed with the mean and standard deviation being -2.825 and 2.229, respectively. This indicates that the effect of pickup truck/utility is not the same across observations. One possible explanation is that it captures unobserved heterogeneity such as safety features, dynamic characteristics of pickup truck/utility and different pickup truck/utility driver behavior.

In the NM model, the snowy/icy road surface indicator in moderate injury outcome is also found uniformly distributed with the mean being –0.806 and standard deviation being 1.570. This implies that 30.3% of the NM crashes happened on snowy/icy road surface result in an increase in the probability of moderate injury and 69.7% of the NM crashes that happened on snowy/icy road surface lead to a decrease in the likelihood of moderate injury. Additionally, wide median indicator, which is defined for moderate injury outcome, is also found normally distributed and the mean and standard deviation are respectively -0.372 and 0.976. For 64.8% of the crashes, wide median decreases the probability of moderate injury; while for 35.2% of the crashes, wide median increases the likelihood of moderate injury. This is probably the outcome from the tradeoff between the improved physical protection and the affected driving behavior due to either "safer" or "more dangerous" interpretations by different drivers.

Average direct pseudo-elasticity results for MT and NM models are presented in Table 5.5. In the following section, detailed observations from Table 5.5 will be discussed.

Table 5.5 Average Direct Pseudo-elasticity for MT and NM models

Variable	**MT Elasticity (%)**			**NM Elasticity (%)**		
	NI	**PI/NII**	**II/F**	**NI**	**PI/NII**	**II/F**
Roadway characteristics						
Wide median (median width>=50ft)	-2.7	16.3	-2.7	8.9	-24.9	8.9
No rut indicator (rut index=100)	-3.1	18.3	-3.1	na	na	na
Deep rut indicator (rut index<=88)	na	na	na	4.0	-12.7	4.0
High speed limit (speed limit=75mph)	na	na	na	-1.6	-1.6	64.4
Depressed median	na	na	na	-5.1	19.6	-5.1
Heavy traffic (AADT/number of lanes>=7.5k)	5.0	-24.8	5.0	1.8	1.8	-34.5
Highway interchange	-3.6	-3.6	132.9*	na	na	na
Low truck percentage (<=4%)	-12.8	79.4	-12.8	na	na	na
Temporal and Environmental characteristics						
Monday	-1.3	-1.3	44.1	-2.0	-2.0	58.9
Snowy/icy road surface	18.7	-53.8	-59.0	21.2	-45.9	-45.9
Snow/sleet/hail	2.2	2.2	-55.3	na	na	na
Wet road surface	na	na	na	2.5	2.5	-57.1
Darkness-road lighted	2.4	2.4	-65.1	na	na	na
Driver characteristics						

Young driver (age<=25)	13.5	-48.8	-23.1	3.7	-11.8	3.7
Old driver (age>=60)	-1.9	-1.9	68.4	-1.8	-1.8	51.9
Female driver	-6.8	44.9	-6.8	-7.5	29.2	-7.5
DUI alcohol/drug use	-15.0	57.9	216.9*	-25.1	58.6	278.8*
Driver was asleep	-17.2	91.1	94.0	-21.3	66.5	78.6
Careless/reckless driving	-8.7	33.9	108.2*	-4.6	16.5	-4.6
Driver was fatigued	-3.3	-3.3	121.6*	-3.7	-3.7	106.4*
Illness/medical	na	na	na	-19.2	72.4	-19.2
Distracted by passenger	na	na	na	-7.8	-7.8	245.8*
Driver had no insurance/no proof of insurance	-13.0	57.9	151.2*	-6.2	12.7	59.1
Driver's license had been denied	na	na	na	-4.9	-4.9	143.4*
Crash characteristics						
Only one vehicle involved	18.5	-43.8	-56.4	21.1	-33.5	-60.0
More than two vehicles involved	-26.6	133.2*	321.1*	-34.0	120.5*	156.1*
Animal caused	13.8	-45.0	-75.3	19.9	-43.4	-73.1
Exceeded legal speed	-22.1	98.1	271.9*	-15.0	-15.0	531.0*
Stopped in traffic	na	na	na	-29.6	75.2	260.8*
Backing	na	na	na	31.6	-90.6	31.6
Overturn	-29.0	168.4*	402.9*	-43.3	173.3*	217.9*
Followed too closely	3.3	3.3	-82.3	2.8	2.8	-63.0
Improper passing	na	na	na	-9.2	-9.2	294.1*
Front to front collision	na	na	na	-4.8	-4.8	142.1*
Front to rear collision	1.8	1.8	-48.2	4.3	4.3	-74.1
Front to side collision	na	na	na	-13.2	48.8	-13.2
Side to side collision with vehicles in same direction	17.3	-52.9	-87.7	27.0	-54.0	-83.9
Collision with parked motor vehicle	-4.4	-4.4	166.8*	na	na	na
Collision with guard rail	-6.9	42.0	-6.9	-10.4	38.3	-10.4
Collision with cable rail	11.4	-57.7	11.4	14.3	-32.8	-54.9
Collision with concrete highway barrier	-6.4	38.5	-6.4	-21.4	84.8	-21.4
Collision with vehicle debris or cargo	22.4	-74.0	-89.0	36.3	-76.4	-94.6
Collision with embankment	-4.6	-4.6	180.5*	-15.3	56.5	-15.3
Collision with delineator post	-16.9	75.7	177.5*	na	na	na
Vehicle characteristics						
Truck (10001 lbs or over)	12.5	-46.3	-48.4	na	na	na
Passenger car/van	16.3	-36.7	-73.4	9.1	-13.3	-52.7
Pickup truck/utility	17.1	-38.6	-93.1	2.1	2.1	-46.4
Pickup truck/utility with trailer	20.0	-64.1	-89.1	22.9	-50.1	-77.4
SUV	10.8	-33.7	-52.6	1.5	1.5	-34.9
SUV with trailer	17.6	-84.6	17.6	na	na	na
Defective tires	na	na	na	-18.0	67.5	-18.0

Note: 1.* indicate significant increase in injury severity probability (elasticity≥100%); 2. na indicates not applicable

5.5.1 Roadway Characteristics

With regard to roadway characteristics, large disparities are found between MT and NM models. Although wide median indicator and heavy traffic indicator are found to be significant in both models, their effects towards crash injury severity are opposite. For MT model, wide median decreases the probability of severe injury by 2.7% while increases the probability of moderate injury by 16.3%. For NM model, however, wide median increases severe injury probability by 8.9% and decreases moderate injury probability by 24.9%. Such findings suggest that complex interactions between crash injury severity and wide median may exist. On non-mountainous highways, the wide median may provoke more aggressive driving behavior while in the meantime, provide more physical protection (Chen and Chen, 2011). The heavy traffic indicator increases severe injury (5%) and decreases moderate injury (24.8%) in the MT model. On the contrary, it reduces the probability of severe injury (34.5%) and increases the likelihood of moderate injury (1.8%) in the NM model. This result may reflect some effects caused by different traffic patterns between mountainous corridor I70 and non-mountainous interstate highway I25 on crash injury severity. Some specific mitigation strategies of severe injury on heavy traffic road sections of mountainous highways may be needed in the future by considering the unique characteristics of MT crashes.

Some variables are found to be only significant in the MT crash model. For instance, no rut indicator and low truck percentage increase the probability of moderate injury by 18.3% and 79.4% respectively, while they both slightly decrease the probability of severe injury on MT crashes. If a crash happens on a mountainous highway interchange, the likelihood of having severe injury is significantly increased by 132.9%. Traffic agencies and research community therefore need to put more efforts on mountainous highway safety by focusing on these unique contributing variables. Some variables are found only significant in the NM crash model. For example, deep rut indicator and high speed limit indicator increase the probability of severe injury (4.0% vs 64.4%), but decrease the likelihood of moderate injury (12.7% vs 1.6%). Depressed median, however, decreases severe injury by 5.1% and increases moderate injury by 19.6%.

5.5.2 Temporal and Environmental Characteristics

Although a variety of temporal indicators are considered, including different hours of a day and different days of a week, only Monday indicator is found significant. If a crash happens on Monday, it is 44.1% and 58.9% more likely to sustain severe injury for the MT and NM models respectively.

As discussed above, snowy/icy road surface condition has been found randomly distributed in both the MT and NM models. According to the elasticity results from Table 5.5, snowy/icy road surface reduces the probability of severe injury by 59.0% and 45.9% and reduces the chance of moderate injury by 53.8% and 45.9% for MT and NM crashes respectively. Wet road surface condition is also found to reduce the probability of severe injury by 57.1% for NM model. These findings are consistent with several previous studies (Chen and Chen, 2011; Christoforou et al., 2010; Malyshkina and Mannering, 2010a; Xie et al., 2009; Yamamoto and Shankar, 2004). Such effect can be partly explained by the fact that drivers tend to drive more cautiously on snowy/icy road surface or wet road surface than on normal surface condition. Besides road surface conditions, the inclement weather indicator (snow/sleet/hail) is found to be only significant for mountainous highway crashes. To be specific, it alleviates the probability of severe injury by 55.3% while slightly aggravates that of moderate injury. Xie et al. (2009) and Paleti et al. (2010) also reported similar findings. The darkness-road lighted indicator also significantly affects the injury severity for mountainous highway crashes (decreases the severe injury likelihood by 65.1%). This finding highlights the importance of lighting on mountainous highways, which may be considered in the future mitigation efforts on some crash hot spots.

5.5.3 Driver Characteristics

Different effects towards injury severity are observed for crashes caused by young drivers. For mountainous highway crashes, it is found that young drivers are less likely to result in severe injury (by 23.1%) and moderate injury (48.8%). For non-mountainous highway crashes, however, Table 5.5 shows that there are 3.7% increase and 11.8% decrease respectively in the probabilities of causing severe injury and moderate injury by young drivers. Careless/reckless driving is usually believed to considerably increase the chance of causing traffic crashes. Nevertheless, its effect toward injury severity has not been fully studied. Based on Table 5.5, on one hand, it is found that careless/reckless driving increases the chance of moderate injury for crashes on both mountainous and non-mountainous highways with varying magnitudes (33.9% vs 16.5%). On the other hand, careless/reckless driving has a 108.2% increase and 4.6% decrease in causing severe injury on mountainous and non-mountainous highways respectively. To the knowledge of the authors, such differences were not observed before, especially for the tremendous influence of careless/reckless driving behavior on severe injury on mountainous highways.

Although differences of driver characteristics' effects between MT and NM models are observed, some variables related to driver characteristics have similar influence towards crash injury severity in both MT and NM crashes. For example, old drivers are more prone to experiencing severe injury (68.4% and 51.9% in MT and NM models). Such a finding echoes with previous studies (Xie et al., 2009; Yasmin and Eluru, 2013), and is perhaps due to longer reaction time compared to other drivers. For female drivers, on the one hand, moderate injury increases for mountainous and non-mountainous highways crashes (44.9% vs 29.2%); on the other hand, severe injury decreases on both types of highways. This finding is in accordance with several studies (for example, Weiss et al. (2014) for young drivers, Malyshkina and Mannering (2010) on design exceptions, Islam and Hernandez (2013b) on heavy vehicles), while different from others (Xie et al., 2009). Besides driver age and gender, if a driver was asleep, it is 94.0% and 78.5% more likely to induce severe injury for MT crashes and NM crashes, respectively. Fatigued driving has long been recognized as hazardous factor in causing crashes. In the present chapter, it is discovered that fatigued driving substantially increases the chance of causing severe injury. Anastasopoulos and Mannering (2011) gave similar findings in their individual crash data models. This observation may be partly attributable to slow reaction time, decreased awareness and impaired judgment from driver fatigue. Aside from above mentioned results, if a driver was influenced by alcohol or drug use or driver was asleep, the MT and NM models lead to increases in both severe injury and moderate injury. Note that under the influence of alcohol and drug, the probability of severe injury increases considerably in both models. This result is supported by several studies (Chimba and Sando, 2009; Chiou et al., 2013; Xie et al., 2009; Yasmin and Eluru, 2013), and confirms the common belief about the higher risk of DUI. In addition, those drivers with no insurance or no proof of insurance are more likely to trigger severe injury and moderate injury in both highways. It implies the need for state patrol to pay more attention to those more vulnerable drivers.

Moreover, two variables are found to only significantly affect NM crashes injury severity. For ill drivers, severe injury is decreased by 19.2%. For the driver being distracted by passenger, it is 245.8% more likely to cause severe injury.

The abovementioned observations have significant implications in driving education, training and police enforcement. These observations highlight similarities and disparities of hazardous factors in MT and NM crashes and thus can be potentially useful in training professional/commercial drivers, and helping police department allocate enforcement resources more efficiently.

5.5.4 Crash Characteristics

Many variables of crash characteristics show similar influences on MT and NM crashes. For single-vehicle crashes, reduction in the probability of both the severe injury and moderate injury are observed. Other variables, which are significant in both MT and NM models, also show plausible trends of influence on the probability of severe injury and moderate injury. For example, animal-caused crashes are less inclined to sustain a severe injury and moderate injury, the probability of severe injury decreases while the probability of moderate injury increases if vehicles collide with the guard rail in both models.

Two variables are noteworthy: more than two vehicles involved indicator (multi-vehicle crash with three or more than three vehicles involved) and overturn indicator. The effects from both indicators on crash injury severity differ greatly in magnitudes for MT and NM crashes. For multi-vehicle crashes, it is found that the probability of severe injury increases by 321.1% on MT crashes and 156.1% on NM crashes. When a vehicle overturns, it is 402.9% more likely to sustain severe injury on MT crashes and 217.9% more likely to sustain severe injury on NM crashes. These results are consistent with previous studies (e.g. Shankar et al. 1996). Both multi-vehicle crash and overturn crash exhibit more critical influence towards crash injury severity on MT crashes. This result may be related to the interactions between complex terrain and driver maneuver difficulties on mountainous highways, and further analysis is needed to fully uncover the mechanism behind this phenomenon.

Except for those indicators that have similar effects on MT and NM crashes, other indicators show substantial differences. One major difference between MT and NM crashes is that the impacts of exceeding legal speed indicator on moderate injury are opposite (98.1% vs. -15.0%). One thing worth mentioning is that exceeding legal speed indicator significantly increases the likelihood of severe injury (271.9% and 531% for the MT and NM crashes respectively) as expected on both types of highways. Christoforou et al. (2010) also found that higher speed is more susceptible to severe injury. This phenomenon emphasizes the importance of speed law enforcement on both mountainous and non-mountainous highways. When a vehicle collides with cable rail, it results in an 11.4% increase and a 54.9% decrease in severe injury for MT and NM crashes, respectively. Collision with embankment increases the severe injury by 180.5% and slightly decreases the moderate injury on MT crashes, while it decreases the severe injury by 15.3% and increases the moderate injury by 56.5% on NM crashes. This probably reflects different effects of roadside design features toward crash injury severity on MN crashes as opposed to NM crashes, leading to some potential improvements over the design of roadside infrastructure of mountainous highways.

Two indicators are found to be exclusively significant in the MT model. Collision with parked motor vehicle increases the severe injury probability substantially (166.8%) on MT crashes. In the meantime, collision with delineator post is inclined to increase both the probability of severe injury and moderate injury (177.5% vs 75.7%). Similar to those two variables that are only significant in the MT model, five indicators are found to be only significant in NM model. For example, improper passing and front-to-front collision are found to substantially increase severe injury probability by 294.1% and 142.1% on NM models, respectively. In addition, if a vehicle stops in traffic, the likelihood of severe injury increases considerably by 260.8% and the likelihood of moderate injury increases by 75.2% in the NM model.

5.5.5 Vehicle Characteristics

If the at-fault vehicle is pickup truck/utility or SUV, there is substantial difference of moderate injury probability between MT and NM models (-38.6% vs 2.1% for pickup truck/utility, -33.7% vs 1.5% for SUV). Passenger car/van and pickup truck/utility with trailer are found to decrease severe injury and moderate injury in both models. Moreover, trucks are found to reduce severe injury and moderate injury on MT crashes. These findings indicate that heavy vehicle occupants are far less likely to sustain severe injury in contrast to light vehicles (for example, motorcycles). Similar conclusions were also drawn by other studies (e.g. Christoforou et al. 2010). One explanation may be that lighter vehicles absorb much more kinetic energy from a collision than heavy vehicles (Christoforou et al., 2010). Note that Abdel-Aty (2003) found that van and pickup occupants suffer less severe injury than passenger car occupants, and Yamamoto and Shankar (2004) found that motorcycle and truck are less likely to suffer from severe injury in urban area. These different findings may reflect the different driver behavior on different regions and terrains. For SUV with trailer, there is a 17.6% increase in severe injury probability and an 84.6% decrease in moderate injury in the MT model. Comparatively, defective tires indicator is found to be only significant in the NM model. It decreases severe injury probability by 18.0% and increases moderate injury probability by 67.5%.

5.5.6 Results Summary

Based on above discussions, critical contributing factors with different directions of the influence in both models and those only significant in one model are summarized in Table 5.6.

Table 5.6 Summary of Variables with Different Effects in MT and NM model

Variable	**MT**		**NM**	
	PI/NII	**II/F**	**PI/NII**	**II/F**
Wide median (median width>=50ft)	↑	↓	↓	↑
Heavy traffic (AADT/number of lanes>=7.5k)	↓	↑	↑	↓
Young driver (age<=25)	↓	↓	↓	↑
Careless/reckless driving	↑	↑	↑	↓
Exceeded legal speed	↑	↑	↓	↑
Collision with cable rail	↓	↑	↓	↓

Collision with embankment	↓	↑	↑	↓
Pickup truck/utility	↓	↓	↑	↓
SUV	↓	↓	↑	↓
No rut indicator (rut index=100)	↑	↓	na	na
Highway interchange	↓	↑	na	na
Low truck percentage (<=4%)	↑	↓	na	na
Snow/sleet/hail	↑	↓	na	na
Darkness-road lighted	↑	↓	na	na
Collision with parked motor vehicle	↓	↑	na	na
Collision with delineator post	↑	↑	na	na
Truck (10001 lbs or over)	↓	↓	na	na
SUV with trailer	↓	↑	na	na
Deep rut indicator (rut index<=88)	na	na	↓	↑
High speed limit (speed limit=75mph)	na	na	↓	↑
Depressed median	na	na	↑	↓
Wet road surface	na	na	↑	↓
Illness/medical	na	na	↑	↓
Distracted by passenger	na	na	↓	↑
Driver's license had been denied	na	na	↓	↑
Stopped in traffic	na	na	↑	↑
Backing	na	na	↓	↑
Improper passing	na	na	↓	↑
Front to front collision	na	na	↓	↑
Front to side collision	na	na	↑	↓
Defective tires	na	na	↑	↓

Note: 1. arrows show an increase (up) or decrease (down) in elasticity. 2. na indicates not applicable

As shown in Table 5.6, there are contributing factors that significantly affect injury severity in both types of highways but with different directions of influence: wide median (width>=50ft), heavy traffic (AADT/number of lanes>=7.5k), young driver, careless/reckless driving, exceeded legal speed, collision with cable rail, collision with embankment, pickup truck/utility and SUV.

