2-1-2016

DATA-DRIVEN BAYESIAN METHOD-BASED TRAFFIC CRASH DRIVER INJURY SEVERITY FORMULATION, ANALYSIS, AND INFERENCE

Cong Chen

Cong Chen
Candidate

Department of Civil Engineering
Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Dr. Guohui Zhang , Chairperson

Dr. Susan Bogus Halter

Dr. Yin Yang

Dr. Rafiqul A. Tarefder

DATA-DRIVEN BAYESIAN METHOD-BASED TRAFFIC CRASH DRIVER INJURY SEVERITY FORMULATION, ANALYSIS, AND INFERENCE

by

CONG CHEN

B.S., Transportation Engineering, Tongji University, 2008

M.S., Highway and Railway Engineering, Tongji University, 2011

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy

Engineering

The University of New Mexico
Albuquerque, New Mexico

December, 2015

ACKNOWLEDGEMENTS

First and foremost, I heartily acknowledge Dr. Guohui Zhang, my advisor and dissertation chair, for his continuous professional and personal help with thoughtful guidance. He has been a great source of knowledge, enthusiasm and encouragement during my entire Ph.D. study. I benefited a lot from his insightful advice on my professional and personal development. His guidance and professional style will remain with me as I continue my career.

I would like to express my gratitude to Dr. Susan Bogus Halter, who is my doctoral committee member. Her expertise and experience in construction and transportation fields are invaluable and I benefited a lot on my major/minor studies. I greatly appreciate her consistent support on me during my dissertation work as well as the entire Ph.D. study.

I would like to express my sincere thanks to Dr. Yin Yang and Dr. Rafiqul Tarefder for their instructions and for serving on my dissertation committee. They spent considerable amount of time on my work and provided a lot of valuable advice. I benefited a lot from their advices and improved my dissertation work significantly.

To my supervisors in New Mexico Department of Transportation (NMDOT), Mr. Timothy Parker, Mr. Antonio Jaramillo, and Mrs. Nancy Perea, and all my colleagues in NMDOT, for their instructions and advice during my internship there, I do thank you from the bottom of my heart.

I wish to express my thanks to my friends and student fellows in the Department of Civil Engineering, University of New Mexico, for their help and advice. Special

thanks are given to Mr. Michael Angel Gonzalez, Mr. Su Zhang, Mr. David Barboza, Ms. Kelly Montoya, Mr. Fei Han, and Mrs. Jielin Pan. I would like to express my gratitude to all of them.

Finally, I would like to thank my family for their persistent support all the time.

DATA-DRIVEN BAYESIAN METHOD-BASED TRAFFIC CRASH DRIVER INJURY SEVERITY FORMULATION, ANALYSIS, AND INFERENCE

by

Cong Chen

B.S., TRANSPORTATION ENGINEERING

M.S., HIGHWAY AND RAILWAY ENGINEERING

DOCTOR OF PHILOSOPHY IN ENGINEERING

with Concentration in Transportation and Traffic Engineering, Civil Engineering

ABSTRACT

Traffic crashes have resulted in significant cost to society in terms of life and economic losses, and comprehensive examination of crash injury outcome patterns is of practical importance. By inferring the parameters of interest from prior information and studied datasets, Bayesian models are efficient methods in data analysis with more accurate results, but their applications in traffic safety studies are still limited. By examining the driver injury severity patterns, this research is proposed to systematically examine the applicability of Bayesian methods in traffic crash driver injury severity prediction in traffic crashes. In this study, three types of Bayesian models are defined: hierarchical Bayesian regression model, Bayesian non-regression model and knowledge-based Bayesian non-parametric model, and a conceptual framework is developed for selecting the appropriate Bayesian model based on discrete research purposes.

Five Bayesian models are applied accordingly to test their effectiveness in traffic crash driver injury severity prediction and variable impact estimation: hierarchical Bayesian binary logit model, hierarchical Bayesian ordered logit model, hierarchical Bayesian random intercept model with cross-level interactions, multinomial logit (MNL)-Bayesian Network (BN) model, and decision table/naïve Bayes (DTNB) model. A complete dataset containing all crashes occurring on New Mexico roadways in 2010 and 2011 is used for model analyses. The studied dataset is composed of three major sub-datasets: crash dataset, vehicle dataset and driver dataset, and all included variables are therefore divided into two hierarchical levels accordingly: crash-level variables and vehicle/driver variables.

From all these five models, the model performance and analysis results have shown promising performance on injury severity prediction and variable influence analysis, and these results underscore the heterogeneous impacts of these significant variables on driver injury severity outcomes. The performances of these models are also compared among these methods or with traditional traffic safety models. With the analyzed results, tentative suggestions regarding countermeasures and further research efforts to reduce crash injury severity are proposed. The research results enhance the understandings of the applicability of Bayesian methods in traffic safety analysis and the mechanisms of crash injury severity outcomes, and provide beneficial inference to improve safety performance of the transportation system.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

1.1 Background

1.1.1 General Background

Traffic crashes have resulted in significant cost to society in terms of fatalities, serious injuries, and property losses. Statistical data show that approximately 1.24 million people are killed and 50 million people are injured each year in traffic crashes worldwide (World Health Organization, 2013). In the U.S., there were 5.6 million reported traffic crashes in 2012, resulting in 33,561 deaths and 2,362,000 injuries (National Highway Traffic Safety Adminstriation(NHTSA), 2013), and each fatality and incapacitating injury, on average, cost approximately $1.42 million and $78,700, respectively (National Health Council, 2013). Specific patterns are also revealed from national crash data. According to NHTSA (2013), 29% of total roadway crashes result in an injury and less than 1% result in a death. 54% of total fatal crashes and 55% of total fatalities occur in US rural areas, where only 19% of the total population is living. With regard to crash types, 61% of fatal crashes are single-vehicle crashes, and these numbers are 32% for injury crashes and 30% for property-damage-only crashes, respectively. Thirty percent of fatal crashes were associated with alcohol-impaired driving.

Significant development in the automotive industry and numerous implementations of national road safety strategies have been made to reduce the frequency and injury severity of traffic crashes by conducting peer research and applying target-oriented countermeasures. At the national strategy level, the Federal Highway Administration (FHWA) proposed numerous traffic safety strategies regarding three

major aspects-management, human resource and technology-to enhance traffic safety and traffic operation efficiency, such as Traffic Safety Management Functions (TSMF), Intelligent Transportation System (ITS), Variable Speed Limits (VSL), etc. (NHTSA, 2001). While at the research level, considerable studies have been conducted to examine crash mechanisms, identify contributing factors to crash frequency and severity, and propose effective countermeasures to reduce both crash occurrences and injury outcomes. Further studies also focus on the characteristics of crashes regarding environmental and geometric features, vehicle situations as well as driver behaviors.

1.1.2 Traffic Safety Analysis

At the beginning of the twentieth century, traffic crashes were believed to be occasional and unpredictable (Riviere et al., 2006). With the development of the automobile industry and statistical modeling techniques in traffic safety analyses, a traffic crash is conventionally considered as a consequence of the complicated interactions of factors related to major components: roadway and environment characteristics, vehicle characteristics and human factors (Hossain and Muromachi, 2012). In recent years, traffic dynamics is proposed to be the fourth contributing component to traffic crashes, suggesting that traffic crashes regularly occur due to sudden formation of disrupted traffic conditions even on roadways that meet design standards and under favorable weather conditions.

Due to the significant economic and emotional burden that traffic crashes impose on social welfare, researchers have persistently sought ways to obtain a better

understanding of the factors that affect the frequency of traffic crashes and the degree of injury suffered by those involved in crashes, and propose implementable improvements in vehicle and roadway design to reduce the number of traffic crashes and traffic injury severity levels. In general, crash data are extracted from standard police reports where some minor collisions are under-reported, and the detailed driving data (acceleration, braking, steering information, driver response to stimuli, etc.) and crash data (for example, what might be available from vehicle black-boxes) that would better assist identification of cause and effect relationships with regard to crash probabilities are typically not available. Therefore, researchers have proposed numerous analytic approaches to study the factors that influence the likelihood of a crash occurring or, given that a crash has occurred, the heterogeneous factors that may mitigate or exacerbate the degree of injury suffered by crash-involved road users. To gain such understanding, safety researchers have applied a wide variety of methodological techniques over the years, addressing traffic safety concerns from multiple aspects, such as crash locations, crash types, driver or vehicle types, weather or road conditions, etc. A summary of the modeling approaches applied in traffic safety analyses is provided in Section 1.1.3.

1.1.3 Modeling Approach and Methodology

Lord and Mannering (2010) summarized the data and methodological issues in crash frequency analyses that should be addressed or taken into account in model development and data analyses in the following eleven aspects: over-dispersion, under-dispersion, time-varying explanatory variables, temporal and spatial correlation, low sample-mean and small sample size, injury-severity and crash-type correlation, under-

reported crashes, omitted-variables bias, endogenous variables, functional form, and fixed parameters. To deal with these data and methodological issues associated with crash-frequency data (many of which could compromise the statistical validity of an analysis if not properly addressed), a wide variety of methods have been applied over the years.

Table 1-1 lists the major existing models applied to crash frequency analysis, with a peer study as an example for each model. The advantage as well as disadvantage of each model was discussed by Lord and Mannering (2010).

Table 1-1 Crash Frequency Research Model Summary.

Model Type	Related Study	Model Type	Related Study
Poisson model	Miaou (1994)	Random-effects model	Wang et al. (2009)
Negative binomial/Poisson-Gamma model	Malyshkina and Mannering (2010)	Negative multinomial model	Caliendo et al. (2007)
Poisson-lognormal model	Lord and Miranda-Moreno (2008)	Random-parameter model	Anastasopoulos and Mannering (2009)
Zero-inflated Poisson and negative binomial model	Lord et al. (2007)	Bivariate/multivariate	Ma and Kockelman (2006)
Conway-Maxwell-Poisson model	Lord et al. (2010)	Finite mixture/Markov switching	Park and Lord (2009)
Gamma Model	Oh et al. (2006)	Duration model	Chung (2010)
Generalized estimating equation	Wang and Abdel-Aty (2006)	Hierarchical/multilevel model	Kim et al. (2007)
Generalized additive model	Xie and Zhang (2008)	Neural network, Bayesian network and support vector machine	Li et al. (2008)

On the other hand, Savolainen et al. (2011) summarized data and methodological issues in crash-injury severity analyses from eight aspects, some of which are similar to

the ones for crash frequency analyses: under-reported crashes, ordinal nature of crash and injury severity data, fixed parameters, omitted variable bias, small sample size, endogeneity, within-crash correlation, and spatial and temporal correlations.

Analysis of crash severity can be conducted in different ways for various purposes. Some studies focus on the crash frequencies at specific traffic sites associated with different severity levels (e.g. fatal, serious, slight) and investigate how geometric, traffic, and environmental factors affect the crash severity. While these kind of studies normally take each crash as the subject unit, analysis can also be undertaken based on the driver-vehicle units involved in crashes to examine individual severity.

Over the years, a wide variety of statistical techniques have been used to study crash-injury severities, such as multinomial logit model, ordered logit or probit model, artificial neural network, etc. Table 1-2 provides the primary models used for crash-injury severity analysis, with an application study for each method.

Table 1-2 Crash Injury Severity Research Model Summary.

Model Type	Related Study	Model Type	Related Study
Artificial neural network	Chimba and Sando (2009)	Mixed joint binary logit-ordered logit	Eluru and Bhat (2007)
Bayesian hierarchical binomial logit	Huang et al. (2008)	Multinomial logit	Islam and Mannering (2006)
Bayesian ordered probit	Xie et al. (2009)	Multivariate probit	Winston et al. (2006)
Binary logit and binary probit	Haleem and Abdel-Aty (2010)	Nested Logit	Savolainen and Mannering (2007)
Bivariate binary probit	Lee and Abdel-Aty (2008)	Ordered logit/probit	Wang and Abdel-Aty (2008)
Bivariate ordered probit	de Lapparent (2008)	Partial proportional odds model	Wang et al. (2009)
Classification and regression tree	Chang and Wang (2006)	Mixed logit	Anastasopoulos and Mannering (2011)
Generalized ordered logit	Quddus et al. (2010)	Mixed ordered logit	Srinivasan (2002)

Table 1-2 (Continued)

Model Type	Related Study	Model Type	Related Study
Heterogeneous outcome model	Quddus et al. (2010)	Mixed ordered probit	Christoforou et al. (2010)
Heteroskedastic ordered logit/probit	Lemp et al. (2011)	Sequential binary logit	Dissanayake and Lu (2002)
Log-linear model	Chen and Jovanis (2000)	Sequential binary probit	Yamamoto et al. (2008)
Markov switching multinomial logit	Malyshkina and Mannering (2009)	Sequential logit	Jung et al. (2010)
Mixed generalized ordered logit	Eluru et al. (2008)	Simultaneous binary logit	Ouyang et al. (2002)

1.1.4 Applications of Bayesian Estimation Methods in Traffic Safety Analyses

Traffic safety engineers are among the early users of Bayesian estimation methods for analyzing crash data (Carriquiry and Pawlovich, 2004). Applications of Bayesian methods in traffic safety analyses are classified into two categories: Bayesian statistical inference and Bayesian network (BN) modeling. Bayesian estimation methods generate a multivariate posterior distribution across all parameters of interest, as opposed to the traditional Maximum Likelihood Estimation (MLE) approach, which emphasizes and offers on the modal values of parameters and relies on asymptotic properties to ascertain covariance. Empirical Bayes (EB) method was the first Bayesian estimation in traffic safety analyses and now has been widely accepted in the field (Cafiso et al., 2010; de Lapparent, 2006; Elvik, 2013, 2008; Hauer, 2001, 1992; Lord and Park, 2008; Persaud and Lyon, 2007; Pulugurtha and Otturu, 2014; Quigley et al., 2011).

However, there are significant drawbacks in the EB approach regarding model assumptions and model time consumption that prevent it from universal applications. Therefore, the Full Bayes (FB) method was proposed for model estimation, in particular for implementations via multi-level hierarchical models. In a full Bayesian analysis, prior

information and all available data are seamlessly integrated into posterior distributions based on expert knowledge, with which all uncertainties are accounted for and there is no need to pre-process data to obtain Safety Performance Functions (SPF) and other such prior estimates of the effect of covariates on the outcome of interest. With these advantages over the EB method, the FB method has been widely applied in traffic safety analysis (Abdalla, 2005; Eksler, 2010; El-Basyouny and Sayed, 2010; Flask and Schneider, 2013; Huang et al., 2008; Ma et al., 2008; MacNab, 2004, 2003; Persaud et al., 2010; Xie et al., 2013; Yanmaz-Tuzel and Ozbay, 2010).

BN is a probability inference method incorporating graphic topology theory and Bayes' Theorem. Gregoriades (2007) highlighted the interest of using BN to model traffic crashes and discussed the need to not consider traffic crashes as a deterministic assessment problem. BNs make it easy to describe crashes that involve many interdependent variables. The relationship and structure of the variables can be studied and trained from crash data, and it is not necessary to know any pre-defined relationships between dependent and independent variables. A BN can be constructed manually, semi-automatically from the data or by a combination of a manual and data driven process (Kjaerulff and Madsen, 2008), and the parameters of the BN is estimated from the database using a learning algorithm, such as the Expectation-Maximization (EM) algorithm. This approach is easily applicable and the learned structure is understandable with expert knowledge involved. Numerous studies have been conducted to analyze traffic crash patterns using BN (Bedeley et al., 2013; Borg et al., 2014; Castillo et al., 2008; Feng and Timmermans, 2013; Goodheart, 2013; Gregoriades and Mouskos, 2013; Jin et al., 2010; Liang and Lee, 2014; Liu et al., 2014; Mbakwe et al., 2014; Mujalli and

de Oña, 2011; Ozbay and Noyan, 2006; Riviere et al., 2006; Zhao et al., 2012). However, searching for an optimal BN classifier in the global space is extremely computation-intensive considering a large amount of independent variables, and it is indispensable to apply a variable selection procedure to find a set of significant contributing variables and screening out redundant ones to achieve feasible and effective network structure estimation. Therefore, several variable selection techniques based on variable correlation or importance ranking are applied to assist BN modeling.

1.2 Problem Statement and Research Objectives

Currently, statistical and mathematical models are major tools used for traffic injury severity analyses. It has been proved in many ways that hierarchical modeling regarding data structure and variable characteristics provides more reliable results in parameter estimates for traffic crash injury analyses. As discussed before, Bayesian estimation methods provide each parameter of interest a posterior density, which is a product of a long series sampling from the posterior distribution and the prior information about the parameter as well as the data. A Bayesian modeling approach provides a considerable interpretive advantage because posterior estimates reflect the probabilities that the analyst is primarily interested in, the probability of the null hypothesis being true (called a posterior credible interval or credible set) (Washington et al., 2005). However, currently studies using hierarchical Bayesian modeling assume crash level heterogeneity to be numerical constants (Huang et al., 2008), rather than established mathematical relationships between crash variables and vehicle variables. Besides, existing studies on traffic crash injury severity, including studies using hierarchical models with FB

estimation, generally consider injury outcome as a binary variable in modeling (Huang et al., 2008; Yu and Abdel-Aty, 2014a), or modified it as an ordered multi-level variable (Huang et al., 2011, 2014), with which the proportional odds assumption are utilized (Congdon, 2005). However, these assumptions may not be suitable for non-monotonic-changing severity data due to the strong model restrictions on the linear relationship between explanatory variables and independent outcomes. Therefore, a more commonly used unordered discrete choice model, hierarchical multinomial logit model, should be used for factor influence examination.

Meanwhile, as mentioned above, it is for most cases assumed that crash driver injury severity or its transformation is a linear regression of its contributing covariates, which may not always be appropriate and universally applied. Non-regression and conditionally probabilistic relationships might exist among driver injury severity and the contributing factors. Hence, Bayesian non-regression models should be applied to investigate dependent relationship between crash driver injury severity and the contributing factors should be assumed and investigated.

Furthermore, several popular knowledge-based non-parametric machine-learning methods, such as artificial neural network (ANN), and classification and regression tree (CART), have been used in traffic safety studies, which is an effective group of methods in crash data analysis. However, no Bayesian concept has been incorporated in this method group. Therefore, this study also aims to propose a knowledge-based Bayesian non-parametric method to examine driver injury severity patterns in traffic crashes.

Overall, this study is proposed to systematically examine the driver injury severity patterns in traffic crashes by developing or applying new Bayesian family models regarding hierarchical regression, non-regression, and non-parametric analyses. The research framework is shown as follows in Figure 1-1.

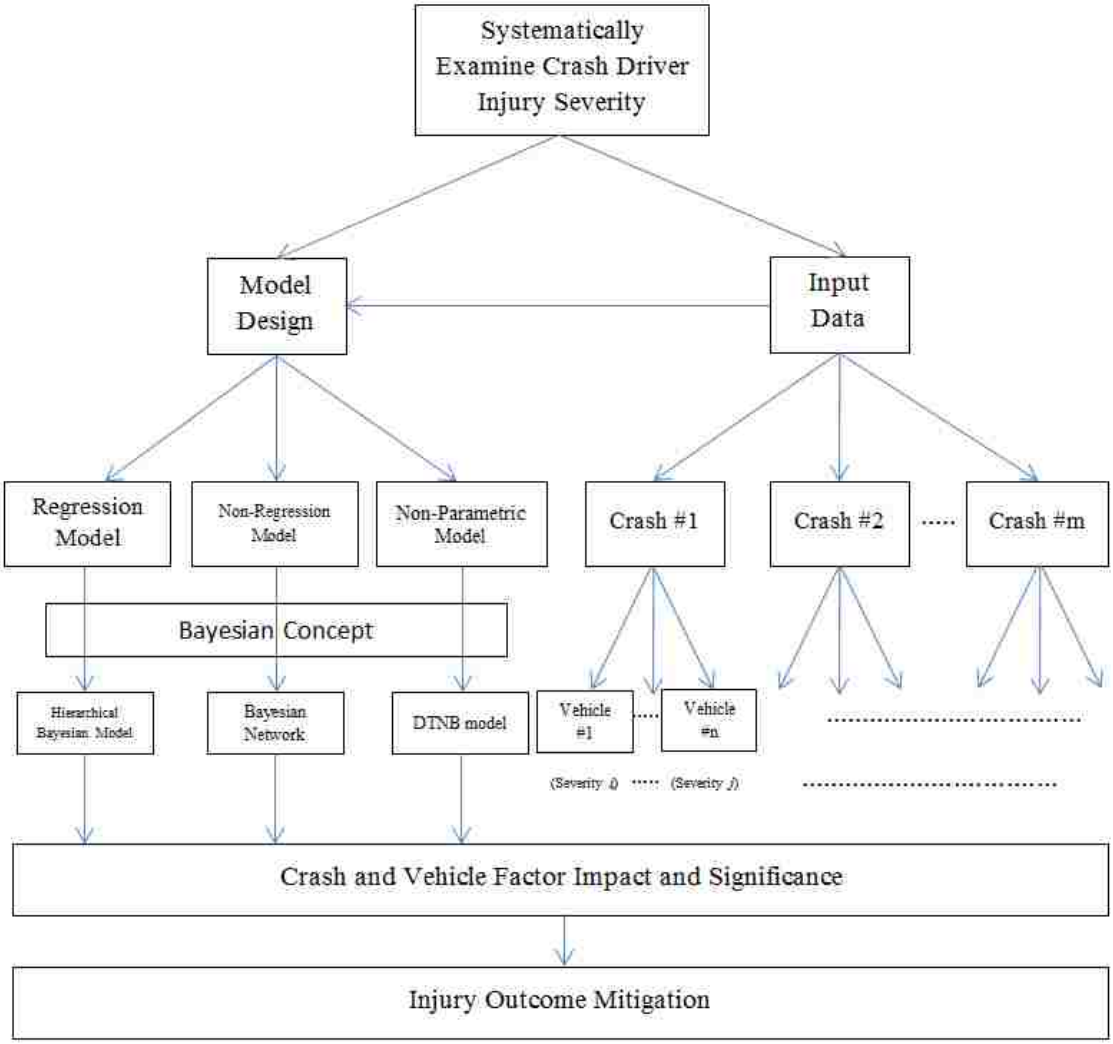

Figure 1-1 Propose Research Framework.

To meet the aim of this study, the following objectives are to be achieved:

1) To develop a methodology framework regarding the appropriate selection of Bayesian family methods on crash data analysis based on distinctive research purpose, data availability and data structure.

2) To summarize existing hierarchical Bayesian regression model structures to better understand and interpret data heterogeneity among crash and vehicle characteristics based on Bayesian inference.

3) To develop and utilize a new hierarchical random intercept model to capture unobserved heterogeneity by systematically examining the cross-level interaction effects between crash-level variables and vehicle/driver-level variables.

4) To develop a new Bayesian non-regression model to predict driver injury severity in traffic crashes and quantify non-regression relationship between significant dependent attributes and independent crash driver injury severity outcomes.

5) To develop a new knowledge-based Bayesian non-parametric model to formulate crash driver injury severity pattern and qualitatively investigate the contributing factors to these injury severity outcomes.

1.3 Dissertation Organization

The remainder of this dissertation is organized in the following manner. Chapter 2 reviews previous work related to this dissertation research. First, the macroscopic and microscopic focuses of traffic safety analysis are introduced and the popular models used in traffic crash frequency analysis are summarized. Then, contemporary work on traffic

injury severity analysis is comprehensively examined. In this section, existing mathematical and machine-learning models that are utilized in traffic injury severity analysis are reviewed, and the contributing factors to crash injury severity, including crash location, crash type, vehicle type, driver characteristics and environment factors, are discussed. Third, peer applications of Bayesian methods in traffic safety analysis are generalized, including Bayesian inference modeling and Bayesian network analysis. Additionally, other data mining techniques such as neural network and classification and regression tree (CART), and their applications in traffic safety analysis, are examined and summarized. Finally, the research explained the unobserved heterogeneity issue in traffic safety research, and examined the popular models and peer studies addressing this issue.

Chapter 3 presents the methodology framework design and the development and specifications of the utilized models in this dissertation. The major aim of this research is to systematically examine the applicability of Bayesian models in traffic crash driver injury severity analysis. Three primary categories of Bayesian methods are defined in this study: hierarchical Bayesian regression models, Bayesian non-regression model, and knowledge-based Bayesian non-parametric model, and a model selection flow chart is developed for the selection of most appropriate model based on discrete data structures and research objectives. Then within each model category, detailed structure design and model specifications of the five utilized models are presented, including hierarchical binary logit model, hierarchical ordered logit model, hierarchical random intercept model with cross-level interactions, MNL-BN hybrid model and decision table/naïve Bayes (DTNB) model.

Using the 2010-2011 New Mexico roadway crash dataset as a base dataset, the applicability and effectiveness of the proposed models are evaluated and the results are discussed in Chapters 4, 5 and 6. Chapter 4 discussed the case studies using three hierarchical Bayesian regression models. A two-year rural interstate crash dataset is modeled by the hierarchical Bayesian binary logit model, where the model fitness is discussed and the influences of heterogeneous contributing factors are assessed based on the estimated posterior coefficients. An extracted rural non-interstate crash dataset is simulated by the hierarchical ordered logit model, where the driver injury severity is defined with 5 monotonically increasing values: no injury, complaint of or possible injury, visible injury, incapacitating injury and fatality. As for the hierarchical random intercept model with cross-level interactions, a dataset of rural truck crashes in 2010 and 2011 is used to examine the applicability of this model and the contributing factors related to truck driver injury severity outcome, which is treated as a 3-level multi-categorical outcome: no injury, non-incapacitating injury, incapacitating injury/death.

Following Chapter 4, Chapter 5 illustrates a case study of the proposed multinomial logit (MNL) -BN hybrid model, where a two-year rear-end crash dataset is used in this analysis to examine driver injury severity patterns. The input variables for BN classifier training are selected through an MNL model and the model performance is evaluated based on classification accuracy, true positive rate, false positive rate, F-measure, receiver operating characteristic (ROC) curve, area under ROC curve (AUC), and classification confusion matrix. The probabilistic influences of contributing factors on driver injury severity are assessed through Bayesian probability inference procedure and are explicitly discussed.

Chapter 6 presents the applicability of the decision table/Naïve Bayes (DTNB) classifier, a representative of the knowledge-based Bayesian non-parametric models, in traffic safety analysis. The same rear-end crash dataset in MNL-BN analysis is also used in this case study. The model performance is also evaluated using the same measurements as used in MNL-BN hybrid analysis, and the variable influences are discussed based on their frequency of values in the extracted decision rules. Additionally, a side-by-side comparison is also conducted to evaluate the performance of the MNL-BN model and the DTNB classifier based on their produced results.