Moreover, there are nine important contributing factors that are only significant in MT model: no rut indicator (rut index=100), highway interchange, low truck percentage (<=4%), snow/sleet/hail, darkness-road lighted, collision with parked motor vehicle, collision with delineator post, truck, SUV with trailer. Comparatively, there are thirteen important contributing factors that are only significant in NM model: deep rut indicator (rut index<=88), high speed limit (speed limit>=75mph), depressed median, wet road surface, illness/medical, distracted by passenger, driver's license had been denied, stopped in traffic, backing, improper passing, front to front collision, front to side collision, defective tires.

5.6 Summary

With four-year detailed crash injury severity data, separate mixed logit models were estimated for one mountainous and one non-mountainous interstate highway in Colorado. To provide scientific insights about potential mitigation efforts, this study comprehensively investigated critical contributing factors on injury severity. Substantial differences in the magnitude and direction of the influence of contributing factors were observed. Out of the factors that significantly affect injury severity, nine are exclusive to MT crashes and thirteen to NM crashes. Additionally, there are nine contributing factors that have opposite effects on injury severity between MT and NM models. Those factors that lead to more severe injury are (elasticity greater than 100%): highway interchange (MT crashes), DUI alcohol/drug use (both MT and NM crashes), careless/reckless driving (MT crashes), driver was fatigued (both MT and NM crashes), driver had no insurance/no proof of insurance (MT crashes), driver's license had been denied (NM crashes), more than two vehicles involved (both MT and NM crashes), exceeded legal speed (both MT and NM crashes), stopped in traffic (NM crashes), overturn (both MT and NM crashes), improper passing (NM crashes), front to front collision (NM crashes), collision with parked motor vehicle (MT crashes), collision with embankment (MT crashes), collision with delineator post (MT crashes).

Those factors that lead to less severe injury are (elasticity greater than 50%): snowy/icy road surface (both MT and NM crashes), snow/sleet/hail (MT crashes), wet road surface (NM crashes), darkness-road lighted (MT crashes), only one vehicle involved (both MT and NM crashes), animal caused (both MT and NM crashes), followed too closely (both MT and NM crashes), front to rear collision (NM crashes), side to side collision with vehicles in the same direction (both MT and NM crashes), collision with cable rail (NM crashes), collision with vehicle debris or cargo (both MT and NM crashes), passenger car/van (both MT and NM crashes), pickup truck/utility (MT crashes), pickup truck/utility with trailer (both MT and NM crashes), SUV (MT crashes).

This chapter is explorative in nature regarding investigating both mountainous and non-mountainous highways from the same region side by side. Rather than offering general findings with the mixed model in most existing studies, it is very helpful in identifying and understanding specific critical factors affecting injury severity on MT and NM crashes respectively. There are, however, some limitations of this chapter, which offer room for future improvements. For example, real-time traffic and weather data were not included in the model due to the incompleteness of the data. Although the adoption of mixed logit model is helpful in capturing heterogeneity in this regard, future work should incorporate the real-time data when it becomes widely available to form decent sample size. Findings based on the comparative study of two typical interstate highways in Colorado can offer valuable information about mountainous highways in general. However, future comparative studies on more highways from various states are still desired to provide more comprehensive insights.

CHAPTER 6 CORRELATED RANDOM EFFECTS BIVARIATE POISSON LOGNORMAL MODEL TO STUDY SINGLE- AND MULTI-VEHICLE CRASHES[5]

6.1 Introduction

Developing crash prediction models remains one of the primary approaches for studying traffic safety. Although these statistical models play a vital role in investigating crash mechanisms, most of them focused on univariate ones (Anastasopoulos and Mannering, 2009; Caliendo et al., 2007; Chen et al., 2014; Ma et al., 2015a; Malyshkina and Mannering, 2010b; Mohammed A Quddus, 2008; Shankar et al., 1998; Usman et al., 2012). In other words, these models studied the total number of crashes over a specified time period for some roadway segments or intersections, without distinguishing possible subgroups of these crashes and capturing correlations between these subgroups. Recently, some researchers have tried to estimate safety performance functions of different crashes simultaneously by dividing total crashes into different categories, such as different levels of injury severity (El-Basyouny and Sayed, 2009a; Ma et al., 2008; Ma and Kockelman, 2006; Park and Lord, 2007), and different collision types (Dong et al., 2014a; El-Basyouny et al., 2014). In addition to these categorizations, crashes can also be categorized by the number of vehicles involved, for example, single-vehicle (SV) and multi-vehicle (MV) crashes (Geedipally and Lord, 2010a; Yu and Abdel-Aty, 2013). Some recent studies have demonstrated that there is a significant difference in the characteristics associated with SV and MV crashes (Geedipally and Lord, 2010a, 2010b; Jonsson et al., 2007; Martensen and Dupont, 2013), highlighting the need to investigate SV and MV crashes separately.

Most of the current studies on SV and MV crashes have only focused on the effects of exposure and geometric features of roadways. Nevertheless, in addition to these two features, weather and traffic conditions have also been found to play a crucial role in crash occurrence (Aguero-valverde and Jovanis, 2007; Caliendo et al., 2007; Chen et al., 2014; Ma et al., 2015a; Usman et al., 2012, 2011, 2010). The present chapter attempts to investigate crash risks by categorizing crashes based on the number of

[5] This chapter is developed based on a research paper by Ma et al. (2016).

vehicles involved, namely SV and MV crashes, and also incorporating the effects of weather and traffic conditions in addition to exposure and geometrics. Moreover, thanks to the advanced monitoring system installed along the highways, detailed weather and traffic data can be adopted in this chapter in order to provide more insightful observations. As a result of adopting detailed data, however, multiple daily observations are generated for SV and MV crashes on each roadway segment, forming a multivariate panel dataset which poses some methodological challenges. Multivariate count data models have been advocated in view of their capability to capture correlations among different categories of crashes (Dong et al., 2015, 2014b; El-Basyouny et al., 2014; El-Basyouny and Sayed, 2009a; Ma et al., 2008; Ma and Kockelman, 2006; Park and Lord, 2007). However, these existing researches mainly focused on exploring cross-sectional multivariate data. The present chapter utilizes a multivariate panel data structure which has rarely been investigated. In contrast to cross-sectional multivariate data, there are three possible sources of correlations within multivariate panel data. First, the unobserved factors across different categories of crashes (in this case, SV and MV) set up a correlation between them. For instance, an increase in the number of SV crashes is also related to some changes in the number of MV crashes. Traditional multivariate data models can readily address this type of correlation. Second, there are correlations across time intervals between observations of each roadway segments for SV and MV crashes respectively. This type of correlation is due to unobserved heterogeneity at segment level. Third, correlations across time intervals between observations of roadway segments for SV and MV crashes may as well be correlated, as the unobserved heterogeneity are shared by both SV and MV crashes. Thus, analyzing multivariate panel data presents great statistical and computational challenges, thus requires more advanced models. To avoid biased parameter estimates and erroneous inferences that ensue, all these types of possible correlation in the data have to be properly accounted for. As explorative efforts, the authors propose a bivariate Poisson-lognormal model with correlated segment level random effects to properly capture those sources of correlation as summarized above.

After the methodology is proposed, it is applied to a mountainous highway in Colorado which is known to be adversely affected by weather conditions. Three models are developed and compared:

traditional bivariate Poisson lognormal (BPL), bivariate Poisson lognormal with uncorrelated segment level random effects (UREBPL), and bivariate Poisson lognormal with correlated segment level random effects (CREBPL). The primary objective of this chapter is to examine whether CREBPL is appropriate for analyzing SV and MV crashes using multivariate panel data. The secondary objective is to study the impact of weather and traffic conditions in addition to exposure and geometric conditions on SV and MV crashes. For a better illustration, this chapter is divided into five sections. The first section provides a literature review of related works on SV and MV crashes and multivariate data models, followed by a description of the data utilized in this chapter. The third section presents the model formulations of the proposed methodology. The fourth section discusses the model results, and the last section provides conclusion remarks as well as future research directions.

6.2 Background

6.2.1 Multivariate and Panel Data Model

Traditionally, traffic crashes are investigated using univariate models without further distinguishing crashes by different types (Chen et al., 2014; Chin and Quddus, 2003; Johansson, 1996; Ma et al., 2015a; Malyshkina and Mannering, 2010b; Shankar et al., 1995). Recognizing the need of accounting for unobserved factors across different types of crashes, Ma and Kockelman (2006) adopted a multivariate Poisson (MVP) model to analyze crashes by injury levels. Their results indicated that a MVP regression is superior to its univariate counterparts. Given that MVP cannot tolerate over-dispersion, Ladrón de Guevara et al. (2004) applied a multivariate negative binomial (MVNB) model to investigate fatal, injury and property-damage crashes simultaneously. However, MVNB assumes a gamma distributed error which is motivated merely by mathematical convenience, and it does not allow negative correlation structure. In order to overcome the drawbacks of MVP and MVNB, Chib and Winkelmann (2001) proposed a multivariate Poisson lognormal (MVPL) model. It is not only capable of addressing over-dispersion, but also allows a full general correlation structure. Given its strengths, MVPL has been widely applied in traffic safety studies (Aguero-Valverde and Jovanis, 2009; El-Basyouny et al., 2014; El-Basyouny and Sayed, 2009a; Ma et al., 2008; Park and Lord, 2007). It is worth noting that most of these past research endeavors have focused on multivariate cross-sectional count data. Panel data, which can capture unobserved factors, has become available in traffic safety studies recently. A variety of panel data models, such as random effects and random parameter models, have been widely applied in univariate models (Chen et al., 2014; Chin and Quddus, 2003; Ma et al., 2015a; Qi et al., 2007; Shankar et al., 1998). However, so far panel data model has been rarely explored in a multivariate setting.

6.2.2 Studies on SV and MV Crashes

Several researchers have investigated different characteristics related to SV and MV crashes. For example, Öström and Eriksson (1993) were among the first to develop different models for SV and MV crashes involving intoxicated drivers in northern Sweden. Their results revealed driver's blood alcohol content has different effects towards SV and MV crash fatality. Ivan et al. (1999) examined causality factors for SV and MV crashes respectively on two-lane rural roads in Connecticut. They found out that explanatory variables are different for SV and MV crashes. Geedipally and Lord (2010b) compared separate SV and MV models with the combined SV and MV model regarding identifying hot spots. They found separate SV and MV models yield fewer false positives and negatives than the combined model, and thus recommended developing separate SV and MV models to predict crashes as well as identify hot spots. Geedipally and Lord (2010a) also discovered that it is beneficial to model SV and MV separately regarding the prediction of confidence levels, and suggested using a joint model to account for correlation. Yu and Abdel-Aty (2013) investigated SV and MV crashes on a mountainous freeway. They developed a bivariate Poisson-lognormal model and hierarchical Poisson model and concluded bivariate Poisson-lognormal model performs better regarding DIC and number of significant variables. These past studies from different aspects have all demonstrated that it is preferred to model SV and MV crashes separately while still accounting for correlations between them using multivariate models. Nevertheless, these previous studies focused mainly on exposure variables and road design geometrics. Although numerous studies have demonstrated that weather and traffic data have significant effects on crash occurrence (Caliendo et al., 2007; Chen et al., 2014; Ma et al., 2015a; Usman et al., 2012, 2011, 2010), few studies have been reported on the effects of weather and traffic data in the context of SV and MV crashes. Two notable studies in this regard were conducted by Yu et al. (2013), and Yu and Abdel-Aty (2013), in which the effects of weather and traffic conditions on SV and MV crashes were investigated using Bayesian hierarchical models. Both studies identified average speed to be one of the most critical factors for SV and MV crashes. However, real-time road surface conditions, which are closely related to the crash occurrence, were not examined.

6.3 Data Description

A mountainous highway I-70 in Colorado, where advanced monitoring system has been installed, was selected for the present chapter. The chosen section of I-70 starts at mile marker (MM) 195.26 and ends at MM 251.32. Because this part of I-70 goes through the Rocky Mountains, it features high elevations and is susceptible to fast-changing weather conditions. This feature makes detailed weather data even more crucial in determining traffic safety than geometrics.

Four sources of accident-related data are incorporated in this chapter, (1) one-year accident data (Jan. 2010- Dec. 2010) provided by Colorado State Patrol (CSP), (2) roadway geometrics from the Roadway Characteristics Inventory (RCI), (3) real-time weather and road surface condition data recorded by the weather stations, and (4) real-time traffic data documented by traffic flow monitoring stations. By combining these data containing rich information, we can perform a thorough analysis of causality factors determining road safety performance on a daily basis. Data provided by RCI contains detailed geometric design features including longitudinal grade, curvature, curve radius, shoulder type and width, lane width, the number of lanes, median type and width, pavement conditions, etc. The ruti index in the RCI database is used to calculate remaining service life for rutting. Its value ranges from 50 to 100, where a value of 50 indicates 0.55-inch average rutting depth or higher and a value of 100 indicates 0.15-inch average rutting depth or less.

There are seven weather stations installed along the selected I-70 section, providing drivers with information about adverse weather conditions. The weather stations record the weather data and surface condition every 20 minutes. Because climate may change dramatically within a short time period (even within a day), indicator variables for surface conditions cannot represent actual surface conditions and thus are not appropriate in this regard. Instead, the percentage of certain surface condition is defined to account for variations within each day. There are 12 traffic stations placed on the west and east bounds respectively which provide real-time traffic data. For the purpose of finding best covariates, a variety of traffic-related variables are defined, among which speed gap is defined as the difference between the speed limit and average traffic speed. Note that if the average traffic speed exceeds the speed limit, it will be truncated to the speed limit in the database. In order to control heterogeneity associated with sudden temperature drop and winter storm onset in the study area during November 2010, special attention is given to November indicator.

The 56-mile freeway section selected for this chapter is divided into 100 homogeneous roadway segments (52 from eastbound and 48 from westbound). First, the roadway segments are defined according to the CDOT traffic station assignment. Then the segments are further divided into homogeneous ones based on changes of geometric features, including lane width, the number of lanes, median type, shoulder type, pavement condition, and speed limit, etc.

During the study period, a total of 29,462 observations were obtained, and there is a total of 340 and 202 accidents recorded for SV and MV crashes respectively. The data is processed into daily records on average 1-mile long segments, which lead to a preponderance of zero observations. About 98.9% and 99.4% of the observations are zeros for SV and MV crashes respectively. It is important, therefore, to check whether the fitted model can properly account for excess zeros. The descriptive statistics results for the data used in this chapter are presented in Table 6.1.

Table 6.1 Descriptive Statistics of Explanatory Variables and Crash Data

Variable	Description	Mean	Std Dev	Minimum	Maximum
Dependent variables					
SV	single-vehicle crash	0.012	0.117	0	3
MV	multi-vehicle crash	0.007	0.107	0	10
Exposure factors					
Length	Roadway segment length (mile)	1.072	0.712	0.368	3.684
Log daily traffic	Logarithm of daily traffic	9.539	0.491	6.579	10.967
Explanatory variables					

Temperature	Average air temperature (°F)	39.625	15.854	-11.232	71.701
Inside shoulder length	Inside shoulder length (feet)	3.932	1.412	0	7
Poor pavement indicator	1 if the condition of the road pavement for the primary direction is poor, 0 otherwise	0.551	0.497	0	1
Wet surface percent	Percentage of wet road surface	0.095	0.192	0	1
Chemically wet surface percent	Percentage of chemically wet road surface	0.059	0.177	0	1
Daily average speed gap	measured as the difference between speed limit and corresponding average traffic speed (mile/hour)	3.431	3.870	0	32.276
November indicator	1 if it is in November, 0 otherwise	0.0938	0.292	0	1
Good rutting indicator	Long remaining service life of rutting (1 if the value of ruti is 100, 0 otherwise)	0.204	0.403	0	1
Two lane indicator	1 if the number of lanes is two, 0 otherwise	0.797	0.403	0	1

6.4 Methodology

To perform multivariate panel data analysis, the Bayesian bivariate Poisson lognormal (BPL) model is developed first to address over-dispersion and account for correlation between SV and MV crash frequencies. Given the hierarchical nature of the data, BPL with uncorrelated random effects and BPL with correlated random effects are then proposed to capture extra unobserved heterogeneity.

6.4.1 Bivariate Poisson Lognormal model

Let Y_{it}^{k} represent the crash frequency on roadway segment *i* (*i*=1,2,…,*n*) during day *t* (*t*=1,2,…,T) which belongs to crash types *k* (*k*=1,2. 1= SV and 2=MV). In BPL, it is assumed that

$$Y_{it}^{k}|\lambda_{it}^{k} \sim Poisson(\lambda_{it}^{k}) \tag{6.1}$$

where λ_{it}^{k} is the Poisson rate.

The probability of Y_{it}^{k} is given by

$$Pr\left(Y_{it}^{k}=y_{it}^{k}|\lambda_{it}^{k}\right)=e^{-\lambda_{it}^{k}}\frac{\lambda_{it}^{k y_{it}^{k}}}{y_{it}^{k}!} \tag{6.2}$$

The Poisson rate is modeled using a log-normal distribution:

$$\log(\lambda_{it}^{k}) = \boldsymbol{X}_{it}^{k}\boldsymbol{\beta}^{k} + \varepsilon_{it}^{k} \tag{6.3}$$

where $\boldsymbol{X}_{it}^{k}$ is a set of explanatory variables for crash type k; $\boldsymbol{\beta}^{k}$ is the corresponding vector of estimable coefficients. The error term ε_{it}^{k} , which captures Poisson-heterogeneity among observations (regardless of which roadway segment it belongs to), is assumed to be multivariate normal distributed with a mean of **0** and a variance-covariance matrix of Σ, where

$$\varepsilon_{it}^{k} = \begin{pmatrix} \varepsilon_{it}^{1} \\ \varepsilon_{it}^{2} \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{21} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \tag{6.4}$$

The diagonal elements σ_{11} and σ_{22} of the variance-covariance matrix $\sum$ denote the variance of ε_{it}^{1} and ε_{it}^{2} respectively, while the off-diagonal elements denote the covariance between ε_{it}^{1} and ε_{it}^{2} . Uninformative proper prior distributions are utilized in the current study due to the lack of prior knowledge. The most common prior, namely diffuse normal distribution with mean 0 and large variance, is specified for the regression coefficients as (El-Basyouny et al., 2014):

$$\beta_j \sim normal(0,10000)(j = 1,2, \dots, J) \tag{6.5}$$

where β_k is the k-th regression coefficient, and J is the total number of regression coefficients. A Wishart prior is specified for the inverse of variance covariance matrix $\sum^{-1}$ (Congdon, 2005; El-Basyouny et al., 2014):

$$\textstyle\sum^{-1} \sim Wishart(I, 2) \tag{6.6}$$

where I is the 2×2 identity matrix.

6.4.2 Bivariate Poisson Lognormal with Independent Segment-level Random Effects

When dealing with cross-sectional multivariate count data, a traditional multivariate Poisson-lognormal model is usually sufficient. In case of panel multivariate count data where temporal correlation within each segment is present, however, the error term ε_{it}^{k} in Eq. (6.3) is not enough to capture unobserved heterogeneity. Thus, a bivariate Poisson lognormal model with independent segment-level random effects is proposed to capture segment-level unobserved factors. In this case, the Eq. (6.3) is reformulated into a dual error structure as (Riphahn et al., 2003):

$$\log(\lambda_{it}^{k}) = \boldsymbol{X}_{it}^{k}\boldsymbol{\beta}^{k} + \mu_{ik} + \varepsilon_{it}^{k} \tag{6.7}$$

where μ_{ik} represents the segment-level random effects for outcome (i.e. crash type) k; and μ_{i1} and μ_{i2} are assumed to be mutually independent. A normal prior is specified for μ_{ik}

$$\mu_{ik} \sim normal\left(0, \frac{1}{\tau_{\mu_k}}\right)\ (k = 1,2) \tag{6.8}$$

τ_{μ_k} is the precision of μ_{ik}, and a gamma hyper-prior is specified for τ_{μ_k}.

$$\tau_{\mu_k} \sim gamma(0.01, 0.01)\ (k = 1,2) \tag{6.9}$$

6.4.3 Bivariate Poisson Lognormal with Correlated Segment-level Random Effects

The preceding model specification imposes an independent assumption between segment-level random effects of different outcomes (crash types). Such a formulation ignores possible correlation of the outcomes at segment level. To overcome such a limitation, a bivariate Poisson lognormal with correlated segment-level random effects is defined as shown in Eq. (6.7), while μ_{ik} are assumed to be correlated which follow a multivariate normal distribution as:

$$\mu_{ik} \sim MVN_2(\mathbf{0}, \Omega)\ \ (k = 1,2) \tag{6.10}$$

where Ω is the variance-covariance matrix of μ_{ik}. A Wishart prior is specified for the inverse of variance covariance matrix Ω^{-1} (Congdon, 2005):

$$\Omega^{-1} \sim Wishart(I, 2) \tag{6.11}$$

where I is the 2×2 identity matrix.

6.4.4 Full Bayesian Estimation

The posterior distributions for the aforementioned Bayesian Poisson-lognormal models can be obtained using MCMC sampling algorithms via WinBUGS software (Spiegelhalter et al., 2005). For each model, the posterior estimates were obtained via two chains run of 200,000 iterations, 50,000 of which were discarded as burn-in period. Convergence of MCMC chains was monitored first by visually inspecting trace plots for the parameters. Then, the Brooks-Gelman-Rubin (BGR) statistic (Brooks and Gelman, 1998) was used to assess the convergence of multiple chains formally. As a rule of thumb, a BGR value less than 1.2 is believed sufficient to claim convergence.

6.4.5 Model Comparison

The Deviance Information Criteria (DIC), a Bayesian generalization of Akaike's Information Criteria (AIC), is used for model comparisons (Spiegelhalter et al., 2002). DIC combines the measure of fit $\overline{D}$ and the measure of model complexity p_D:

$$DIC = \overline{D} + p_D = D(\overline{\theta}) + 2p_D \tag{6.12}$$

where $D(\overline{\theta})$ is the deviance evaluated at the posterior means of the parameters of interest $\overline{\theta}$, p_D is the effective number of parameters in the mode, and $\overline{D}$ is the posterior mean of the deviance $D(\theta)$.

Same as AIC, models with smaller DIC are usually favored. It should be noted that the differences in DIC between models are more important than the absolute values of DIC. According to Spiegelhalter et al. (2005), differences in DIC of more than 10 might definitely exclude the model with higher DIC, and differences between 5 and 10 are considered substantial. If the difference in DIC is less than 5 and the models yield considerably different inferences, there is uncertainty about the choice of model, and as a result, it could be misleading to report only the model with the smallest DIC.

6.4.6 Model Checking

Although an optimal model may be determined simply based on DIC, it is prudent to assess its fit to the data further, because it is possible that the optimal model among the alternatives may still fit the data poorly concerning certain important aspects. In order to evaluate the fit of a model to the observed data, Gelman et al. (1996) proposed a posterior predictive checking diagnostic, in which the observed data is compared with the replicated one from the posterior predictive distribution. If the model fits the data well, the replicated data y^{rep} should resemble satisfactorily certain important aspects of the observed data y. A discrepancy measure $D(y;\theta)$ can be employed to quantify the similarity between the observed data and the replicated one. Various forms of $D(y;\theta)$ can be calculated, including skewness measure and residual-based measure. For a chosen $D(y;\theta)$, its reference distribution is obtained from the joint posterior distribution of y^{rep} and θ:

$$P_R(y^{rep},\theta) = P(y^{rep}|\theta)P(\theta|y) \tag{6.13}$$

The Bayesian p-value (P_B) represents the probability that the observed discrepancy measure $D(y;\theta)$ is more extreme than the replicated discrepancy measure $D(y^{rep};\theta)$, and is defined as:

$$P_B = P[D(y^{rep}) \geq D(y)|y] \tag{6.14}$$

A Monte Carlo estimate of P_B can be readily calculated through evaluating the proportion of draws in which $D(y^{rep}) \geq D(y)$. With regard to the chosen criteria, a p-value close to 0.5 indicates adequate model fit, while a p-value near 0 or 1 (above 0.8 or below 0.2) indicates discrepancy between the observations and the fitted model (Neelon et al., 2010).

Since the data utilized in this chapter contains a preponderance of zeros, it is of paramount importance to evaluate how well the model accounts for excess zeros. A corresponding discrepancy measure, i.e. the proportion of observations equaling to zero, is therefore adopted in this chapter.

6.5 Results

This section discusses the estimation results of BPL, UREBPL, and CREBPL models for SV and MV crashes. The BGR statistics of the model parameters presented herein are less than 1.2, and the corresponding trace plots show no sign of periodicity and tendency.

6.5.1 BPL Model Results

Table 6.2 presents the parameter estimates, 95% credible intervals and goodness of fit measure for BPL model. As shown in Table 6.2, different sets of explanatory variables are identified for SV and MV crashes. The correlation coefficient between SV and MV crashes is 0.62, implying that these two crash types are highly correlated. The Bayesian p-value for the proportion of zeros is 0.55, which indicates that about 55% of the replicated datasets produce higher or same amounts of zeros as compared to the observed data, while the rest 45% of the replicated datasets produce fewer amounts of zeros. Thus, it can be concluded that the model can properly account for excess zeros in the observed data.

Table 6.2 Parameter Estimates for BPL

Variable	Mean	Std	2.5%	97.5%
Single-vehicle				
Intercept	-9.2880	0.6751	-10.630	-7.9880
Length	0.4840	0.0705	0.3452	0.6221

Log daily traffic	0.3163	0.1316	0.0661	0.5722
Inside shoulder length	0.3725	0.0546	0.2660	0.4796
Poor pavement indicator	-0.4421	0.1242	-0.6873	-0.1997
Temperature	-0.0177	0.0045	-0.0266	-0.0089
Wet surface percent	0.9659	0.2466	0.4738	1.4430
Chemically wet surface percent	1.0730	0.2227	0.6341	1.5070
Daily average speed gap	0.0852	0.0110	0.0634	0.1066
σ_{11}	2.2120	0.3546	1.5630	2.9510
Multi-vehicle				
Intercept	-38.490	4.3220	-46.990	-30.200
Length	0.6146	0.1000	0.4174	0.8102
Log daily traffic	1.4820	0.2152	1.0690	1.9070
November indicator	0.8377	0.2298	0.3796	1.2810
Good rutting indicator	0.5992	0.2190	0.1629	1.0230
Two lane indicator	0.4540	0.2181	0.0361	0.8924
Poor pavement indicator	-0.5854	0.1854	-0.9498	-0.2243
Wet surface percent	1.0500	0.3513	0.3440	1.7190
Daily average speed gap	0.1446	0.0128	0.1195	0.1698
σ_{22}	3.8360	0.5755	2.8020	5.0700
σ_{12}	1.8180	0.3851	1.0450	2.5290
Correlation ρ_{12}		0.62		
P_B		0.55		
DIC		4860.87		
MAD		0.01580		
MSPE		0.018530		

6.5.1.1 Single-vehicle crashes

Both the exposure factors, i.e. segment length and logarithmic daily traffic, are found significant with positive signs. This implies that longer segments and larger daily traffic increase the likelihood of SV crashes. This finding is somewhat different from some existing observations, in which for example, single-vehicle crashes were found to be more likely to occur when traffic volume is low (Ivan et al., 1999). A further discussion on this finding is given in the following section. Inside shoulder length is also significant with a positive sign, which implies that segments with wider inside shoulder length are more likely to incur crashes. Poor pavement indicator, on the other hand, is found to be negatively associated with SV crashes. In addition to exposure factors and geometrics discussed above, the variables related to weather and traffic conditions are found to play a vital role in crash occurrence. In terms of temperature, different measures were tried, including minimum temperature and average temperature. Average temperature is included in the model as it yielded significant results in the model. Average temperature is found to be significant with a negative sign, meaning SV crashes decrease with an increase in average temperature. This result is consistent with a previous study in which an increase of temperature was found to decrease crash occurrence (El-Basyouny et al., 2014). Two surface characteristics variables are found significant in the model: wet surface percent and chemically wet surface percent, representing respectively the percentage of wet road surface and chemically wet road surface for some segment within a day. Both surface characteristics are positively related to crash occurrence, suggesting the higher the wet surface percent and chemically wet surface percent, the more SV crashes likely to occur. As for traffic conditions, comparisons among average speed gap, maximum speed gap, average speed, standard deviation of speed, maximum speed and minimum speed suggest the inclusion of daily average speed gap in the model is preferred according to goodness of fit. Daily average speed gap is significant with a positive sign, which suggests that an increase in speed gap could result in an increase in crash occurrence, which corroborates some existing studies (Chen et al., 2014; Ma et al., 2015a).

6.5.1.2 Multi-vehicle crashes

Similar to single-vehicle crashes, both segment length and logarithmic daily traffic are positively associated with MV crashes. This result is consistent with a previous study (Yu and Abdel-Aty, 2013) which found that more MV crashes are likely to occur on longer segments and segments with higher traffic volume. November indicator is found to be positively correlated with MV crashes, implying November tends to sustain more MV crashes than other months of the year. This is possibly due to unobserved effects associated with early winter storms in Colorado during November 2010. Good rutting indicator, which represents whether the rutting depth is 0.15 inches or less, is found significant with a positive sign. This indicates that MV crashes are more likely to happen on segments with rutting of 0.15 inches or less than those with higher rutting depth. Although one might expect that higher rutting poses threats on driving safety such as lane changing maneuver, this seemingly counterintuitive finding may be attributed to risk compensation theory (Anastasopoulos and Mannering, 2009). Two lane indicator is significant with positive sign, which indicates that two-lane segments are likely to have higher crash counts for MV crashes. This finding is consistent with previous study that was conducted on the same highway (Yu and Abdel-Aty, 2013). Poor pavement indicator is again found significant with negative sign, meaning fewer MV crashes are likely to occur on segments with poor pavement. This result is similar to several previous studies (Anastasopoulos et al., 2008; Anastasopoulos and Mannering, 2009). It is possible that poor pavement indicator is picking up some unobserved characteristics associated with driver behavior, like a previous study (Mannering, 2007) which showed that drivers tend to drive faster when they believe pavement quality is good. Furthermore, wet surface percent is found to be significant with a positive sign, suggesting the higher the wet surface percent, the more MV crashes likely to occur. Same as in SV crashes, daily average speed gap is positively associated with MV crashes, which indicates an increase in speed gap leads to an increase in MV crash frequency.

6.5.2 UREBPL Model Results

In order to account for unobserved heterogeneity across time within segments, a bivariate Poisson-lognormal model with independent segment-specific random effects (UREBPL) is developed. The segment-specific random effects ($\mu_{ik}, k = 1,2$) in UREBPL for SV and MV crashes are assumed to be independent. The parameter estimates, 95% credible intervals and goodness of fit measure for UREBPL model are presented in Table 6.3.

Table 6.3 Parameter Estimates for UREBPL

Variable	Mean	Std	2.5%	97.5%
Single-vehicle				
Intercept	-8.8680	0.9270	-10.720	-7.0900
Length	0.5441	0.1355	0.2819	0.8167
Log daily traffic	0.1291	0.1469	-0.1510	0.4262
Inside shoulder length	0.3010	0.0864	0.1335	0.4736
Poor pavement indicator	-0.2573	0.2070	-0.6612	0.1532
Temperature	-0.0189	0.0045	-0.0278	-0.0099
Wet surface percent	0.8436	0.2493	0.3475	1.3260
Chemically wet surface percent	0.9087	0.2380	0.4400	1.3730

Daily average speed gap	0.0941	0.0117	0.0711	0.1169
σ_{11}	1.7130	0.3389	1.0730	2.4460
σ_{μ_1}	0.7276	0.1071	0.5353	0.9553
Multi-vehicle				
Intercept	-39.140	4.5090	-47.880	-30.000
Length	0.6452	0.1428	0.3689	0.9316
Log daily traffic	1.5120	0.2263	1.0550	1.9500
November indicator	0.8243	0.2293	0.3653	1.2640
Good rutting indicator	0.6480	0.2855	0.0867	1.2130
Two lane indicator	0.5351	0.2757	0.0081	0.9989
Poor pavement indicator	-0.7176	0.2409	-1.2050	-0.2558
Wet surface percent	1.0270	0.3541	0.3146	1.7030
Daily average speed gap	0.1440	0.0133	0.1182	0.1703
σ_{22}	3.4770	0.5641	2.5140	4.7560
σ_{12}	1.7660	0.3914	0.9536	2.5020
σ_{μ_2}	0.6039	0.1520	0.2924	0.8512
Correlation ρ_{12}		0.72		
P_B		0.54		
DIC		4830.77		
MAD		0.01579		
MSPE		0.018525		

It can be seen that after accounting for segment specific random effects, two explanatory variables become insignificant, namely log daily traffic and poor pavement indicator for SV crashes. One possible explanation is that both explanatory variables may capture some unobserved heterogeneity in the BPL model. It is noteworthy that some previous studies have demonstrated that SV crashes are not associated with high traffic volume (Ivan et al., 1999; Yu and Abdel-Aty, 2013). When traffic volume gets large, there will be more interactions between vehicles, increasing the risk of MV crashes while decreasing SV crashes. That said, it is expected that log daily traffic becomes insignificant for SV crashes. The fact that log daily traffic became insignificant in the UREBPL model for SV crashes somewhat serves as an indicator that UREBPL performs better than BPL by accounting for segment-level heterogeneity. As for poor pavement indicator, it is possible that it was picking up some unobserved characteristics in BPL (where segment-level unobserved heterogeneity was not accounted for). These unobserved characteristics may include geographic locations of segments, traffic characteristics, and driver behavior. Some or all of these may happen to turn out to vary in a systematic way without properly accounting for the hierarchical structure of the data. It can be concluded that without properly accounting for segment-specific unobserved heterogeneity, the model (BPL) estimates could produce misleading results. Hence, when dealing with panel multivariate count data, it is crucial to address unobserved heterogeneity appropriately.

Concerning model comparison, UREBPL model has a considerably smaller DIC value than BPL model (4860.87 vs. 4830.77). Moreover, the Bayesian p-value for the proportion of zeros is 0.54, indicating UREBPL can also properly address excess zeros. We can, therefore, conclude that UREBPL model is superior to BPL model.

6.5.3 CREBPL Model Results

Although the preceding UREBPL model can properly account for segment-specific unobserved heterogeneity, it overlooks possible correlations between segment-specific unobserved heterogeneity for SV and MV crashes. To overcome such a limitation, a bivariate Poisson lognormal with correlated segment-specific random effects is estimated. The parameter estimates, 95% credible intervals and goodness of fit measure for CREBPL model, are listed in Table 6.4. It can be seen that CREBPL model yields consistent parameter estimates as compared to UREBPL model. Note that although two lane indicator becomes insignificant at 95% interval for MV crashes in CREBPL model, it is still significant at 90% interval.

Table 6.4 Parameter Estimates for CREBPL

Variable	Mean	Std	2.5%	97.5%
Single-vehicle				
Intercept	-8.9530	1.4260	-11.810	-6.2110
Length	0.5500	0.1365	0.2853	0.8235
Log daily traffic	0.1566	0.1434	-0.1192	0.4436
Inside shoulder length	0.3050	0.0850	0.1402	0.4742
Poor pavement indicator	-0.2654	0.2096	-0.6765	0.1486
Temperature	-0.0189	0.0046	-0.0279	-0.0099
Wet surface percent	0.8498	0.2512	0.3490	1.3340

Chemically wet surface percent	0.9114	0.2395	0.4391	1.3770
Daily average speed gap	0.0949	0.0117	0.0717	0.1177
σ_{11}	1.7910	0.3286	1.1750	2.4680
σ_{μ_1}	0.5508	0.1577	0.3012	0.9145
Multi-vehicle				
Intercept	-23.380	2.3230	-28.030	-18.930
Length	0.6525	0.1470	0.3672	0.9460
Log daily traffic	1.5030	0.2255	1.0670	1.9510
November indicator	0.8197	0.2296	0.3611	1.2610
Good rutting indicator	0.6602	0.2887	0.0937	1.2270
Two lane indicator	0.5216	0.2789	-0.0119	1.0830
Poor pavement indicator	-0.7065	0.2459	-1.1980	-0.2342
Wet surface percent	1.0210	0.3559	0.3043	1.7000
Daily average speed gap	0.1454	0.0134	0.1194	0.1718
σ_{22}	3.5220	0.5468	2.5660	4.7000
σ_{12}	1.8190	0.3792	1.0670	2.5510
σ_{μ_2}	0.4322	0.1698	0.1754	0.8305
$\sigma_{\mu_1\mu_2}$	0.0823	0.1115	-0.1262	0.3169
$\rho_{\mu_1\mu_2}$		0.17		
Correlation ρ_{12}		0.72		
P_B		0.55		
DIC		4815.59		
MAD		0.01578		
MSPE		0.018505		

In addition to the correlation between SV and MV crashes, the correlation between segment-specific random effects for SV and MV crashes is 0.17, which indicates both segment-specific random effects are moderately correlated and thus should be appropriately addressed. In terms of goodness of fit, Bayesian predictive p-value is 0.55 for the proportion of zeros, suggesting adequate fit of CREBPL model. Moreover, after allowing for correlation between segment-specific random effects, DIC value drops significantly from 4830.77 to 4815.59. According to above discussion, it is concluded that CREBPL model is superior to UREBPL model, and provides the best fit among all the three candidate models.