Finally, Chapter 7 provides conclusions of this research effort and recommendations for future research.

Chapter 2 State Of The Art

2.1 General Traffic Safety Analysis and Traffic Frequency Analysis

Traffic safety analyses are conventionally composed of two major parts: traffic crash frequency analyses and traffic crash severity analyses. Traffic crash frequency analyses, partly overlapping with traffic crash severity analyses, help either at a macroscopic level to examine traffic crash frequency on roadway segments for different crash injury severity levels (i.e. property-damage-only, injury and fatality), or at a microscopic level to identify the contributing factors and their respective influences on the probability of each injury severity level in a crash.

With respect to traffic crash frequency analyses, Lord and Mannering (2010) summarized a variety of methodological alternatives that are used in crash frequency studies; strengths and weaknesses of these modeling techniques have been assessed. According to existing crash frequency studies, the major modeling techniques applied are:

Random effect models (including Poisson and negative binomial models) (Chin and Quddus, 2003; Lord, 2006; Shankar et al., 1995, 1998; Yaacob et al., 2010). For example, Shankar et al. (1998) compared Random Effects Negative Binomial (RENB) model with cross-sectional Negative Binomial (NB) model in predicting crash occurrence.

Hierarchical Bayesian models (Huang and Abdel-Aty, 2010; Shively et al., 2010; Xie et al., 2013; Yu and Abdel-Aty, 2014b, 2013). For instance, Huang and Abdel-Aty (2010) argued that traffic safety studies frequently contain multilevel data structures, e.g. [Geographic region level-Traffic site level – Traffic crash level – Driver and vehicle unit level – Occupant level] × Spatio-temporal level.

Tobit model (Anastasopoulos et al., 2012b, 2008; Farah et al., 2009; Lord and Mannering, 2010): For instance, Anastasopoulos et al. (2008) firstly introduced the tobit model to analyze crash rates instead of focusing on crash counts of roadway segments. Crash rates were treated as a continuous variable with left-censored at zero. The authors concluded that tobit regression models had substantial potentials in analyzing crash rate data.

Weather and traffic flow conditions are two major factors related to crash occurrence frequency. Weather conditions are relevant to crash occurrence and researchers have developed several ways to consider weather influences in the crash frequency models (Caliendo et al., 2007; Jung et al., 2010; Malyshkina et al., 2009; Usman et al., 2010; Yaacob et al., 2010). For instance, Caliendo et al. (2007) used hourly rainfall data and transformed them into binary indicators of daily status of the pavement surface ("dry" or "wet"). Traffic variables also play a vital role in crash occurrence (Chang and Chen, 2005; Das and Abdel-Aty, 2011; Kononov et al., 2011; Noland and Quddus, 2004). For example, Kononov et al. (2011) related traffic flow parameters (speed and density) with different functional forms of safety performance functions (SPF) and concluded that (1) on un-congested freeway segments, the numbers of crashes increase only moderately with an increase in traffic; (2) once some critical traffic density was reached, the numbers of crashes would increase at a much faster rate as the increase of traffic.

Efforts were also made to identify factor influence across crash types. Qin et al. (2006) utilized a hierarchical Bayesian framework to predict crash occurrence in relation to the hourly exposure according to four crash types: single-vehicle, multi-vehicle same

direction, multi-vehicle opposite directions, and multi-vehicle intersecting directions. Other previous studies (Jonsson et al., 2009, 2007) have also addressed the crash types' propensity through developing safety performance functions for highway intersections. Results demonstrated that the relationship between traffic flow and crash frequency vary by different crash types; better model fit could be achieved by modeling different crash types separately.

2.2 Traffic Injury Severity Analysis

2.2.1 Traffic Injury Severity Models

A variety of methodological techniques have been applied to analyze crash-severity data, shown in Table 1-2. These methods are affiliated to two major types: statistical regression models, or non-regression data-mining methods. The dependent variables of existing crash severity models are typically either a binary response outcome (e.g., injury or non-injury, or severe or non-severe) or a multiple-response outcome (e.g., fatality, disabling injury, evident injury, possible injury, or no injury). Dependent variables with multiple-response outcomes have been treated as either ordinal (accounting for the ordinal nature of injury data) or nominal (i.e., unordered).

Traffic crash injury severity analyses such as severe vs. non-severe crashes or fatal vs. non-fatal crashes have natural discrete outcomes. Binary logit or probit models (fixed parameter) have been widely employed to analyze crash injury severity (Bedard et al., 2002; Farmer and Lund, 2002; Lee and Abdel-Aty, 2008). However, although modeling procedures and result interpretations of fixed parameter logit models are

straightforward, it is not sufficient to describe relationships between explanatory variables and crash injury severity outcomes. Extensions of the binary logit models (e.g. hierarchical logit model (Huang et al., 2008)) and other non-parametric models (e.g. BN models (de Oña et al., 2011)) were introduced to account for unobserved heterogeneity and non-linearity.

Random parameter logit models (also called mixed logit model) have been extensively used in crash injury severity analyses. Compared to the fixed parameter models, random parameter models account for the unobserved heterogeneity by allowing parameters to vary across observations (Hensher and Greene, 2003). Milton et al. (2008) utilized a random parameter model to investigate the crash severities along with the frequency model. The model allows some variables to vary across different roadway segments and in this way the methodology could account for the unobserved effects (roadway characteristics, environmental factors and driver behavior) on crash severity. Gkritza and Mannering (2008) employed a mixed logit model (model with both fixed and random parameters) to achieve better understandings of the effects of safety belts usages in single- and multi-occupant vehicles. The mixed logit models were used to account for vehicle-specific variations of the independent variables' effects on safety-belt use probabilities. The authors claimed that this approach has its flexibility to capture individual-specific heterogeneity. Kim et al. (2013) also utilized a random parameter model to analyze single-vehicle crash injury severity data in California. Xiong and Mannering (2013) utilized a more general approach to develop the random parameter model. The random parameter vector was set to follow a multivariate normal distribution

with an unrestricted variance-covariance matrix. Correlation effects of the guardian indicator on other explanatory variables were able to be captured.

Further developed models were also utilized to improve the performance of traditional regression models. Malyshkina and Mannering (2009) developed a two-state Markov switching multinomial logit model to study crash-injury severity under the assumption that there exist two unobserved states of roadway safety. Yamamoto et al. (2008) showed that sequential models could provide superior performance to traditional ordered-response pro-bit models, which assume the same factors correlate across all levels of severity.

Besides, non-regression statistical models, such as BN and neural network, and non-parametric data mining techniques, such as CART, decision tree, support vector machine, etc., have been increasingly applied to crash injury severity analysis. Simoncic (2004) utilized a BN to examine crash injury patterns in two-vehicle crashes. Chimba and Sando (2009) utilized a neural network to predict highway crash injury severity. Kashani and Mohaymany (2011) applied classification tree models to predict injury severity patterns of two-lane rural roadway traffic crashes. While a CART provides an efficient data mining technique, it does not provide the interpretive capabilities of discrete outcome models.

2.2.2 Contributing Factors to Crash Injury Severity

Numerous studies have been conducted through different models to investigate the contributing factors related to crash injury severity regarding weather, traffic flow,

roadway condition, crash location, crash type, and vehicle and driver characteristics. Detailed analyses were also performed to further examine the crash injury patterns with respect to a particular factor.

2.2.2.1 Crash Location Analyses

Particular roadway locations have been identified as crash hotspots, for which significant studies were conducted to address the crash severity patterns at these locations. According to the FHWA (2010), people killed in crashes on rural highways accounted for nearly 57 % of total crash-related fatalities in the U.S in 2009, while the annual Vehicle Miles Traveled (VMTs) on rural highways are only approximately 34% of these on entire highway networks. Besides, 72% fatal crashes in the United States occurred on two-lane highways (NHTSA, 2011). These data indicate that it is critical to investigate the unique characteristics and attributes associated with rural crashes, especially those occurring on rural two-lane highways. Cafiso et al. (2010) developed synthetical analysis models to investigate two-lane rural highway crash characteristics taking into account the factors associated with safety performance, such as exposure and context variables. De Oña et al. (2011b) studied the impacts of a variety of causal factors, such as crash type, driver age and lighting condition on crash injury severity on Spanish rural highways. Weiss et al. (2001) compared rural and urban ambulance crashes regarding the frequency, speed, vehicle damage and personal injury patterns, and found that rural ambulance and its people are more likely to suffer severe injuries. Czech et al. (2010) evaluated the corresponding costs induced by alcohol related crashes in rural and urban areas and found that the attributable cost in rural areas is four times higher than that in urban environment.

Intersection is a hazardous location type on roadways, accounting for a substantial portion of traffic crashes. Inappropriate acceleration, insufficient deceleration, less driver reaction and perception time, etc. may dramatically contribute to severe crash outcomes. Kim et al. (2007) investigated crash outcome potential for different crash types at rural intersections, concluding that the variance of outcome probabilities in these crashes is closely associated with the heterogeneous nature of different intersection structures. Huang et al. (2008) examined the crash injury severity patterns at urban intersections and found that X type intersections may have an averagely positive effect on reducing the crash severity. Haleem and Abdel-Aty (2010) applied multiple approaches to the analysis of crash injury severity at three- and four-legged un-signalized intersections, and concluded that having a 90-degree intersection design is the most appropriate safety design. Abdel-Aty and Keller (2005) applied probit model to examine the overall and specific crash severity levels at signalized intersections and identified that a combination of crash-specific information and intersection characteristics results in the highest prediction rate of injury level.

2.2.2.2 Crash Type Analyses

Generally, there are two ways to classify crashes: vehicle-number-based and vehicle-action-based. Based on vehicle numbers, crashes are usually defined as Single-Vehicle (SV) crashes and Multi-Vehicle (MV) crashes. Based on vehicle actions, crashes could be classified as rollover, rear-end, side-swipe, angle collision, etc.

SV and MV crashes show different crash injury severity patterns. For example, according to NHTSA (2013), there were 1,661,000 single-vehicle crashes and 3,677,000 multi-vehicle crashes in the US in 2011, of which 17,991 and 11,766 were fatal crashes, respectively, indicating that there was a higher probability for severe injuries or deaths in SV crashes. Ulfarsson and Mannering (2004) discovered that SV and two-vehicle crashes should be modeled separately since their differences could not be accurately captured by one model. Therefore, researchers started to explore SV and MV crash characteristics separately to better understand the unique contributing factors for SV and MV crash injury outcomes. Savolainen and Mannering (2007) developed a nested logit model and an MNL model to analyze motorcyclists' injury severities in SV and MV crashes respectively. Geedipally and Lord (2010) employed Poisson-gamma models to explore the separate modeling effect of SV and MV crashes on predicting confidence intervals. They proved the necessity of the separation of SV and MV crashes in highway crash analysis. Ivan et al. (1999) analyzed the distinctiveness of contributing factors in determining SV and MV crash severities on rural roads.

Rear-end crashes and rollover crashes are two major types of traffic crashes resulting in significant injury outcomes. Li and Bai (2008) analyzed crashes occurred in highway construction zones and concluded that rear-end crash is the most frequent type of injury crashes. Duan et al. (2013) investigated the minimum safe vehicle headways between consecutive vehicles for rear-end crash prevention and developed car-following strategies under different weather and traffic conditions. Davis and Swenson (2006) conducted a freeway rear-end collision analysis and found that insufficient headway and long reaction time are important causes. Hu and Donnell (2011) proposed severity

models to examine rollover crashes on rural divided highways, and found that the highest probability of a fatal or major injury in rollover crashes was found to occur in cases when a driver was not using a seatbelt. Dobbertin et al., (2013) estimated the association between vehicle roof crash and head, neck and spine injury in rollover crashes, and discovered that increasing roof crush measurements were statistically associated with higher odds of injury on head, neck and spine. Conroy et al. (2006) investigated occupant, vehicle, and crash characteristics in predicting serious injury during rollover crashes. The results indicate that intrusion (especially roof rail or B-pillar intrusion) at the occupant's position, the vehicle interior side and roof as sources of injury, and improper safety belt use are significantly associated with serious injuries.

2.2.2.3 Driver Characteristic Analyses

Special care has also been taken in traffic safety analyses to address the impacts of driver characteristics on crash injury severity patterns among particular driver characteristics, such as age, gender, drug use, etc.

Driver age has been found to be a significant factor related to crash injury severity in many studies. Hilakivi et al. (1989) and Huang et al. (2008) showed that young drivers as well as senior drivers are more at risk of being involved in severe crashes. Kockelman and Kweon (2002) proposed that senior drivers are less likely to make appropriate and immediate responses when facing crash risks due to their relative slow reactions, while young drivers are more likely to conduct careless driving or speeding, resulting in a considerable potential of severe injuries. Existing studies indicate that teenage drivers tend to maintain shorter headways and higher speed when there are two or more

passengers in their vehicle (Lambert-Bélanger et al., 2012; Simons-Morton et al., 2005). For senior drivers, Rifaat and Chin (2005) found that decrease of visual power, deterioration of muscle strength and reaction time may be responsible for the aged drivers to be involved in severe crashes. Moreover, Abdel-Aty et al. (1998) comprehensively evaluated the effects of driver age across different traffic-related factors on traffic crash involvement, indicating the importance of interactive effects between driver age and crash-related factors.

Driver gender is also found to be statistically significant in predicting crash injury severities. Kockelman and Kweon (2002) also discovered that male drivers are associated with lower driver injury severities comparing to female drivers. Islam and Mannering (2006) identified that female drivers have more interacting factors to increase the likelihood of injuries and deaths comparing to male drivers. There are also contradictory studies with opposite findings. Massie et al. (1995) concluded that vehicles with male drivers are more likely to be involved in fatal crashes than female drivers. Kim et al. (2013) found that male drivers are a contributing factor to fatal injuries in single-vehicle crashes. To be more specific and accurate, Ulfarsson and Mannering (2004) examined the distinctive effects of males and females and their respective interactive effects with other factors on injury severities.

It is well known that driver drunk driving or drug usage is significantly associated with traffic crashes and casualties, which have been proved in many authentic papers. Weiss et al. (2014) concluded that alcohol use is one of the fatal causes in single-vehicle crashes. Poulsen et al. (2014) testified the independent effect of cannabis and the combined effect of alcohol and cannabis in increasing crash potential. Using a case-

control experiment design, Hels et al. (2013) verified the close association between high risk of severe driver injury and high concentration of alcohol in bodies. Siskind et al. (2011) evaluated the impacts of factors containing information on environmental, vehicle and operation on fatal crashes in rural Australian area, and found that alcohol involvement is one of the major factors for fatal crashes.

2.2.2.4 Vehicle Type Analyses

Trucks and motorcycles are two major types of vehicles on roads beside passenger cars. Trucks induce more impact in traffic crashes and cause more severe damage to other vehicles due to their relative large weight and size. The impact of trucks on crash injury patterns have been examined from different aspects. Chen and Chen (2011) examined the difference between injury severity patterns of truck drivers in rural single and multi-vehicle crashes in terms of the impacts of their respective contributing factors. Khorashadi et al. (2005) assessed the difference of driver injuries between rural and urban highway crashes with large truck involvement, and identified unique variables for predicting driver injuries in rural and urban crashes, respectively. As was found by Rifaat and Chin (2005), truck crashes in single-vehicle crashes are more likely to result in serious injuries and fatalities. However, heavy vehicles, such as trucks and semi-trailers, reduce the odds of drivers driving them being severely injured. It is not surprising that as the vehicle weight increases, the risks of being injured or damaged decrease substantially, even though other driver-vehicle units involved in the same crash may be more vulnerable to be injured or damaged. Levine et al. (1999) who reported that every 454 kg

(1000 lbs) increase in vehicle weight was equivalent to the driver's ability to withstand front impact crashes of 10 more kph (6 mph) before being fatally injured.

Motorcyclists are more exposed to open traffic environment and are more vulnerable in crashes, compared with drivers of other vehicles. The number of fatalities for motorcycle crashes is about 12% of the total fatalities for road traffic crashes, although motorcycle crashes account for only 5% of road traffic crashes (Chung et al., 2014). Support for these findings has been offered from other related studies. Huang et al. (2008) discovered that two-wheel vehicles, most of which are motorcycles, are a major factor related to severe injuries in traffic crashes at intersections. Kockelman and Kweon (2002) discovered that motorcyclists are expected to suffer more severe injuries comparing with vehicle drivers. Chiang et al. (2014) found that motorcyclists are the most vulnerable driver group on roadways. More detailed research found that head injury is the main cause of motorcyclist deaths and helmet use is an effective prevention of driver trauma (Hefny et al., 2012; Kelly et al., 1991).

2.2.2.5 Environment Factor Analyses

Weather condition has been identified as a significant factor to crash injury severities. Yu and Abdel-Aty (2014a) incorporated weather data into crash injury severity analysis, and found that real-time traffic and weather variables have substantial influences on crash injury severities. Weather is highly related to road conditions and therefore road surface condition is often used in crash injury severity analyses as an alternative. Shaheed et al. (2013) discovered that dry pavement condition significantly

increases the potential of fatal and major injuries in motorcycle-involved crashes. Through probabilistic modeling, Savolainen and Mannering (2007) found that crashes occurring under wet road surface conditions tend to be less severe. Other studies generate composite conclusions regarding the safety effect of wet pavement conditions. Morgan and Mannering (2011) found that there is significant recognition difference for drivers of different age groups and genders on wet pavement conditions, with which wet or snow/ice road surfaces tend to decrease the probability of severe injury for male drivers less than 45 years old and while increase that for the other driver groups.

Other factors are related to road geometry, lighting condition, etc. For example, Huang et al. (2008) found that right-most driving lane was identified to be significant on increasing the odds of severe crashes by 26%, compared with central lane. Khorashadi et al. (2005) found that for right driving behavior, if the location of collision is on the left lane, the likelihood of injury severity increased by 268.1%. The higher severity risk may be caused by higher speed on left-most lane. According to Bedard et al. (2002), traveling at speeds exceeding 112 kilometers per hour (kph) is independently associated with a 164% increase in the odds of a fatality compared with speeds less than 56 kph. Huang et al. (2008) also discovered that a bad street lighting condition can increase the odds of severe crashes by nearly 69%. Yau (2004) found that street lighting condition affects the crash severity for the SV crashes in Hong Kong.

2.3 Bayesian Method Applications in Traffic Safety Analyses

2.3.1 Bayesian Inference Modeling in Traffic Safety Analyses

As discussed before, EB was firstly applied in traffic safety analysis, and has been widely used as an inference method to address different traffic safety issues of interest. De Lapparent (2006) studied the probability distribution of different socio-demographic elements for four levels of motorcycle crash severity via EB model and found that females aged from 30 to 50 riding powerful motorcycles are the most vulnerable group for injury. Elvik (2013) provides a discussion on the influence of speed limit on traffic crashes with the application of an EB method, and found that the speed limit could decrease injury crashes around 30%.

Due to the internal limitations of EB method, the FB approach was proposed and utilized to facilitate the consistent consideration of aleatory and epistemic uncertainties, non-linear dependencies amongst the indicator variables and the updating of the developed risk models based on new available data. Yanmaz-Tuzel and Ozbay (2010) estimated the impact of various road safety countermeasures in reducing crash frequency with FB models, and concluded that enhancement in vertical and horizontal alignments brought highest crash rate decrease. Persaud et al. (2010) did similar evaluation via comparison of FB and EB models, and proved their effectiveness in road safety assessment. Flask and Schneider IV (2013) modeled SV motorcycle crash data with FB binomial model and discussed its spatial correlation at town and county levels. FB inference was proposed to work on hierarchical models for posterior probability inference for parameters of interest, and therefore is increasingly known as hierarchical Bayesian model. Yu and Abdel-Aty (2013a) employed hierarchical Bayesian model to investigate the characteristics of SV and MV crashes on mountainous freeways via aggregate and disaggregate modeling procedures. Xie et al. (2013) proposed a Bayesian hierarchical

negative binomial model to examine significant factors from both intersection and corridor levels for crash frequency prediction at signalized intersections, and concluded that the proposed model was superior to regular Bayesian negative binomial models and Bayesian random effect models in traffic risk factor analysis. Deublein et al. (2013) proposed a hierarchical Bayesian approach for road crash prediction by grouping gamma distribution, multivariate Poisson-lognormal regression and Bayesian inference together and proved its robustness in forecasting crash occurrences. MacNab (2003) proposed a Bayesian hierarchical Poisson regression model to facilitate crash monitoring and prevention in both spatial and temporal domains. Using Bayesian hierarchical Poisson model with tolerance of autoregressive dependence, Haque et al. (2010) explored the significant factors contributing to motorcycle crash frequencies at signalized T-intersections and four-way intersections.

2.3.2 BN in Traffic Safety Analyses

BN method, as a non-regression method in Bayesian family, has been increasingly utilized in traffic safety analysis. Ozbay and Noyan (2006) employed a BN model to estimate time needed for crash clearance and identify the stochastic characteristics of incidents. Gregoriades and Mouskos (2013) proposed an approach to identity roadway traffic conditions by measuring traffic crash risks through BN models. Goodheart (2013) applied Bayesian belief network to extract the causal rules and predict runway crash risks in aviation operations. De Oña et al. (2013) applied Latent Class Cluster (LCC) and BN into traffic crash severity classification and analysis, and identified the most contributing factors to severe injuries and fatalities. Bedeley et al.

(2013) applied BN to examine factors affecting pedestrian crossing patterns, and concluded that internal motives were more decisive than external elements in affecting pedestrian behavior. Mbakwe et al. (2014) developed a BN model to analyze highway safety performance by estimating traffic crash data, assisted by Delphi Process.

2.4 Variable Selection Summery

Various variable selection approaches have been proposed for applications in different research fields. Variable selection methods are either "performance based" or "test-based". Performance-based approach is to repeatedly fit models to the data in order to determine the best performing one in terms of prediction accuracy. Svetnik et al. (2004) produced several orderings of variables via the computation of importance measures on each training set of a 5-fold cross-validation. Jiang et al. (2004) introduced a method in which they claim to combine the unsupervised 'gene shaving' approach (Hastie et al., 2000) and the supervised random forests. Similar approach was also proposed by Díaz-Uriarte and Alvarez de Andrés (2006). It uses the "Out-Of-Bag (OOB)" error and computes variable importance only once. The best model is chosen to be the smallest one with an error rate within the standard errors of the best performing model.

Test-based approach applies a permutation test framework to estimate the significance of variable importance. Altmann et al. (2010) presented a method that uses a permutation test framework to produce unbiased importance measures (Strobl et al., 2007). An almost identical approach was introduced earlier by Rodenburg et al. (2008) whereas these authors directly aimed at introducing of a variable selection approach.

They repeated the procedure several times and combine the selected variables in a final set. Another related work of Wang et al. (2010) was based on a different kind of importance measure called the 'maximal conditional chi-square importance' to identify relevant Single-Nucleotide Polymorphisms (SNPs) in Genome-Wide Association Studies (GWAS). Following the same research goal, Tang et al. (2009) simultaneously permuted entire sets of SNPs which belong to the same gene.

2.5 Applications of Other Data-mining Techniques in Traffic Safety Analyses

Data mining has been an active analytical technique in many scientific areas for years. In the field of safety analysis, some studies applied tree-based models to analyze crash rates and injury severity problems. Kuhnert et al. (2000) employed logistic regression (also called logit regression), CART and Multivariate Adaptive Regression Splines (MARS) to analyze motor vehicle injury data. By comparing the analysis results with logit regression, they demonstrated that CART and MARS can graphically display the analysis results and identify the groups of people with higher crash risk, making them attractive for motor vehicle crash analysis. Sohn and Shin (2001) applied classification tree, neural network and logit regression models to identify crash severity-related factors using road traffic crash data from Korea. The findings indicated that protective device (i.e., seatbelt or helmet) is the most important factor in the cash severity variation. Other factors include collision type, speed before crash, violent driving, road width and car shape (i.e., with or without bonnet). Karlaftis and Golias (2002) applied Hierarchical Tree-Based Regression (HTBR) to analyze the effects of road geometry and traffic characteristics on crash rates for rural two-lane and multilane roads. The analysis results

by HTBR indicated that Annual Average Daily Traffic (AADT), lane width, serviceability index, pavement friction and pavement type are critical in determining crash rates for rural two-lane highways, while the factors for multilane highway crash rates are AADT, median width, and access control.

Artificial Neural Network (ANN) is another non-parametric model frequently applied to analyze traffic safety problems. Abdel-Aty and Abdelwahab (2004) applied multilayer perceptron and fuzzy adaptive resonance theory neural networks to analyze driver injury severity in traffic crashes. The results indicated that gender, vehicle speed, seatbelt use, vehicle type, point of impact and area type of crash location can affect injury severity likelihood. By comparing the prediction performance with an ordered logit model, the study shows that ANN models have more accurate prediction capability over traditional statistical models. Mussone et al. (1999) employed ANN modeling approach to analyze vehicular crashes in Italy. A three-layer neural network model was proposed to estimate crash index (defined as the ratio of the number of crashes in the *i*th intersection to the number of crashes in the most dangerous intersection) of urban intersections. Their results shows that the ANN model can identify the degree to which factors contribute to intersection crashes and demonstrates that ANN is a good alternative method for traffic safety analysis.

2.6 Unobserved Heterogeneity Issue in Traffic Crash Modeling

Unobserved heterogeneity has been recognized as a critical issue in traffic safety research. Unobserved heterogeneity is defined as the unobservable factors or data that

affect crash potential or severity, and they may generate biased estimations if their correlations with observed variables are not accounted for in model design (Mannering and Bhat, 2014). The unobserved heterogeneity could be attributed from different types of factors, including roadways (Flask et al., 2014; Haleem and Gan, 2013; Malyshkina and Mannering, 2010; Morgan and Mannering, 2011), drivers' demographic and behavior characteristics (Haleem and Gan, 2013; Islam and Mannering, 2006; Kim et al., 2013, 2010; Morgan and Mannering, 2011; Ulfarsson and Mannering, 2004), spatial and temporal variations (Malyshkina and Mannering, 2009; Malyshkina et al., 2009; Ukkusuri et al., 2011; Xiong et al., 2014; Xu and Huang, 2015), etc. For instance, Kim et al. (2010) evaluated pedestrian injury severity patterns in pedestrian-vehicle crashes considering the unobserved pedestrian heterogeneity regarding health, strength and behavior. Anastasopoulos et al. (2012) investigated traffic accident rate patterns accounting for the unobserved heterogeneity effects of highway segments. Xiong et al. (2014) examined crash injury severity patterns based on the heterogeneous temporal influence of roadway segment features.

Thanks to the recent development in crash data organization and mathematical model design, numerous advanced models have been proposed and applied into traffic accident research to account for unobserved heterogeneity within crash data, of which random parameters models and finite-mixture (latent-class) models are two major approaches. Random parameters models are a group of models that simulate individual unobserved heterogeneity by assuming a distribution for parameters of interest to allow them vary across observations or (group of observations) and/or determine observation groups, and include popular models such as random parameter logit (mixed logit) model

(Anastasopoulos and Mannering, 2011; Gkritza and Mannering, 2008; Haleem and Gan, 2015, 2013; Kim et al., 2010, 2008; Malyshkina and Mannering, 2010; Milton et al., 2008; Moore et al., 2011; Pai et al., 2009; Shaheed et al., 2013; Wu et al., 2014), random parameter probit model (Christoforou et al., 2010; Russo et al., 2014; Tay, 2015), random parameter negative binomial models (Chen and Tarko, 2014; Dong et al., 2014; Flask et al., 2014; Venkataraman et al., 2014, 2013; Wu et al., 2013), random parameter Tobit model (Anastasopoulos et al., 2012a; Chen et al., 2014; Yu et al., 2015) and Markov switching models (Malyshkina and Mannering, 2009; Malyshkina et al., 2009; Xiong et al., 2014). Milton et al., (2008) were the first to apply random parameter model in traffic crash analysis, and verified its effectiveness in traffic crash data modeling. With that, random parameters models, including the popular models listed above, have been increasingly used recently to address unobserved heterogeneity relating to multiple factors. Shaheed et al. (2013) utilized a mixed logit model to investigate the construing factors to crash severities in the collisions between a motorcycle and other automotive. Tay (2015) applied a random parameter probit model to assess the difference between urban and rural intersection crashes regarding road, traffic, environment and driver behavior characteristics. Venkataraman et al. (2014) developed a random parameter negative binomial model to crash occurrence patterns based on different interchange types with the assumption that the estimated random parameters are heterogeneous in their means. Yu et al. (2015) estimated the influence of weather conditions on mountain freeway crash potential using a correlated random parameter tobit model. Malyshkina and Mannering (2009) modeled unobserved heterogeneity by assuming that the variance between two unobserved roadway safety statuses follows a Markov switching pattern on

injury severity. But the disadvantage of random parameters models is that it may not be able to capture the heterogeneity across different data groups, and therefore result in biased estimations.

Finite-mixture (latent-class) models are another major type of models addressing unobserved heterogeneity in crash data, and are developed by relaxing the assumption of random parameters models and assuming discrete distribution with a limited number of latent classes to identify homogeneous groups in crash data. Finite-mixture (latent-class) models are presented with different model structures and have been gaining their popularity in traffic safety analysis (Eluru et al., 2012; Lemp et al., 2011; Shaheed and Gkritza, 2014; Xie et al., 2012; Xiong and Mannering, 2013; Zou et al., 2014, 2013). For example, Shaheed and Gkritza (2014) utilized an MNL model with two latent crash data classes to investigate crash severities in single-vehicle motorcycle crashes. Zou et al. (2013) advocated that weight parameter configuration is preferred in finite mixture negative binomial models to better assess heterogeneity effects in crash data analysis, and they further developed different functional forms for weight parameter estimation (Zou et al., 2014). Studies were conducted to compare random parameters models and latent class models in crash data analysis regarding their pre-assumption, applicability and effectiveness (Cerwick et al., 2014; Mannering and Bhat, 2014). For instance, Cerwick et al. (2014) comprehensively compared the advantage and disadvantage of random parameters models and finite-mixture (latent-class) crash severity analysis, concluding that latent class models illustrate slight superiority to mixed logit models in model fit and parameter estimation when modeling unobserved heterogeneity. However, a disadvantage of finite mixture models is that they neglect the observation heterogeneity within each

data group due to the assumption of observation homogeneity in each group. Therefore, a hybrid model by combining random parameters and finite mixture models was proposed by Xiong and Mannering (2013) to account for both group-specific heterogeneity in crash data and individual-observation heterogeneity within each group.

As discussed before, unobserved heterogeneity may result from correlations between unobserved and observed factors, and it may contribute considerably in crash injury severity patterns. Previous traffic crash injury severity studies primarily focused on the main effects of crash-level and vehicle-level variables, but omitted the potential interactions between the cross-level interactions between crash-level and vehicle-level variables. These cross-level interaction effects are generally unobservable in the dataset, as is aforementioned, but may play an important role in driver injury severity outcomes. For example, the variable driver age (vehicle/driver level), as was discussed in Mannering and Bhat (2014), is associated with many unobservable factors such as physical health and reaction time, while these unobservable factors may affect the drivers' operations on roadway segments with special geometric features (crash level), such as curvature or grade. Therefore, there may be an interaction effect between driver age and roadway geometry that contributes to driver injury severities. By defining the hierarchical Bayesian random intercept model with cross-level interaction configuration, this research aims to comprehensively examine the unobserved heterogeneity, represented by cross-level interactions effects between crash level and vehicle/driver level variables, and provides more in-depth findings to supplement contemporary traffic crash injury severities studies.

Chapter 3 Research Methodology Design

3.1 Research Methodology Design

This research aims to comprehensively examine the applicability and effectiveness of Bayesian methods in traffic crash driver injury severity analyses. Bayesian methods, including Bayesian inference and BN methods, have emerged as a powerful framework in traffic safety analyses to identify the contributing factors and their relative impact on injury severity outcomes. The Bayesian inference method is generally used in parameter estimation in regression models, and BN is a powerful model to extract variable statistical dependence using graphic topology and probability inference.

Because the injury severity outcomes of traffic crashes can be regarded as a random event, statistical models, particularly regression analyses, have been extensively employed to explore the factors contributing to fatal or injurious crashes. Among these regression models, logit regression models and ordered outcome models have been the most commonly applied techniques. Meanwhile, hierarchical models are able to capture the hierarchical nature of crash data using random parameter estimation and therefore provide more reliable results than traditional logit models. But in these hierarchical models, the parameters of crash level and vehicle/driver level variables are often estimated independently, although their potential connection have been suggested (Snijders and Bosker, 2000), and the crash heterogeneity is generally assumed to be constants sampled from prior distribution (Huang et al., 2008). Besides, the ordered outcome assumption on crash driver injury severity levels may not be always valid since non-monotonic-changing effect of contributing factors on severity levels may exist

(Moore et al., 2011; Patil et al., 2012). Moreover, it is assumed in these models that driver injury severity variable or its transformation is a linear regression of its contributing covariates, which may not always be appropriate and universally applied. Therefore, in this research, a hierarchical multinomial logit model will be applied to examine the significant factors and their impact on injury severity levels. A random slope model will be applied and the parameters of vehicle/drier-level variables would be assumed to be a function of crash level variables, which enriches existing studies (Huang et al., 2008; D.-G. Kim et al., 2007). The Bayesian inference method is superior to traditional point estimations by being able to model parameter estimates with posterior distributions and predict new observations from a given sample of data, and therefore is used in this proposed hierarchical model for point estimation, given informative and non-informative priors. Other non-linear relationships between the independent and dependent variables should also be assumed and investigated.

Most regression models have their own model assumptions and pre-defined underlying relationships between dependent and independent variables. If these assumptions are violated, the model could lead to erroneous estimations of the likelihood of injury severities. BN is a non-regression method able to model the statistical dependence between dependent variable and independent variables based on graphic presentation and probability inference without any pre-defined assumptions on these variables. Also, BN is capable of capturing the interactions among the independent variables, which outperforms regression models. To fully assess the applicability of Bayesian method, a BN model would be trained in this study to extract the relationships among injury severity and contributing factors based on pre-defined training algorithms

and model quality measurements. To overcome the internal disadvantage of BN, that it is not able to select the most important variables and remove redundant ones for model training, multiple correlation-based and machine-learning methods would be applied for variable selection.

Bayesian statistical methods could also work with non-parametric machine-learning methods, such as tree-based models, DTs, etc. A DTNB model proposed by Hall and Frank (2008) by combining DT and NB classifier provides a connection between these two categories. In this research, the DTNB classifier would be used to extract the significant variables and the associated decision rules for crash driver injury severity prediction, and the performance of this model would be compared with the abovementioned hierarchical Bayesian models and BN model. Figure 3-1 illustrates the conceptual framework of the appropriate selection of Bayesian models for traffic crash driver injury severity analyses, where available Bayesian models, including the proposed models in this study, are the boxes highlighted in blue. In Figure 3-1, the green boxes represent available model types during the decision making procedure, and the red diamonds show the decision points.

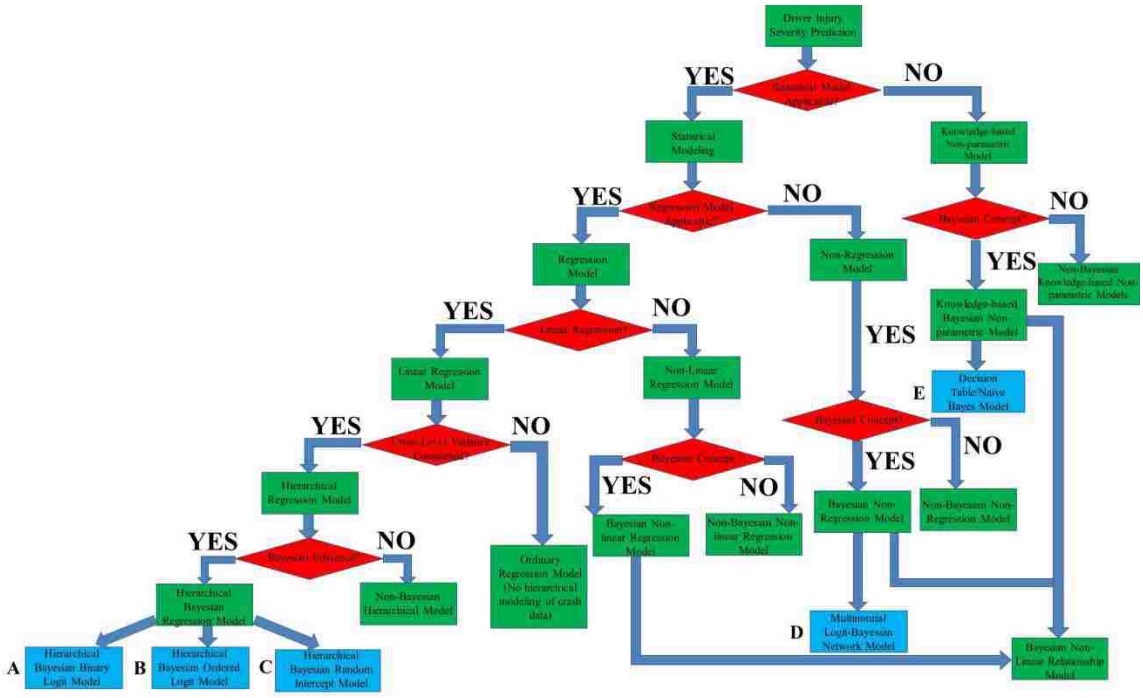

Figure 3-1 Conceptual Framework for Appropriate Selection of Bayesian Models for Driver Injury Severity Analysis.

3.2 Hierarchical Model Development with Bayesian Inference

3.2.1 Hierarchical Bayesian Binary Logit Model

3.2.1.1 Model Design

A two-level hierarchical Bayesian logit model with binary response (indicated as Box A in Figure 3-1) was developed to estimate the effects of crash-level variables and vehicle/driver-level variables on driver injury severities, with the consideration of within-crash correlations. In the lower level (vehicle/driver level), the injury severity of driver i in crash j, denoted as S_{ij}, is a binary variable with $S_{ij} = 0$ indicating no injury or slight injury, and $S_{ij} = 1$ representing incapable injury or death. The probability of $S_{ij} = 1$, denoted as $P_{ij} = \Pr(S_{ij} = 1)$, is assumed to follow a binomial distribution,

$$logit(P_{ij}) = \log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_{0j} + \sum_{k=1}^{K} \beta_{kj} V_{kij} \quad (3\text{-}1)$$

where, V_{kij} is the *kth* vehicle/driver-level variable for the *i*th vehicle/driver unit in the *j*th rural interstate crash, and β_{kj} is the corresponding coefficient for V_{kij} to be estimated; K is the number of vehicle/driver-level variables; β_{0j} is the intercept to be estimated in this regression model. β_{0j} and β_{kj} are summarized from the regression modeling of crash-level variables in the upper level to represent the within-crash correlations, and are defined as,

$$\beta_{0j} = \gamma_{00} + \sum_{m=1}^{M} \gamma_{0m} C_{mj} + \mu_{0j} \quad (3\text{-}2)$$

$$\beta_{kj} = \gamma_{k0} + \sum_{m=1}^{M} \gamma_{km} C_{mj} + \mu_{kj} \quad (3\text{-}3)$$

where, C_{mj} is the *m*th crash-level variable for the *j*th rural interstate crash, and M is the number of crash-level variables; γ_{0m} and γ_{km} are coefficients for C_{mj} corresponding to β_{0j} and β_{kj} respectively; γ_{00} and γ_{k0} are intercepts for β_{0j} and β_{kj}; μ_{0j} and μ_{kj} are random effects representing between-crash variance, which are consistent for vehicles in the same crash but vary across different crashes. Equations (3-2) and (3-3) allow to model within-crash correlation as well as between-crash variations (D.-G. Kim et al., 2007).

The combination of Equations (3-1)-(3-3) produces a random slope model with high complexity (Snijders and Bosker, 2000). To avoid excessive model complexity resulting in intensive model computation while retaining model reasonableness and accuracy, it was assumed that the between-crash variance only works on the intercepts γ_{k0} in Equation (3-3) and the crash-level regression for the *k*th vehicle/driver-level

variable $\sum_{m=1}^{M}\gamma_{km}C_{mj}$. The random effect part μ_{kj} was ignored, forming a random intercept model (Huang et al., 2008),

$$\beta_{kj}=\gamma_{k0} \tag{3-4}$$

Therefore, the full hierarchical binary logit model is formulated as follows,

$$logit(P_{ij}) = \log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \gamma_{00} + \sum_{m=1}^{M}\gamma_{0m}C_{mj} + \sum_{k=1}^{K}\gamma_{k0}V_{kij} + \mu_{0j} \tag{3-5}$$

where μ_{0j} are generally assumed to follow a normal distribution, $\mu_{0j}\sim(0,\ \sigma_0^2)$ (Snijders and Bosker, 2000).

In this research, an ordinary logit regression model was also derived from Equation (3-5) and provided as a reference for model performance comparisons. An ordinary logit regression model was formulated by removing the random effect term μ_{0j} in Equation (3-5),

$$logit(P_{ij}) = \log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \gamma_{00} + \sum_{m=1}^{M}\gamma_{0m}C_{mj} + \sum_{k=1}^{K}\gamma_{k0}V_{kij} \tag{3-6}$$

To examine the between-crash variance, the intra-class correlation coefficient (ICC) was employed (Jones and Jørgensen, 2003; D.-G. Kim et al., 2007; Kutner et al., 2004), which is defined as

$$ICC = \frac{\sigma_c^2}{(\sigma_v^2+\sigma_c^2)} \tag{3-7}$$

where, σ_c^2 is the between-crash variance which is equal to σ_0^2 in this research; σ_v^2 is the vehicle/driver-level variance, which is equal to $\frac{\pi^2}{3}$ = 3.29 for a hierarchical logit distribution (Huang et al., 2008; D.-G. Kim et al., 2007). The ICC is defined in this analysis to evaluate the portion of total variance explained by between-crash variance with a range from 0 to 1. An ICC value close to 0 suggests that between-crash variance is

a small portion of the total variance and the ordinary logit model is more suitable for the analysis. A large ICC value close to 1 indicates the significance of between-crash variance in explaining total variance and demonstrates that a hierarchical model is preferable in the study (Huang et al., 2008; Kutner et al., 2004).

3.2.1.2 Model Specification

Comparing to MLE, the Bayesian inference method is able to model parameter estimates with posterior distributions and predict new observations from a given sample of data. Besides, based on the given dataset, the prior information for fixed and random effects could both be updated during Bayesian inference procedure and revealed in posterior distributions, which are more reliable than regular MLE results. In this research, non-informative priors are defined due to limited historical crash data availability for the unknown parameters, which are estimated based on previous studies (Huang et al., 2008; MacNab, 2003; Yu and Abdel-Aty, 2014b). The intercept term γ_{00}, the coefficients of crash-level variables, γ_{0m}, and the coefficients of vehicle/driver-level variables, γ_{k0}, are all assumed to follow a normal distribution (0,1000). As stated before, the random effects μ_{0j} are assumed as normally distributed$(0,\ \sigma_0^2)$, and σ_0^2 is following an inverse Gamma distribution (0.001, 0.001). The model simulation procedure was conducted with a Monte Carlo Markov Chain (MCMC) algorithm in the WinBUGS platform (Gilks et al., 1995).

For modeling result interpretation, the odds ratio rather than the estimated mean was utilized to explain the influence of the identified variables on driver injury severity. The odds ratio is the exponential output of the estimated mean for γ, $\exp(\gamma)$. An odds

ratio equal to 1 means no effect for the studied variable on driver injury severity, which is corresponding to $\gamma = 0$; an odds ratio larger than 1.0 indicates that an increase of one unit on the studied variable would increase the odds of drivers being incapably injured or killed in a rural interstate crash by $100(\exp(\gamma) - 1)\%$ compared with the base case. An odds ratio less than 1.0 implies that an increase of one unit on the studied variable would decrease the odds of drivers being incapably injured or killed in a rural interstate crash by $100(1 - \exp(\gamma))\%$. The 95% Bayesian Credible Interval (BCI) is provided to indicate the significance of the variables (Gelman et al., 2013), and 90% BCI is also calculated as an additional reference. A variable is considered significant in affecting driver injury severity if the 95% BCI of its odds ratio does not cover 1 and is not significant if otherwise. Experience and consensus on traffic safety analyses are also referred to for result reasonableness examination.

3.2.2 Hierarchical Bayesian Ordered Logit model

3.2.2.1 Model Design

The hierarchical Bayesian binary logit model treats driver injury severity outcome as a binomial variable, which is reasonable but may not be able to fully excavate the influence of factors on different injury severities. It is understandable that driver injury severity is ordinal in nature, and an ordered response model may provide better model fit and estimation results. A hierarchical ordered logit model (indicated as Box B in Figure 3-1) is utilized in this study. Let S_{ij} =*k* be the driver injury severity equal to level *k* for the *j*th vehicle in the *i*th crash, which is a response variable with five ordered categories:

no injury ($k = 1$), complaint of injury/possible injury ($k = 2$), visible injury ($k = 3$), incapacitating injury ($k = 4$) and death ($k = 5$). In this ordered-response model, a latent variable, S_{ij}^*, associated with the actual driver injury severity S_{ij}, is proposed to establish the mathematical relationship between driver injury severity and the predicting covariates. A set of four thresholds (h_{im}, m=1, 2 ,3, 4) are defined to divide the virtual injury severity line into the five abovementioned categories. The actual injury severity variable S_{ij} is associated with the latent variable, S_{ij}^*, as follows:

$$S_{ij}=k = \begin{cases} 1, if\ -\infty < S_{ij}^* < h_{i1}, \\ m, if\ h_{i(m-1)} < S_{ij}^* < h_{im}, m=2, 3, 4 \\ 5, if\ h_{i4} < S_{ij}^* < +\infty \end{cases} \quad (3\text{-}8)$$

The latent variable S_{ij}^* is a prediction of the crash risk factors and could be written as follows

$$S_{ij}^* = \eta_{ij} + \varepsilon_{ij} = \sum_{p=1}^{P} \beta_p \times V_{ijp} + \varepsilon_{ij} \quad (3\text{-}9)$$

where V_{ijp} is the pth covariate for the jth vehicle/driver unit in the ith crash; P is the total number of variables in model estimation; β_p is the corresponding coefficient; ε_{ij} is the error term and is assumed to follow a logit distribution. Therefore, the cumulative response probability for the five ordinal injury severity categories is expressed as,

$$P_{ij,k} = Pr(S_{ij} \le k) = Pr(S_{ij}^* \le h_{im}) = Pr(S_{ij}^* - \eta_{ij} \le h_{im} - \eta_{ij}) = F(h_{im} - \eta_{ij}) = \frac{\exp(h_{im}-\eta_{ij})}{1+\exp(h_{im}-\eta_{ij})} \text{ for } m=k=1,2,3,4 \quad (3\text{-}10)$$

where F is the cumulative density function. Therefore,

$$logit\ (P_{ij,k}) = log\left(\frac{P_{ij,k}}{1-P_{ij,k}}\right) = h_{im} - \eta_{ij}, \text{ for m=1, 2, 3, 4.} \quad (3\text{-}11)$$

In this model, h_{im} is specified as a random variable associated with crash-level variance,

$$h_{im} = h_m + u_i \tag{3-12}$$

where h_m represents the mean of the threshold for all crashes, and u_i is a random effect component indicating the variance among different crashes, and is assumed to follow a normal distribution with a mean of zero and a variance of σ^2.

An ordinary ordered logit model dismissing the between-crash variance term u_i was also employed to examine the same dataset.

3.2.2.2 Bayesian Inference Specification

In this study, Bayesian non-informative priors were also applied to infer the unknown parameters of interest. The latent threshold mean, h_m, and all the coefficients of binary response variables, such as driver gender and driver under impairment, were assumed to follow a normal distribution (0,1000). The coefficients of each categorical value of multi-categorical variables were assumed to follow a normal distribution (0, 10000). The between-crash variance u_i was assumed to follow a normal distribution (0, σ^2), where σ^2 is inversely gamma distributed (0.01, 0.01).

3.2.3 Hierarchical Random Intercept Model with Cross-Level Interactions

3.2.3.1 Model Design

Many discrete choice modeling techniques have been applied to formulate crash driver injury severity outcomes, such as MNL models, nested-logit models, ordered probit models, etc. Ordered logit models may not be suitable for non-monotonic-changing severity data due to their strong restrictions on the linear relationship between explanatory variables and independent outcomes. For example, the steep roadway grade may increase crash driver injury severities when its absolute value is small or moderate. When its absolute value continuously increases beyond a certain range, crash driver injury severities tend to decrease due to the facts that drivers will travel much slower and pay more attention to handle abrupt grade changes in these situations. The application restriction of ordered logit models indicates that changing explanatory variables shall either increase or decrease crash severities in a monotonic manner across all the possible outcomes, which is not always supported by the severity data. Therefore, a more commonly used unordered discrete modeling approach, MNL model, is utilized for hierarchical modeling development.

As discussed in Chapter 2, unobserved heterogeneity may result from correlations between unobserved and observed factors, and the unobserved heterogeneity may contribute considerably in crash injury severity patterns. In the development of the two previous models, unobserved heterogeneity was only modeled with the random error terms, but neglected the potential interactions between crash-level variables and vehicle/driver level variables, and these interactions may play an important role in driver injury severity outcomes. For example, the variable driver age (vehicle/driver level), as was discussed in Mannering and Bhat (2014), are associated with many unobservable factors such as physical health and reaction time, while these unobservable factors may

affect the drivers' operations on roadway segments with special geometric features (crash level), such as curvature or grade. Therefore, there may be an interaction effect between driver age and roadway geometry that contributes to driver injury severities. Thus, a hierarchical Bayesian MNL model with random intercept setting (indicated as Box C in Figure 3-1) is utilized, and the cross-level interactions between crash-level and vehicle/driver-level variables are examined based on the assumption of linear regression.

In this model design, a MNL model is developed to estimate the probability of three driver injury outcomes in rear-end crashes, and the response variable, driver injury severity, is considered as a multi-categorical variable. It is assumed that for any attribute changes, the marginal costs for each severity outcome (no injury, injury, and fatality) are different. Suppose that the response variable $Y_{ij} = (Y_{ij1}, \dots, Y_{ijK})$ has *K* level, where K=3 in this study. The multinomial logit regression can be written as

$$Y_{ij} \sim categorical(P_{ij1,} \dots, P_{ijK}) \tag{3-13}$$

and

$$log \frac{P_{ijk}}{P_{ijK}} = \eta_{ijk} = \beta_{0jk} + \sum_{p=1}^{P} \beta_{pjk} V_{pij} \tag{3-14}$$

where P_{ijk}=$Pr(Y_{ijk} = 1)$ is the probability of the driver injury severity of vehicle *i* in crash *j* being in category *k* (*k=1,…,K-1*), V_{pij} is the *pth* vehicle/driver-level variable for the *i*th vehicle/driver unit in the *j*th crash, and β_{pjk} is the corresponding coefficient for V_{kij} to be estimated; *P* is the number of vehicle-level predictor variables; β_{0jk} is the intercept to be estimated in this regression model. The *K*th category is set as the reference category and therefore the coefficients of the *K*th category are zero. β_{0jk} and β_{pjk} are

summarized from the regression modeling of crash-level variables in the upper level to represent the within-crash correlations, and are defined as,

$$\beta_{0jk} = \gamma_{000} + \sum_{m=1}^{M} \gamma_{0m} C_{mj} + \mu_{0j} \quad (3\text{-}15)$$

$$\beta_{pjk} = \gamma_{p00} + \sum_{m=1}^{M} \gamma_{pm} C_{mj} + \mu_{pj} \quad (3\text{-}16)$$

where, C_{mj} is the mth crash-level variable for the jth rural interstate crash, and M is the number of crash-level variables; γ_{0m} and γ_{pm} are coefficients for C_{mj} corresponding to β_{0jk} and β_{pjk} respectively; γ_{000} and γ_{p00} are intercepts for β_{0jk} and β_{pjk}; μ_{0jk} and μ_{kjk} are random effects representing between-severity level variance, which are consistent for vehicles with the same severity level in the same crash.

The total of *(K-1)* equations are solved simultaneously to estimate the coefficients. The coefficients in the model express the effects of the predictor variables on the relative risk or the log odds of being in category j versus the reference category, here K. In this model, linear relationships are assumed for them with the crash level covariates C_{mj}, which is reasonable since the various crash features may result in different severity levels. Besides the fixed parts which depend on crash level covariates, random effects are assumed to simulate potential random variance across different crashes (μ_{0j} and μ_{pj}) and different severity levels ($\varepsilon_{0k}$and ε_{pk}).