6.6 Summary

This chapter presented a novel approach to analyze and identify different hazardous factors for SV and MV crashes. A bivariate Poisson-lognormal model with correlated segment-specific random effects was proposed to characterize both the multivariate and panel nature of the data. The proposed model structure readily addresses three types of serial correlations within the multivariate panel data used in this chapter: (1) correlation between SV and MV, (2) temporal correlations across time within each segment for SV and MV respectively, (3) possible connection between temporal correlations for SV and MV crashes. Moreover, by incorporating real-time weather and traffic data, the different effects of weather and traffic conditions, as well as geometrics towards SV and MV crashes, were able to be comprehensively examined. In order to provide more insightful and rigorous comparative results with the proposed model, two other candidate models were also developed including the proposed one: (1) bivariate Poisson-lognormal model, and (2) bivariate Poisson-lognormal model with independent segment specific random effects. The estimated results show that the proposed bivariate Poisson-lognormal model with correlated segment-specific random effects outperforms the other two candidate models by addressing as much unobserved heterogeneity as possible, dealing with excessive zeros in the observed data, and bearing the smallest DIC value.

After the proposed methodology had been developed, it was further applied to a mountainous freeway section on I-70 in Colorado, where the climate is subjected to rapid change due to high elevation, requiring the inclusion of weather conditions in traffic safety analysis. With the help of advanced monitoring system, real-time weather and traffic data were incorporated in the present chapter in addition to exposure and geometrics, which can provide a more comprehensive understanding of the factors affecting SV and MV crashes than most existing studies. Results showed that weather and traffic related explanatory variables, especially surface conditions, play a significant role in affecting the occurrence of SV and MV crashes. Moreover, differences between hazardous factors for SV and MV crashes are also investigated. All these findings may benefit future engineering practices on traffic designs and active traffic management. For instance, the results can be used to quantify potential safety benefits of road maintenance, help design more appropriate traffic control measures and improve road infrastructure during adverse weather conditions.

It should be noted that there are still some limitations of the present chapter: (1) it mainly focused on basic linear effects of the factors and possible non-linear relations between those factors and crash frequencies may be investigated in the future; (2) only one freeway was investigated in the present chapter. In order to evaluate the transferability of the proposed model and provide more general insights, more studies using the proposed methodology on various sites with different traffic, environmental characteristics and jurisdictions should be carried out; (3) only one year of data was studied, and such data is limited in capturing long-term temporal trends (yearly variation). More years of data are desired to overcome this limitation in the future.

CHAPTER 7 MULTIVARIATE SPACE-TIME MODELING OF CRASH FREQUENCIES BY INJURY SEVERITY LEVELS[6]

7.1 Introduction

Road traffic crashes impose enormous emotional and economic burdens on human society due to the associated physical suffering, losses in life and financial burden. To reduce the number and mitigate the impacts of traffic crashes, numerous studies have been conducted to improve the understanding and prediction of traffic crashes. As relatively rare events, highway traffic crashes are usually assessed by aggregating the crash counts over a certain time period (week, month, etc.) and within a specific geographical range (e.g. roadway segment, intersection). Since traffic crash data is always associated with certain spatial and temporal dimensions, both spatial and temporal correlations/heterogeneities are often present within the data. In addition, when crash frequencies of different injury severity levels are to be modeled, dependence among the counts for a specific injury severity should also be accounted for. Despite the methodological advances during the past years, following methodological challenges remain in terms of predicting crash frequencies of different injury severity levels: 1) spatial correlation and/or heterogeneity; 2) temporal correlation and/or heterogeneity; and 3) correlations between crash frequencies of different injury severity levels. Although the last decade has witnessed substantial methodological improvements in crash prediction modeling, methods that can appropriately address all the challenges mentioned above are still not available. The model estimation results and the following inferences rely heavily on the underlying assumptions about crash prediction model and data structure. It, therefore, becomes critical to develop more advanced crash prediction models that can address these challenges appropriately.

Despite that crash data is in nature spatially and temporally distributed, most existing studies aggregated data over an extended time period (e.g. one or several years). These studies usually addressed the over-dispersion problem but assumed independence of observations from different roadway entities

---

[6] This chapter is developed based on a research paper by Xiaoxiang Ma, Suren Chen and Feng Chen, which was submitted to a refered journal Analytic Methods in Accident Research.

(Ma et al., 2008; Malyshkina and Mannering, 2010a; Shankar et al., 1995). Neither spatial nor temporal correlations were considered in these literature. Some studies attempted to use data with multiple temporal observations generated for each roadway entity and treated the data as panel data. In that case, it results in a temporal correlation because unobserved heterogeneity associated with a specific roadway entity may be similar from time to time. Different panel data models were proposed to address temporal correlation, such as random effects models (Chen et al., 2014; Shankar et al., 1998), negative multinomial models (Shankar et al., 2003) and generalized estimating equations (Lord and Persaud, 2000). By the same token, these models were also applied to address spatial correlation (e.g. Wang and Abdel-Aty, 2006). However, a major drawback of these methods is the lack of correlation structure which can explicitly define the spatial and/or temporal correlations. As a result, these models do not easily lend themselves to explicitly analyzing temporal trend and spatial pattern of crash risks.

Recently, spatial modeling has gained wide recognition in evaluating traffic safety risks on various types of roadway entities. Spatial modeling deals with spatial correlation which may reflect unmeasured confounders. Different approaches were proposed for spatial modeling, including intrinsic conditional autoregressive (CAR) model (Aguero-Valverde and Jovanis, 2008; Lee et al., 2015; MacNab, 2004; Wang and Kockelman, 2013; Xie et al., 2014), spatial autoregressive model (Mohammed A. Quddus, 2008), spatial error model (Mohammed A. Quddus, 2008), and geographic weighted Poisson regression (Hadayeghi et al., 2010). Most of these studies adopted CAR model, which enables “spatial smoothing” by borrowing strength from neighboring sites. This was possibly because CAR model takes advantage of the flexibility of Bayesian hierarchical framework to incorporate spatial correlation and can be readily extended to address more complicated models. The ‘Besag–York–Mollie’ (BYM) model, an extension of CAR model, was proposed to address spatial correlation from neighboring effects as well as spatial heterogeneity (Besag et al., 1991). Many studies demonstrated that the BYM model is a proper tool for spatial modeling of traffic crashes, where study units span a large geographical area or comprise highways of varying functional classes (Aguero-Valverde and Jovanis, 2010, 2008; Barua et al., 2014; Mohammed A. Quddus, 2008). Aguero-Valverde (2013) explored the multivariate CAR (MCAR) model and the result indicated MCAR is superior over its univariate counterparts. Barua et al. (2014) developed three MCAR models and showed that the multivariate model with both spatial correlation and heterogeneity provides a better fit than other ones. Similar MCAR models were also applied for crash modeling at varying spatial units (segment, intersection, zones, etc.) in some other literature (Aguero-Valverde et al., 2016; Lee et al., 2015; Wang and Kockelman, 2013). Although abovementioned studies contributed to state-of-the-art methodologies for spatial modeling and offered valuable insights with regard to crash risk, none of them has accounted for temporal variation at the same time.

In contrast to the abundance of literature in spatial models, space-time studies in the traffic safety analysis are very limited. The first reported attempt to explore space-time model in traffic crash analysis was conducted by Miaou et al. (2003). In their study, Miaou et al. (2003) developed various county-level space-time crash risk models for Texas, and they argued that temporal component was better modeled with fixed effects than with first-order autoregressive (AR(1)). Wang et al. (2013) developed a spatio-temporal model to study the impact of congestion on traffic safety. Their results showed that models with fixed time effects and first order random walk (RW1) time effects produced similar results for fatal and serious injury accidents. Aguero-Valverde and Jovanis (2006) adopted a space-time model proposed by Bernardinelli et al. (1995) to study county-level injury crashes in Pennsylvania. This model defines a spatio-temporal interaction that allows for different temporal trends for different locations. However, it restricts the temporal trends to be linear, which is clearly unrealistic for traffic safety studies especially those with fine temporal resolutions. Dong et al. (2016) also used the same space-time model for hotspot identification at the scale of traffic analysis zone (TAZ) to account for possible space-time interaction. There are, however, two major limitations of the space-time studies discussed above. First, they were all conducted over an extended time period (i.e. a year) with limited time points and thus suffer from loss of important time-varying information. Space-time study in fine temporal scales is yet to be conducted to bring new insights into the crash analysis. Secondly, only separate univariate analyses were carried out even when crashes of different injury levels were studied. It has been well established in the literature that the correlation among injury severities is important and thus needs to be considered to avoid the potential statistical problem (Mannering and Bhat, 2014). This problem is likely to be carried over to space-time setting, creating a need for joint models of different injury severity levels with great complexity. Multivariate space-time analyses in traffic safety, which bear the potential to offer better insights into crash risks, are clearly desired to fill the gap remaining in the current literature.

The objective of this chapter is to propose a framework of Bayesian multivariate space-time model that can address spatial correlation/heterogeneity, temporal correlation/heterogeneity, and the correlation between different injury severities. To this end, a series of alternative models are also presented to compare different structures of spatial and temporal random effects, and these alternative models extend the space-time approach by Knorr-Held and Besag (1998) and Abellan et al. (2008). The proposed methodology is demonstrated using daily crash frequency data (as opposed to annual crash data) from the interstate highway I70 in Colorado. As the first attempt to explore multivariate space-time modeling in crash analysis with fine temporal scale, the current study has the potential to bring new insights into the crash analysis.

7.2 Data Description

In this chapter, datasets were collected from a mountainous portion of the interstate highway I70 in the state of Colorado. This portion of I70 under study starts at mile marker (MM) 195.26 and ends at MM251.32. Crash data were collected by Colorado State Patrol (CSP) over a one-year period (Jan. 2010 – Dec. 2010), and were originally coded as property damage only, possible injury, non-incapacitating injury, incapacitating injury and fatal injury. Due to the sparseness of severe injury crashes, two severity levels are considered herein: (1) injury crash, which consists of possible injury, non-incapacitating injury, incapacitating injury and fatal injury; and (2) no injury crash (property damage only). Roadway geometry and pavement condition data were provided by Roadway Characteristics Inventory (RCI), which contains detailed information such as curve radius, vertical grade, pavement condition, etc. To facilitate space time modeling in fine temporal scale, real-time data were also collected and processed into daily records. Real-time traffic data and real-time weather/ road surface data were monitored by traffic stations and weather stations respectively. A variety of traffic and weather related variables were defined to represent covariate effects. Then the preliminary multicollinearity tests and backward elimination procedures were conducted to find the most promising set of covariates.

The chosen portion of I70 was divided into 100 homogeneous roadway segments, including 52 from the eastbound and 48 from the westbound. After combining above mentioned data sources together and processing into daily records, a total of 29,462 observations were obtained. The descriptive statistics of the dependent variables and the explanatory variables are shown in Table 7.1.

Table 7.1 Summary Statistics.

Variable	**Description**	**Mean**	**Standard deviation**	**Min**	**Max**
Dependent Variables					
No injury crash	property damage only crash	0.022	0.164	0	6
Injury crash	Injury crash (all crash except property damage only)	0.005	0.075	0	4
Explanatory Variables					
Log of vehicle miles traveled	Natural logarithm of vehicle miles traveled	9.432	0.730	5.580	11.911
Daily average speed gap	Measured as the difference between posted speed limit and corresponding average traffic speed (miles/hour)	3.431	3.870	0	32.276
November indicator	1 if it is in November, 0 otherwise	0.094	0.292	0	1
Inside shoulder width	Inside shoulder width (feet)	3.932	1.412	0	7
Number of entering ramp per mile per lane	Number of entering ramp per mile per lane	0.201	0.312	0	1.225
Poor pavement indicator	1 if the overall road pavement condition for the primary direction is poor, 0 otherwise	0.551	0.497	0	1
Two lane indicator	1 if the number of lanes is two, 0 otherwise	0.797	0.403	0	1
Wet surface percent	Percentage of wet road surface of the day	0.095	0.192	0	1
Steep downgrade slope indicator	1 if the slope is downgrade and grade is greater than 5%	0.115	0.319	0	1

7.3 Methodology

In this section, specifications of alternative Bayesian hierarchical models are presented including multivariate Poisson-lognormal model, multivariate spatial Poisson-lognormal model, and multivariate spatiotemporal Poisson-lognormal model.

Poisson distribution is usually used to model crash frequency data due to their count data nature. Specifically, let Y_{it}^{k} denote the crash frequency of injury severity k (=1, 2) on roadway segment i (=1,2,…,100) during t-th (t=1,2,…,358) day of the year. In the first level of the model hierarchy, it is assumed that

$$Y_{it}^{k} \,|\lambda_{it}^{k} \sim Poisson(\lambda_{it}^{k}) \tag{7.1}$$

where λ_{it}^{k} is the Poisson rate. Then the probability of observing y_{it}^{k} crashes can be given by

$$Prob\left(Y_{it}^{k} = y_{it}^{k}\right) = \frac{e^{-\lambda_{it}^{k}}(\lambda_{it}^{k})^{y_{it}^{k}}}{y_{it}^{k}!} \tag{7.2}$$

However, Poisson distribution requires that the variance equals to the mean, an assumption which is often violated in crash frequency data. Moreover, such specification does not allow dependence between crash frequencies of different injury severity levels. To overcome such limitations, multivariate Poisson-lognormal model (MVPLN) is formulated by specifying Poisson rate at the second level of the hierarchy models as following:

$$\log(\lambda_{it}^{k}) = \boldsymbol{X}_{it}^{k}\boldsymbol{\beta}^{k} + \varepsilon_{it}^{k} \tag{7.3}$$

where $\boldsymbol{X}_{it}^{k}$ is a set of observed risk factors (vehicle miles traveled, geometric features, traffic conditions, etc.) for crash severity level *k*; $\boldsymbol{\beta}^{k}$ is the corresponding vector of coefficients to be estimated. At the third level, appropriate priors are assigned. Owing to the lack of prior information, a highly non-informative normal prior is assigned to $\boldsymbol{\beta}^{k}$'s with a mean of 0 and variance of 10000. The error term ε_{it}^{k} , which captures unstructured over-dispersion, is assumed to be multivariate normally distributed with a mean vector $\mathbf{0}$ and a variance-covariance matrix of $\boldsymbol{\Sigma}$, where

$$\varepsilon_{it}^{k} = \begin{pmatrix} \varepsilon_{it}^{1} \\ \varepsilon_{it}^{2} \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{21} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \tag{7.4}$$

The diagonal elements σ_{11} and σ_{22} of the variance covariance matrix $\boldsymbol{\Sigma}$ denote the variance of ε_{it}^{1} and ε_{it}^{2} respectively, while the off-diagonal elements denote the covariance between ε_{it}^{1} and ε_{it}^{2}. A non-informative Wishart distribution is specified for the precision matrix $\boldsymbol{\Sigma}^{-1}$ (Congdon, 2005):

$$\Sigma^{-1} \sim Wishart(I, 2) \tag{7.5}$$

where I is the 2×2 identity matrix.

7.3.1 Heterogeneity Random Effects

Although the above-presented MVPLN model addresses over-dispersion as well as correlations between crashes of different injury severity levels (Ma et al., 2008), it largely ignores the hierarchy of data structure. To address possible unobserved heterogeneity resulted from the data hierarchy; segment-specific random effects are incorporated to form a multivariate random effects Poisson-lognormal model (REMVPLN):

$$\log(\lambda_{it}^{k}) = \boldsymbol{X}_{it}^{k}\boldsymbol{\beta}^{k} + \upsilon_{i}^{k} + \varepsilon_{it}^{k} \tag{7.6}$$

where υ_i^k is assumed to be an exchangeable normal prior with mean 0 and precision τ_υ^k (reciprocal of variance).

$$\upsilon_i^k \sim N\left(0, 1/\tau_\upsilon^k\right), k = 1,2 \tag{7.7}$$

τ_υ^k is assigned a gamma prior $\tau_\upsilon^k \sim Gamma(0.5, 0.0005)$.

7.3.2 Spatial Component

Above model specification is capable of capturing unstructured heterogeneity; however, it does not capture spatial correlations due to 'spillover' effects between these neighboring roadway segments. Numerous previous traffic safety studies have applied the convolution prior for modeling spatial random effects (Aguero-Valverde et al., 2016; Aguero-Valverde and Jovanis, 2010, 2009; Barua et al., 2014; Dong et al., 2016; Wang et al., 2013b). Initially introduced by Besag et al. (1991), this approach is also known as a BYM model. In a BYM model, spatial component is decomposed into two parts: structured spatial effects and unstructured spatial effects, as shown below:

$$\log(\lambda_{it}^{k}) = \boldsymbol{X}_{it}^{k}\boldsymbol{\beta}^{k} + \mu_{i}^{k} + \upsilon_{i}^{k} + \varepsilon_{it}^{k} \tag{7.8}$$

The unstructured spatial effects (spatial heterogeneity effects) υ_i^k that control for spatial (segment) heterogeneity are again assigned exchangeable normal priors across different severity level k as shown in Eq. (7.7). The structured spatial effects (spatial correlation effects) μ_i^k which capture the spatial correlation between neighboring segments are assigned an intrinsic conditional autoregressive (CAR) prior for each severity level k:

$$\mu_i^k|\mu_{-i}^k \sim N(\frac{\sum_{j\sim i} w_{ij}\mu_j^k}{\sum_{j\sim i} w_{ij}}, \frac{1/\tau_\mu^k}{\sum_{j\sim i} w_{ij}}) \tag{7.9}$$

where μ_{-i} refer to the neighbors of segment i; $j\sim i$ denote all the neighbors of segment i; w_{ij} denotes the weight that segment j has on segment i, and τ_μ^k is the precision for each severity level k. Both adjacency-based and distance-based measures can be used for neighboring structures. As shown by Aguero-Valverde and Jovanis (2008), first order adjacency-based proximity structure is preferable for modeling crash frequency on roadway segment level. Thus, first order adjacency-based neighboring structure is adopted for the present chapter. $w_{ij} = 1$ if segment j is adjacent to segment i (share vertex) and $w_{ij} = 0$ otherwise.

Despite the fact that BYM models were extensively applied in previous studies, the BYM specification suffers from several potential limitations. First, the BYM construct is only identifiable up to an additive constant of μ_i^k and υ_i^k. Additional constraint $\sum_i \mu_i^k = 0$ must be specified to overcome the identifiability problem. Second, with the structured spatial effects incorporated in the model, adding an unstructured heterogeneity term may be redundant (Banerjee et al., 2014). The latter issue was largely overlooked by previous studies in which BYM specification was usually presumed to be the "go to" specification for spatial effects. There is a lack of exploration of whether the data at hand adequately support both spatial correlation and spatial heterogeneity. It is possible that one type of spatial effect may dominate the other, and such a possibility is worth investigating. Moreover, for the multivariate model, correlations may arise between spatial structured effects as well as spatial unstructured effects across injury severity levels. To accommodate such multivariate setting, multivariate conditional autoregressive (MCAR) prior (Mardia, 1988) and multivariate normal (MVN) prior can be assigned to μ_i^k and υ_i^k respectively:

$$\boldsymbol{U}_i|(\boldsymbol{U}_{\mathbf{1}(-\boldsymbol{i})}, \boldsymbol{U}_{\mathbf{2}(-\boldsymbol{i})}) \sim MCAR(\overline{\boldsymbol{M}_{\boldsymbol{\iota}}}, \frac{\boldsymbol{\Omega}}{n_i}) \tag{7.10}$$

$$\boldsymbol{V}_{\boldsymbol{i}} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_V) \tag{7.11}$$

where $(\boldsymbol{U}_{\mathbf{1}(-\boldsymbol{i})}, \boldsymbol{U}_{\mathbf{2}(-\boldsymbol{i})})$ denote the elements of $2 \times n$ roadway segments matrix $\boldsymbol{U}$ excluding the i-th segment. $\overline{\boldsymbol{M}_{\boldsymbol{\iota}}} = (\mu_i^1, \mu_i^2)^T$, n_i is the number of neighboring segments, and $\boldsymbol{\Omega}$ is a 2×2 variance covariance matrix with diagonal elements representing conditional variances of μ_i^1 and μ_i^2, and off-diagonal elements representing the spatial conditional covariance between different severity levels. $\boldsymbol{\Sigma}_V$ is the variance covariance matrix for correlated spatial heterogeneity effects, the inverse of which is assigned a Wishart prior as defined in Eq. (7.5).