In this study, the random intercept model without the cross-level interaction part $\sum_{p=1}^{P} \sum_{m=1}^{M} \gamma_{pmk} C_{mj} V_{pij}$, was used for model comparison purpose and the deviance information criterion (DIC) is utilized as a Bayesian measurement for model performance measurement.

3.2.3.2 Model Calibration Using Bayesian Inference and Prior Information Specification

Bayesian inference method is applied in this research for model parameter estimation and non-informative priors are used. The main intercept term γ_{00k}, the severity-specified coefficients γ_{0mk}, γ_{p0k} and γ_{pmk}, are all assumed to follow a normal distribution (0,1000). The crash-level random effect μ_{0j} is assumed to be normally distributed (0, σ_0^2), and σ_0^2 follows an inverse Gamma distribution (0.001, 0.001). The model simulation procedure was performed with the Gibbs sampler, a Monte Carlo Markov Chain (MCMC) algorithm in WINBUGS, and the 95% BCI was also used to indicate the significance of examined covariates.

3.2.3.3 Pseudo-Elasticity Analysis

According to Kim et al. (2007), for discrete choice models with multiple categories in the response variable, the positivity or negativity the coefficients could not be freely interpreted as the increase or decrease on the probability of injury severity levels. This is because the rate of change in the probability is not a simple linear function of the coefficient specific to that particular injury severity, but is also a function of its effect and the effects of all the other coefficients in all other injury severities. To accurately assess the influence of contributing factors on multi-categorical injury outcome, the pseudo-elasticity analysis needs to be performed. To properly evaluate the influence of contributing factors on injury severity outcomes, a direct pseudo-elasticity analysis is necessary by altering the values of each contributing factor and examining the probability change. In this study, the variables were all converted to 0-1 indicator

variables for logit modeling. The pseudo-elasticity is defined by the percentage change in probability when an indicator variable is changed from 0 to 1 (and 1 to 0), and is calculated as follows:

$$E_{x_{nk}}^{P_{ni}} = \frac{P_{ni}[x_{nk}=1]-P_{ni}[x_{nk}=0]}{P_{ni}[x_{nk}=0]} \quad (9)$$

where $E_{\boldsymbol{x}_{nk}}^{P_{ni}}$ is the direct pseudo-elasticity of the *k*th variable from the vector $\boldsymbol{x}_n$. P_{ni} is the probability of driver *n* suffering injury severity level *i* and is defined as

$$P_{ni} = \frac{e^{\beta_i x_n}}{\sum_{i'} e^{\beta_{i'} x_n}} \quad (10)$$

where β_i is the vector of coefficients estimated specific to injury severity level *i* and $\boldsymbol{x}_n$ is a vector of exogenous variables for driver *n*. This pseudo-elasticity method has been utilized in several authentic traffic safety studies (Shankar and Mannering, 1996; Ulfarsson and Mannering, 2004), and therefore is also used in this study to evaluate the marginal effects of the contributing factors. In this study, the percentage change in probability by altering variable values is evaluated for each driver/vehicle record using the estimated mean of each coefficient, and the pseudo-elasticity is summarized by averaging the result for each observation.

3.2.4 Model Performance Comparison

In the development of these three hierarchical regression models, corresponding control models are developed and used to examine the same datasets for performance comparison purpose, in which the Deviance Information Criterion (DIC) is used as model performance measurement. DIC is proposed by Spiegelhalter et al.(2002) as a Bayesian

measurement for model performance comparison in Bayesian model selection procedure. It is a generalization of two hierarchical modeling measurements: Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The DIC is defined as

$$DIC = D(\bar{\gamma}) + 2pD = \overline{D(\gamma)} + pD \quad (3\text{-}17)$$

where, $D(\bar{\gamma})$ is the deviance obtained at the posterior means of estimated parameter γ, and is specified as $D(\bar{\gamma}) = -2\log\big(p(y|\bar{\gamma})\big) + C$, where y is the response value, $\bar{\gamma}$ is the posterior mean of estimated parameter γ, and C is a constant term that could be canceled out in model comparison calculation. $\overline{D(\gamma)}$ is the posterior mean of the deviance, and is defined as $\overline{D(\gamma)} = E^{\gamma}(D(\gamma))$, which could be considered as a measurement of model suitability. pD is the effective number of parameters and is generally considered as a model complexity measurement, $pD = \overline{D(\gamma)} - D(\bar{\gamma})$. In model comparison problems, a lower DIC value indicates a preferable model for parameter estimation and response prediction.

3.3 MNL-BN Hybrid Model

Most regression models have their own model assumptions regarding data structure and underlying relationships between dependent and independent variables, and violation of these assumptions could lead to erroneous estimations of the likelihood of injury severities. Therefore, non-regression models relaxing these restrictions are needed in this study for model applicability examination. A MNL-BN hybrid model (indicated as Box D in Figure 3-1) is proposed as a non-regression machine-learning method in this

section for driver injury severity prediction, and the detailed model development procedure is presented below.

3.3.1 BN Definition

BN is employed as a classifier to analyze driver injury severity outcomes based on the given variables. BN is capable of quantifying conditional probability relationships among variables via graphic presentation, known as a Directed Acyclic Graph (DAG) (Bouckaert, 2008). A BN can be represented by a network structure B_s over a set of variables, $V = \{x_1, x_2, \dots, x_v\}$, $v > 1$. The DAG is portrayed to show cause-effect relationships among variables. A set of probability tables $B_p = \{p(x_i|pr(x_i)), x_i \in V\}$ are provided to quantitatively interpret these cause-effect relationships depicted by the graphical structure, B_s, where $pr(x_i)$ is the set of parent variables of x_i in B_s and i=1,2,...,v. Technically speaking, A BN over a set of variables, V, represents joint probability distribution, $P(V) = \prod_{x_i \in V} p(x_i|pr(x_i))$ for i=1,2,..., v. Using BN to analyze crash driver injury severities is to classify a potential driver injury outcome, $y = y_0$ (e.g. no injury, injury, fatality), given a set of significant variables identified in the MNL model, $X = \{x_1, x_2, \dots, x_k\}$, $k = v - 1$. The driver injury outcome, y, and the attribute variables, X, constitute the overall variable set $V = (X, y)$. The classifier is a function mapping a case of X to an outcome of y, which could be trained from a given dataset D that contains sample instances of (X, y). To use BN as a classifier, we need to calculate $argmax_y P(y|X)$, the value of y that maximizes $P(y|X)$, using the distribution $P(V)$, where

$$P(y|X) = \frac{P(X,y)}{P(X)} = \frac{P(V)}{P(X)}$$

$$\propto P(V)$$

$$= \prod_{x_i \in V} p(x_i | pr(x_i)) \quad (3\text{-}18)$$

The BN structure graphically represents various interactions among variables. The variables are denoted as nodes and their interactions are represented by directional arcs and edges between two nodes. Unconnected nodes signify direct independence between the variables represented by the corresponding nodes.

BN forms a complete probabilistic model so that it represents the joint probability distributions of all variables involved. Theoretically speaking, a BN can use both continuous and discrete variables. However, in most approaches to learning BN structures from data, one common assumption is made that all the input variables are discrete variables to circumvent practical problems in the implementation of the BN specification and estimation theory (Buntine, 1991; Cooper and Herskovits, 1992, 1991; Heckerman et al., 2013). There are two ways of discretize numeric variable. First, numeric variables could be discretized by several discretization algorithms, such as Equal Width Interval Binning, Holte's 1R Discretizer, Recursive Minimal Entropy Partitioning, etc. (Dougherty et al., 1995). Besides, numeric variables could also be categorized based on the accepted standards or experience in the studied area, such as traffic related experience or engineering experience in traffic safety analysis, as shown in many authentic studies (Ahmed et al., 2011; de Oña et al., 2011; Huang et al., 2008). In this study, numeric variables are discretized based on these previous studies as well as engineering experience, rather than relying on discretization algorithms, for the reason that engineering experience-based discretization produces more reasonable categories. For example, driver age is a popular variable in traffic safety analyses. Based on previous

studies or engineering experience, researchers often divide driver age into three exclusive categories: Young drivers (16-25), mid-age drivers (26-64), senior drivers(65 or older), with which we can examine the age effect on crash frequency or crash severity.

3.3.2 BN Structure Quality Measurement-Scoring Metric

To find a globally optimal BN structure, the searching algorithm needs to test all possible DAG options in the structure space. The number of possible DAGs with n nodes is (Mujalli, 2011; Robinson, 1977),

$$F(n) = \sum_{i=1}^{n}(-1)^{i+1}\frac{n!}{i!(n-i)!}2^{(n-i)}F(n-i) \quad (3\text{-}19)$$

Generally, searching an optimal BN structure is a Non-deterministic Polynomial-time hard (NP-hard) problem defined in computational complexity theory. Therefore, it is necessary to employ effective training algorithms to find an approximately optimal DAG in a heuristic way. In this study, prior knowledge and BN scoring metrics are combined to achieve an efficient BN structure estimation. Several BN scoring metrics are commonly used as structure quality measurements, such as Minimum Description Length (MDL), AIC, Bayes metric, structure entropy, and Bayesian metric with Dirichlet priors and equivalence (BDe).

To describe these metrics, the following terms are defined (Bouckaert, 2008):N is the number of instances in a dataset, D; r_i is the cardinality of a variable, x_i; $pr(x_i)$ denotes the set of the parent variables of x_i in B_s; q_i is the cardinality of $pr(x_i)$, and $q_i=\prod_{x_j \in pr(x_i)} r_j$; N_{ij} represents the number of cases in the dataset that $pr(x_i)$ takes its jth value; N_{ijk} is the number of cases in the dataset where $pr(x_i)$ takes its jth value and x_i takes its kth value,

and N_{ij}=$\sum_{k=1}^{r_i} N_{ijk}$. $P(B_s)$ represents the prior information for BN structure, B_s; N'_{ij} and N'_{ijk} are the prior knowledge on N_{ij} and N_{ijk}, restricted by $N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$. Thus, for a BN structure, B_s, over a database D:

The entropy metric $H(B_s, D)$ is defined as

$$H(B_s, D) = -N \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} \frac{N_{ijk}}{N} log \frac{N_{ijk}}{N} \tag{3-20}$$

and the number of parameters T as

$$T = \sum_{i=1}^{n} (r_i - 1) * q_i \tag{3-21}$$

The AIC metric $Q_{AIC}(B_s,D)$ is expressed as

$$Q_{AIC}(B_s, D) = H(B_s, D) + T \tag{3-22}$$

The MDL metric $Q_{MDL}(B_s,D)$ is defined as

$$Q_{MDL}(B_s, D) = H(B_s, D) + \frac{T}{2} \log N \tag{3-23}$$

The Bayes metric is

$$Q_{Bayes}(B_s, D) = P(B_s) \prod_{i=0}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij}+N_{ij})} \prod_{k=1}^{ri} \frac{\Gamma(N'_{ijk}+N_{ijk})}{\Gamma(N'_{ijk})} \tag{3-24}$$

when $N'_{ijk} = 1$, $Q_{Bayes}(Bs, D)$ is converted to K2 metric as follows. K2 metric is a entropy-based score metric proposed by Cooper and Herskovits (1992) for BN heuristic learning.

$$Q_{K2}(B_s, D) = P(B_s) \prod_{i=0}^{n} \prod_{j=1}^{q_i} \frac{(r_i-1)!}{(r_i-1+N_{ij})!} \prod_{k=1}^{r_i} N_{ijk}! \tag{3-25}$$

with $N'_{ijk} = \frac{1}{r_i * q_i}$, we have the Bayesian BDe metric as follows (Bouckaert, 2008; Heckerman et al., 1995):

$$Q_{BDe}(B_s, D) = P(B_s)\prod_{i=0}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(\frac{1}{q_i})}{\Gamma(\frac{1}{q_i}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(\frac{1}{r_i*q_i}+N_{ijk})}{\Gamma(\frac{1}{r_i*q_i})} \quad (3\text{-}26)$$

where, $\Gamma(*)$ is the Gamma function. Based on these metrics, optimal BN structures could be determined in the BN learning and model specification development as detailed in the following sections.

3.3.3 BN Structure Learning Algorithm

Various structure learning algorithms have been proposed for the optimal BN structure, such as hill climbing, simulated annealing, genetic algorithm, etc. In this study, a popular hill-climbing based K2 algorithm would be used for BN structure training.

The K2 algorithm is a type of greedy hill climbing search algorithm, and based on this staring point, all the neighboring DAGs are established by adding, removing, and reversing an existing arc of the initial DAG. The scoring metrics are used to evaluate each DAG performance. A new DAG with a higher score will replace the current DAG, and new neighboring DAGs are generated to enable search processes to iterate until a DAG is found with the highest score (Cooper and Herskovits, 1992). A DAG with the highest score is the optimal network structure. The less restricted version of K2 algorithm in this study can allow no predefinition of nodes order but greedily add or remove edges between random node pairs and even examine the inversion of existing directed arcs, which produced more reliable results (Witten et al., 2011).

In the training procedure, the initial DAG guided by prior knowledge may potentially lead to an optimal model structure specification with reasonable cause-effect elaborations. However, the identified BN structure may be greatly impacted by the initial

knowledge-based DAG and it could be a locally optimal solution based on this type of greedy hill climbing search algorithm. To address this problem, different initial DAGs are developed as the starting points for multiple search iterations to ensure at least a globally suboptimal DAG will be generated.

In this study, a simple estimator is used to estimate the conditional probability table of a node after the BN structure is determined, taking appropriate prior knowledge into account (Bouckaert, 2008). It calculates the conditional probabilities directly as follows,

$$P(x_i = k|pr(x_i) = j) = \frac{Q_{ijk}+Q'_{ijk}}{Q_{ij}+Q'_{ij}} \quad (3\text{-}27)$$

where, as defined before, Q_{ij} represents the number of cases in the dataset that $pr(x_i)$ takes its jth value; Q_{ijk} represents the number of cases in the dataset where $pr(x_i)$ takes its jth value and x_i takes its kth value, and $Q_{ij}=\sum_{k=1}^{r_i} Q_{ijk}$. Q'_{ij} and Q'_{ijk} are the prior knowledge on Q_{ij} and Q_{ijk}, restricted by $Q'_{ij} = \sum_{k=1}^{r_i} Q'_{ijk}$, and could be set. When Q'_{ijk} is set as 0, the Maximum Likelihood Estimation (MLE) would be obtained.

3.3.4 BN Input Variable Selection Procedures

The ordinary MNL model is a test-based model to identify significant variables for a target variable. It is assumed that for any attribute changes, the marginal costs for different severity outcomes are different. P_{is}, the probability of driver, s, being involved in a crash with injury severity level, i, is determined by the utility function U_{is}:

$$P_{is} = P\left(U_{is} \geq U_{js}, \forall i,j \in C, i \neq j\right) = P(u_{is} + \varepsilon_{is} \geq u_{js} + \varepsilon_{js}, \forall i,j \in C, i \neq j) \quad (3\text{-}28)$$

where u_{is} is the deterministic component that is only modeled by significant variables describing the instance; ε_{is} is the random component representing the hidden effect on driver injury severity; C is the choice set of possible driver injury severity outcomes. u_{is} is defined as a linear function for driver s,

$$u_{is} = \beta_i \times V_{is} + \alpha_{is} \tag{3-29}$$

where V_{is} is the exogenous variable vector influencing injury severity, i, for driver, s, and β_i is a coefficient vector to be estimated for measuring the influence of V_{is} on driver injury severity, i; α_{is} is the constant term. ε_{is} is normally assumed to follow a Generalized Extreme Value (GEV) distribution, and a MNL model can be derived as

$$P_{is} = \frac{e^{u_{is}}}{\sum_{j \in C} e^{u_{js}}} = \frac{e^{\beta_i \times V_{is} + \alpha_{is}}}{\sum_{j \in C} e^{\beta_j \times V_{js} + \alpha_{js}}} \tag{3-30}$$

where, P_{is} is the probability of driver, s, suffering injury outcome, i, in a crash. The coefficients β_i and α_{is} are estimated via MLE method. All the variables are used for MNL model development and significant ones are identified based on their T-ratios and P-values at the confidence level of p=0.05. These identified significant variables will be used for BN structure establishment and probabilistic parameter learning to explicitly formulate cause-effect relationships between injury severity outcomes and explanatory attributes.

3.4 Knowledge-Based Bayesian Non-parametric Method

In order to comprehensively investigate the feasibility of applying Bayesian methods in crash driver injury severity analyses, non-parametric machine-learning

models should also be included. In this section, a Decision Table/Naïve Bayes (DTNB) hybrid classifier (indicated as Box E in Figure 3-1) is proposed as a representative model for driver injury severity pattern investigation, as discussed in the following sections.

3.4.1 Decision Table (DT)

DT is a scheme-specific learning algorithm modeling and presenting complicated logics (Witten et al., 2011). It is defined as a table representing a complete set of decision rules under all mutually exclusively conditional scenarios in a pre-defined problem (Witlox et al., 2009). A standard DT consists of four parts. In a DT, the upper left part is a list of all the conditions, denoted as C_i for i=1,…, c, where c is the number of conditions in the problem. A condition-state set CS_i contains all the possible alternative states that C_i is able to attain within a particular pre-defined problem:

$$CS_i = \{S_{i1}, S_{i2},\ldots, S_{iti}\} \tag{3-31}$$

where t_i is the number of alternative states for the ith condition C_i in the pre-defined problem.

The upper right part of a DT is its condition space, which is a Cartesian product of all the condition-state sets CSi (i=1,…, c), as shown below:

$$SP(C) = CS_1 \times CS_2 \times\ldots\times CS_c \text{ for } c>1 \tag{3-32}$$
$$= CS_1 \qquad \text{for } c=1$$

Each element in the condition space is a condition entry (CE) with ordered c dimensions (also known as an ordered c-tuple) (Witlox et al., 2009), and the whole set of these condition entries in the DT is defined as the domain of a DT, denoted as DOM(DT).

The lower left part in a DT includes all the possible action subjects used to express the decisions, represented as A_j for j=1,…,a, where a is the number of all possible actions. Similar to CS_i, an action-state set AT_j includes all the attainable states for action A_t within a particular pre-defined problem, defined as:

$$AT_j = \{T_{j1}, T_{j2}, \ldots, T_{jmj}\} \tag{3-33}$$

where m_j is the number of alternatives for A_j in the pre-defined problem.

The lower right part of a DT is its action space, which is also a Cartesian product of the all the action sets AT_j for j=1,…,a,

$$SP(A) = AT_1 \times AT_2 \times \ldots \times AT_a \text{ for } a>1 \tag{3-34}$$
$$= AT_1 \qquad \text{for } a=1$$

Similar to the condition space, each element in the action space is an *a*-dimensional Action Entry (AE).

The presentation of a complete DT is a matrix and could be written as follows: Let *n* be the number of decision rules (columns) and *c* be the number of conditions (rows). The condition part of a DT is then expressed as,

$$D = (d_{ir}), i = 1, \ldots, c \text{ and } r = 1, \ldots, n \tag{3-35}$$

where $d_{ir} \in CS_i$

The action part could be expressed as:

$$E = (e_{jr}), j = 1, \ldots, a \text{ and } r = 1, \ldots, n \tag{3-36}$$

where a is the number of actions (rows) and $e_{jr} \in AT_j$. Therefore, A DT specifies the relations between condition space and action space as,

$$DT = (dt_{qr}) = \binom{D}{E} \tag{3-37}$$

where $dt_{qr} = d_{qr}$, for $q = 1, \ldots, c$ and $r = 1, \ldots, n$

$= e_{(q-c)r}$, for $q = c+1, \ldots, c+a$ and $r = 1, \ldots, n$

In application, a DT is used as a lookup table based on the selected attributes. Each entry in the DT is associated with the class probability estimated based on the observed frequencies in the original dataset. The critical procedure of learning a DT is the selection of highly discriminative attributes, given the class variable, and it is normally conducted by maximizing cross-validated performance (Hall and Frank, 2008). Cross-validation is efficient for DT learning since the learned structure would not change with the addition or deletion of instances, and only the class counts vary according to the entries. Detailed explanation of the cross-validation procedure is discussed below in Section 3.4.3.

Lew (1991) and Witlox et al. (2009) concluded the advantages using DTs: they can be used for algorithm design in a schematic way; they provide a compact visual presentation of the classification results; they are flexible algorithms that enable automated error-checking and formal verification; they are also a compatible type of model easy to be embedded in other models and computing languages.

3.4.2 Naïve Bayes (NB) Model

In a classification task, assume Y is the class variable and $X=(X_1, X_2, \ldots, X_n)$ is the set of attribute variables. A Bayes classifier prediction for the value y of class variable Y is a process to find y that $P(Y=y_i)$ has the highest posterior conditional probability given $x=(x_1,x_2,\ldots x_n)$, shown in Equation (3-38).

$$P(Y=y_i|X=x=(x_1,x_2,\ldots x_n))>P(Y=y_j \mid X=x=(x_1,x_2,\ldots x_n)),\ \forall j, j \neq i \qquad (3\text{-}38)$$

Using Bayes' Theorem, it can be expressed as Equation (3-39),

$$P(Y = y_i | X = x = (x_1, x_2, \dots, x_n)) = \frac{P(X = x = (x_1, x_2, \dots, x_n) | Y = y_i) P(Y = y_i)}{P(X = x = (x_1, x_2, \dots, x_n))} \quad (3\text{-}39)$$

An NB classifier is a Bayesian model with the "naïve" conditional independence assumption that the presence or absence of an attribute is independent from the presence or absence of other attributes in the attribute set, given the class variable value. Therefore, the predicting probability of class variable $Y=y_i$ conditioned on X= $x=(x_1,x_2,\dots x_n)$ is as follows (Domingos and Plazzani, 1997):

$$\begin{aligned} & P\big(Y = y_i | X = x = (x_1, x_2, \dots, x_n)\big) \\ & = \frac{P(X = x = (x_1, x_2, \dots, x_n) | Y = y_i) P(Y = y_i)}{P\big(X = x = (x_1, x_2, \dots, x_n)\big)} \\ & \propto P(X = x = (x_1, x_2, \dots, x_n) | Y = y_i) \mathrm{P}(Y = y_i) \\ & = \mathrm{P}(Y = y_i) \prod_{j=1}^{n} \mathrm{P}\big(x_j | Y = y_i\big) \end{aligned} \quad (3\text{-}40)$$

NB classifier demonstrates preferable performance in analyzing many real datasets which do not strictly follow the "naïve" independent assumption. Specifically, the impact of the "naïve" assumption on the classification performance of an NB classifier would be insignificant if the classification tool is evaluated by zero-one loss or accuracy (Domingos and Plazzani, 1997). For attributes X_j with discrete values, the probability $p(x_j/y_i)$ was estimated by the proportion of the training instances with both $X_j=x_j$ and the class variable $Y= y_i$ over the number of all instances with the class variable $Y=y_i$ in the training dataset. Continuous or numeric attributes X_j are usually categorized with discretization techniques to enhance model performance, which will also be conducted in this research. The probability inference method for discrete variables is also applicable for the categorized continuous and numeric variables.

Similar to DT learning procedure, NB model learning procedure with cross-validated performance measurement is also very efficient since the frequency of each class could be updated in constant time (Hall and Frank, 2008).

3.4.3 Decision Table/Naïve Bayes (DTNB) Hybrid Model

As a hybrid classification model, a DTNB is an incorporation of a DT and an NB classifier (Hall and Frank, 2008). The learning algorithm for a DTNB is similar to learning a stand-alone DT. At each point of attribute search, the learning algorithm assesses the merit of splitting the entire attribute set into two disjoint attribute subsets, with one modeled by the DT model and the other by a NB classifier. The standard method to choose an optimal attribute set for a DT is to maximize cross-validated performance. In a typical cross-validation procedure, the entire dataset is divided into two segments: one for model learning and the other for model validation, and the training and validation sets must cross-over successively so that each data in the entire dataset is validated (Refaeilzadeh et al., 2009). A commonly used cross-validation method is Leave-One-Out Cross-Validation (LOO-CV) (Witten et al., 2011), which is also applied in this study. LOO-CV is a special type of *n*-fold cross validation, in which *n* is equal to the number of instances in the dataset. In each step of the cross-validation, a single instance in the dataset would be put aside and the rest of the dataset would be used for the training procedure. The trained classifier is tested by its prediction on the left instance, with 1 for success and 0 for failure. This procedure repeats *n* times and ends till each instance in the dataset is used at least once for validation (Kohavi, 1995). Numerous evaluation measurements are generally used for cross-validation, including Root Mean-

Squared Error (RMSE) for numeric classes, accuracy for discrete classes, and AUC. Starting with all attributes modeled by the DT, a greedy search algorithm with forward selection approach is used for attribute splitting procedure in this study, where the selected attributes are modeled with NB classifier and the remaining ones are modeled by DT model in each step. LOO-CV accuracy is applied as an evaluation measurement to assess the quality of attribute split based on probability estimation produced by the hybrid model.

The classification results and probability estimations of response classes from the DT and NB classifier are combined to generate overall modeling results (Hall and Frank, 2008). Let X^{DT} be the attribute set in the DT and X^{NB} be the one in the NB model, where X^{DT} and X^{NB} are complementary with each other. The overall class probability is calculated as follows,

$$P(y|X) = \alpha \times P_{DT}(y|X^{DT}) \times \frac{P_{NB}(y|X^{NB})}{P(y)} \quad (3\text{-}41)$$

where $P_{DT}(y|X^{DT})$ and $P_{NB}(y|X^{NB})$ are the class probabilities estimated by the DT and NB model respectively, α is a normalization constant, and $P(y)$ is the prior probability of the class. The Laplace-corrected observed counts are used in the estimation of all probabilities.

3.5 Conclusions

Traffic Crashes result in significant cost and induce considerable casualties and property losses. Investigating traffic crash data and examining the casual mechanisms is of practical importance. Statistical models and machine-learning methods are two major types of methods that have been extensively used in traffic crash injury severity analysis,

and regression models are the mostly developed and used techniques. Compared with traditional estimation methods extracting parameters of interest solely from the studied dataset, Bayesian methods provides posterior parameter estimations by incorporating parameter prior distribution information and evidence from the studied dataset, which is more reliable and increasingly used in traffic safety studies. This research aims to comprehensively examine the applicability and effectiveness of Bayesian method in traffic crash driver injury severity analyses, including hierarchical Bayesian regression models, Bayesian non-regression models and knowledge-based Bayesian non-parametric method. At the beginning, a Bayesian model selection decision chart is developed based on certain research purpose, crash data availability and data structure, where researchers could select the most appropriate Bayesian model for their studies.

Regression models are the mostly applied research models in traffic crash injury severity analysis, and it is found that hierarchical Bayesian models are more robust and produce more accurate results due to the hierarchical structure of crash data (i.e. road section, crash, vehicle/driver). With Bayesian inference method, hierarchical Bayesian regression models are an indispensable component in Bayesian method family. In this study, three hierarchical Bayesian regression models are considered: hierarchical Bayesian binary logit model, hierarchical Bayesian ordered logit model, and hierarchical random intercept model with cross-level interactions based on the difference in driver injury categorization and model development. In hierarchical Bayesian binary logit model, driver injury severity is assumed to be a binary outcome and the unobserved heterogeneity is simulated by a random error term representing hidden variance among crashes. For hierarchical ordered logit models, driver injury severity is defined as a

variable with 5 increasing severity levels, and the unobserved heterogeneity is modeled by a crash-level variance random error term. These models both have their own disadvantages and could be improved to better excavate driver injury severity patterns. To overcome these disadvantages, a random intercept model is developed in this study, where the driver injury severity is defined as a three-level multinomial variable, and the cross-level interactions between crash and vehicle/driver level variables are considered to better illustrate unobserved heterogeneity in crash data. In all three models, due to the limited crash data availability, non-informative prior are all used for parameter posterior estimation, and the model simulation procedure are conducted in WinBUGS via a Gibbs Sampler, a MCMC algorithm. Traditional regression models are used for model comparison purpose and the DIC measurement is utilized to evaluate model performance. The posterior parameter coefficients are summarized to indicate variable influence on driver injury outcome and 95% BCI are employed to indicate variable significance.

Most regression models have their own model assumptions and pre-defined underlying relationships between dependent and independent variables, which may not hold universally, and violation of these assumptions could lead to erroneous estimations of the likelihood of injury severities. A MNL-BN hybrid model is utilized as a non-regression machine-learning method in this study by relaxing certain hierarchical model in model assumptions to predict driver injury severities, where the multinomial logit model is utilized to select significant variables for driver injury prediction and the BN model is used to train an optimal classifier. The model performance is evaluated in terms of classification measurements such as prediction accuracy, F-measure, Receiver Operating Characteristic (ROC) curve, the area under ROC curve (AUC) and

classification confusion matrix. The variable influences on driver injury severities are evaluated through Bayesian network probability inference procedure.

In order to comprehensively investigate the feasibility of applying Bayesian methods in crash driver injury severity analyses, a DTNB hybrid classifier, which is an incorporation of a decision table and a naïve Bayes classifier and has never been used in traffic safety analysis before, is utilized to identify the deterministic attribute set that best predicts driver injury severities and extract the corresponding decision rules based on these attributes. The model performance would also be evaluated in terms of prediction accuracy, F-measure, ROC curve, AUC value and the confusion matrix. The variable influences on driver injury severities are evaluated based on extracted decision rules for each injury severity.

The applicability and effectiveness of these models are verified using different crash datasets from a complete New Mexico roadway crash dataset collected in 2010 and 2011. These case studies are described and the analysis results are discussed in the following chapters.

Chapter 4 Hierarchical Bayesian Modeling Results

4.1 Hierarchical Bayesian Binary Logit Modeling Analysis

4.1.1 Case Study Data

Although it is less populated in rural areas, traffic crashes occurring in rural locations result in more severe injuries and fatalities (NHTSA,2013). Rural highways are major corridors carrying a significant portion of high speed traffic and are prone to inducing traffic accidents with severe injuries. Therefore, a dataset including 3,939 driver injury records from 3,137 rural interstate crashes is used for model development and estimation in this study. The entire dataset is composed of three major sub-datasets: crash dataset, vehicle dataset and driver dataset, revealing explicit information regarding crash occurrence time and locations, crash types, weather condition, roadway geometry features, vehicle characteristics, driver injury severity, demographic and behavior characteristics. After the variable selection procedure for collinearity avoidance, 12 variables were used as the initial input for hierarchical Bayesian modeling, and significant test was conducted to remove insignificant variables. The significant variables and their impacts are illustrated in Table 4-1.

Table 4-1 Rural Interstate Crash Dataset Description and Statistics.

Hierarchy	Variable	Codes/Values	Driver Injury Severity				Total
			No or light Injury	Percentage	Incapacitating Injury or Fatality	Percentage	
Crash Level Variables	Light	Daylight	2120	87.14%	313	12.86%	2433
		Dawn/Dusk	170	88.08%	23	11.92%	193
		Dark	1121	85.38%	192	14.62%	1313
	Curve	Curve Road	281	84.38%	52	15.62%	333
		Straight Road	3130	86.80%	476	13.20%	3606
	Grade	Grade(including hill, dip, etc)	824	87.47%	118	12.53%	942
		Level	2587	86.32%	410	13.68%	2997
	Number of vehicles in crash	Single vehicle	1705	81.74%	381	18.26%	2086
		Two vehicles	1464	91.96%	128	8.04%	1592
		Multiple vehicles	242	92.72%	19	7.28%	261
	Crash Location	Short distance(less than 0.1 mile)	2096	85.17%	365	14.83%	2461
		Medium distance (between 0.1 mile and 1 mile)	272	80.95%	64	19.05%	336
		Far distance (more than 1 mile)	1043	91.33%	99	8.67%	1142
Vehicle Level Variables	Number of lanes per vehicle direction	One lane	318	86.89%	48	13.11%	366
		Two lanes	2597	86.02%	422	13.98%	3019
		Multiple lanes (three or more)	496	89.53%	58	10.47%	554
	Vehicle Types	Light vehicles (passenger car and van)	1944	84.97%	344	15.03%	2288
		Heavy vehicles (pickup, semi-trucks, bus, trailers, etc.)	1328	88.53%	172	11.47%	1500
		Motorcycles(motorcycle and scooter)	1	50.00%	1	50.00%	2
	Driver Age	Young driver(less than 25)	727	85.73%	121	14.27%	848
		Mid-aged driver(25-63)	2357	87.59%	334	12.41%	2691
		Senior drivers(64 or older)	327	81.75%	73	18.25%	400
	Traffic Control	Traffic Control (no passing zone, stop/yield sign, signal control, railroad gate)	776	85.65%	130	14.35%	906
		No Control	2635	86.88%	398	13.12%	3033
	Wet Road Surface	Wet surface (water, ice/snow, slush, etc)	990	90.49%	104	9.51%	1094
		Dry Road	2421	85.10%	424	14.90%	2845
	Driver Alcohol or Drug Involvement	Driver Alcohol/drug involved	89	61.81%	55	38.19%	144
		Sober Driver	3322	87.54%	473	12.46%	3795
	Driver Gender	Male	2383	88.23%	318	11.77%	2701
		Female	1028	83.04%	210	16.96%	1238

A preliminary statistical analysis was conducted to examine the existence of within-crash correlation based on the assumption that vehicles/drivers involved in the same crash share the same crash characteristics, which may result in a high probability of same driver injury severities. In the studied dataset, 729 crashes were multi-vehicle

crashes, and 644 of them have all the drivers in a same crash suffering the same injury severities, accounting for 88.3% of all multi-vehicle crashes. Therefore, within-crash correlation is assumed to exist in this dataset and should be considered in model development and specifications.

4.1.2 Model Fit and Estimation Results

After checking potential variable collinearity and removing less important variables, 12 variables were used as the initial input for hierarchical Bayesian binary logit modeling. In the finalized model, two crash-level variables and four vehicle/driver-level variables were retained for posterior distribution learning, and three chains with different initial value settings were simulated for 20000 iterations in which the first 5000 iterations were discarded as "burn-ins". The trace plots of three iteration chains revealed a good mixing, and Brooks, Gelman and Rubin (BGR) convergence diagnostics illustrated satisfied modeling convergence (Brooks and Gelman, 1998). In order to reduce the autocorrelation among the sampled data, the posterior samples in every fifth iteration were extracted and retained for result generalization (Spiegelhalter et al., 2003), with a storage of 6000 samples in total. The model estimation results are summarized in Table 2.

Table 4-2 Hierarchical Bayesian Binary Logit Model Posterior Estimation Results.

Parameter	Posterior Point Estimate			95% BCI of Odds Ratio	
	Mean	Standard Deviation	Odds Ratio	2.50%	97.50%
Number of Vehicles in a crash					
Single Vehicle	1.50	0.23	4.48	2.91	7.22
Multiple Vehicle	-0.18	0.49	0.83	0.32	2.13
Two Vehicle*	0.00	0.00	1.00	1.00	1.00

Table 4-2 (Continued)

Parameter	Posterior Point Estimate			95% BCI of Odds Ratio	
	Mean	Standard Deviation	Odds Ratio	2.50%	97.50%
Wet Road Surface	-0.88	0.22	0.42	0.26	0.62
Vehicle Type					
Heavy Vehicle	-0.09	0.19	0.91	0.63	1.30
Motorcycle**	4.46	2.61	86.75	0.62	16865.07
Light Vehicle*	0.00	0.00	1.00	1.00	1.00
Driver Age					
Young driver (16-25)	-0.005	0.20	0.99	0.67	1.47
Senior driver (>63)	0.91	0.27	2.47	1.49	4.29
Mid-Age (25-63)*	0.00	0.00	1.00	1.00	1.00
Driver alcohol or drug involvement	2.51	0.42	12.35	5.84	30.05
Driver Gender	-0.64	0.18	0.52	0.36	0.75
Random Effects					
Between-crash Variance(σ_0^2)	6.45	1.97			
Within-crash Variance(σ_v^2)	3.29				
ICC	0.662				

Note: *reference category for a multinomial variable
**Significant at 90% BCI

As illustrated in Table 4-2, the ICC value is 0.662, indicating that 66.2% of the total variance in the response variable was explained by the variance among different crashes. This is consistent with the fact that most of all rural interstate crashes (2408 of 3137) are single vehicle crashes, generating significant between-crash variance contributing to overall data variance. The relatively large ICC value indicates the preference of the hierarchical Bayesian model in this analysis.

In this study, an ordinary binary logit model was also used as a control model for performance comparison on the same dataset. The DIC values for both models are listed in Table 4-3. The overall DIC value of the proposed hierarchical Bayesian model is 2522.69, which is lower than that for ordinary logit model (2928.25), verifying that the

hierarchical Bayesian logit model is superior to the control model in model fit, and that including between-crash variance into the proposed model could sustainably improve model performance.

Table 4-3 DIC Results for Model Comparison.

	$\overline{D(\gamma)}$	$D(\bar{\gamma})$	pD	DIC
Hierarchical Bayesian binary logit model	1726.6	930.517	796.086	2522.69
Ordinary binary logit model	2917.98	2907.71	10.272	2928.25

4.1.3 Model Analysis results

Six variables were considered significant in predicting driver injury severities in rural interstate crashes, including two crash-level variables and four vehicle/driver-level variables: number of vehicles in a crash, wet road surface, vehicle type, driver age, driver alcohol or drug involvement and driver gender. The variables are listed in Table 4-2, and explicit discussions of these variables occur below.

There are three discrete levels categorizing the number of vehicles in a crash: single vehicle, two vehicles and multiple vehicles. In this analysis, two-vehicle crashes are treated as the reference category. The estimated odds ratio for single-vehicle crashes, 4.48, suggests that the probability for drivers suffering incapable injuries or deaths is 3.48 times higher in single-vehicle crashes than that in two-vehicle crashes. The 95% BCI of its odds ratio (2.91, 7.22) verifies its statistical significance. Compared to rural interstate crashes with two vehicles involved, multi-vehicle rural interstate crashes tend to induce less severe driver injuries, indicated by the estimated mean odds ratio 0.83. However, the effect is not significant based on its 95% BCI (0.32, 2.13). This discovery is consistent with previous studies. According to the NHTSA (2013), there were 1,661,000 single-

vehicle crashes and 3,677,000 multi-vehicle crashes in the US in 2011, of which 17,991 and 11,766, respectively, were fatal crashes. This indicates that there was a higher probability for severe injuries or deaths in single-vehicle crashes. Further analyses also indicate that single-vehicle crashes and multi-vehicle crashes should be examined separately due to their distinctive mechanisms in causing traffic casualties. For instance, Savolainen and Mannering (2007) applied two different models to analyze motorcyclists' casualties in single-vehicle and multi-vehicle crashes separately.

Wet road surface condition was found to be a significant variable in predicting driver injury severities in rural intestate crashes (95% odds ratio BCI (0.26, 0.62)). Its estimated mean odds ratio (0.42) indicates that wet road surface could reduce the probability of drivers being incapably injured or killed by 58% compared to dry road surface conditions. This finding is also reinforced by previous research. Haque et al. (2012) found that wet surface leads to decrease of motorcycle crash risks and concluded that motorcycle drivers tend to be more careful when driving on wet road surfaces. Shaheed et al. (2013) discovered that dry pavement conditions significantly increase the potential of fatal and major injury in motorcycle-involved crashes. Through probabilistic modeling, Savolainen and Mannering (2007) found that crashes occurred under wet road surface conditions tend to be less severe. However, some studies also draw seemingly contradictory conclusions, indicating that wet conditions are a contributory factor to traffic crashes. For example, Caliendo et al. (2007) found that wet road conditions are a significant factor increasing crash frequency. The contradiction in research findings is explainable. Although the crash frequency is increasing due to low skid resistance, road users tend to be more aware of the adverse pavement surface condition and drive at

relatively low speeds. However, in dry and clear road conditions, the odds of traffic safety might be reduced by the propensity of speeding. Other studies generate composite conclusions regarding the safety effect of wet pavement conditions. Morgan and Mannering (2011) found that there is significant recognition difference for drivers of different age groups and genders on wet pavement conditions. Wet or snowy/icy road surfaces tend to decrease the probability of severe injury for male drivers less than 45 years old but increase for the other driver groups. Mayora and Piña (2009) investigated the impact of skid resistance of both wet and dry road surfaces on traffic safety and summarized that the increase of skid resistance is negatively associated with crash rates regardless of pavement surface conditions. This indicates that pavement surface condition is a complex factor related to crash risks and injury severities, and it needs to be comprehensive examined.

Vehicle type is not a statistically significant factor affecting driver injury severities in rural interstate crashes based on the 95% BCI, but the motorcycle category is significant at 90% BCI. Compared to drivers in light vehicles, drivers in heavy vehicles tend to suffer less severe injuries in rural interstate crashes, indicated by the estimated mean odds ratio 0.91. This severity probability reduction is small (9%) and insignificant (95% BCI of odds ratio (0.63, 1.30)). Motorcycle drivers are more likely to get incapably injured or killed in crashes, with a probability increase of more than 80 times (odd ratio=86.75). This effect is not significant based on the 95% BCI (0.62, 16868.07) but significant based on the 90% BCI of odds ratio (1.41, 6891.20). The large variance and insignificance in the estimation for motorcycles are possibly due to the limited number of motorcycle records in the dataset (Table 4-1) in which insufficient information on injury

mechanism and pattern has been provided. However, as an important vehicle type on highways, motorcycles should not be ignored in this study, and more comprehensive data are desired to enhance the reliability of the estimation. Although not significant, the driver injury patterns and tendencies for heavy vehicles and motorcycles revealed in this research are understandable. Heavy vehicles are of significant size and weight where drivers are more protected, while motorcyclists are more exposed to open traffic environments and more vulnerable in crashes. Support for these findings has been offered from other related studies. Kockelman and Kweon (2002) discovered that motorcyclists are expected to suffer more severe injuries compared to vehicle drivers. Chiang et al. (2014) found that motorcyclists are the most vulnerable driver group on roadways. More detailed research found that head injury is the main cause of motorcyclist deaths and helmet use is effective prevention of driver trauma (Hefny et al., 2012; Kelly et al., 1991). For heavy vehicles, Levine et al. (1999) discovered that vehicle weight increase could enhance the driver's capability of enduring the front impact from crashes, and therefore reduce driver injuries. Overall, motorcycles and heavy vehicles are important factors for driver injuries severities. Hence, law enforcement on these vehicles and defensive driving training for the corresponding drivers are recommended.

Driver age is found to be significant in affecting driver injury severities, especially for senior drivers that are over 64 years old (95% odds ratio BCI (1.49, 4.29)). The estimated mean of odds ratio for senior drivers is 2.47, suggesting that the odds of senior drivers sustaining incapable injuries or deaths in rural interstate crashes are 147% higher than that for mid-age drivers. This finding has been proven by earlier studies. Kim et al. (2013) discovered that older drivers (>63 years old) are a significant factor

increasing the odds of fatal injuries in crashes. Kockelman and Kweon (2002) proposed that senior drivers are less likely to make appropriate and immediate responses when facing crash risks due to their relative slow reactions. On the other hand, young drivers are more likely to engage in careless driving or speeding, resulting in a considerable potential for severe injuries. Huang et al. (2008) showed that both senior driver and young driver groups are more likely to suffer severe injuries in traffic crashes. In this analysis, however, young age (16-25) is not significantly associated with driver injury severities compared to mid-age drivers based on the 95% BCI of odds ratio (0.67, 1.47), which does not seem to be supported by previous studies. This could be explained by the fact that in this analysis, all types of rural interstate crashes were bundled in the study dataset and the driver age impact was not examined by crash type. As is discovered by Yasmin et al. (2014), young drivers are more likely to be involved in rear-end crashes due to insufficient driving experience and inferior distance judgment, while older drivers are more associated with angular collisions due to their relative slow reaction and inability to maneuver quickly to complete turning actions. Moreover, Abdel-Aty et al. (1998) comprehensively evaluated the effects of driver age across different traffic-related factors of traffic accident involvement, which indicated the importance of interactive effects between driver age with crash-related factors. As a result, there should be further investigations to enrich this research.

Driver involvement of alcohol or drugs is found to significantly increase the probability of drivers with incapable or fatal injuries, illustrated by the 95% BCI of odds ratio (5.84, 30.05). It is shown that drivers with drug or alcohol usage have a probability of being incapably or fatally injured that is 11.35 times (odds ratio=12.35) higher than

that of drivers without any use of alcohol or drugs. This is reasonable since alcohol and drugs have significant effects in impairing drivers' judgment and visibility. This discovery echoes our common sense and receives unanimous proof from other studies. Weiss et al. (2014) concluded that alcohol use is one of the fatal causes in single-vehicle crashes. Poulsen et al. (2014) testified to the independent effect of cannabis and combined effect of alcohol and cannabis in increasing crash potential. Using a case-control experiment design, Hels et al. (2013) verified the close association between high risk of severe driver injury and high concentration of alcohol in bodies. Therefore, law enforcement of blood alcohol concentration (BAC) testing and drunken driving prohibition should be enhanced.

Driver gender is statistically significant in predicting driver injury severities in rural interstate crashes, illustrated by the 95% BCI of its odds ratio (0.36, 0.75). The estimated mean of odds ratio (0.52) indicates that the probability of male drivers with incapable or fatal injuries is 48% less than that for female drivers in rural interstate crashes. Kockelman and Kweon (2002) also discovered that male drivers are associated with lower driver injury severities compared to female drivers. Islam and Mannering (2006) identified that female drivers have more interacting factors to increase the likelihood of injuries and deaths comparing to male drivers. However, other studies provided opposite findings. Massie et al. (1995) concluded that vehicles with male drivers are more likely to be involved in fatal crashes than female drivers. Kim et al. (2013) found that male drivers are a contributing factor to fatal injuries in single-vehicle crashes. To be more specific and accurate, Ulfarsson and Mannering (2004) examined the distinctive effects of males and females and their respective interactive effects with other

factors on injury severities. Other previous studies also addressed the interactive effects of driver gender with other factors on traffic crashes rather than examining the gender effect alone (Hels et al., 2013; Morgan and Mannering, 2011; Poulsen et al., 2014). The gender effect on driver injury severity in this research is a general conclusion for rural interstate crashes and a detailed examination of gender effects across other crash-related factors should be conducted.

4.2 Hierarchical Bayesian Ordered Logit Modeling Results

4.2.1 Case Study Data

According to NMDOT (2012), among rural fatalities, 73.9% happened at rural non-interstate locations, despite rural interstate highways carrying the primary portion of rural traffic volume. Based on the complete dataset including all reported crashes in New Mexico in 2010 and 2011, a rural non-interstate crash dataset is extracted for this case study**.** Special effort was taken to examine and remove incomplete and erroneous records, such as records with driver gender information as "unknown." Overall, the studied dataset contains 10,770 vehicles involved in 8,580 crashes occurring at rural non-interstate locations, with an average of 1.26 vehicles in each crash. Each record in the studied dataset indicates a vehicle/driver unit in a crash. The response variable representing driver injury severity is ordinal with five injury levels: no injury, complaint of injury/possible injury, visible injury, incapacitating injury and death, denoted by integer numbers from 1 to 5, respectively. The detailed descriptive statistics of the dataset are illustrated in Table 4-4 below.

Table 4-4 Rural Non-interstate Crash Dataset Description and Statistics.

Hierarchy	Variable	Codes/Values	Driver Injury Severity										Total
			No Injury	Percentage	Possible Injury	Percentage	Visible Injury	Percentage	Incapacitating Injury	Percentage	Fatality	Percentage	
Crash Level Variables	Light	Daylight	5077	73.80%	910	13.23%	550	8.00%	262	3.81%	80	1.16%	6879
		Dawn/Dusk	485	78.35%	67	10.82%	43	6.95%	18	2.91%	6	0.97%	619
		Dark	2406	73.53%	404	12.35%	283	8.65%	136	4.16%	43	1.31%	3272
	Curve	Curve Road	1442	69.46%	287	13.82%	220	10.60%	98	4.72%	29	1.40%	2076
		Straight Road	6526	75.06%	1094	12.58%	656	7.55%	318	3.66%	100	1.15%	8694
	Grade	Grade(including hill, dip, etc)	1906	74.28%	320	12.47%	216	8.42%	98	3.82%	26	1.01%	2566
		Level	6062	73.89%	1061	12.93%	660	8.04%	318	3.88%	103	1.26%	8204
	Number of vehicles in a crash	Single vehicle	3682	71.34%	635	12.30%	542	10.50%	231	4.48%	71	1.38%	5161
		Two vehicles	4013	77.25%	673	12.95%	294	5.66%	164	3.16%	51	0.98%	5195
		Multiple vehicles	273	65.94%	73	17.63%	40	9.66%	21	5.07%	7	1.69%	414
	Crash Location	Short distance(less than 0.1 mile)	4968	72.40%	922	13.44%	590	8.60%	299	4.36%	83	1.21%	6862
		Medium distance (between 0.1 mile and 1 mile)	470	67.72%	92	13.26%	92	13.26%	16	2.31%	24	3.46%	694
		Far distance (more than 1 mile)	2530	78.72%	367	11.42%	194	6.04%	101	3.14%	22	0.68%	3214
	Maximum Vehicle Damage in Crash	No/Slight	2612	90.26%	214	7.39%	45	1.55%	16	0.55%	7	0.24%	2894
		Functional (affecting vehicle normal operation)	2080	87.10%	197	8.25%	81	3.39%	29	1.21%	1	0.04%	2388
		Disabled (Vehicle can't be driven)	3276	59.69%	970	17.67%	750	13.67%	371	6.76%	121	2.20%	5488
Vehicle Level Variables	Number of lanes per vehicle direction	One lane	4562	73.81%	777	12.57%	521	8.43%	238	3.85%	83	1.34%	6181
		Two lanes	2992	74.39%	520	12.93%	317	7.88%	150	3.73%	43	1.07%	4022
		Multiple lanes (three or more)	419	73.12%	84	14.66%	39	6.81%	28	4.89%	3	0.52%	573

Table 4-4 (Continued)

Hierarchy	Variable	Codes/Values	Driver Injury Severity										Total
			No Injury	Percentage	Possible Injury	Percentage	Visible Injury	Percentage	Incapacitating Injury	Percentage	Fatality	Percentage	
Vehicle Level Variables	Vehicle Types	Light vehicles (passenger car and van)	4898	71.96%	973	14.29%	571	8.39%	290	4.26%	75	1.10%	6807
		Heavy vehicles (pickup, semi-trucks, bus, trailers, etc.)	3069	77.85%	405	10.27%	295	7.48%	119	3.02%	54	1.37%	3942
		Motorcycles(motorcycle and scooter)	1	4.76%	3	14.29%	10	47.62%	7	33.33%	0	0.00%	21
	Driver Age	Young driver(less than 25)	2092	71.16%	401	13.64%	298	10.14%	117	3.98%	32	1.09%	2940
		Mid-aged driver(25-63)	5129	75.23%	864	12.67%	493	7.23%	260	3.81%	72	1.06%	6818
		Senior drivers(64 or older)	747	73.81%	116	11.46%	85	8.40%	39	3.85%	25	2.47%	1012
	Traffic Control	Traffic Control (no passing zone, stop/yield sign, signal control, railroad gate)	3258	73.33%	587	13.21%	367	8.26%	184	4.14%	47	1.06%	4443
		No Control	4710	74.44%	794	12.55%	509	8.04%	232	3.67%	82	1.30%	6327
	Wet Road Surface	Wet surface (water, ice/snow, slush, etc)	1697	77.38%	288	13.13%	134	6.11%	55	2.51%	19	0.87%	2193
		Dry Road	6271	73.11%	1093	12.74%	742	8.65%	361	4.21%	110	1.28%	8577
	Driver Seatbelt Use	Seatbelt used	7890	75.17%	1346	12.82%	826	7.87%	357	3.40%	77	0.73%	10496
		Seatbelt not used	78	28.47%	35	12.77%	50	18.25%	59	21.53%	52	18.98%	274
	Driver Alcohol or Drug Involvement	Driver Alcohol/drug involved	396	51.03%	121	15.59%	140	18.04%	87	11.21%	32	4.12%	776
		Sober Driver	7572	75.77%	1260	12.61%	736	7.36%	329	3.29%	97	0.97%	9994
	Driver Gender	Male	5147	76.58%	706	10.50%	523	7.78%	249	3.70%	96	1.43%	6721
		Female	2821	69.67%	675	16.67%	353	8.72%	167	4.12%	33	0.82%	4049

4.2.2 Model Fit and Estimation Results

The model simulation procedure was conducted with a Monte Carlo Markov Chain (MCMC) algorithm in WinBUGS. With the first 5000 iterations as "burn-ins," sufficient iterations have been simulated and model convergence was achieved. Table 4-5 illustrates the analyses results of the proposed hierarchical ordered logit model. Compared to generalized ordered-response models, this paper applies hierarchical model structure specification by taking between-crash variance into consideration and utilizes 95% BCI to illustrate the significance of the estimated parameter. As discussed before, the random effect u_i follows a normal distribution $(0,\sigma^2)$. It is shown in Table 2 that the estimated mean of σ^2 is 3.091, with its 95% BCI (2.548, 3.797) indicating that it is significantly different from 0. This verifies the existence of between-crash variance, which should be considered in crash data modeling. Sufficient sample values of u_i were randomly selected for model assumption checking, and it was found that these values are normally distributed, which verifies the appropriateness of the model utilized in this study.