7.3.3 Temporal Component

When multiple observations are made for each location over time, it introduces possible temporal correlation and/or heterogeneity because unobserved factors may change from time to time. Thus, it is reasonable to incorporate a temporal component to capture temporal variation. In the similar vein as BYM specification for the spatial component, the temporal component is added to Eq. (7.8) and decomposed into temporal correlation and temporal heterogeneity terms:

$$\log(\lambda_{it}^k) = \boldsymbol{X}_{it}^k \boldsymbol{\beta}^k + \mu_i^k + \upsilon_i^k + \theta_t^k + \varphi_t^k + \varepsilon_{it}^k \tag{7.12}$$

where θ_t^k is the temporal correlation term and a random walk prior of first order (RW1) is used (Li et al., 2012), the underlying assumption is that temporal effects for neighboring time points tend to be similar. It is defined as:

$$\theta_t^k|\theta_{(-t)}^k \sim \begin{cases} N\left(\theta_{t+1}^k, \sigma_\theta^{2(k)}\right) & \text{for } t = 1, \\ N\left(\dfrac{\theta_{t-1}^k + \theta_{t+1}^k}{2}, \dfrac{\sigma_\theta^{2(k)}}{2}\right), & \text{for } t = 2,3, \dots, T-1 \;\; (k = 1,2) \\ N\left(\theta_{t-1}^k, \sigma_\theta^{2(k)}\right) & \text{for } t = T, \end{cases} \tag{7.13}$$

where $\theta_{(-t)}^k$ denote all elements of θ_t^k except for time point t, and T=358. $\sigma_\theta^{2(k)}$ is the variance of the temporal effects for severity level k, the inverse of which is assigned a gamma hyper-prior distribution $Gamma(0.5, 0.0005)$. Again, a sum to zero constraint is placed on vector θ^k to ensure identifiability. To implement RW1 temporal prior, an analogous spatial CAR prior is used where neighboring time points of t is defined as $t-1$ and $t+1$. φ_t^k is the temporal heterogeneity term, and an exchangeable normal prior with mean 0 and precision τ_φ^k is assigned:

$$\varphi_i^k \sim N\left(0, \frac{1}{\tau_\varphi^k}\right), k = 1,2 \tag{7.14}$$

Again, a gamma hyperprior distribution $Gamma(0.5, 0.0005)$ is adopted for precision τ_φ^k. Above specification assumes independent temporal correlation and temporal heterogeneity across different injury severity levels. A natural extension to multivariate setting is to allow both temporal correlation and temporal heterogeneity effects to be dependent across injury severity levels, as shown below.

$$\mathbf{\Theta}_i|(\mathbf{\Theta}_{\mathbf{1}(-i)}, \mathbf{\Theta}_{\mathbf{2}(-i)}) \sim MCAR(\overline{\mathbf{N}_i}, \frac{\mathbf{R}}{n_i}) \tag{7.15}$$

$$\mathbf{\Phi}_i \sim MVN(\mathbf{0}, \mathbf{\Sigma}_\Phi) \tag{7.16}$$

The parameters defined in Eq. (7.15) and Eq. (7.16) are similar to those defined in Eq. (7.10) and Eq. (7.11). Similar to the BYM specification for spatial effects, the temporal analogy of BYM as discussed above may be redundant for some datasets.

7.3.4 Model Specifications

The multivariate spatiotemporal specification outlined in Eq. (7.12) is decomposed into four components: (1) the covariate effect $\boldsymbol{X}_{it}^k \boldsymbol{\beta}^k$ that captures systematic trend caused by spatially and/or temporally varying risk factors; (2) a common spatial effect component $\mu_i^k + \upsilon_i^k$ for all time points that captures spatial correlation and/or heterogeneity; (3) a common temporal effect component $\theta_t^k + \varphi_t^k$ for every segment that captures temporal trend and/or heterogeneity; and (4) an extra spatio-temporal component ε_{it}^k that captures additional variability that is not explained by other model components. Such a specification is a multivariate extension to some previous separable space-time model studies (Abellan et al., 2008; Knorr-Held and Besag, 1998). As much as it is tempting to fit one model specified by Eq. (7.12) and make inferences based on the model results, it may lead to confirmation bias without a proper comparison between different model specifications. Focusing on the presence of spatial structured and unstructured effects, temporal structured and unstructured effects, as well as the correlations of those effects among different severity levels, the authors proposed a two-step modeling framework to find the most suitable model for the data at hand. At the first step, with the temporal component excluded, the most plausible specification for spatial component is identified by comparing different spatial prior specifications. At the second step, different temporal specifications for multivariate space-time models are compared based on the best spatial specification identified from the first step. The proposed model specifications are shown in Table 7.2.

Table 7.2 Proposed Model Specifications

Model	Spatial Component		Temporal Component		Extra ST
	$\mu_i^{(k)}$	$v_i^{(k)}$	$\theta_i^{(k)}$	$\varphi_i^{(k)}$	$\varepsilon_{it}^{(k)}$
Spatial models					
S1	-	normal	-	-	MN
S2	CAR	-	-	-	MN
S3	CAR	normal	-	-	MN
S4	MCAR	-	-	-	MN
S5	MCAR	normal	-	-	MN
S6	MCAR	MN	-	-	MN
Spatio-temporal models					
ST1	MCAR	-	-	normal	MN
ST2	MCAR	-	-	MN	MN
ST3	MCAR	-	CRW1	-	MN
ST4	MCAR	-	RW1	-	MN
ST5	MCAR	-	CRW1	normal	MN
ST6	MCAR	-	RW1	normal	MN
Competing spatio-temporal models					
CST1	CAR	-	-	normal	MN
CST2	MCAR	normal	-	normal	MN
CST3	MCAR	MN	-	MN	MN
CST4	MCAR	-	-	normal	-

Note: “-” indicates not specified, “CRW1” denotes correlated RW1 temporal priors as defined in Eq. (7.15), “normal” refers to independent normal prior and “MN” refers to multivariate normal prior.

7.3.5 Model Comparison and Checking

The Deviance Information Criteria (DIC) is used for model comparisons (Spiegelhalter et al., 2002). The DIC is a generalization of Akaike Information Criteria (AIC), which trades off the model fit against a measure of model complexity. It is defined as follows:

$$DIC = D(\overline{\theta}) + 2p_D = \overline{D} + p_D \tag{7.17}$$

where $D(\overline{\theta})$ is the deviance evaluated at the posterior means of the parameters of interest $\overline{\theta}$; p_D is the effective number of parameters in the mode, and $\overline{D}$ is the posterior mean of the deviance $D(\theta)$. Similar to AIC, a model with smaller DIC value is preferred.

Although DIC has been used extensively for comparisons between Bayesian hierarchical models, it is never intended as an absolute measure of model fit. In order to further evaluate the adequacy of a model fit to the observed data, Gelman et al. (1996) proposed a posterior predictive checking diagnostic, in which the observed data is compared with the replicated ones from the posterior predictive distribution. If the model fits the data adequately, the replicated data y^{rep} should be similar to the observed data y. A discrepancy measure $D(y;\theta)$ can be employed to quantify the similarity between the observed data and the replicated ones. For a chosen $D(y;\theta)$, its reference distribution is obtained from the joint posterior distribution of y^{rep} and θ:

$$P_R(y^{rep},\theta) = P(y^{rep}|\theta)P(\theta|y) \tag{7.18}$$

The Bayesian p-value (P_B) represents the probability that the observed discrepancy measure $D(y;\theta)$ is more extreme than the replicated discrepancy measure $D(y^{rep};\theta)$, and is calculated as:

$$P_B = P[D(y^{rep}) \geq D(y)|y] \tag{7.19}$$

A Monte Carlo estimate of P_B can be readily calculated through evaluating the proportion of draws in which $D(y^{rep}) \geq D(y)$. With regard to the chosen criteria, a p-value close to 0.5 implies adequate model fit, while a p-value near 0 or 1 (above 0.8 or below 0.2) indicates discrepancy between the observations and the fitted model (Neelon et al., 2010). Since the data adopted in the present chapter contains excessive zeros, it is therefore reasonable to evaluate how well the model can account for such characteristic. The proportion of zero observations is therefore employed as a discrepancy measure in the present chapter.

7.4 Results

7.4.1 Model Comparisons

The proposed Full Bayesian models were estimated with the freeware WinBUGS package (Spiegelhalter et al., 2005), which implements Markov Chain Monte Carlo (MCMC) method using the Metropolis-Hastings algorithm to sample from the full unnormalized posterior distribution. These samples were then used to summarize quantities of interest for the parameters such as means and 95% credible intervals. For each model, two MCMC chain runs of 80,000 iterations were initialized at dispersed starting values with the first 10,000 iterations discarded as burn-in period. Convergences were first monitored by visual inspection of trace plots and autocorrelation plots for the model parameters. Then, the Brooks-Gelman-Rubin (BGR) statistics (Brooks and Gelman, 1998) were calculated to assess the convergence of multiple chains formally.

Table 7.3 Goodness of Fit Measures of Proposed Models

Models	$\overline{D}$	P_D	DIC	${p_{no}}^1$	${p_{inj}}^2$
Spatial models					
S1	5672.0	855.8	6527.8	0.6	0.6
S2	5633.7	880.0	6513.7	0.71	0.51
S3	5660.4	868.6	6529.1	0.63	0.63
S4	***5619.6***	***873.5***	***6493.1***	***0.71***	***0.60***
S5	5686.5	895.2	6581.7	0.63	0.55
S6	5679.6	952.3	6631.9	0.66	0.59
Spatio-temporal models					
ST1	***5546.6***	***770.1***	***6316.6***	***0.65***	***0.68***
ST2	5587.5	785.3	6372.8	0.62	0.63
ST3	5693.4	828.1	6521.5	0.61	0.59
ST4	5686.0	827.3	6513.3	0.62	0.61
ST5	5596.3	860.8	6457.2	0.64	0.69
ST6	5590.3	790.6	6380.9	0.62	0.63
Competing spatio-temporal models					
CST1	5585.1	770.0	6355.2	0.66	0.57
CST2	5554.8	773.3	6328.1	0.71	0.67
CST3	5619.2	839.3	6458.5	0.63	0.64
CST4	6389.0	328.4	6717.4	0.31	0.39

[1]: Bayesian Posterior predictive check P-value for no injury crash

[2]: Bayesian Posterior predictive check P-value for injury crash

The goodness of fit measures for the proposed models are presented in Table 7.3. Focusing first on the structure of spatial random effects, six spatial models were estimated. Model S1 includes only a spatial heterogeneity term, which is equivalent to the random effects multivariate model proposed in Ma et al. (2016). Random parameter multivariate model was also estimated. However, it did not provide any improvement as compared to model S1 and thus is not presented. Model S2 to S6 are different variations of the spatial CAR and BYM models. A comparison between model S1 and S2 highlights the importance of including spatial correlations (DIC value decreases from 6527.9 to 6513.7). It is shown in Table 7.3 that model S4 provides the best model fit among all proposed spatial models, as implied by the smallest DIC value (6493.1). Since model S4 includes only structured spatial effects in the spatial component and allows multivariate structure in the spatial effects, this result suggests not only the significance of spatial correlations but also a non-negligible dependence in structured spatial effects between no injury and injury crashes.

Comparing between S4 and S5 or S6 implies that adding spatial heterogeneity effects, be they independent or correlated, actually lead to worse model fit (DIC values increase drastically by 88.3 and 138.8 respectively). Such a finding is seemingly contradictory to those indicated in previous studies (Aguero-Valverde, 2013b; Aguero-Valverde et al., 2016; Barua et al., 2014). For example, Barua et al. (2014) developed three models and concluded that the multivariate model with both spatial correlation and heterogeneity effects provides better model fit than the multivariate model with only spatial correlation effects. A close investigation into the differences between those previous studies and the current study offers a reasonable explanation. Data used in those previous studies were either based on TAZ or road segments from highways of various functional classes. Therefore, the unobserved spatial heterogeneities were presumed to be large and could not be properly captured by observed covariates and spatial smoothing. As compared to those data, the data adopted in the present chapter are much more homogenous in space. All the roadway segments are obtained from a mountainous part of I70, which are highly consistent in terms of geometric design standards, terrains, transportation operational features, and most importantly driver population characteristics. As a result, it can be concluded that spatial correlation dominates spatial heterogeneity for the present data given that the pure MCAR specification (without spatial heterogeneity term) is the favored spatial structure.

Based on the most plausible specification for spatial component (i.e. pure MCAR) identified above, six spatio-temporal models (i.e. ST1, ST2, ST3, ST4, ST5, and ST6) were proposed with varying structures for the temporal component. Model ST1 and ST2 incorporate only temporal heterogeneity term, and model ST3 and ST4 incorporate only temporal correlation term. Model ST5 and ST6 incorporate both temporal heterogeneity and correlation terms (BYM equivalent of the spatial component). Turning to the performance of those spatio-temporal models as shown in Table 7.3, several inferences can be made. The benefit of introducing unstructured temporal effects is revealed by the significant lower DIC values of model ST1 (6316.6) and model ST2 (6372.8) as compared to that of model S4 (6493.1). This finding indicates that there are significant temporal heterogeneities in the data which need to be incorporated into the model. On the other hand, including structured temporal effects in the model leads to poorer model fits, which suggests there is no significant temporal correlation to be addressed in the data. Such a finding is not surprising, given that the present chapter adopts fine temporal scale (daily) as opposed to extended ones (annual). It is presumed that even after controlling for the observable explanatory variables (including some time-varying variables), the unobserved temporal factors are still significant and vary abruptly from one day to the next for the present data. This finding points out the importance of time-varying information in the crash analysis. Only first order random walk prior is considered in the present study. Although the higher order of random walk priors could be considered, they tend to impose stronger smoothing in time, which contradicts the nature of the present data. Moreover, the fact that model ST1 is preferred over ST2 indicates that the unobserved temporal trends are quite different between no injury and injury crashes. Fig. 7.1 presents the mean temporal effects along with corresponding 95% intervals under models ST1 and ST3, and no clear patterns can be observed. The temporal trends tend to be smoothed under model ST3 as compared to those under ST1. Under model ST1, the temporal trend of no injury crash differs from that of injury crash, and they both tend to fluctuate significantly from one day to the next over the studied period.

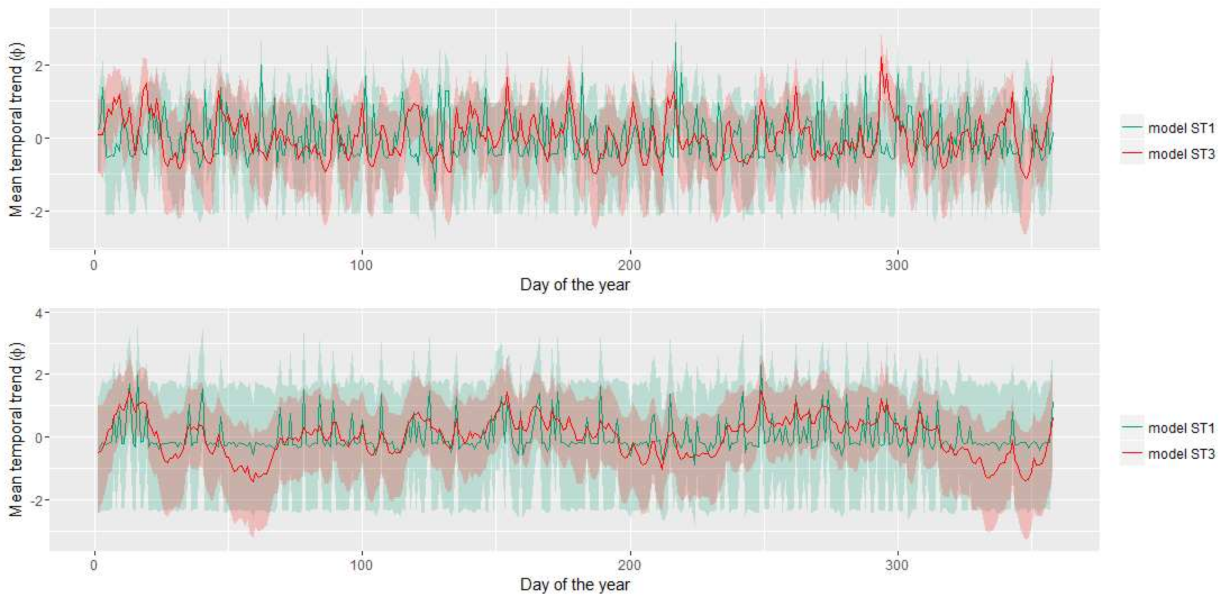

Figure 7.1 Mean temporal trends and 95% intervals under model ST1 and ST3 (top: no injury crash. bottom: injury crash)

Overall, the multivariate spatio-temporal model ST1 is clearly preferred over other models regarding DIC values. The Bayesian P-values produced from posterior predictive check indicate that model ST1 is consistent with the data in terms of zero proportions. To ensure that the pure MCAR spatial structure can still be warranted after introducing temporal dimension, several variations of ST1 are estimated, including model CST1, CST2, and CST3. As shown in Table 7.3, these models provide poorer fits than ST1, suggesting the superiority of ST1 over the rest models again. Another issue of interest for the current study relates to the extra space-time term (ε_{it}^{k}). We propose model CST4, which has the same model specification as ST1 except for the extra space-time term. As compared to model ST1, the effective number of parameters in model CST4 drops significantly by 441.7 signifying that CST4 is much more parsimonious. However, the sharp increase in DIC value implies that CST4 provides the worst model fit throughout. This finding suggests that there is significant space-time variability in the data which needs to be accounted for.

According to the above results about the model comparison, it is justifiable to draw conclusions regarding the source of variability based on model ST1. DIC’s strong preference for ST1 suggests that the current data have a complex structure and also the variability in the data comes mainly from several sources. Correspondingly, covariates, spatially structured effects, temporal unstructured effects, extra space-time effects, and the correlation between injury levels need to be considered in order to explain the variability properly.