Table 4-5 Hierarchical Bayesian Ordered Logit Model Posterior Estimation Results.

Variables		Estimated Mean	Standard Deviation	95% BCI of Mean	
				2.50%	97.50%
Driver Age	Young Drivers	-0.044	0.070	-0.180	0.095
	Elder Drivers	**0.256**	**0.119**	**0.021**	**0.484**
	Mid-Aged Drivers*	0.000	0.000	0.000	0.000
Crash Location	0.1-1.0mile	0.205	0.126	-0.040	0.453
	Larger than 1.0mile	**-0.326**	**0.083**	**-0.492**	**-0.165**
	Less than 0.1 mile*	0.000	0.000	0.000	0.000
Lighting Condition	**Dark**	**-0.323**	**0.080**	**-0.480**	**-0.166**
	Dawn and Dusk	**-0.564**	**0.161**	**-0.881**	**-0.244**
	Daylight*	0.000	0.000	0.000	0.000

Table 4-5 (Continued)

Variables		Estimated Mean	Standard Deviation	95% BCI of Mean	
				2.50%	97.50%
Number of Vehicles in a Crash	**Two Vehicles**	**-0.306**	**0.077**	**-0.460**	**-0.157**
	Three Vehicles or More	0.154	0.198	-0.239	0.532
	Single Vehicle*	0.000	0.000	0.000	0.000
Vehicle Type	**Heavy Vehicles**	**-0.337**	**0.071**	**-0.478**	**-0.200**
	Motorcycles	**4.379**	**0.549**	**3.316**	**5.472**
	Light Vehicles*	0.000	0.000	0.000	0.000
Maximum Vehicle Damage in Crash	**Functional Damage**	**0.475**	**0.123**	**0.235**	**0.721**
	Disabled Damage	**2.612**	**0.115**	**2.390**	**2.842**
	No/Slight Damage*	0.000	0.000	0.000	0.000
Road Curve		**0.190**	**0.086**	**0.022**	**0.362**
Wet Road Surface		**-0.248**	**0.091**	**-0.425**	**-0.072**
Seat Belt Use		**-3.146**	**0.200**	**-3.556**	**-2.767**
Driver with Impairment		**1.083**	**0.115**	**0.859**	**1.311**
Male Driver		**-0.599**	**0.070**	**-0.738**	**-0.462**
Latent Thresholds	***h[1]***	**-0.559**	**0.208**	**-0.961**	**-0.115**
	h[2]	**0.849**	**0.208**	**0.427**	**1.299**
	h[3]	**2.459**	**0.214**	**2.028**	**2.909**
	h[4]	**4.457**	**0.234**	**3.999**	**4.922**
σ^2		**3.091**	**0.335**	**2.548**	**3.797**

*Reference Category for the associated multinomial variable

An ordinary ordered logit model dismissing the between-crash variance term u_i was also employed in this study to examine the same dataset, and Table 4-6 demonstrates the DIC values for the two models indicating model fit. It shows that the simulation procedure through hierarchical Bayesian ordered logit model produces a lower DIC value, suggesting that the proposed model is superior in analyzing the selected dataset.

Table 4-6 DIC Result for Model Comparison.

Model Design	DIC
Hierarchical Bayesian ordered logit model	15708.3
Ordinary ordered logit model	16716.5

4.2.3 Factor Impact Analysis

The significant variables extracted for driver injury severity prediction are highlighted in Table 4-5, including the significant categorical values in multinomial variables. 11 variables were identified to be significant, including three variables describing crash features, three variables characterizing environmental conditions at occurrence, two variables describing driver demographic features, two variables explaining driving status and one variable representing vehicle type, some of which are were also found significant for driver injury severity prediction in Section 4.1. The positivity or negativity of the estimated mean indicates the increasing or decreasing effect on driver injury severity. The detailed effects of these variables are discussed below.

The three factors regarding crash features found to be significant in predicting driver injury severity are crash location, number of vehicles in the crash and maximum vehicle damage in the crash, and the last two were also found significant in previous studies. Crash location, represented by the distance to the nearest intersection, is an important factor associated with driver injury severity. Compared to intersection-related locations (less than 0.1 mile to intersection), far crash locations (larger than 1.0 mile) are prone to inducing less driver injury severities, indicated by the estimated mean value (-0.326) and 95% BCI (-0.492, -0.165). Medium crash locations (between 0.1 and 1.0 mile) are likely to cause higher driver injury severities in crashes, but the increasing effect is ambiguous, as illustrated by its 95% BCI across zero. These results signify that intersection-related locations are crash hotspots for the more severe injury severity outcomes. This is reasonable since intersections or intersection-related locations are characterized with complicated traffic movements, and any inappropriate acceleration,

deceleration or unattended driving may lead to crash occurrence and injuries. On the other hand, traffic flow at further locations is relatively stable without much fluctuation, which reduces the potential of crash occurrence and severe injuries. This result justifies the purpose of traffic safety studies regarding intersections (Huang et al., 2008; D.-G. Kim et al., 2007; Wang and Abdel-Aty, 2006).

The number of vehicles involved in a crash, which has been used to define crash types in some studies (Chen and Chen, 2011; Geedipally and Lord, 2010), is an important risk factor for driver injury severity prediction in rural non-interstate crashes. Using single-vehicle crashes as the reference category, the analyses results show that two-vehicle crashes reduce driver injury severities significantly (Mean=-0.306, 95% BCI (-0.460, -0.157)), while multiple-vehicle crashes tend to increase driver injury severities (Mean=0.154). However, the impact is not significant (95% BCI (-0.239, 0.532)).

It is to be expected that the maximum vehicle damage in a crash is closely associated with driver injury severity. As shown in the results, both functional and disabled vehicle damages have positive correlations with driver injury outcomes, where a higher posterior mean for disabled vehicle damage (Mean=2.612) is estimated. This indicates a larger impact on increasing injury severity than functional vehicle damage (Mean=0.475). As discussed before, it is reasonable since maximum vehicle damage is the deformation caused by the crash impact produced in collisions, and it is a reflection of the transferrable impact from vehicles to drivers.

Three variables describing environment elements were found to be significantly associated with driver injury severity： road curvature, road surface condition and

lighting condition. It was found in this study that road curvature is significantly associated with higher driver injury severity. This is to be expected as drivers on road curves usually have restricted visibilities on further road conditions. Road curvatures also require drivers to take particular care in order to maneuver vehicles properly. Both of these factors associated with road curves increase the risk of higher driver injuries in crashes. More specifically, van Petegem and Wegman (2014) concluded that road curvature is a major factor that increases the potential for run-off-road crashes.

The research results also suggest that wet road surfaces, contrary to the common expectations, are prone to reducing driver injury severity in traffic crashes at rural non-interstate locations, as inferred from the estimated mean (-0.248) and 95% BCI (-0.425, -0.072). An explanation for this result is that drivers on rural wet or icy roads tend to be more cautious in order to avoid crashes. However, on comfortable road conditions, drivers are more likely to engage in reckless driving where the driving safety and comfort are compromised. This factor was also found significant in Section 4.1 but with the exacerbating influence on driver injury severity. As discussed before, road surface condition provides complex influence under different traffic and vehicle conditions, and should be further investigated. Similar results were also obtained regarding lighting conditions at crash occurrence. The analyses results demonstrated that drivers in rural non-interstate crashes occurring under dawn, dusk and dark nighttime conditions, are less likely to get severely injured, compared with those occurring during daylight conditions. This is probably because drivers with inferior light conditions are aware of the limited visibility of the external traffic environment and drive more carefully than when driving under daylight conditions.

Two driver demographic variables were identified to have significant influence on driver injury outcomes in rural non-interstate crashes: age and gender, which were both found significant in hierarchical Bayesian binary logit results as well. Taking mid-aged drivers as the base category, senior drivers are significantly more vulnerable to higher injury severity in these crashes, suggested by the estimated mean (0.256) and 95% BCI (0.021, 0.484). This is understandable since senior drivers are less agile in maneuvering a vehicle and it takes more time for them to make appropriate and timely responses to deal with traffic emergencies. Young drivers tend to be less severely injured than mid-aged drivers (mean=-0.044), but with an insignificant trend (95% BCI=(-0.180, 0.095)). Compared with female drivers, male drivers, whose estimated mean and 95% BCI are -0.599 and (-0.738, -0.462), respectively, are significantly more likely to suffer less severe injury in rural non-interstate crashes. In other words, females are more likely to suffer severe injuries or death in rural non-interstate crashes.

Seatbelt use and driver with alcohol/drug impairment, two variables describing driving status, are significantly correlated with driver injury severity, and they were found in Section 4.1 as well. In this analysis, driver seatbelt use is found to provide effective protection for drivers from being severely injured, suggested by the negative estimation (-3.146) and 95% BCI (-3.556, -2.767). This finding verifies the protective effect of seatbelt use while driving. The estimated results illustrate that driver with impairment, beyond no expectation, is significantly and positively associated with driver injury severities, with an estimated mean of 1.083 and 95% BCI (0.859, 1.311). It implies that driver use of alcohol and/or drugs significantly increases driver injury outcomes in crashes due to the fact that alcohol and drugs considerably undermine drivers' normal

vision and judgment, making it difficult for drivers to perform appropriately when driving. Therefore, even though current blood alcohol concentration (BAC) tests sets a certain threshold for permitted alcohol absorption, zero-tolerance of driver alcohol and drug use while driving should be advocated.

In terms of vehicle type, it is not surprising that motorcyclists are more vulnerable in rural non-interstate crashes than light vehicle drivers, as showed by the estimated results (Mean=4.379, 95% BCI (3.316, 5.472)). This is to be expected since motorcyclists are the most exposed to traffic environments and their driving behaviors are more agile and unpredictable than other road users. On the other hand, drivers of heavy vehicles, such as pickup trucks, semi-trailers, buses, etc., are less related with severe injuries, indicated by the negative estimated coefficient. This result supports the finding by Levine et al. (1999) that heavy vehicles are able to withstand higher crash impacts than other vehicles due to their relative large size and weight, which could provide more protection for the drivers from being severely injured. However, it should be noted that heavy vehicles impose more impact on other vehicles and drivers in the crash, resulting in more damage and higher injury severity. Therefore, more specific restrictions regarding vehicle size, weight and speed should be enforced on heavy vehicles.

4.3 Hierarchical Random Intercept Model with Cross-Level Interaction Analysis

4.3.1 Case Study Dataset

In this analysis, we utilized a dataset containing all truck records in rural crashes extracted from two-year crash records in the State of New Mexico in 2010 and 2011,

provided by the New Mexico Department of Transportation (NMDOT) and Geospatial and Population Studies (GPS) at the University of New Mexico (UNM). In this study, the five-level driver injury severity (as was discussed in Section 4.1.2.1) is simplified to three categorical levels based on the data size for each category and the similarities among driver injury severity levels: no injury (original Category O, coded as **N**), non-incapacitating injury (original Categories B and C, coded as **I**), and incapacitating injury and fatality (original Categories A and K, coded as **F**). In total, there are 5,398 eligible truck records from 4,868 rural crashes included in the studied dataset, which results in 1.11 vehicles per crash on average. Each record in the studied dataset represents a truck unit involved in a rural crash, accompanied with detailed driver and crash information. Detailed information of the studied dataset is shown in Table 4-7.

Table 4-7 Rural Truck Crash Dataset Description and Statistics.

Variable Description	Driver Injury Severity						Total
	N	Percentage	**I**	Percentage	FATALITY	**F**	
Driver Injury Severity	4231	78.38%	926	17.15%	241	4.47%	5398
Crash-Level Variables							
Intersection Related							
Intersection related	506	81.48%	97	15.62%	18	2.90%	621
Not intersection related	3725	77.98%	829	17.35%	223	4.67%	4777
First Harmful Event Location							
On road	3289	79.93%	651	15.82%	175	4.25%	4115
Off road*	942	73.42%	275	21.43%	66	5.14%	1283
Lighting Condition							
Dark	1279	75.77%	314	18.60%	95	5.63%	1688
Dawn/dusk	245	81.67%	43	14.33%	12	4.00%	300
Daylight*	2707	79.38%	569	16.69%	134	3.93%	3410
Road Curvature							
Curve road	640	74.16%	166	19.24%	57	6.60%	863
Straight road*	3591	79.18%	760	16.76%	184	4.06%	4535

Table 4-7 (Continued)

Variable Description	Driver Injury Severity						Total
	N	Percentage	**I**	Percentage	FATALITY	**F**	
Road Grade							
Road with grade	982	78.56%	214	17.12%	54	4.32%	1250
Level road*	3249	78.33%	712	17.16%	187	4.51%	4148
Number of Vehicles in Crash							
One vehicle	1877	73.49%	527	20.63%	150	5.87%	2554
Two vehicles*	2139	83.29%	353	13.75%	76	2.96%	2568
Three or more	215	77.90%	46	16.67%	15	5.43%	276
Hazard Material Involvement							
Hazard material involved	19	76.00%	5	20.00%	1	4.00%	25
Otherwise*	4212	78.39%	921	17.14%	240	4.47%	5373
Maximum Vehicle Damage in Crash							
Slight damage*	1312	93.05%	90	6.38%	8	0.57%	1410
Functional damage	1062	90.77%	91	7.78%	17	1.45%	1170
Disabled damage	1857	65.90%	745	26.44%	216	7.67%	2818
Vehicle-Level Variables							
Driver Residency							
Non New Mexico driver	1563	80.19%	302	15.50%	84	4.31%	1949
New Mexico driver*	2668	77.36%	624	18.09%	157	4.55%	3449
Road Pavement							
Road paved	4002	78.39%	880	17.24%	223	4.37%	5105
Road not paved*	229	78.16%	46	15.70%	18	6.14%	293
Wet Road Surface							
Wet road	1126	80.77%	224	16.07%	44	3.16%	1394
Dry road*	3105	77.55%	702	17.53%	197	4.92%	4004
Traffic Control							
Traffic control	1449	78.62%	307	16.66%	87	4.72%	1843
No traffic control*	2782	78.26%	619	17.41%	154	4.33%	3555
Number of Lanes Available for That Car's Travel							
One lane*	2378	78.25%	518	17.05%	143	4.71%	3039
Two lanes	1549	77.53%	361	18.07%	88	4.40%	1998
Three or more	309	84.20%	48	13.08%	10	2.72%	367
Vehicle Action							
Go straight*	3606	77.67%	817	17.60%	220	4.74%	4643
Acceleration or deceleration	302	86.04%	37	10.54%	12	3.42%	351
Turn	323	79.95%	72	17.82%	9	2.23%	404
Driver Seatbelt Use							
Seatbelt is used	4193	79.55%	895	16.98%	183	3.47%	5271
Seatbelt not used*	38	29.92%	31	24.41%	58	45.67%	127

Table 4-7 (Continued)

Variable Description	Driver Injury Severity						Total
	N	Percentage	**I**	Percentage	FATALITY	**F**	
Driver Age							
Young: 24 or younger	634	73.46%	175	20.28%	54	6.26%	863
Mid-aged: between 25 to 63*	3231	79.48%	671	16.51%	163	4.01%	4065
Senior: 64 or older	366	77.87%	80	17.02%	24	5.11%	470
Driver Under Influence							
Driver under influence	157	51.99%	97	32.12%	48	15.89%	302
Driver not under* influence	4074	79.95%	829	16.27%	193	3.79%	5096
Driver Gender							
Male	3567	80.12%	692	15.54%	193	4.34%	4452
Female*	664	70.19%	234	24.74%	48	5.07%	946

Note: * reference category used in the model.

4.3.2 Model Fit and Estimation Results

All the crash-level and vehicle/driver-level variables listed in Table 4-7 were used as inputs for model development. Due to the relative high complexity of this model, a single chain was simulated for 65,000 iterations, and the trace plot of the iteration chain was examined to ensure reasonable model convergence. The convergence was reached after 50,000 interactions, and therefore the first 50,000 iterations were discarded as "burn-ins" (Cowles, 2003). To reduce auto-correlation of the extracted samples, every fifth sample after "burn-ins" was extracted as posterior samples, with a total of 3000 samples for parameter estimation. The significant variables and their impacts on driver injury severity in terms of estimated mean, standard deviation, and 95% BCI for the estimated mean are summarized in Table 4-8.

Table 4-8 Hierarchical Bayesian Random Intercept Model Estimation Results.

Variable	Specific to	Estimated Mean	Standard Deviation	95% BCI of Mean 2.50%	97.50%
Constant(Intercept)*	**I**	**-4.349**	**0.940**	**-6.369**	**-2.729**
Constant(Intercept)	**F**	**-9.427**	**2.512**	**-14.230**	**-5.750**
Intersection Related	I	1.783	1.808	-2.149	5.145
First Harmful Event on Road	I	-0.736	0.808	-2.626	0.678
Curve Road	I	1.732	0.996	-0.233	3.647
Road Grade	**I**	**3.304**	**1.294**	**0.926**	**5.954**
Lighting Condition					
Dark	I	-0.824	1.863	-4.495	2.897
Dawn/Dusk	I	-0.864	0.905	-2.694	0.868
Daylight**	I	0.000	0.000	0.000	0.000
Maximum Vehicle Damage in Crash					
Slight damage**	I	0.000	0.000	0.000	0.000
Functional damage	I	2.225	1.420	-0.476	5.194
Disabled damage	**I**	**3.301**	**0.851**	**1.735**	**4.969**
Road Pavement	**I**	**2.256**	**0.717**	**0.728**	**3.648**
Wet Road	I	0.668	0.363	-0.037	1.357
Vehicle Action					
Go straight**	I	0.000	0.000	0.000	0.000
Acceleration or deceleration	**I**	**-1.748**	**0.898**	**-3.766**	**-0.142**
Turn	**I**	**-3.826**	**1.065**	**-6.164**	**-1.897**
Driver Seatbelt Use	I	0.432	1.032	-1.066	2.745
Driver Under Influence	**I**	**-1.935**	**0.948**	**-4.006**	**-0.253**
Driver Gender	**I**	**-1.270**	**0.372**	**-1.989**	**-0.553**
Intersection Related	F	4.597	2.839	-0.796	10.460
Road Grade	**F**	**5.015**	**1.697**	**2.036**	**8.696**
First Harmful Event on Road	F	-0.285	1.010	-2.302	1.514
Number of Vehicles in Crash					
One vehicle	**F**	**4.733**	**1.715**	**1.960**	**8.532**
Two vehicles**	F	0.000	0.000	0.000	0.000
Three or more	F	-5.124	7.376	-20.010	6.563
Maximum Vehicle Damage in Crash					
Slight damage**	F	0.000	0.000	0.000	0.000
Functional damage	F	4.273	2.532	-0.254	9.923
Disabled damage	**F**	**6.054**	**1.986**	**3.102**	**10.660**
Traffic Control	F	-0.937	0.881	-2.646	0.701
Vehicle Action					
Go straight**	F	0.000	0.000	0.000	0.000
Acceleration or deceleration	F	-1.363	2.001	-5.986	2.120
Turn	**F**	**-16.350**	**7.880**	**-32.240**	**-4.110**

Table 4-8 (Continued)

Variable		Specific to	Estimated Mean	Standard Deviation	95% BCI of Mean 2.50%	97.50%
Driver Seatbelt Use		F	0.918	1.999	-2.340	5.460
Driver Under Influence		F	0.672	1.431	-2.237	3.316
Driver Age						
Young		F	-1.836	1.483	-4.950	0.920
Mid-Aged**		F	0.000	0.000	0.000	0.000
Senior		F	-11.810	7.462	-25.620	0.535
Interactive Effects						
Vehicle/Driver Level	**Crash Level**					
Road Pavement	**Curve Road**	**I**	**-1.439**	**0.473**	**-2.367**	**-0.533**
Road Pavement	**One vehicle**	**I**	**-1.339**	**0.582**	**-2.395**	**-0.170**
Wet Road	**Disabled vehicle damage**	**I**	**-0.945**	**0.278**	**-1.495**	**-0.398**
Traffic Control	**Intersection Related**	**I**	**-0.529**	**0.244**	**-1.032**	**-0.058**
Traffic Control	**Dawn**	**I**	**0.736**	**0.379**	**0.002**	**1.491**
Traffic Control	**Curve Road**	**I**	**0.475**	**0.213**	**0.065**	**0.889**
Turn	**First Harmful Event on Road**	**I**	**3.183**	**0.805**	**1.649**	**4.853**
Turn	**Dawn**	**I**	**1.758**	**0.667**	**0.465**	**3.038**
Turn	**One vehicle**	**I**	**1.701**	**0.611**	**0.525**	**2.967**
Driver Seatbelt Use	**Road Grade**	**I**	**-3.074**	**1.221**	**-5.642**	**-0.690**
Drive Under Influence	**First harmful event on road**	**I**	**1.565**	**0.430**	**0.727**	**2.434**
Drive Under Influence	**One vehicle in crash**	**I**	**1.153**	**0.500**	**0.184**	**2.174**
Driver Gender	**Curve Road**	**I**	**0.625**	**0.281**	**0.110**	**1.195**
Traffic Control	**Dawn**	**F**	**1.961**	**0.805**	**0.481**	**3.608**
Turn	**First harmful event on road**	**F**	**4.436**	**1.969**	**0.753**	**8.604**
Driver Seatbelt use	**Intersection Related**	**F**	**-3.817**	**1.722**	**-7.478**	**-0.455**
Driver Seatbelt use	**Road Grade**	**F**	**-4.452**	**1.392**	**-7.367**	**-1.871**
Driver Seatbelt use	**One vehicle in crash**	**F**	**-2.850**	**1.077**	**-4.965**	**-0.850**
Young driver	**First harmful event on road**	**F**	**1.179**	**0.513**	**0.178**	**2.212**

Table 4-8 (Continued)

Variable		Specific to	Estimated Mean	Standard Deviation	95% BCI of Mean 2.50%	97.50%
Young driver	**Three or more vehicles**	**F**	**2.580**	**0.968**	**0.705**	**4.498**
Driver Under Influence	**First harmful event on road**	**F**	**1.597**	**0.617**	**0.392**	**2.809**
Driver Under Influence	**Road Grade**	**F**	**1.266**	**0.651**	**0.019**	**2.505**
Driver Under Influence	**One vehicle in crash**	**F**	**1.741**	**0.829**	**0.212**	**3.484**
	$\boldsymbol{\sigma_0^2}$		**3.850**	**0.761**	**2.340**	**5.458**

* Significant variables are marked in bold
** Reference category for the multi-categorical variable

As is shown in Table 4-8, the estimated σ_0^2 is 3.850. Therefore, the ICC for this study is calculated and is equal to 53.92%, indicating that between-crash variance accounts for 53.92% of the total unobserved variance and verifying the appropriateness of the proposed model structure.

A generalized random intercept model without cross-level interactions was also utilized to analyze the same dataset for model performance comparison, and the DIC values for these two compared models are illustrated in Table 4-9. It is shown by the close DIC values of these two models that, although with significant higher model complexity in model structure (pD =252.040), the proposed random intercept model produces comparable performance in model fit and in parameter estimation, indicating the appropriateness of including cross-level interactions in model development.

Table 4-9 DIC Result for Model Comparison.

Model Design	$\overline{D(\gamma)}$	$D(\bar{\gamma})$	pD	DIC
Hierarchical Random intercept model with cross-level interactions	5840.410	5574.360	252.040	6092.450
Hierarchical random intercept model without cross-level interactions	5959.770	5914.120	45.650	6005.420

4.3.3 Factor Impact Analysis

Average pseudo-elasticity analysis was conducted for the proposed model to quantify the influence of contributing factors on driver injury severity outcome, and the pseudo-elasticity results are shown in Table 4-10.

Table 4-10 Average Direct Pseudo-elasticity Analysis Result for Proposed Model.

Variable		Injury Severity		
		N	**I**	**F**
Curve Road		-43.85%	223.15%	-43.85%
Road Grade		-72.93%	1028.17%	6200.32%
Maximum Vehicle Damage in Crash				
Disabled Damage		-63.20%	2054.92%	13199.89%
Road Pavement		-28.21%	419.93%	-28.21%
Number of Vehicles in crash				
One Vehicle		-16.30%	-16.30%	4246.86%
Vehicle Action				
Acceleration or deceleration		76.00%	-70.02%	-64.11%
Turn		5967.46%	21.59%	-100.00%
Driver Under Influence		54.16%	-75.27%	204.30%
Driver Gender		73.75%	-52.17%	73.75%
Interactive Effects				
Road Pavement	Curve Road	54.86%	-62.19%	54.86%
Road Pavement	One vehicle	52.54%	-54.96%	52.54%
Wet Road	Disabled vehicle damage	34.22%	-48.09%	34.22%
Traffic Control	Intersection Related	17.44%	-30.88%	17.44%
Traffic Control	Dawn	-32.65%	41.17%	354.84%
Traffic Control	Curve Road	-14.72%	39.20%	-14.72%
Turn	First Harmful Event on Road	-74.80%	530.63%	1972.94%
Turn	Dawn	-47.91%	211.96%	-47.91%
Turn	One vehicle	-46.29%	199.93%	-46.29%

Table 4-10 (Continued)

Variable		Injury Severity		
		N	**I**	**F**
Driver Seatbelt Use	Road Grade	629.44%	-77.30%	-94.00%
Drive Under Influence	First harmful event on road	-47.90%	145.47%	165.92%
Drive Under Influence	One vehicle in crash	-40.42%	84.45%	249.80%
Driver Gender	Curve Road	-18.24%	55.05%	-18.24%
Driver Seatbelt use	Intersection Related	37.56%	37.56%	-97.51%
Driver Seatbelt use	One vehicle in crash	71.77%	71.77%	-86.85%
Young driver	First harmful event on road	-10.31%	-10.31%	194.82%
Young driver	Three or more vehicles	-25.95%	-25.95%	857.89%
Driver Under Influence	Road Grade	-11.52%	-11.52%	218.23%

Since driver incapacitating injury/fatality (**F**) is the chief concern in traffic safety analyses, this discussion would primarily focus on variables that are significantly increasing or decreasing driver incapacitating injury and fatality (**F**), and the influences of risk factors on complaint of injury and visible injury (**I**) would be discussed in an accompanying way or could be interpreted accordingly. Similar to the results in Section 4.2, some of these factors were also found significant in the previous two hierarchical models to predict driver injury severities.