Table 7.4 Parameter Estimates for Model S4

Variables	No injury crash				Injury crash			
	Mean	SD	(2.5%	97.5%)	Mean	SD	(2.5%	97.5%)
Constant	-6.078	0.319	-6.733	-5.477	-8.302	0.543	-9.439	-7.312
Log of vehicle miles travelled	0.581	0.088	0.409	0.755	0.585	0.150	0.294	0.880
Daily average speed gap	0.146	0.008	0.130	0.161	0.125	0.017	0.092	0.158
November indicator	0.492	0.129	0.235	0.742	0.662	0.257	0.147	1.150
Inside shoulder width	0.105	0.069	-0.030	0.241	0.312	0.111	0.095	0.534
Number of entering ramp per mile per lane	0.061	0.236	-0.403	0.523	-1.273	0.485	-2.256	-0.358
Poor pavement indicator	0.118	0.228	-0.326	0.570	-0.044	0.328	-0.685	0.601
Two lane indicator	0.646	0.322	0.044	1.317	0.791	0.424	-0.028	1.640
Wet surface percent	0.724	0.200	0.328	1.110	1.458	0.363	0.730	2.153
Steep downgrade slope indicator	0.518	0.268	-0.004	1.051	0.474	0.393	-0.302	1.239
Variance (σ_{ε}^2)	1.793	0.229	1.391	2.275	2.884	0.662	1.837	4.367
Spatial variance (σ_{μ}^2)	0.414	0.145	0.195	0.754	0.351	0.169	0.126	0.773

Table 7.5 Parameter Estimates for Model ST1

Variables	No injury crash				Injury crash			
	Mean	SD	(2.5%	97.5%)	Mean	SD	(2.5%	97.5%)
Constant	-6.094	0.310	-6.714	-5.505	-8.788	0.619	-10.08	-7.650
Log of vehicle miles travelled	0.636	0.095	0.453	0.824	0.596	0.158	0.290	0.910
Daily average speed gap	0.138	0.010	0.119	0.158	0.135	0.021	0.095	0.176
November indicator	0.631	0.235	0.169	1.094	0.729	0.384	-0.024	1.488
Inside shoulder width	0.094	0.068	-0.041	0.228	0.321	0.113	0.101	0.546
Number of entering ramp per mile per lane	0.112	0.234	-0.349	0.571	-1.265	0.493	-2.264	-0.334
Poor pavement indicator	0.107	0.225	-0.332	0.555	-0.022	0.337	-0.685	0.643
Two lane indicator	0.617	0.313	0.014	1.241	0.796	0.422	-0.019	1.635
Wet surface percent	0.261	0.238	-0.214	0.721	1.596	0.443	0.719	2.456
Steep downgrade slope indicator	0.515	0.267	-0.003	1.049	0.513	0.400	-0.272	1.301
Variance (σ_{ε}^2)	0.970	0.186	0.648	1.379	2.420	0.728	1.229	4.051

Spatial variance (σ_{μ}^2)	0.417	0.143	0.197	0.751	0.361	0.177	0.128	0.806
Temporal variance(σ_{φ}^2)	0.932	0.083	0.777	1.104	1.171	0.202	0.794	1.589

Table 7.6 Parameter Estimates for Model ST6

Variables	No injury crash				Injury crash			
	Mean	SD	(2.5%	97.5%)	Mean	SD	(2.5%	97.5%)
Constant	-5.891	0.294	-6.484	-5.333	-8.469	0.564	-9.675	-7.431
Log of vehicle miles travelled	0.627	0.099	0.437	0.827	0.583	0.159	0.278	0.904
Daily average speed gap	0.140	0.010	0.119	0.160	0.133	0.020	0.093	0.173
November indicator	0.625	0.235	0.167	1.083	0.702	0.374	-0.032	1.435
Inside shoulder width	0.091	0.069	-0.043	0.226	0.319	0.113	0.100	0.544
Number of entering ramp per mile per lane	0.119	0.236	-0.351	0.578	-1.262	0.485	-2.240	-0.346
Poor pavement indicator	0.027	0.237	-0.435	0.504	-0.100	0.347	-0.789	0.583
Two lane indicator	0.434	0.315	-0.163	1.061	0.616	0.424	-0.203	1.459
Wet surface percent	0.253	0.239	-0.229	0.719	1.577	0.426	0.723	2.399
Steep downgrade slope indicator	0.492	0.271	-0.037	1.016	0.470	0.400	-0.318	1.247
Variance (σ_{ε}^2)	0.946	0.160	0.655	1.292	2.294	0.657	1.256	3.819
Spatial variance (σ_{μ}^2)	0.413	0.142	0.196	0.745	0.354	0.172	0.124	0.784
Temporal variance(σ_{φ}^2)	0.933	0.084	0.778	1.108	1.114	0.214	0.712	1.530
Temporal variance(σ_{θ}^2)	0.027	0.009	0.015	0.048	0.033	0.013	0.016	0.067

7.4.2 Parameter Interpretation

For brevity, only relevant model results are presented. Table 7.4, Table 7.5, and Table 7.6 summarize the model estimation results for model S4, ST1, and ST6 respectively. The results of estimated covariate coefficients appear to be very robust across these models. This finding indicates when being used to identify risk contributing factors, these models would produce consistent results.

Turning to the specific results of parameter estimates, the following parameter inferences are mainly based on model ST1 as presented in Table 7.5. The model results reveal that logarithm of vehicle miles traveled (VMT) is statistically significant at 95% confidence interval and positively associated with both no injury and injury crashes. This is intuitive as VMT is regarded as the main risk exposure to crashes and is in line with previous studies (Ma and Kockelman, 2006; Wang et al., 2013b). Daily average speed gap is found to be significant (at a 95% confidence interval) and is positively related to no injury and injury crashes. This result indicates that higher speed gap is associated with more crashes and is consistent with existing studies (Chen et al., 2014; Ma et al., 2016). As for November indicator, it is found to be significant with a positive sign at both injury levels. This variable may be capturing unobserved effects associated with early winter storms in Colorado during November 2010. Inside shoulder width is statistically significant and positively associated with injury crash, suggesting more injury crashes would occur on roadway segments with wider inside shoulder. Two lane indicator is found to be statistically significant at 90% confidence interval with a positive sign for both injury levels. This implies that crashes are more likely to occur on roadways with two lanes as opposed to those with three lanes. Wet road surface percent is found to be significant at 95% confidence interval and is positively associated with injury crash in all three models. This finding implies that injury crashes are more likely to occur on roads with a higher wet surface percent. For no injury crash, however, wet surface percent is found to be significant only in model S4. Steep downgrade slope indicator is found to be significant at 90% confidence interval with a positive sign for no injury crash. Adding temporal effects in the model leads to the result that wet surface percent becomes insignificant with no injury crashes. This is possibly due to that wet surface percent variable captures some of the unobserved temporal heterogeneity in model S4. By comparing the relative magnitudes of spatial variance (σ_{μ}^2) and temporal variance (σ_{φ}^2) (0.417 and 0.932 for no injury crash, and 0.361 and 1.171 for injury crash), it appears that more variation comes from temporal component as opposed to spatial component. Therefore, it is important to incorporate the temporal component in the model.

Table 7.7 Variance-covariance Matrix for Model ST1

ε_{it}^k		
	No injury crash	**Injury crash**
No injury crash	0.970 (0.186)	1.263 (0.249)
Injury crash	1.263 (0.249)	2.420 (0.728)
μ_i^k		
No injury crash	0.417 (0.143)	0.258 (0.129)
Injury crash	0.258 (0.129)	0.361 (0.177)

Concerning the estimation of random effects, the variance σ_ε^2 is noticeably smaller in model ST1 and ST6 than those in model S4 for both no injury and injury crashes. This result is expected since model ST1 and ST6 add temporal random effects and they capture some of the variability that is captured by σ_ε^2 in S4. Focusing on temporal effects presented in Table 7.6, a comparison between variance of temporal effects indicates that temporal heterogeneity is much stronger than temporal correlations, since σ_φ^2 is much larger (0.933 and 1.114 for no injury crash and injury crash respectively) and σ_θ^2 (0.027 and 0.033 for no injury crash and injury crash respectively). This finding again indicates that the structured temporal effects are negligible because that the unstructured temporal effects capture most of the temporal variability. The variance-covariance matrix for model ST1 is presented in Table 7.7. The correlation between injury levels (ε_{it}^k) is estimated to be 0.82, and the correlation between spatial correlation effects (μ_i^k) is estimated to be 0.66. This implies there is non-negligible dependence between injury levels and the spatial correlation effects.

7.5 Summary

This chapter investigates the application of multivariate space-time models to jointly analyze crash frequency by injury severity levels in fine temporal scale. A multivariate space-time modeling framework is proposed within a Full Bayesian paradigm which focuses on finding best specifications for spatial and temporal random effects. In addition to the ability to consider both temporal and spatial trends, the proposed model framework is also capable of addressing complex correlations between crash types. It allows the underlying unobserved heterogeneity to be captured more comprehensively and enables borrowing strength across spatial units and over crash types, thus offering new insights into crash data analysis.

The proposed methodology is illustrated using one-year daily traffic crash data from the mountainous interstate highway I70 in Colorado. Apart from exposure factor (VMT), some variables are identified as important predictors, including geometric characteristics such as inside shoulder width, two-lane indicator, and steep downgrade slope indicator, and time-varying variables such as daily average speed gap and wet road surface. These results are intuitive and in line with existing findings. The model estimation results also highlight the importance of including a temporal component in the model.

In addition to offering some insights concerning the effects of predictors, some important conclusions in terms of the source of variation are also made. The empirical evidence provided herein suggests that the ST1 specification is the most appropriate one among all the models being compared. Since this model consists of multivariate spatial structured effects and independent unstructured temporal effects, it can disclose the complex structure and sources of variation in the present data. Hence, it is important to include temporal heterogeneity, spatial correlations, as well as dependence between spatial correlations on different injury levels in the model to explain such variation. It is also interesting to note that both spatial heterogeneity and temporal correlation appear to be insignificant in the model. Through identifying the source of variation in the data, the proposed framework helps researchers to understand the characteristics of crash data better.

Lastly, this study is not without limitations. The proposed multivariate space-time model framework was illustrated with only one specific dataset. Therefore, some specific observations and conclusions associated with the dataset are suited to galvanize future research rather than provide general unequivocal evidence with regard to the characteristics of crash data. Before more extensive studies with different datasets are made, discretion should be exercised when it comes to generalization regarding common characteristics of crash data. The next step following this line of research would be to carry out multivariate space-time modeling on other road entities, such as intersection, TAZ and between different crash types (heads on, rear end, etc.). Another limitation of the present study is that the proposed models do not take dynamic space-time interaction into consideration. Incorporating space-time interaction may entail much more complex model structure and bring further insights of the data, although it can also be computationally expensive or even cost-prohibitive, along with other technical challenges.

CHAPTER 8 CONCLUSIONS AND FUTURE RESEARCH

8.1 Summary and Conclusions

The objective of the current dissertation is to develop advanced multilevel regression models to address existing knowledge gaps and further utilize these models to investigate important methodological and empirical problems in crash data. This dissertation intends to advance state of the art in modeling crash data, which in return extend our understanding towards underlying crash mechanisms.

The contributions and findings of this dissertation are summarized in the following, which correspond to chapter two to chapter seven:

(1) This dissertation reports so far the first research attempt on developing random effects tobit models for both daytime and nighttime crash rates based on disaggregate modeling approach with panel data in refined temporal and spatial scales. By adopting tobit model with random effects, not only the censoring effects of crash rates data can be accounted for, but the unobserved heterogeneity across observations can also be potentially captured. Comprehensive road geometry, real-time traffic, weather and road surface data in refined temporal and spatial scales (hourly record and 1-mile roadway segments on average) was integrated into the crash rate model with panel data structure. The utilization of panel data in refined temporal scales enables capturing time-varying nature of variables (e.g. hourly traffic volume, visibility, wet road surface, etc.), which was usually ignored in traditional traffic crash modeling through data aggregation. In addition to the refined scales, for the first time, crash rate was studied with two separate models developed for nighttime and daytime, which led to some new findings and more refined information than traditional pooled data model. The results showed that there was major difference in contributing factors towards crash rate between daytime and nighttime, implying the considerable needs to consider daytime and nighttime crashes separately when refined-scale data (e.g. hourly) is studied. Based on these findings, more effective and adaptive mitigation efforts may become possible to save more people’s lives from crashes. These efforts include improvements in vehicle design, highway design, traffic management or law enforcement based on these new findings. Along this line,

some future studies can be carried out such as risk-based optimal route selection for adverse driving conditions, active law enforcement/traffic control intervention and advanced resource allocation and planning for trucking industry, etc.

(2) A random parameter tobit model was developed to analyze accident rates on mountainous highways. By facilitating the data in refined temporal scale (hourly basis), this study differs from previous accident rates studies in adopting disaggregated data. In addition to handling left-censoring effects and explicitly capturing unobserved heterogeneity, the random parameter tobit model can also properly deal with serial-correlations that are usually present in panel data. Moreover, by incorporating weather and traffic data in refined temporal scale (hourly basis) in our study, detailed phenomena were observed including those related to time-varying factors often being masked in traditional studies with aggregated data. A typical mountainous interstate highway I-70 was selected to demonstrate the proposed methodology. A random parameter tobit model was estimated by combining different data into one comprehensive data set. The likelihood ratio test result showed the superiority of random parameter tobit model over its fixed parameter counterpart. Model results demonstrated that various factors related to traffic and weather/surface conditions, roadway geometry and pavement play significant roles in crash rates. Poor visibility was found to increase accident rate. Similarly, adverse road surface conditions, including wet road surface, chemically wet road surface and ice warning, were all found to increase accident rate. Traffic-related factor, speed gap, was also found to significantly affect accident rate. Some factors were also found to possibly have a mixed influence on accident rate, such as speed limit, and inside shoulder width indicator. In addition, speed gap was found to produce random influence on accident rates in the model. The proposed model was developed based on real-time weather and traffic data from RWIS, a common data source around the country. Therefore, the technique developed in this chapter is easily transferrable to other highway safety studies with refined scales, exhibiting great engineering application potentials.

(3) A correlated random parameter two-part (CRPMTP) model is developed to study crash rates using refined-scale longitudinal data with excess zeros. Marginalized two-part model was adopted for the

first time to study crash rates. By comparing the model performance between marginalized two-part model and Tobit model, it shows that marginalized two-part model outperforms Tobit model. In addition, a correlated random parameter model as opposed to the uncorrelated one employed in most existing literature was developed to avoid the inappropriate independence assumption on random parameters. Likelihood ratio test along with AIC and BIC measures indicated correlated random parameter model is superior to corresponding uncorrelated one. This finding supports our hypothesis that correlated random parameter model could account for both the unobserved heterogeneities across roadway segments and the correlations between those unobserved heterogeneities. Likelihood ratio tests showed that the CRPMTP model was the best among the competing models in terms of goodness of fit. It was also found that Tobit model was not the preferred model choice. By facilitating a multivariate normal distribution of the random parameters, the CRPMTP model not only accounts for unobserved heterogeneity but also captures the correlations between random parameters. This study demonstrated that two-part model may be a better alternative to tobit model in analyzing crash rates when the data is right-skewed with large proportion of zero values. Moreover, speed gap in the binary part, curvature in the binary part, speed gap in the continuous part and percentage of snow occurring in the continuous part were found to have mixed effects on crash occurrence. Correlations between those random parameters were also revealed by adopting CRPMTP model. These finding can improve the understanding of the relationship between crash occurrence and contributing factors. Developing crash models that incorporate time-varying variables on a daily basis not only contributes to the improved understanding of crash occurrence, but also bears the potential to provide road users and policy makers with more detailed and relevant crash risk information.

(4) With four-year detailed crash injury severity data, separate mixed logit models were estimated for one mountainous and one non-mountainous interstate highway in Colorado. To provide scientific insights about potential mitigation efforts, critical contributing factors were comprehensively investigated. Substantial differences in the magnitude and direction of the influence of contributing factors were observed. Out of the factors that significantly affect injury severity, nine are exclusive to MT crashes and

thirteen to NM crashes. Additionally, nine contributing factors that have opposite effects on injury severity between MT and NM models are identified. Factors that contribute to increase injury severity and factors that contribute to mitigate injury severity are also identified. This study is explorative in nature in terms of investigating both mountainous and non-mountainous highways from the same region side by side. Rather than offering general findings with the mixed model in most existing studies, it is very helpful in identifying and understanding specific critical factors affecting injury severity on MT and NM crashes respectively. There are, however, some limitations of this chapter, which offer room for future improvements. For example, real-time traffic and weather data were not included in the model due to the incompleteness of the data. Although the adoption of mixed logit model is helpful in capturing heterogeneity in this regard, future work should incorporate the real-time data when it becomes widely available to form decent sample size. Findings based on the comparative study of two typical interstate highways in Colorado can offer valuable information about mountainous highways in general.

(5) A novel approach to analyze and identify different hazardous factors for SV and MV crashes was presented. A bivariate Poisson-lognormal model with correlated segment-specific random effects was proposed to characterize both the multivariate and panel nature of the data. In order to provide more insightful and rigorous comparative results with the proposed model, two other candidate models were also developed including the proposed one. The estimated results show that the proposed bivariate Poisson-lognormal model with correlated segment-specific random effects outperforms the other two candidate models by addressing as much unobserved heterogeneity as possible, dealing with excessive zeros in the observed data, as evidenced by the smallest DIC value. With the help of advanced monitoring system, real-time weather and traffic data were incorporated in the present chapter in addition to exposure and geometrics, which can provide more comprehensive understanding of the factors affecting SV and MV crashes than most existing studies. Results showed that weather and traffic related explanatory variables, especially surface conditions, play a significant role in affecting the occurrence of SV and MV crashes. Moreover, differences between hazardous factors for SV and MV crashes are also investigated. All these findings may benefit future engineering practices on traffic designs and active traffic

management. For instance, the results can be used to quantify potential safety benefits of road maintenance, help design more appropriate traffic control measures and improve road infrastructure designs during adverse weather conditions.

(6) A multivariate space-time modeling framework was proposed within Full Bayesian paradigm which focuses on finding best specifications for spatial and temporal random effects. In addition to the ability to consider both temporal and spatial trends, the proposed model framework is also capable of addressing complex correlations between crash types. It allows the underlying unobserved heterogeneity to be captured more comprehensively and enables borrowing strength across spatial units and over crash types, thus offering new insights into crash data analysis. The proposed methodology was illustrated using one-year daily traffic crash data from the mountainous interstate highway I70 in Colorado. Apart from exposure factor (VMT), some variables are identified as important predictors, including geometric characteristics such as inside shoulder width, two lane indicator, and steep downgrade slope indicator, and time-varying variables such as daily average speed gap and wet road surface. These results are intuitive and in line with existing findings. The model estimation results also highlight the importance of including temporal component in the model. In addition to offering some insights with regard to the effects of predictors, some important conclusions in terms of the source of variation are also made. The empirical evidence provided herein suggests that the ST1 specification is the most appropriate one among all the models being compared. Since this model consists of multivariate spatial structured effects and independent unstructured temporal effects, it can disclose the complex structure and sources of variation in the present data. Hence, it is important to include temporal heterogeneity, spatial correlations, as well as dependence between spatial correlations on different injury levels in the model to explain such variation. Through identifying the source of variation in the data, the proposed framework helps researchers to better understand the characteristics of crash data, which in turn can potentially improve prediction accuracy.