Road grade is estimated to be significantly related to truck driver incapacitating injury and fatality in Table 4-8. The elasticity analysis regarding road grade illustrates that the presence of road grade would increase the average probabilities of injury levels **I** and **F** by approximately 1,000% and 6,200%, respectively. This is expected since truck drivers need to apply brakes more frequently to keep vehicle speeds stable during driving,

which increases the risk of brake failure and loss control of vehicle. Besides, trucks with high speeds are more likely to get longitudinal rollovers when sudden hard brakes are applied on graded roadways. Therefore, it is necessary for truck drivers to inspect and maintain brakes on a regular basis and drive more cautiously if there is a significant portion of trips on mountainous or hilly roads.

Maximum vehicle damage in a crash is also a critical predictor of truck driver injury severity outcome, as is revealed in Table 4-8 that disabled vehicle damage in a crash is significant in predicting truck driver incapacitating and fatal injuries (**F**). The pseudo-elasticities for disabled vehicle damage are 2,054.92% for complaint of injury and visible injury (**I**) and 13,199.89% for incapacitating and fatal injuries (**F**), indicating that disabled vehicle damage is closely associated with high probabilities of driver injury and fatalities. These results are understandable since the maximum vehicle damage in a crash could be treated as a visible and qualitative measurement of the impact energy generated from the crash, and disabled vehicle damage indicates massive crashing impact that passes onto drivers' bodies and causes severe injury outcomes. Additionally, it is also found in the pseudo-elasticity analysis that the interaction effect of wet road and disabled vehicle damage increase the probability of driver incapacitating injury and fatality by 34.22%. Several studies have proposed proper indices or methods to evaluate crash impact energy. Riviere et al. (2006) developed Energy Equivalent Speed (EES) to measure the impact energy a vehicle receives in a crash, and applied it to retrieving crash scene. Therefore, detailed examination of vehicle damage in crashes would be beneficial to reconstruct crash scene and facilitate crash investigation.

It is expected in the results that the number of vehicles in a crash plays an important role in predicting truck driver incapacitating injury and fatality. It is revealed in Table 4-8 that single-vehicle rural truck crashes are significantly associated with truck driver incapacitating and fatal injuries (**F**) (4.733, 95%BCI (1.960, 8.532)). The direct pseudo-elasticity for single-vehicle truck crash is 4,246.86%, indicating an extremely significant increase of fatality probability comparing with multi-vehicle rural truck-involved crashes. According to NHTSA (2006), the primary type of single-truck crashes are roadside departures, accounting for 61% of all single-truck crashes; while there are only 3% of multi-vehicle truck-involved crashes resulting from road departure. For single-vehicle truck crashes, the top critical events are vehicle loss of control and vehicle traveling, and these crashes are primarily caused by improper drivers' recognition, physical and decision factors, such as fatigue driving, driving under the influence, and driving while on the phone, etc. Preventive countermeasures, such as retroreflective signs, dynamic message signs (DMS) or rumble strips, should be recommended along rural low volume roadways to remind drivers to maintain vehicle operation and therefore enhance traffic safety.

Vehicle actions at crash occurrence are found to be significant in truck driver injury prediction. It is found that vehicle turning actions (left turn or right turn) is significantly associated the potential of driver injury and fatalities ((-3.826, 95% BCI (-6.164, -1.897)) specific to **I**, and (-16.350, 95% BCI (-32.240, -4.110)) specific to **F**), and the interactive effect of turning action and first harmful event on road is significant in predicting the risk of truck driver injuries (**I**) and fatalities (**F**). By examining the estimated average pseudo-elasticity, it is found that the interaction effect of turning

movement and first harmful event has pseudo-elasticities of 530.63% for complaint of injury and visible injury (**I**) and 1,972.94% for incapacitating and fatal injuries (**F**). Previous studies examined the impacts of left turn and right turn on crash risk and injury severity outcomes. Wang and Abdel-Aty (2008) applied partial proportional odds model to examine injury severity of left-turn crashes with different collision patterns. Zador et al. (1982) found that right-turn on red regulation increased right turn crashes by more than 20 percent, and their conclusion was supplemented (Frith, 1984) that a 0.7% decrease in incapacitating injury crash occurrence was discovered.

Driver age is also a significant factor predicting truck driver injury severities in this study. Although it is found in this study that none of the main effects for the three age groups is significant in predicting truck driver injury severities, two interaction effects associated with young drivers are found to be significantly contributing to driver incapable injuries and fatalities (**F**): young driver and first harmful event on road (1.179, 95% BCI (0.178, 2.212)), and young driver and three more vehicles in a crash (2.580, 95% BCI (0.705, 4.498)). Furthermore, the average pseudo-elasticities of these two interaction effects on driver incapable injuries and fatalities (**F**) are 194.82% and 857.89%, respectively, indicating their considerable effects on introducing more severe injuries and fatalities on truck drivers in rural crashes. These are reasonable since young drivers generally lack driving experience, proper recognition and decision skills, and they are more likely to conduct inattentive or risky driving behavior but less likely to take proper actions in road emergencies such as crash occurrence on roadways or multivehicle crashes, and therefore suffer higher injury severities. Hence, more effective traffic safety

countermeasures regarding young drivers, such as defensive driving course, should be recommended and enforced to enhance youth driving safety.

It is found in this study that driver seatbelt use is an effective way of protecting truck drivers from being injured or killed. The interactive impacts of driver seatbelt use with several other crash-level factors, including intersection-related crash ((-3.817, 95% BCI (-7.478,-0.455)) specific to **F**), road with grade ((-4.452, 95% BCI (-7.367, -1.871)) specific to **F**, and (-3.074, 95% BCI (-5.642, -0.690)) specific to **I**) and single-vehicle crash ((-2.850, 95% BCI(-4.965, -0.850)) specific to **F**) are found significant in truck driver injury severity prediction, although the main effects of driver seatbelt use ((0.432, 95% BCI (-1.006, 2.745)) specific to **I**, and (0.918, 95% BCI(-2.340, 5.460)) specific to **F**) are not found to be significant. The estimated pseudo-elasticities of these interactive effects verified the effects of seatbelt in reducing driver injury severities, especially on incapable injuries and fatalities (**F**). As is shown in Table 4-10, the pseudo-elasticities with respect to incapable injury and fatality (**F**) are -94.00% for the interaction between driver seatbelt use and road grade, -97.51% for the interaction between driver seatbelt use and intersection-related crash, and -86.85% for the interaction between driver seatbelt use and single-vehicle crash, all of which verify the protective effect of seatbelt use in driving. These findings also indicate that the protective effects of seatbelt use vary across different crash scenarios and should be examined associatively with other factors regarding road geometric design, environmental conditions and other crash, vehicle, or driver related characteristics. These findings are expected since the interactive effects of seatbelt use and other risk factors have been examined by peer studies from multiple aspects. Gross et al. (2007) discovered that alcohol consumption is closely associated

with insufficient seatbelt usage for Native Americans and contributes to significant trauma outcomes in traffic crashes. Chliaoutakis et al. (2000) examined seatbelt usage of young drivers and found that lengthy trips and driver discomfort tend to reduce the seatbelt usage rate, making drivers less protected.

Driver under the influence, either alcohol influence or drug influence, is a factor describing drivers' state of consciousness that is expected to be significantly associated with truck driver injury severity. It is found in Table 4-8 that its main effect is insignificant in predicting driver incapable injuries and fatalities (**F**) (0.672, 95% BCI(-2.237, 3.316)), but it is illustrated that the driver under influence interactively works with road with grade (1.266, 95% BCI(0.019, 2.505)), first harmful event on road (1.597, 95% BCI(0.392, 2.809)), and single-vehicle crash (1.741, 95% BCI(0.212, 3.484)) and contributes to severe driver injuries and deaths. The estimated average pseudo-elasticity for the variable "driver under influence" is 204.30% for injury severity **F**, and those for these three interactions with respect to **F** are 218.23%, 165.92% and 249.80%, respectively, verifying the considerable impacts of alcohol or drug influence on driver incapable injury and fatality outcome. Similar effects are also revealed regarding complaint injury/visible injury pattern (**I**) for its interactions with first harmful event on road (145.47%) and with single-vehicle crashes (84.45%), as indicated in Table 6. These results are expected since it takes more effort for truck drivers to maneuver properly due to the sizes and weights of trucks, especially on grade or bumpy roadways, and alcohol and drugs compromise drivers' capabilities for vehicle operations and judgment, leading to single-vehicle crashes such as overturn or run-off-road and multi-vehicle collisions on roadways, as well as severe body injuries on truck drivers. Hence, it is necessary for law

enforcement to perform driver with impairment (DWI) tests on roadways on a regular basis. The other pseudo-elasticity values could be interpreted similarly. These findings are helpful to understand the respective or joint impacts of heterogeneous attributes on truck driver injury patterns in rural truck-involved crashes.

4.3.4 Unobserved Heterogeneity Simulation Comparison

Random parameter logit (mixed logit) models are a major type of models to address unobserved heterogeneity issue in traffic safety research, and therefore have been utilized in the same dataset in this study for model performance comparison. The mixed logit model estimation results are shown in Table 4-11.

Table 4-11 Mixed Logit Model Estimation Results.

Variable	Specific to	Value	Stddev	t-test	p-value
Constant	F	-2.31	0.549	-4.21	0
Constant	I	-0.754	0.301	-2.5	0.01
	N	0	--fixed--		
Acceleration or Deceleration	F	0.948	0.448	2.12	0.03
Driver Seatbelt Use	F	-3.33	0.262	-12.72	0
Disabled Damage	F	2.97	0.373	7.96	0
Driver Gender	F	-0.4	0.182	-2.2	0.03
Driver Under Influence	F	1.04	0.218	4.75	0
Functional Damage	F	1.1	0.443	2.49	0.01
Go Straight	F	0.876	0.334	2.62	0.01
Two Vehicles	F	-0.462	0.159	-2.91	0
Wet Road	F	-0.561	0.186	-3.01	0
Driver Seatbelt Use	I	-1.21	0.282	-4.3	0
Disabled Damage	I	1.79	0.106	16.91	0
Driver Gender	I	-0.681	0.101	-6.74	0
Driver Under influence	I	0.759	0.152	5	0
Two vehicles	I	-0.428	0.0857	-4.99	0
Wet Road	I	-1.15	0.397	-2.91	0
Wet Road*	I	2.12	0.507	4.18	0

*Identified random parameter in the model.

Table 4-12 Mixed Logit Model Pseudo-elasticity Analysis Results.

Variable	Injury Severity		
	N	**I**	**F**
Two Vehicles	9.03%	-28.94%	-31.31%
Disabled Damage	-27.65%	333.32%	1310.19%
Wet Road	18.49%	-62.48%	-32.39%
Functional Damage	-6.92%	-6.92%	179.62%
Go Straight	-2.97%	-2.97%	133.00%
Acceleration or Deceleration	-5.80%	-5.80%	143.09%
Driver Seatbelt Use	133.69%	-30.31%	-91.64%
Driver under Influence	-17.43%	76.38%	133.61%
Driver Gender	15.37%	-41.61%	-22.66%

By comparing the pseudo-elasticity results from the mixed logit model (Table 4-12) and from the proposed random intercept model (Table 4-10), it is shown that minor differences exist between these results. A side-by-side comparison of the results from these two models was conducted and some agreements as well as discrepancies could be identified. For instance, it is found in Table 4-12 that the pseudo-elasticity of "Driver Seatbelt Use" is -91.64% for severity level **F**, and it is shown in Table 4-10 that the pseudo-elasticities with respect to incapable injury and fatality (**F**) are -94.00% for the interaction between "Driver Seatbelt Use" and "Road Grade", -97.51% for the interaction between "Driver Seatbelt Use" and "Intersection Related", and -86.85% for the interaction between "Driver Seatbelt Use" and "One Vehicle in Crash", even though the main effect of driver seatbelt use is found insignificant in Table 4-8. All of these verified the protective effect of seatbelt in reducing driver incapacitating injuries and fatalities. Besides, the elasticity for "Driver under Influence" in Table 6 is 133.61% specific to **F**, and it is shown in Table 4-10 that the elasticity of "Driver under Influence" is 204.30% specific to **F**, which is very close to the results in Table 4-12. Even though it is found that the elasticities of "Driver under Influence" specific to **I** are different (76.38% in Table 4-

12, and -75.27% in Table 4-10), but the elasticities of its interaction effects are consistent with the results in Table 4-10 for both severity levels **I** and **F**. These results illustrate the capabilities of the proposed random intercept model with cross-level interactions in examining driver injury severity patterns and variable marginal impacts. Also some discrepancies are revealed by comparing these results, regarding significant variable detection and pseudo-elasticity estimation. These differences are attributed from model structure design and specification, and both of these models have their own advantage in modeling crash injury outcomes and examining variable impacts. Therefore, the proposed model provides competent performance in parameter estimation comparing with mixed logit model and shed more light on understandings of these cross-level interaction effects on driver injury severity outcomes in rural interstate crashes.

4.4 Conclusions

Hierarchical regression models are proved to be effective in predicting traffic crash frequency and injury severity outcomes by capturing the hierarchical crash data structure and Bayesian inference method produces more accurate estimating results from parameter prior information and the studied dataset. To examine the applicability and effectiveness of the hierarchical Bayesian regression models in predicting driver injury severity in traffic crashes and the heterogeneous influence of contributing factors on these outcomes, three representative models are presented in this study: hierarchical binary logit model, hierarchical ordered logit model and hierarchical random intercept model with cross-level interaction effects. These models are developed and applied in the above order by overcoming its predecessor. On the other hand, traffic crashes result in

significant life and economic loss, and compared to urbanized areas, rural areas have a higher potential of inducing more severe driver injuries in traffic crashes in spite of a lower crash frequency. Therefore, three different rural crash datasets are examined respectively by the three hierarchical Bayesian models as case studies.

Hierarchical binary logit model is the simplest model configuration in these three hierarchical Bayesian models where the driver injury severity outcome is assumed to be a binary response: 0 indicating no injury or slight injury and 1 denoting incapable injury or death. A rural interstate crash dataset is utilized for this case study to investigate the impacts of crash-level and vehicle/driver-level variables on driver injury severity. Research results indicate that the proposed hierarchical Bayesian logit model outperforms the ordinary binary logit model in model fit and estimation effectiveness, based on the DIC criteria. Variables of crash and vehicle/driver levels are included in this research, and their effects on driver injury severities are reported in terms of odds ratio, with 95% BCI (or 90% BCI) indicating the statistical significance of the effects. Research results show that two crash-level variables (including the number of vehicles in a crash and wet road surface) and four vehicle/driver-level variables (including vehicle type, driver age, gender and alcohol/drug involvement) are significant in predicting driver injury severities.

Hierarchical ordered logit model overcomes the disadvantage of binary response configuration in the previous model by assuming driver injury severity with 5-level monotonic increasing values, and it is used to examine driver injury severity patterns in rural non-interstate crashes and variable impacts on driver injury severities. Similarly, the research results illustrate that the proposed model structure is superior in analyzing the selected dataset to an ordinary ordered logit model dropping off the between-crash

variance term, according to the DIC model performance measurement. 11 variables regarding crash, vehicle and driver characteristics were identified to be significant in driver injury severity prediction in rural non-interstate crashes. In this analysis, road segments far from intersections, wet road surface conditions and driver seatbelt use tend to reduce driver injury severity levels. Single-vehicle crashes, daylight driving conditions, severe vehicle damage in a crash and driver with alcohol or drug impairment increased the potential of higher driver injury severities and fatalities. In terms of vehicle type, motorcyclists are most vulnerable in traffic crashes, and heavy vehicle drivers receive best protection from their vehicles. It was also found that females and senior drivers are two driver groups that are prone to higher injury severities than their counterparts. Overall, this study provides reasonable results and deep insights for better understanding the internal mechanism of rural non-interstate crashes.

Hierarchical random intercept model with cross-level interaction effects is developed based on the two models and by overcoming the disadvantages of them in driver injury outcome configuration and model structure, where the driver injury severity is considered a multinomial variable with three exclusive level, and the interaction effects between crash-level variables and vehicle/driver level variables are systematically examined. The results demonstrate that the proposed model could effectively identify significant variables contributing to driver injury outcome and extract cross-level interactions among crash-level and vehicle-level attributes, and produces comparable performance with traditional random intercept model and the mixed logit model in model fit and analyses, even after penalized by the high complexity in model structure. A direct pseudo-elasticity analysis is conducted to evaluate the influence of the heterogeneous

contributing factors and their interaction effects on driver injury severity outcomes. Research results indicate that roadways with grades are a contributing factor to incapacitating injury and fatality of truck drivers; compared with two-vehicle truck-involved crashes, single-vehicle rural truck crashes are more likely to result in driver incapacitating injuries and deaths; maximum vehicle damage in truck-involved crashes is a significant factor positively related to truck driver injuries. Vehicle turning actions tend to reduce driver injury severities, but its interactive effects with other factors are inclined to produce severe injury outcome. The protective effect of driver seatbelt use is verified from its interactive effects with intersection-related crashes, roadways with grades, and single-vehicle truck crashes. Young truck drivers tend to be severely or fatally injured when the first harmful event was on road or they are involved in multi-vehicle truck-related crashes. The adverse effects of drivers using alcohol or drug also work interactively with crash-level features to induce serious injuries and fatalities.

These three models are proposed with increasing model configuration complexity by overcoming the disadvantage of the previous one and utilized in rural crash driver injury severity patterns in rural traffic crashes. Although each model has its distinctive model assumption and limitations, by utilizing Bayesian inference method, they all provide reliable analysis results in driver injury prediction, and constitute an important component in Bayesian method family. Some although addressing different types of crash datasets, some common contributing factors are found, including road surface condition, crash type (SV and MV), driver age, maximum vehicle damage in crash, seatbelt use, and driver drug or alcohol use. However, because these models all have certain regression assumptions with respect to data structure and parameter distribution,

non- regression Bayesian method by relaxing these rigid limitations are needed, whose applicability and usefulness are discussed in the following chapters.

Chapter 5 MNL-BN Hybrid Model Case Study

5.1 Case Study Dataset

We applied the BN into a rear-end crash dataset to predict driver injury severity in rear-end crashes. In total, 23,433 driver injury records from 11,383 rear-end crashes are used for model development and parameter estimation, where 2010 crash dataset (11,486 records) was used as training dataset, and 2011 crash dataset (11,947 records) was used as testing dataset. Table 5-1 shows the definitions of variables in this dataset in this research.

Table 5-1 Rear-end Crash Dataset Descriptions and Statistics.

Attribute		Value		SEV						Total
				NO INJURY	Percentage	INJURY	Percentage	FATALITY	Percentage	
DAY	Day	MON	Monday	2286	64.09%	1275	35.74%	6	0.17%	3567
		TUE	Tuesday	2516	61.76%	1550	38.05%	8	0.20%	4074
		WED	Wednesday	2610	63.88%	1470	35.98%	6	0.15%	4086
		THU	Thursday	2525	62.45%	1512	37.40%	6	0.15%	4043
		FRI	Friday	2712	61.04%	1710	38.49%	21	0.47%	4443
		SAT	Saturday	1259	61.96%	760	37.40%	13	0.64%	2032
		SUN	Sunday	711	59.85%	471	39.65%	6	0.51%	1188
RDREL	First Harmful Event Location	ONWAY	on roadway	14567	62.38%	8719	37.34%	66	0.28%	23352
		OFFWAY	off roadway	52	64.20%	29	35.80%	0	0.00%	81
LIGHT	Lighting Condition	DAYLIGHT	daylight	12600	62.84%	7420	37.01%	31	0.15%	20051
		DARK	dark	1547	58.58%	1061	40.17%	33	1.25%	2641
		DAWN/DUSK	dawn or dusk	472	63.70%	267	36.03%	2	0.27%	741
CURVE	Curvature	CURVE	curve road	616	67.62%	295	32.38%	0	0.00%	911
		STAIGHT	straight road	14003	62.17%	8453	37.53%	66	0.29%	22522
		LEVEL	level	12755	62.84%	7584	37.36%	59	0.29%	20298
RDGRD	Road Grade	HCRST	hillcrest	365	57.21%	273	42.79%	0	0.00%	638
		ONGRADE	On grade	1434	59.50%	969	40.21%	7	0.29%	2410
		DIP	dip	45	73.77%	16	26.23%	0	0.00%	61
		OTHER	other road grade	20	76.92%	6	23.08%	0	0.00%	26
DRESD	Driver Residency	ST	NM residency	12679	63.01%	7423	36.89%	21	0.10%	20123
		NST	other state residency	1940	58.61%	1325	40.03%	45	1.36%	3310
NVEH	Number of Vehicles Involved	TWO	two vehicles	11872	67.51%	5671	32.25%	43	0.24%	17586
		THREE	three vehicles	2192	49.58%	2215	50.10%	14	0.32%	4421
		MORE	more than three vehicles	555	38.92%	862	60.45%	9	0.63%	1426
RDFUNC	Road Function	URBN	urban road	13306	63.21%	7719	36.67%	26	0.12%	21051
		RINT	rural interstate	460	52.27%	404	45.91%	16	1.82%	880
		RNINT	rural non-interstate	853	56.79%	625	41.61%	24	1.60%	1502
PEDINV	Pedestrian Involvement	Y	involved	5	38.46%	8	61.54%	0	0.00%	13
		N	not involved	14614	62.40%	8740	37.32%	66	0.28%	23420
MCINV	Motorcycle Involvement	Y	involved	116	29.74%	265	67.95%	9	2.31%	390
		N	not involved	14503	62.94%	8483	36.81%	57	0.25%	23043
HEVINV	Heavy Vehicle Involvement(including bus, pickup, semi-truck and lorries)	Y	involved	388	52.29%	311	41.91%	43	5.80%	742
		N	not involved	14231	62.72%	8437	37.18%	23	0.10%	22691
HZINV	Hazard Material Involvement	Y	involved	6	46.15%	5	38.46%	2	15.38%	13
		N	not involved	14613	62.40%	8743	37.33%	64	0.27%	23420
DTINC	Distance from Crash Location to Intersection	NEAR	<0.1mile	4624	58.87%	3191	40.62%	40	0.51%	7855
		MID	0.1-1.0 mile	596	52.37%	536	47.10%	6	0.53%	1138
		FAR	>1.0 mile	9399	65.09%	5021	34.77%	20	0.14%	14440

Table 5-1 (Continued)

Attribute		Value		SEV						Total
				NO INJURY	Percentage	INJURY	Percentage	FATALITY	Percentage	
DLRST	Driver License Restriction	RST	with restriction	3625	62.52%	2153	37.13%	20	0.34%	5798
		NORST	no restriction	10994	62.34%	6595	37.40%	46	0.26%	17635
RDPV	Road Paving Condition	PAVED	paved surface	14552	62.34%	8724	37.37%	66	0.28%	23342
		UNPAVED	unpaved surface	67	73.63%	24	26.37%	0	0.00%	91
TRFCTL	Traffic Control	NCTL	no control	11132	61.59%	6899	38.17%	42	0.23%	18073
		SYSIGN	Stop/yield sign control	124	73.37%	45	26.63%	0	0.00%	169
		SGCTL	signal control	507	64.18%	279	35.32%	4	0.51%	790
		RRGATE	railroad gate	6	85.71%	1	14.29%	0	0.00%	7
		OTHER	other control measures, such as passing zones, detours,etc.	2850	64.86%	1524	34.68%	20	0.46%	4394
NLANE	Number of Lanes with Same Direction at Crash Location	ONE	one lane	3363	64.40%	1846	35.35%	13	0.25%	5222
		TWO	two lanes	6024	62.65%	3550	36.92%	41	0.43%	9615
		MORE	more than two lanes	5232	60.87%	3352	38.99%	12	0.14%	8596
VACT	Vehicle Action	STRT	straight	9176	62.69%	5416	37.00%	46	0.31%	14638
		BACK	backup	35	89.74%	4	10.26%	0	0.00%	39
		SLOW	slow	1411	62.24%	854	37.67%	2	0.09%	2267
		LTURN	left turn	484	65.58%	250	33.88%	4	0.54%	738
		RTURN	right turn	432	70.70%	176	28.81%	3	0.49%	611
		UTURN	U-turn	20	66.67%	10	33.33%	0	0.00%	30
		OTK	overtaking	130	64.68%	71	35.32%	0	0.00%	201
		OTHER	other action	2931	59.71%	1967	40.07%	11	0.22%	4909
VTYPE	Vehicle Type	LVEH	light vehicle, including passenger car or van	10747	62.35%	6463	37.49%	27	0.16%	17237
		HVEH	heavy vehicle, including bus, pickup, semi-truck and lorries	3454	63.53%	1950	35.87%	33	0.61%	5437
		MC	motorcycle	59	29.65%	136	68.34%	4	2.01%	199
		OTHER	other	359	64.11%	199	35.54%	2	0.36%	560
DBELT	Driver Seatbelt Use	Y	seatbelt used	13698	62.03%	8332	37.73%	52	0.24%	22082
		N	seatbelt not used	921	68.17%	416	30.79%	14	1.04%	1351
DAGE	Driver Age	YOUNG	16~25	4744	65.28%	2510	34.54%	13	0.18%	7267
		MID	25~64	8814	60.91%	5608	38.75%	49	0.34%	14471
		OLD	64 or older	1061	62.60%	630	37.17%	4	0.24%	1695
DALC	Driver Alcohol Involvement	Y	involved	115	44.40%	139	53.67%	5	1.93%	259
		N	not involved	14504	62.59%	8609	37.15%	61	0.26%	23174
DSEX	Driver Sex	M	male	7967	63.89%	4454	35.72%	49	0.39%	12470
		F	female	6652	60.68%	4294	39.17%	17	0.16%	10963
MAXDAM	Most Serious Vehicle Damage	NSLT	no damage or slight damage	6147	72.65%	2312	27.33%	2	0.02%	8461
		FUNC	functional damage that affects operations of vehicle	4284	68.61%	1960	31.39%	0	0.00%	6244
		DSABL	disabled damage that vehicles can't be driven	4188	47.98%	4476	51.28%	64	0.73%	8728

5.2 BN Input Variable Selection

In this analysis, an ordinary MNL model was used for the BN input variable selection procedure. Individual-specific model specifications are established so that each variable has different marginal costs for different driver injury severity levels. Three different coefficients β_{ki}(i=1, 2, 3) specified for NO INJURY, INJURY and FATALITY for the *k*th variable. 18 variables are selected as inputs for BN structure learning, model specification development, and conditional probability inference: *DALC(driver alcohol involvement), DBELT(driver seatbelt use), DSABL(vehicle disabled damage), LIGHT(lighting condition), MCINV(motorcycle involvement), NST(non-local driver), NVEH(number of vehicles in a crash), TKINV(truck involvement), DINTC(distance of crash location to nearest intersection), FUNC(functional vehicle damage), HCRST(hillcrest terrain), NCTL(no traffic control), SGCTL(signal control), WIND(windy weather), CURVE(road curve), DSEX(driver gender), EVE(evening time),* and *URBN (urban roads).* The detailed estimation results from the MNL model are shown below in Table 5-2.