8.2 Directions for Future Research

This dissertation is not free of limitations, which leaves some room for future improvement. Below are some of the directions future researches can focus on:

(1) The data are limited to several interstate highways from Colorado. It is well known that crash data from different jurisdictions may exhibit varying characteristics. To generalize the findings from this dissertation, data from other states or countries with varying functional classes are desired.

(2) Researches conducted in this dissertation mainly focused on data in fine temporal scale (i.e. hourly or daily crash) by addressing methodological issues associated with time-varying variables. Often times significant research interests also lie on modeling annual crash frequency (for site-ranking or network screening purposes). To appropriately aggregate the findings in refined scales to annual results deserves further explorations.

REFERENCES

Aarts, L., Van Schagen, I., 2006. Driving speed and the risk of road crashes: A review. Accident Analysis and Prevention 38, 215–224. doi:10.1016/j.aap.2005.07.004

AASHTO, 2010. Highway Safety Manual. Washington, D.C.

Abdel-Aty, M., 2003. Analysis of driver injury severity levels at multiple locations using ordered probit models. Journal of Safety Research 34, 597–603. doi:10.1016/j.jsr.2003.05.009

Abdel-Aty, M.., Radwan, A.., 2000. Modeling traffic accident occurrence and involvement. Accident Analysis and Prevention 32, 633–642. doi:10.1016/S0001-4575(99)00094-9

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, 167–174. doi:10.1109/TITS.2006.874710

Abdel-Aty, M., Uddin, N., Pande, A., Abdalla, F., Hsia, L., 2004. Predicting Freeway Crashes from Loop Detector Data by Matched Case-Control Logistic Regression. Transportation Research Record: Journal of the Transportation Research Board 1897, 88–95. doi:10.3141/1897-12

Abellan, J.J., Richardson, S., Best, N., 2008. Use of space time models to investigate the stability of patterns of disease. Environmental Health Perspectives 116, 1111–1119. doi:10.1289/ehp.10814

Aguero-Valverde, J., 2013a. Full Bayes Poisson gamma, Poisson lognormal, and zero inflated random effects models: Comparing the precision of crash frequency estimates. Accident Analysis and Prevention 50, 289–297. doi:10.1016/j.aap.2012.04.019

Aguero-Valverde, J., 2013b. Multivariate spatial models of excess crash frequency at area level: Case of Costa Rica. Accident Analysis and Prevention 59, 365–373. doi:10.1016/j.aap.2013.06.014

Aguero-Valverde, J., Jovanis, P.P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. Accident Analysis and Prevention 38, 618–625. doi:10.1016/j.aap.2005.12.006

Aguero-valverde, J., Jovanis, P.P., 2007. Identifying Road Segments with High Risk of Weather-Related Crashes Using Full Bayesian Hierarchical Models, in: Transportation Research Board 86th Annual

Meeting, Transportation Research Board of the National Academies. Washington, D.C.

Aguero-Valverde, J., Jovanis, P.P., 2008. Analysis of Road Crash Frequency with Spatial Models. Transportation Research Record: Journal of the Transportation Research Board 2061, 55–63. doi:10.3141/2061-07

Aguero-Valverde, J., Jovanis, P.P., 2009. Bayesian Multivariate Poisson Lognormal Models for Crash Severity Modeling and Site Ranking. Transportation Research Record: Journal of the Transportation Research Board 2136, 82–91. doi:10.3141/2136-10

Aguero-Valverde, J., Jovanis, P.P., 2010. Spatial Correlation in Multilevel Crash Frequency Models. Transportation Research Record: Journal of the Transportation Research Board 2165, 21–32. doi:10.3141/2165-03

Aguero-Valverde, J., Wu, K.-F. (Ken), Donnell, E.T., 2016. A multivariate spatial crash frequency model for identifying sites with promise based on crash types. Accident Analysis and Prevention 87, 8–16. doi:10.1016/j.aap.2015.11.006

Ahmed, M., Huang, H., Abdel-Aty, M., Guevara, B., 2011. Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. Accident Analysis and Prevention 43, 1581–1589. doi:10.1016/j.aap.2011.03.021

Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41, 153–159. doi:10.1016/j.aap.2008.10.005

Anastasopoulos, P.C., Mannering, F.L., 2011. An empirical assessment of fixed and random parameter logit models using crash- and non-crash-specific injury data. Accident Analysis and Prevention 43, 1140–1147. doi:10.1016/j.aap.2010.12.024

Anastasopoulos, P.C., Mannering, F.L., Shankar, V.N., Haddock, J.E., 2012a. A study of factors affecting highway accident rates using the random-parameters tobit model. Accident Analysis and Prevention 45, 628–633. doi:10.1016/j.aap.2011.09.015

Anastasopoulos, P.C., Shankar, V.N., Haddock, J.E., Mannering, F.L., 2012b. A multivariate tobit analysis of highway accident-injury-severity rates. Accident Analysis and Prevention 45, 110–119. doi:10.1016/j.aap.2011.11.006

Anastasopoulos, P.C., Shankar, V.N., Haddock, J.E., Mannering, F.L., 2012c. A multivariate tobit analysis of highway accident-injury-severity rates. Accident Analysis and Prevention 45, 110–119. doi:10.1016/j.aap.2011.11.006

Anastasopoulos, P.C., Tarko, A.P., Mannering, F.L., 2008. Tobit analysis of vehicle accident rates on interstate highways. Accident Analysis and Prevention 40, 768–775. doi:10.1016/j.aap.2007.09.006

Arabmazar, A., Schmidt, P., 1982. An Investigation of the Robustness of the Tobit Esitmator to Non-Normality. Econometrica: Journal of the Econometric Society 50, 1055–1063.

Assum, T., Bjørnskau, T., Fosser, S., Sagberg, F., 1999. Risk compensation--the case of road lighting. Accident Analysis and Prevention 31, 545–553.

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2014. Hierarchical modeling and analysis for spatial data, 2nd ed. CRC Press.

Barua, S., El-Basyouny, K., Islam, M.T., 2014. A Full Bayesian multivariate count data model of collision severity with spatial correlation. Analytic Methods in Accident Research 3–4, 28–43. doi:10.1016/j.amar.2014.09.001

Belasco, E.J., Ghosh, S.K., 2012. Modelling semi-continuous data using mixture regression models with an application to cattle production yields. The Journal of Agricultural Science 150, 109–121. doi:10.1017/S0021859611000608

Bera, A.K., Jarque, C.M., Lee, L., 1984. Testing the Normality Assumption in Limited Dependent Variable Models. International Economic Review 25, 563–578.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., Songini, M., 1995. Bayesian analysis of space-time variation in disease risk. Statistics in Medicine 14, 2433–2443. doi:10.1002/sim.4780142112

Bertness, J., 1980. Rain-Related Impacts on Selected Transportation Activities and Utility Services in the Chicago Area. Journal of Applied Meteorology 19, 545–556. doi:10.1175/1520-

0450(1980)019<0545:RRIOST>2.0.CO;2

Besag, J., York, J., Mollie, A., 1991. Bayesian image-restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics 43, 1–20. doi:10.1007/BF00116466

Bhat, C.R., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. Transportation Research Part B: Methodological 37, 837–855. doi:10.1016/S0191-2615(02)00090-5

Brooks, S.P.B., Gelman, A.G., 1998. General Methods for Monitoring Convergence of Iterative Simulations. Journal Of Computational And Graphical Statistics 7, 434–455.

Bullough, J.D., Donnell, E.T., Rea, M.S., 2013. To illuminate or not to illuminate: roadway lighting as it affects traffic safety at intersections. Accident Analysis and Prevention 53, 65–77. doi:10.1016/j.aap.2012.12.029

Caliendo, C., De Guglielmo, M.L., Guida, M., 2013. A crash-prediction model for road tunnels. Accident Analysis and Prevention 55, 107–115. doi:10.1016/j.aap.2013.02.024

Caliendo, C., Guida, M., Parisi, A., 2007. A crash-prediction model for multilane roads. Accident Analysis and Prevention 39, 657–670. doi:10.1016/j.aap.2006.10.012

Calzolari, G., Magazzini, L., Mealli, F., 2001. Simulation-based estimation of Tobit model with random effects, in: Econometric Studies: A Festschrift in Honour of Joachim Frohn. p. 349.

Carson, J., Mannering, F., 2001. The effect of ice warning signs on ice-accident frequencies and severities. Accident Analysis and Prevention 33, 99–109.

Centers for Disease Control and Prevention, 2015. Leading Causes of Death. (Web-based Injury Statistics Query and Reporting System database). Atlanta, Available at http://www.cdc.gov/injury/wisqars/leading_causes_death.html.

Cerwick, D.M., Gkritza, K., Shaheed, M.S., Hans, Z., 2014. A comparison of the mixed logit and latent class methods for crash severity analysis. Analytic Methods in Accident Research 3–4, 11–27. doi:10.1016/j.amar.2014.09.002

Chang, L.Y., Chen, W.C., 2005. Data mining of tree-based models to analyze freeway accident frequency.

Journal of Safety Research 36, 365–375. doi:10.1016/j.jsr.2005.06.013

Chang, L.Y., Mannering, F., 1999. Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accidents. Accident Analysis and Prevention 31, 579–592.

Chen, F., Chen, S., 2011. Injury severities of truck drivers in single- and multi-vehicle accidents on rural highways. Accident Analysis and Prevention 43, 1677–1688. doi:10.1016/j.aap.2011.03.026

Chen, F., Ma, X., Chen, S., 2014. Refined-scale panel data crash rate analysis using random-effects tobit model. Accident Analysis and Prevention 73, 323–332. doi:10.1016/j.aap.2014.09.025

Chen, S., Chen, F., Wu, J., 2011. Multi-scale traffic safety and operational performance study of large trucks on mountainous interstate highway. Accident Analysis and Prevention 43, 429–438. doi:10.1016/j.aap.2010.09.013

Chib, S., Winkelmann, R., 2001. Markov Chain Monte Carlo Analysis of Correlated Count Data. Journal of Business & Economic Statistics 19, 428–435. doi:10.1198/07350010152596673

Chimba, D., Sando, T., 2009. The Prediction of Highway Traffic Accident Injury Severity with Neuromorphic Techniques. Advances in Transportation Studies 19, 17–26.

Chin, H.C., Quddus, M.A., 2003. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. Accident Analysis and Prevention 35, 253–259. doi:10.1016/S0001-4575(02)00003-9

Chiou, Y.-C., Hwang, C.-C., Chang, C.-C., Fu, C., 2013. Modeling two-vehicle crash severity by a bivariate generalized ordered probit approach. Accident Analysis and Prevention 51, 175–184. doi:10.1016/j.aap.2012.11.008

Christoforou, Z., Cohen, S., Karlaftis, M.G., 2010. Vehicle occupant injury severity on highways: an empirical investigation. Accident Analysis and Prevention 42, 1606–1620. doi:10.1016/j.aap.2010.03.019

Congdon, P., 2005. Bayesian Models for Categorical Data. John Wiley & Sons, Chichester, UK.

Conway, K.S., Kniesner, T.J., 1991. The important econometric features of a linear regression model with cross-correlated random coefficients. Economics Letters 35, 143–147. doi:10.1016/0165-

1765(91)90161-D

Cragg, J.G., 1971. Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods. Econometrica 39, 829–844.

Davis, G. a., 2004. Possible aggregation biases in road safety research and a mechanism approach to accident modeling. Accident Analysis and Prevention 36, 1119–1127. doi:http://dx.doi.org/10.1016/j.aap.2004.04.002

Dias, C., Miska, M., 2009. Relationship between congestion and traffic accidents on expressways: an investigation with Bayesian belief networks. Proceedings of 40th Annual Meeting of Infrastructure Planning (JSCE).

Dinu, R.R., Veeraragavan, A., 2011. Random parameter models for accident prediction on two-lane undivided highways in India. Journal of Safety Research 42, 39–42. doi:10.1016/j.jsr.2010.11.007

Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2014a. Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. Accident Analysis and Prevention 70, 320–329. doi:10.1016/j.aap.2014.04.018

Dong, C., Nambisan, S.S., Richards, S.H., Ma, Z., 2015. Assessment of the effects of highway geometric design features on the frequency of truck involved crashes using bivariate regression. Transportation Research Part A: Policy and Practice 75, 30–41. doi:10.1016/j.tra.2015.03.007

Dong, C., Richards, S.H., Clarke, D.B., Zhou, X., Ma, Z., 2014b. Examining signalized intersection crash frequency using multivariate zero-inflated Poisson regression. Safety Science 70, 63–69. doi:10.1016/j.ssci.2014.05.006

Dong, N., Huang, H., Lee, J., Gao, M., Abdel-Aty, M., 2016. Macroscopic hotspots identification: A Bayesian spatio-temporal interaction approach. Accident Analysis and Prevention 92, 256–264. doi:10.1016/j.aap.2016.04.001

Donnell, E.T., Porter, R.J., Shankar, V.N., 2010. A framework for estimating the safety effects of roadway lighting at intersections. Safety Science 48, 1436–1444. doi:10.1016/j.ssci.2010.06.008

Duan, N., Manning, W.G., Morris, C.N., Newhouse, J.P., 1983. A Comparison of Alternative Models for the Demand for Medical Care. Journal of Business & Economic Statistics 1, 115–126.

Dulisse, B., 1997. Methodological issues in testing the hypothesis of risk compensation. Accident Analysis and Prevention 29, 285–292. doi:10.1016/S0001-4575(96)00082-6

Eisenberg, D., 2004. The mixed effects of precipitation on traffic crashes. Accident Analysis and Prevention 36, 637–647. doi:10.1016/S0001-4575(03)00085-X

El-Basyouny, K., Barua, S., Islam, T., 2014. Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson lognormal models. Accident Analysis and Prevention 73, 91–99.

El-Basyouny, K., Sayed, T., 2009a. Collision prediction models using multivariate Poisson-lognormal regression. Accident Analysis and Prevention 41, 820–828. doi:10.1016/j.aap.2009.04.005

El-Basyouny, K., Sayed, T., 2009b. Accident prediction models with random corridor parameters. Accident Analysis and Prevention 41, 1118–1123. doi:10.1016/j.aap.2009.06.025

Eluru, N., Bagheri, M., Miranda-Moreno, L.F., Fu, L., 2012. A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings. Accident Analysis and Prevention 47, 119–127. doi:10.1016/j.aap.2012.01.027

Freedman, D.A., 1999. Ecological Inference and the Ecological Fallacy, Proceedings of the National Academy of Sciences (USA). Berkeley, CA. doi:10.1073/pnas.96.19.10578

Geedipally, S.R., Lord, D., 2010a. Investigating the effect of modeling single-vehicle and multi-vehicle crashes separately on confidence intervals of Poisson-gamma models. Accident Analysis and Prevention 42, 1273–1282. doi:10.1016/j.aap.2010.02.004

Geedipally, S.R., Lord, D., 2010b. Identifying Hot Spots by Modeling Single-Vehicle and Multivehicle Crashes Separately. Transportation Research Record: Journal of the Transportation Research Board 2147, 97–104. doi:10.3141/2147-12

Gelman, A., Meng, X.-L., Stern, H., 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica 6, 733–807. doi:10.1.1.142.9951

Gkritza, K., Mannering, F.L., 2008. Mixed logit analysis of safety-belt use in single- and multi-occupant vehicles. Accident Analysis and Prevention 40, 443–451. doi:10.1016/j.aap.2007.07.013

Greene, W., 2008. Econometric Analysis, 6th ed. Prentice Hall.

Greene, W., 2012. Limdep, Version 10.0. Econometric Software, Inc., Plainview, NY.

Hadayeghi, A., Shalaby, A.S., Persaud, B.N., 2010. Development of planning level transportation safety tools using Geographically Weighted Poisson Regression. Accident Analysis and Prevention 42, 676–688. doi:10.1016/j.aap.2009.10.016

Halton, J.H., 1960. On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. Numerische Mathematik 2, 84–90. doi:10.1007/BF01386213

Hossain, M., Muromachi, Y., 2013. Understanding crash mechanism on urban expressways using high-resolution traffic data. Accident Analysis and Prevention 57, 17–29. doi:10.1016/j.aap.2013.03.024

Hosseinpour, M., Yahaya, A.S., Sadullah, A.F., 2014. Exploring the effects of roadway characteristics on the frequency and severity of head-on crashes: case studies from Malaysian Federal Roads. Accident Analysis and Prevention 62, 209–222. doi:10.1016/j.aap.2013.10.001

Huang, H., Chin, H.C., 2010. Modeling road traffic crashes with zero-inflation and site-specific random effects. Statistical Methods & Applications (Journal of the Italian Statistical Society) 19, 445–462.

Islam, M., Hernandez, S., 2013. Large Truck–Involved Crashes: Exploratory Injury Severity Analysis. Journal of Transportation Engineering 139, 596–604. doi:10.1061/(ASCE)TE.1943-5436.0000539.

Islam, M., Hernandez, S., 2013. Modeling injury outcomes of crashes involving heavy vehicles on Texas highways. Transportation Research Record: Journal of the Transportation Research Board 2338, 28–36.

Ivan, J.N., Pasupathy, R.K., Ossenbruggen, P.J., 1999. Differences in causality factors for single and multi-vehicle crashes on two-lane roads. Accident Analysis and Prevention 31, 695–704. doi:10.1016/S0001-4575(99)00030-5

Johansson, P., 1996. Speed limitation and motorway casualties: a time series count data regression approach. Accident Analysis and Prevention 28, 73–87. doi:10.1016/0001-4575(95)00043-7

Jones, S., Hensher, D. a., 2007. Evaluating the Behavioural Performance of Alternative Logit Models: An

Application to Corporate Takeovers Research. Journal of Business Finance & Accounting 34, 1193–1220. doi:10.1111/j.1468-5957.2007.02049.x

Jonsson, T., Ivan, J.N., Zhang, C., 2007. Crash Prediction Models for Intersections on Rural Multilane Highways: Differences by Collision Type. Transportation Research Record: Journal of the Transportation Research Board 2019, 91–98. doi:10.3141/2019-12

Keay, K., Simmonds, I., 2005. The association of rainfall and other weather variables with road traffic volume in Melbourne, Australia. Accident Analysis and Prevention 37, 109–124. doi:10.1016/j.aap.2004.07.005

Knorr-Held, L., Besag, J., 1998. Modeling risk from a disease in time and space. Statistics in Medicine 17, 2045–2060.

Kononov, J., Bailey, B., Allery, B.K., 2008. Relationships Between Safety and Both Congestion and Number of Lanes on Urban Freeways. Transportation Research Record: Journal of the Transportation Research Board 2083, 26–39. doi:10.3141/2083-04

Ladrón de Guevara, F., Washington, S., Oh, J., 2004. Forecasting Crashes at the Planning Level: Simultaneous Negative Binomial Crash Model Applied in Tucson, Arizona. Transportation Research Record: Journal of the Transportation Research Board 1897, 191–199. doi:10.3141/1897-25

Lao, Y., Zhang, G., Wang, Y., Milton, J., 2014. Generalized nonlinear models for rear-end crash risk analysis. Accident Analysis and Prevention 62, 9–16. doi:10.1016/j.aap.2013.09.004

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

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

Li, G., Best, N., Hansell, A.L., Ahmed, I., Richardson, S., 2012. BaySTDetect: Detecting unusual temporal patterns in small area data via Bayesian model choice. Biostatistics 13, 695–710.

doi:10.1093/biostatistics/kxs005

Liu, L., Strawderman, R.L., Johnson, B. a., O'Quigley, J.M., 2012. Analyzing repeated measures semi-continuous data, with application to an alcohol dependence study. Statistical Methods in Medical Research. doi:10.1177/0962280212443324

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, 291–305. doi:10.1016/j.tra.2010.02.001

Lord, D., Persaud, B.N., 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. doi:10.3141/1717-13

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, 35–46. doi:10.1016/j.aap.2004.02.004

Ma, J., Kockelman, K.M., 2006. Bayesian Multivariate Poisson Regression for Models of Injury Count , by Severity. Transportation Research Record: Journal of the Transportation Research Board 1950, 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 and Prevention 40, 964–975. doi:10.1016/j.aap.2007.11.002

Ma, X., Chen, F., Chen, S., 2015a. Modeling Crash Rates for a Mountainous Highway Using Refined-Scale Panel Data. Transportation Research Record: Journal of the Transportation Research Board 2515, 10–16.

Ma, X., Chen, F., Chen, S., 2015b. Empirical Analysis of Crash Injury Severity on Mountainous and Nonmountainous Interstate Highways. Traffic Injury Prevention 16, 715–723. doi:10.1080/15389588.2015.1010721

Ma, X., Chen, S., Chen, F., 2016. Correlated Random-Effects Bivariate Poisson Lognormal Model to

Study Single-Vehicle and Multivehicle Crashes. Journal of Transportation Engineering. doi:10.1061/(ASCE)TE

MacNab, Y.C., 2004. Bayesian spatial and ecological models for small-area accident and injury analysis. Accident Analysis and Prevention 36, 1019–1028. doi:10.1016/j.aap.2002.05.001

Maddala, G.S., Lahiri, K., 1992. Introduction to econometrics, 2nd ed. Macmillan, New York.

Malyshkina, N. V., Mannering, F.L., Tarko, A.P., 2009. Markov switching negative binomial models: An application to vehicle accident frequencies. Accident Analysis and Prevention 41, 217–226. doi:10.1016/j.aap.2008.11.001

Malyshkina, N. V, Mannering, F.L., 2010a. Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents. Accident Analysis and Prevention 42, 131–139. doi:10.1016/j.aap.2009.07.013

Malyshkina, N. V, Mannering, F.L., 2010b. Zero-state Markov switching count-data models: an empirical assessment. Accident Analysis and Prevention 42, 122–130. doi:10.1016/j.aap.2009.07.012

Mannering, F., 2007. The effects of interstate speed limits on driving speeds: Some new evidence, in: Compendium of Papers CD-ROM, Transportation Research Board 86th Annual Meeting. Washington, D.C.

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. doi:10.1016/j.amar.2013.09.001

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. doi:10.1016/j.amar.2016.04.001

Mansky, C., McFadden, D., 1981. Econometric Models for Probabilistic Choice, in: Structural Analysis of Discrete Data with Econometric Applications. MIT Press, Cambridge, MA, pp. 198–272.

Mardia, K.., 1988. Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. Journal of Multivariate Analysis 24, 265–284. doi:10.1016/0047-259X(88)90040-1

Martensen, H., Dupont, E., 2013. Comparing single vehicle and multivehicle fatal road crashes: A joint analysis of road conditions, time variables and driver characteristics. Accident Analysis and Prevention 60, 466–471. doi:10.1016/j.aap.2013.03.005

McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. Journal of Applied Econometrics 15, 447–470.

Miaou, S.P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis and Prevention 26, 471–482. doi:10.1016/0001-4575(94)90038-8

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, 33–57.

Miaou, S.S.-P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. Transportation Research Record: Journal of the Transportation Research Board 1840, 31–40. doi:10.3141/1840-04

Milton, J., Mannering, F., 1998. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. Transportation 25, 395–413. doi:10.1023/a:1005095725001

Milton, J.C., Shankar, V.N., Mannering, F.L., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accident Analysis and Prevention 40, 260–266. doi:10.1016/j.aap.2007.06.006

Mohammadi, M. a., Samaranayake, V. a., Bham, G.H., 2014. Crash frequency modeling using negative binomial models: An application of generalized estimating equation to longitudinal data. Analytic Methods in Accident Research 2, 52–69. doi:10.1016/j.amar.2014.07.001

Morgan, A., Mannering, F.L., 2011. The effects of road-surface conditions, age, and gender on driver-injury severities. Accident Analysis and Prevention 43, 1852–1863. doi:10.1016/j.aap.2011.04.024

Neelon, B.H., O'Malley, a. J., Normand, S.-L.T., 2010. A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. Statistical Modelling 10, 421–439. doi:10.1177/1471082X0901000404

NHTSA, 2010. Traffic Safety Facts 2010: A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System, USDOT. doi:DOT HS 811 659

NHTSA, 2014. Traffic Safety Facts 2013. National Highway Traffic Safety Administration, National Center for Statistics and Analysis, U.S. Department of Transportation, Washington, D.C.

Noland, R.B., 2003. Traffic fatalities and injuries: the effect of changes in infrastructure and other trends. Accident Analysis and Prevention 35, 599–611. doi:10.1016/S0001-4575(02)00040-4

Noland, R.B., Oh, L., 2004. The effect of infrastructure and demographic change on traffic-related fatalities and crashes: a case study of Illinois county-level data. Accident Analysis and Prevention 36, 525–532. doi:10.1016/S0001-4575(03)00058-7

Öström, M., Eriksson, A., 1993. Single-vehicle crashes and alcohol: A retrospective study of passenger car fatalities in Northern Sweden. Accident Analysis and Prevention 25, 171–176. doi:10.1016/0001-4575(93)90057-4

Paleti, R., Eluru, N., Bhat, C.R., 2010. Examining the influence of aggressive driving behavior on driver injury severity in traffic crashes. Accident Analysis and Prevention 42, 1839–1854. doi:10.1016/j.aap.2010.05.005

Parisi, M., Sembenelli, A., 2003. Is Private R & D Spending Sensitive to Its Price? Empirical Evidence on Panel Data for Italy. Empirica 30, 357–377.

Park, B.-J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accident Analysis and Prevention 41, 683–691. doi:10.1016/j.aap.2009.03.007

Park, E.S., 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. doi:10.3141/2019-01

Pei, X., Wong, S.C., Sze, N.N., 2012. The roles of exposure and speed in road safety analysis. Accident Analysis and Prevention 48, 464–471. doi:10.1016/j.aap.2012.03.005

Pinheiro, J.C., Bates, D.M., 1995. Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. Journal Of Computational And Graphical Statistics 4, 12–35.

doi:10.2307/1390625

Poch, M., Mannering, F.L., 1996. Negative Binomial Analysis of Intersection-Accident Frequencies. Journal of Transportation Engineering 122, 105–113.

Qi, Y., Smith, B.L., Guo, J., 2007. Freeway Accident Likelihood Prediction Using a Panel Data Analysis Approach. Journal of Transportation Engineering 133, 149–156. doi:10.1061/(ASCE)0733-947X(2007)133:3(149)

Quddus, M.A., 2008. Time series count data models: an empirical application to traffic accidents. Accident Analysis and Prevention 40, 1732–1741. doi:10.1016/j.aap.2008.06.011

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

Riphahn, R.T., Wambach, A., Million, A., 2003. Incentive effects in the demand for health care: A bivariate panel count data estimation. Journal of Applied Econometrics 18, 387–405. doi:10.1002/jae.680

Savolainen, P., Mannering, F., 2007. Probabilistic models of motorcyclists’ injury severities in single- and multi-vehicle crashes. Accident Analysis and Prevention 39, 955–963. doi:10.1016/j.aap.2006.12.016

Savolainen, P.T., Mannering, F.L., 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, 1666–1676. doi:10.1016/j.aap.2011.03.025

Shankar, V., Albin, R., Milton, J., Mannering, F., 1998. Evaluating Median Crossover Likelihoods with Clustered Accident Counts: An Empirical Inquiry Using the Random Effects Negative Binomial Model. Transportation Research Record: Journal of the Transportation Research Board 1635, 44–48. doi:10.3141/1635-06

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

Shankar, V., Mannering, F., Barfield, W., 1996. Statistical Analysis of Accident Rural Freeways. Accident Analysis and Prevention 28, 391–401.

Shankar, V., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. Accident Analysis and Prevention 29, 829–837. doi:10.1016/S0001-4575(97)00052-3

Shankar, V.N., Ulfarsson, G.F., Pendyala, R.M., Nebergall, M.B., 2003. Modeling crashes involving pedestrians and motorized traffic. Safety Science 41, 627–640. doi:10.1016/S0925-7535(02)00017-6

Sigelman, L., Zeng, L., 1999. Analyzing censored and sample-selected data with Tobit and Heckit models. Political Analysis 8, 167–182. doi:10.1093/oxfordjournals.pan.a029811

Sittikariya, S., Shankar, V., 2009. Critical Modeling Issues in Transportation Safety Programming: Comprehensive View from Academia and Practitioners. Journal of the Asian Transportation Society 1, 39–45.

Smith, V. a., Preisser, J.S., Neelon, B., Maciejewski, M.L., 2014. A marginalized two-part model for semicontinuous data. Statistics in Medicine 33, 4891–4903. doi:10.1002/sim.6263

Spiegelhalter, D., Thomas, A., Best, N., Lunn, D., 2005. WinBUGS User Mannual. 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, 583–639. doi:10.1111/1467-9868.00353

Su, L., Tom, B.D.M., Farewell, V.T., 2009. Bias in 2-part mixed models for longitudinal semicontinuous data. Biostatistics 10, 374–389. doi:10.1093/biostatistics/kxn044

Theofilatos, A., Yannis, G., 2014. A review of the effect of traffic and weather characteristics on road safety. Accident Analysis and Prevention 72, 244–256. doi:10.1016/j.aap.2014.06.017

Tobin, J., 1958. Estimation of Relationships for Limited Dependent Variables. Econometrica 26, 24–26. doi:10.2307/1907382

Tooze, J. a, Grunwald, G.K., Jones, R.H., 2002. Analysis of repeated measures data with clumping at zero.

Statistical Methods in Medical Research 11, 341–355. doi:10.1191/0962280202sm291ra

Train, K., 1999. Halton sequences for mixed logit. Working Paper. University of California, Department of Economics, Berkley.

Train, K., 2003. Discrete Choice Methods with Simulation, 1st ed. Cambridge University Press.

Treat, J.R., 1980. A study of pre crash factors involved in traffic accidents, Research Review.

Ulfarsson, G., Shankar, V., 2003. Accident Count Model Based on Multiyear Cross-Sectional Roadway Data with Serial Correlation. Transportation Research Record: Journal of the Transportation Research Board 1840, 193–197. doi:10.3141/1840-22

Ulfarsson, G.F., Mannering, F.L., 2004. Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents. Accident Analysis and Prevention 36, 135–147.

Usman, T., Fu, L., Miranda-Moreno, L.F., 2010. Quantifying safety benefit of winter road maintenance: accident frequency modeling. Accident Analysis and Prevention 42, 1878–1887. doi:10.1016/j.aap.2010.05.008

Usman, T., Fu, L., Miranda-Moreno, L.F., 2011. Accident Prediction Models for Winter Road Safety. Transportation Research Record: Journal of the Transportation Research Board 2237, 144–151. doi:10.3141/2237-16

Usman, T., Fu, L., Miranda-Moreno, L.F., 2012. A disaggregate model for quantifying the safety effects of winter road maintenance activities at an operational level. Accident Analysis and Prevention 48, 368–378. doi:10.1016/j.aap.2012.02.005

Venkataraman, N., Shankar, V., Ulfarsson, G.F., Deptuch, D., 2014. A heterogeneity-in-means count model for evaluating the effects of interchange type on heterogeneous influences of interstate geometrics on crash frequencies. Analytic Methods in Accident Research 2, 12–20. doi:10.1016/j.amar.2014.01.001

Venkataraman, N., Ulfarsson, G.F., Shankar, V.N., 2013. Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type. Accident Analysis and Prevention 59, 309–18. doi:10.1016/j.aap.2013.06.021

Vuong, Q.H., 1989. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica 57, 307–333.

Wang, C., Quddus, M. a., Ison, S.G., 2013a. The effect of traffic and road characteristics on road safety: A review and future research direction. Safety Science 57, 264–275. doi:10.1016/j.ssci.2013.02.012

Wang, C., Quddus, M., Ison, S., 2013b. A spatio-temporal analysis of the impact of congestion on traffic safety on major roads in the UK. Transportmetrica A: Transport Science 9, 124–148. doi:10.1080/18128602.2010.538871

Wang, X., Abdel-Aty, M., 2006. Temporal and spatial analyses of rear-end crashes at signalized intersections. Accident Analysis and Prevention 38, 1137–1150. doi:10.1016/j.aap.2006.04.022

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

Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2011. Statistical and Econometric Methods for Transportation Data Analysis, 2nd ed. CRC Press.

Weiss, H.B., Kaplan, S., Prato, C.G., 2014. Analysis of factors associated with injury severity in crashes involving young New Zealand drivers. Accident Analysis and Prevention 65, 142–155. doi:10.1016/j.aap.2013.12.020

Winston, C., Maheshri, V., Mannering, F., 2006. An exploration of the offset hypothesis using disaggregate data: The case of airbags and antilock brakes. Journal of Risk and Uncertainty 32, 83–90. doi:10.1007/s11166-006-8288-7

World Health Organization, 2013. Global status report on road safety, World Health Organization. Geneva, Switzerland. doi:http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/

Xie, K., Wang, X., Ozbay, K., Yang, H., 2014. Crash frequency modeling for signalized intersections in a high-density urban road network. Analytic Methods in Accident Research 2, 39–51.

doi:10.1016/j.amar.2014.06.001

Xie, Y., Zhang, Y., Liang, F., 2009. Crash Injury Severity Analysis Using Bayesian Ordered Probit Models. Journal of Transportation Engineering 135, 18–25. doi:10.1061/(ASCE)0733-947X(2009)135:1(18)

Xie, Y., Zhao, K., Huynh, N., 2012. Analysis of driver injury severity in rural single-vehicle crashes. Accident Analysis and Prevention 47, 36–44. doi:10.1016/j.aap.2011.12.012

Xiong, Y., Mannering, F.L., 2013. The heterogeneous effects of guardian supervision on adolescent driver-injury severities: A finite-mixture random-parameters approach. Transportation Research Part B: Methodological 49, 39–54. doi:10.1016/j.trb.2013.01.002

Xu, C., Wang, W., Liu, P., 2013. Identifying crash-prone traffic conditions under different weather on freeways. Journal of Safety Research 46, 135–144. doi:10.1016/j.jsr.2013.04.007

Xu, X., Kouhpanejade, A., Šarić, Ž., 2013. Analysis on Influencing Factors Identification of Crash Rates Using Tobit Model with Endogenous Variable. PROMET - Traffic & Transportation 25, 217–224. doi:10.7307/ptt.v25i3.291

Yaacob, W., Lazim, M., Wah, Y., 2010. Evaluating spatial and temporal effects of accidents likelihood using random effects panel count model, in: Proceedings of the 2010 International Conference on Science and Social Research. Kuala Lumpur, Malaysia.

Yamamoto, T., Shankar, V.N., 2004. Bivariate ordered-response probit model of driver's and passenger's injury severities in collisions with fixed objects. Accident Analysis and Prevention 36, 869–876. doi:10.1016/j.aap.2003.09.002

Yasmin, S., Eluru, N., 2013. Evaluating alternate discrete outcome frameworks for modeling crash injury severity. Accident Analysis and Prevention 59, 506–521. doi:10.1016/j.aap.2013.06.040

Ye, F., Lord, D., 2014. Comparing three commonly used crash severity models on sample size requirements: Multinomial logit, ordered probit and mixed logit models. Analytic Methods in Accident Research 1, 72–85. doi:10.1016/j.amar.2013.03.001

Yu, R., Abdel-Aty, M., 2013. Multi-level Bayesian analyses for single- and multi-vehicle freeway crashes.

Accident Analysis and Prevention 58, 97–105. doi:10.1016/j.aap.2013.04.025

Yu, R., Abdel-Aty, M., 2014a. Using hierarchical Bayesian binary probit models to analyze crash injury severity on high speed facilities with real-time traffic data. Accident Analysis and Prevention 62, 161–167. doi:10.1016/j.aap.2013.08.009

Yu, R., Abdel-Aty, M., 2014b. Analyzing crash injury severity for a mountainous freeway incorporating real-time traffic and weather data. Safety Science 63, 50–56. doi:10.1016/j.ssci.2013.10.012

Yu, R., Abdel-Aty, M.A., Ahmed, M.M., Wang, X., 2013a. Crash-Type Propensity Analysis with Bayesian Models Using Microscopic Traffic and Weather Data, in: Transportation Research Board 92nd Annual Meeting, Transportation Research Board of the National Academies.

Yu, R., Abdel-Aty, M., Ahmed, M., 2013b. Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. Accident Analysis and Prevention 50, 371–376. doi:10.1016/j.aap.2012.05.011

Yu, R., Xiong, Y., Abdel-Aty, M., 2015. A correlated random parameter approach to investigate the effects of weather conditions on crash risk for a mountainous freeway. Transportation Research Part C: Emerging Technologies 50, 68–77. doi:10.1016/j.trc.2014.09.016

APPENDIX

This dissertation is based on six research papers which have already been published, accepted or currently under review.

a. paper published in refered journals

Ma, X., Chen, S., Chen, F. (2016). Correlated Random-Effects Bivariate Poisson Lognormal Model to Study Single-Vehicle and Multivehicle Crashes. Journal of Transportation Engineering (Accepted).

Ma, X., Chen, F., Chen, S. (2015). Modeling Crash Rates for a Mountainous Highway Using Refined-Scale Panel Data. Transportation Research Record: Journal of the Transportation Research Board 2515, 10–16.

Ma, X., Chen, F., Chen, S. (2015). Empirical Analysis of Crash Injury Severity on Mountainous and Nonmountainous Interstate Highways. Traffic Injury Prevention 16, 715–723.

Chen, F., **Ma, X.**, Chen, S. (2014). Refined-scale panel data crash rate analysis using random-effects tobit model. Accident Analysis and Prevention 73, 323–332.

b. paper submitted to refered journals

Ma, X., Chen, S., Chen, F. (2016). Correlated Random Parameter Marginalized Two-Part Model: an Application to Refined-Scale Longitudinal Crash Rates. Under Review.

Ma, X., Chen, S., Chen, F. (2016). Multivariate Space-Time Modeling of Crash Frequencies by Injury Severity Levels. Under Review.