Table 5-2 MNL Model Estimation Results and Significant Variable Identification.

Variable	Coef.[a]	Std. Err.[b]	T-Ratio	P-Value
Constant (Specific to INJURY)	5.89	0.93	6.33	0.00
Constant (Specific to NO INJURY)	7.53	0.93	8.09	0.00
DALC (Specific to FATALITY)	1.36	0.53	2.56	0.01
DALC (Specific to INJURY)	0.63	0.13	4.75	0.00
DBELT (Specific to FATALITY)	-0.84	0.39	-2.16	0.03
DSABL (Specific to FATALITY)	3.70	0.73	5.07	0.00
DSABL (Specific to INJURY)	0.89	0.03	26.22	0.00
LIGHT (Specific to FATALITY)	-0.65	0.14	-4.75	0.00
MCINV (Specific to FATALITY)	3.02	0.51	5.94	0.00
MCINV (Specific to INJURY)	1.43	0.12	12.33	0.00
NST (Specific to FATALITY)	0.95	0.36	2.64	0.01
NVEH (Specific to FATALITY)	0.26	0.12	2.12	0.03
NVEH (Specific to INJURY)	0.44	0.02	19.67	0.00
TKINV (Specific to FATALITY)	3.19	0.37	8.61	0.00
DINTC (Specific to INJURY)	-0.11	0.02	-6.94	0.00
FUNC (Specific to INJURY)	0.18	0.04	4.83	0.00
HCRST (Specific to INJURY)	0.21	0.09	2.45	0.01
NCTL (Specific to INJURY)	0.14	0.04	3.67	0.00
SGCTL (Specific to INJURY)	-1.53	0.53	-2.91	0.00
SGCTL (Specific to NO INJURY)	-1.61	0.53	-3.07	0.00
WIND (Specific to INJURY)	0.33	0.15	2.24	0.03
CURVE (Specific to NO INJURY)	0.19	0.08	2.49	0.01
DSEX (Specific to NO INJURY)	0.22	0.03	7.61	0.00
EVE (Specific to NO INJURY)	0.11	0.03	3.4	0.00
URBN (Specific to NO INJURY)	0.20	0.05	4.22	0.00
Log-likelihood for estimation		-14736.40		
Likelihood ratio test		22014.77		
Likelihood ratio index, ρ^2		0.43		

[a]Estimated Coefficient, [b]Standard Error.

5.3 BN Model Performance on Rear-end Traffic Crash Driver Injury Severity Prediction

This research employed Weka (Waikato Environment for Knowledge Analysis) software (Bouckaert, 2008), developed by University of Waikato, New Zealand, to

establish the BN structure and estimate the parameters. The whole dataset was divided into two approximately equal-sized subsets, classified by the years of crash occurrences. The 2010 crash dataset was used for BN structure learning and the 2011 dataset was utilized for model validation and performance test. Three major BN scoring metrics, AIC, MDL and BDe, were used for classifier training, and it is found that the MDL score controlled training procedure produced BN structures with least variance and achieves the best classification performance. Therefore, the MDL score criterion is employed in this study for BN structure learning. Based on prior knowledge on the 18 significant variables identified by the MNL model, the initial DAGs are developed. In order to avoid locally optimal solution, different initial DAGs are used to ensure at least a globally suboptimal network structure is generated.

Table 5-3 MNL-BN Estimation Results.

	Training		Testing	
Correctly Classified Instances	Number	Percentage	Number	Percentage
	7677	66.84%	7856	65.76%
Incorrectly Classified Instances	Number	Percentage	Number	Percentage
	3809	33.16%	4091	34.24%
Total Number of Instances	11486		11947	

Table 5-4 MNL-BN Classification Performance by Driver Injury Severities.

Driver Injury Severity	TP Rate		FP Rate		Precision		F-Measure		ROC Area: AUC	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing
NO INJURY	0.856	0.852	0.634	0.661	0.689	0.683	0.764	0.759	0.679	0.659
INJURY	0.360	0.332	0.144	0.148	0.600	0.569	0.450	0.419	0.676	0.654
FATALITY	0.333	0.273	0.002	0.002	0.355	0.281	0.344	0.277	0.987	0.956
Weighted Average	0.668	0.658	0.448	0.469	0.655	0.640	0.645	0.631	0.679	0.658

As can be seen in Table 5-3, the overall estimation accuracies of this trained BN are 66.84% and 65.76% for training and testing datasets, respectively. Compared to the

model accuracies ranging from 60% to 65% for testing and training datasets in Abdelwahab and Abdel-Aty (2001)'s study and from 61% to 62% in de Oña et al (2011) 's study, the results obtained are reasonably acceptable. The variance between the estimation accuracies for training and testing datasets is around 1%, indicating that the trained network is transferable and able to explain and model the testing data fairly well.

The true positive (TP) rates range from 0.273 (FATALITY) to 0.852 (NO INJURY) with a weighted average of 0.658 for the testing dataset. The result indicates that the BN is capable of classifying 85.2% of no injuries correctly, but its ability to classify fatalities is relatively poor. This implies that the BN is able to better classify no injuries than injuries and fatalities as expected since the majority of the crash data records are no injuries.

F-measure ranges from 0 to 1 and can be used as an effective performance measure for the built BN. F-Measure=0 means extremely poor model classification results and F-Measure=1 represents perfect model classification performance. The trained BN performs best of the instances of no injuries using the test dataset and its F-Measure is equal to 0.759. Overall, for the entire test dataset, the average F-Measure is 0.631, indicating an acceptable model predication performance.

ROC curve is another important indicator to evaluate the overall performance of the BN model. Figure 5-1 shows an example of a typical ROC curve. An ROC curve above the diagonal line indicates a model performance better than random guess. The ROC curves are demonstrated for three driver injury outcomes in Figures 5-2, 5-3 and 5-4. We can see that all three ROC curves locate above the diagonal lines, indicating that

the trained BN performs reasonably well for three injury severity classifications. The AUC value is a quantitative index that assesses the overall performance of model classification estimation with a maximum value of 1.00, which indicates a perfect classification prediction. A value of 0.5 indicates poor model prediction performance so that random classification is produced by the model. Figures 5-2, 5-3, 5-4, and Table 5-4 show the AUCs produced by the trained BN for the test dataset. The highest AUC is 0.956 showing the best performance achieved by the BN for fatal injury outcome classification. The AUCs are 0.659 and 0.654 for classifying severity outcomes of no injuries and injuries, respectively. For the entire dataset, the overall AUC can be calculated as a weighted average for each injury outcome classification as follows (Provost and Domingos, 2001),

$$AUC_{overall} = \sum_{i=1}^{n} AUC_{c_i}\, p(c_i) \tag{5-1}$$

where AUC_{c_i} is the AUC for injury severity class c_i, $p(c_i)$ is the probability of occurrences for injury severity class c_i, and n is the number of classes, which is equal to 3 in this study. For the testing dataset, $AUC_{overall}$=0.658, indicating that the trained BN is able to effectively discover the classification patterns and the performance is acceptable based on Tape (2001)'s criteria.

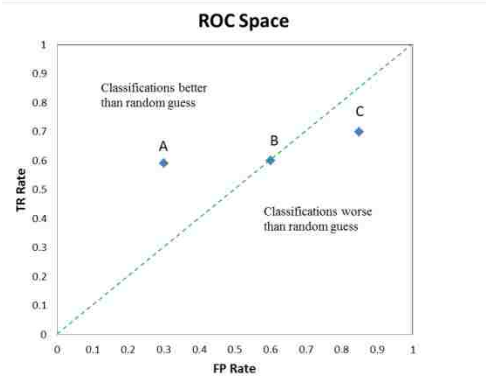

Figure 5-1 An Example of ROC Space.

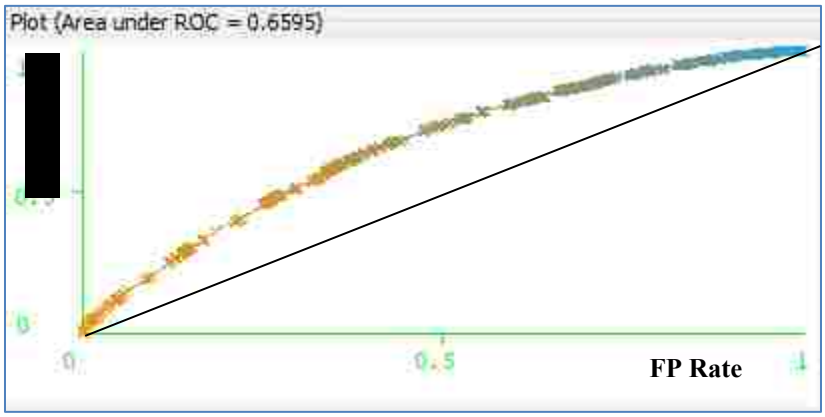

Figure 5-2 ROC Curve for the Category of NO INJURY.

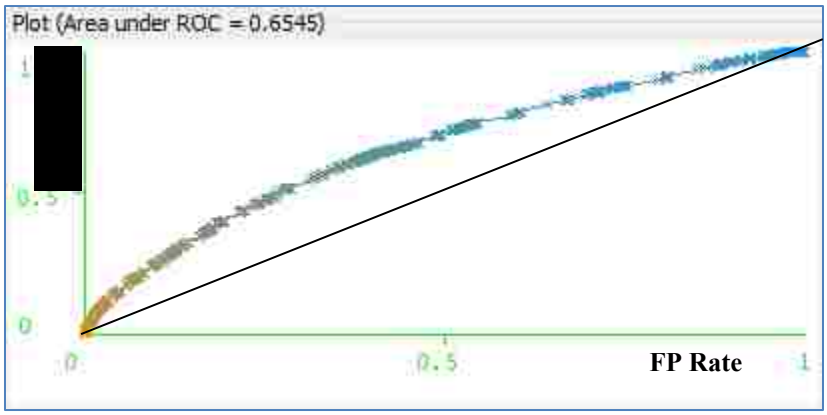

Figure 5-3 ROC Curve for the Category of INJURY.

Figure 5-4 ROC Curve for the Category of FATALITY.

Table 5-5 illustrates the classification confusion matrix for the testing dataset, where, each row represents the actual number of observed instances for each injury

severity category and each column denotes the number of predicted instances for each injury severity category. The diagonal cells indicate the correct predictions and non-diagonal cells are erroneously predicted instances. As can been seen, the BN tends to overestimate the number of instances of no injury and underestimates the number of instances of injury. The overall match rate for the test dataset is 65.8%.

Table 5-5 BN Classification Confusion Matrix for the Test Dataset.

		Predicted Instances Classified by Severity		
		NO INJURY (9327)	INJURY (2588)	FATALITY (32)
Observed Instances Classified by Severity	NO INJURY (7477)	6374	1093	10
	INJURY (4437)	2951	1473	13
	FATALITY (33)	2	22	9

5.4 BN Model Structure and Most Probable Explanation (MPE) Analysis

The BN is trained and its final network structure is illustrated in Figure 5-5, with a conditional probability table for each node, which is used for MPE calculation and evidence inference.

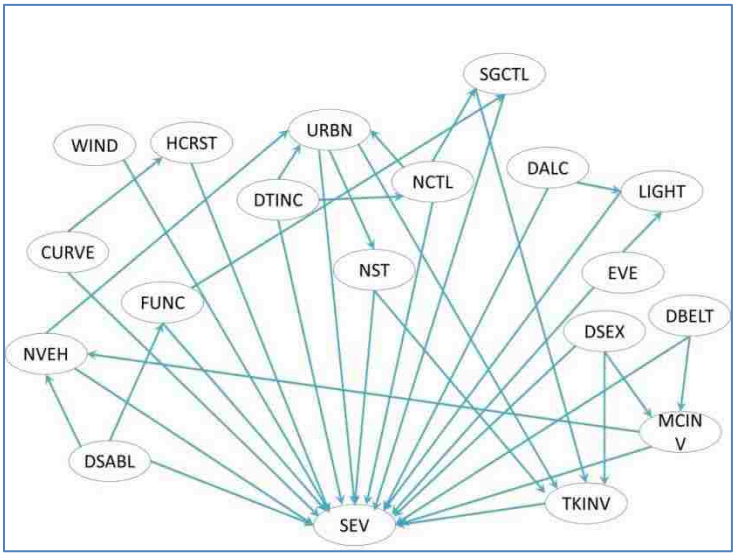

Figure 5-5 BN Classifier Structure with MDL Score.

Table 5-6 MPE Configuration for Training and Testing Datasets.

Variable	MPE Value	Percentage	
		Training	Testing
SEV	1	62.18%	62.58%
LIGHT	3	85.60%	85.54%
WIND	0	98.91%	99.34%
CURVE	0	95.73%	96.48%
HCRST	0	97.27%	97.28%
NST	0	86.31%	85.45%
NVEH	1	74.01%	76.04%
URBN	1	90.62%	89.08%
MCINV	0	98.36%	98.31%
TKINV	0	96.68%	96.98%
EVE	0	77.58%	76.13%
DTINC	3	61.27%	61.96%
NCTL	1	76.73%	77.51%
SGCTL	0	94.21%	94.89%
DBELT	1	94.35%	94.12%
DALC	0	98.89%	98.90%
DSEX	1	53.12%	53.31%
DSABL	0	62.19%	63.30%
FUNC	0	77.02%	69.83%

MPE analysis is an effective way to examine the graphical performance of the trained BN structure (de Oña et al., 2011). MPE can be calculated based on the most probable configuration of values for all the variables given the dataset. By comparing MPE to the relative frequency computed based on the dataset, the statistical quality of the trained BN structure could be quantitatively measured (Simoncic, 2004). In this study, the most probable values and the corresponding frequencies of the variables are illustrated in Table 5-6 for both training and testing datasets. Given the trained BN structure and conditional probabilities for each node, MPE for the testing dataset can be calculated using the MPE formula with the most possible value for each variable illustrated in Table 5-7. The relative frequency for the testing dataset, P(Test), is computed also for the comparison purposes. Obviously, the MPE is a small probability, approximately 0.02829 for the testing dataset, although it represents the most likely

explanations. The difference between the MPE and the relative frequency is about 0.29% for the testing dataset. Such small difference further verifies that the BN structure performs reasonably well.

Table 5-7 MPE Results for Training and Testing Datasets.

MPE formula	MPE_{Test}	P(Test)
P(WIND=0)P(CURVE=0)P(HCRST=0|CURVE=0) P(NVEH=1|DSABL=0,MCINV=0)P(DSABL=0) P(FUNC=0|DSABL=0)P(DTINC=3) P(URBN=1|DTINC=3,NCTL=1,NVEH=1) P(NST=0|URBN=1)P(NCTL=1) P(SGCTL=0|FUNC=0,NCTL=1) P(DALC=0)P(EVE=0)P(LIGHT=3|DALC=0,EVE=0) P(DSEX=1)P(TKINV=0|NST=0,URBN=1, SGCTL=0, DSEX=1)P(MCINV=0|DSEX=1,DBELT=1)P(DBELT=1) P(SEV=1|WIND=0,CURVE=0,HCRST=0,NVEH=1,DSABL=0,FUNC=0, DTINC=3,URBN=1,NST=0,SGCTL=0,DALC=0,EVE=0,LIGHT=3,DSEX=1, TKINV=0,MCINV=0,NCTL=1,DBELT=1)	0.028290	0.028207

5.5 Influence of Contributing Factors on Driver Injury Severity

Two aspects should be included in the results for non-regression influence estimation: the learned optimal BN structure, and the influence of variables on crash driver injury severities in terms of probability change. In the learned BN structure, the nodes indicate the included variables, and the arcs represent the statistical dependence among these variables. The BN structure explicitly formulates the interdependency among the variables and is capable of providing probability inference analyses based on the conditional probability tables for each node. Through setting evidences for the related variables, their contributions to crash occurrences with certain severity outcomes can be quantified.

Table 5-8 BN Probability Inference Results for the Variables Increasing Driver Injury Severities.

Variable		Severity		
		NO INJURY	INJURY	FATALITY
Proportion Distribution		0.626	0.371	0.003
WIND		0.491	0.503	0.006
DALC		0.425	0.535	0.031
DSABL		0.490	0.502	0.007
TKINV		0.490	0.449	0.061
NST		0.575	0.413	0.012
DTINC	NEAR	0.410	0.586	0.004
	MID	0.534	0.453	0.014
	FAR	0.654	0.345	0.001
LIGHT	DAYLIGHT	0.629	0.370	0.001
	DAWN/DUSK	0.650	0.344	0.007
	DARK	0.595	0.390	0.014
NVEH	2	0.671	0.327	0.002
	3	0.510	0.486	0.004
	≥ 4	0.601	0.391	0.008

Table 5-8 illustrates the inference results for the variables which significantly increase the likelihoods of driver suffering injury and fatality given rear-end crash occurrences. For each variable, the probability of a predetermined value is set as 1.0 in the first column during evidence setting processes, and the probabilities for driver injury severity outcomes are inferred in other columns showing the impact of these variables with specific values on the likelihoods of various driver injury severities. In the first row, Proportion Distribution, indicates the corresponding proportion of each driver injury severity extracted directly in the testing dataset. In the second row, a probability of 1.0 is assigned to the variable, *WIND*, with the value, 1, (e.g. *WIND*=TURE) as evidence, and the probabilities of INJURY and FATALITY increase from 0.371 to 0.503 and from

0.003 to 0.006, respectively, comparing to the original distributions. This implies that windy weather conditions will increase the propensity of driver injury and fatality in rear-end crashes. If a vehicle is involved in a rear-end crash under windy conditions, the likelihoods for drivers suffering injury and fatality increase from 37.1% to 50.3% and from 0.3% to 0.6%, respectively. Alcohol influence also significantly affects the probabilities of driver injury and fatality in rear-end crashes, supported by the inference results of the variable, *DALC*. When alcohol influence is set as evidence, the probabilities of INJURY and FATALITY increase from 37.1% to 53.5% and from 0.3% to 3.1%, respectively. This implies that drivers under the influence of alcohol are more likely to be seriously and fatally injured in rear-end crashes, which is consistent with the conclusions in the previous studies (Hels et al., 2013; Weiss et al., 2014).

As can be expected, the inference results of the variable, *DSABL*, indicate that drivers are more likely to suffer serious injury and fatality when vehicles involving in rear-end crashes are disabled. The corresponding probabilities increase up to 53.5% and 0.7%. It is understandable that severe vehicle damage is normally associated with high probabilities of driver injury and fatality since vehicle damage level is a reflection of the impact produced in crashes. Truck involvement (*TKINV*) is a significant factor substantially contributing to serious driver injuries and fatalities in rear-end crashes. As shown in Table 10, when truck is involved in rear-end crashes, the probability of driver fatality increases by 20 times from 0.3% to 6.1% comparing to its probability under regular conditions. The likelihood of driver injury also increases from 37.1% to 44.9%. These results underscore that the significant impacts of trucks on driver injury and fatality in rear-end crashes, which is consistent with the previous study (Chang and

Mannering, 1999) that large trucks have the significant impact on the most severely injured occupants. Large trucks account for 8% of all vehicles involved in fatal crashes although they are only 4% of total registered vehicles in the U. S. in 2010 (NHTSA, 2012). These findings emphasize that special research efforts should be undertaken to address truck involvements in severe rear-end crashes, such as investigations on effective countermeasures to improve truck drivers' visibility.

Drivers coming outside of the state are more likely to be seriously and fatally injured in rear-end crashes. The probabilities of nonlocal drivers being injured and killed increase up to 41.3% and 1.2%, respectively, due to their unfamiliarity to local roadway networks. In addition, a dependent relationship is also observed in Figure 5-5 between the variables, *NST* and *TKINV*. It could be explained by the fact that many nonlocal drivers are from trucking industry to transporting a large amount of goods and materials through the states. Normally, they are less familiar with local roadway network, environment characteristics, traffic regulations, and driver behavior. They are more likely to involve in severe rear-end crashes.

The variable, *DTINC*, denotes the distance between the crash location to the nearest intersection with three values: NEAR (less than 0.1 mile), MID (between 0.1 and 1.0 mile), and FAR (more than 1.0 mile). The reference results indicate when this distance increases, the likelihood of drivers being injured in read-end crashes decreases. When the distance between the crash location to the nearest intersection is less than 0.1 mile, the probability of drivers being injured is 58.6% given rear-end crash occurrences. As the distance increases up to 1.0 mile, the probability of driver fatality increases by more than three times from 0.4% to 1.4%. These findings imply that driver injury

severities increase in rear-end crashes around intersections due to the complex conflicting movements. However, fatal rear-end crashes are more likely to occur when vehicles are approaching intersections at relatively high speeds. Inappropriate acceleration, insufficient deceleration, less driver reaction and perception time, etc. may dramatically contribute to severe crash occurrences. These unique injury distribution patterns should be considered when developing the countermeasures to mitigate intersection-related crash severities. This variable was also found significant in hierarchical regression modeling in Chapter 4. Lighting condition (*LIGHT*) is significantly contributing to driver injury outcomes in rear-end crashes. The probability inference results indicate that the probabilities of FATALITY consistently increase when lighting conditions become inferior from DAYLIGHT to DARK. However, under unfavorable lighting conditions at dusk and dawn, the likelihood of drivers being injured decreases slightly from 37.0% to 34.4%. This could be attributed to the facts that drivers become more cautious to prepare for unfavorable lighting conditions at dusk and dawn, so the probability of driver injury may decreases relative to daylight conditions. Similar analyses can be conducted for the variable, *NVEH,* representing the number of vehicles involved in rear-end crashes. The probability of drivers suffering fatality consistently increases when the number of vehicles involved increases. Interestingly, drivers are most likely to be injured when three vehicles are involved in a rear-end crash. When four or more vehicles are involved a crash, the probability of driver injury decreases. These findings are helpful to understand the attributes of multi-vehicle involved rear-end crashes.

5.6 Model Performance Comparison with Linear Statistical Models

In this case study, performance comparison is conducted between the proposed approach and the MNL model in this study. Table 5-9 shows the MNL classification confusion matrix for the testing dataset. As illustrated in Table 5-9, the total number of correctly predicted observations by the MNL model is 6664 (6664= 4881+1782+1), and the overall correction rate of this MNL prediction is 55.78%, which is considerably lower than the classification accuracy from the proposed logit-based BN hybrid approach (65.76%). It reveals that the proposed approach is more effective and accurate in predicting driver injury severities in rear-end crashes.

Table 5-9 MNL Classification Confusion Matrix for the Testing Dataset.

		Predicted Instances Classified by Severity		
		NO INJURY (7533)	INJURY (4377)	FATALITY (37)
Observed Instances Classified by Severity	NO INJURY (7477)	4881	2580	15
	INJURY (4437)	2635	1782	20
	FATALITY (33)	16	16	1

5.7 Conclusions

Rear-end crash is one of the major traffic accidents and has been investigated in the past decades. A good understanding of significant attributes affecting driver injury severities and their contributions in rear-end crashes is of practical importance to develop cost-effective countermeasures against serious driver injury and fatality in rear-end crashes. This case study applies the proposed MNL-BN hybrid model to examine rear-end crash dataset and investigate the impacts of significant contributing attributes on driver injury severity outcomes in rear-end crashes.

In model development procedure, the MNL model is developed to identify significant variables, and the BN is employed to explicitly formulate statistical associations between driver injury severity outcomes and explanatory attributes, including driver behavior, demographic features, vehicle factors, geometric and environmental characteristics, etc. The BN structure is trained based on prior domain knowledge and performance scoring metric using state-wide crash data collected in New Mexico from 2010 to 2011. Various statistical model performance measures, such as F-Measure, ROC curve, AUC, and MPE, are used to quantify the BN model performance. The results demonstrate that the trained BN model can effectively discover the interdependency among variables and the proposed hybrid approach performs reasonably well.

The inference analyses are conducted to quantify the contributions of the most significant variables to driver injuries and fatalities in rear-end crashes. The factors including truck involvement, inferior lighting conditions, windy weather conditions, the number of vehicles involved, etc. can significantly increase driver injury severities in rear-end crashes. For example, when truck is involved in rear-end crashes, the probability of driver fatality increases by 20 times from 0.3% to 6.1% comparing to its probability under regular conditions. The likelihood of driver injury also increases from 37.1% to 44.9%. These results underscore the considerable impacts of these significant variables on driver injury and fatality in rear-end crashes. The proposed methodology and research findings provide insights for developing effective countermeasures to reduce rear-end crash injury severities and improve traffic system safety performance.

While studies in previous chapters introduced the applicability of hierarchical regression models with Bayesian inference in traffic crash injury severity analysis, this study introduced Bayesian network model, an indispensable component of Bayesian family methods, and assessed its effectiveness in predicting driver injury severity outcomes and evaluating variable impacts on injury outcomes, which enhances our understanding regarding Bayesian model applications in traffic crash injury severity analysis.

Chapter 6 DTNB Classifier Case Study

6.1 Case Study Dataset

In this case study with the DTNB classifier, the same rear-end dataset described in Section 4.2 was used for driver injury severity analysis, and therefore the variable description table is omitted in this section. Correlation analyses were also conducted to avoid significant correlations among the explanatory variables, according to the "naïve" assumption of inter-independence for NB models. For variables with significant correlations, variables most related to driver injury severities were kept for model estimation and less significant ones were removed, based on traffic engineering experience. Overall, the studied dataset includes 23,433 driver/vehicle records from 11,383 rear-end crashes on New Mexico roadways.

By applying the DTNB model as a classifier into traffic injury severity analyses, it is expected to obtain analyses results composed of four aspects: 1) model performance analysis; 2) A set of most contributable variables to driver injury severity from the trained classifier; 3) variable influence and decision rule analysis, 4) model performance comparison with statistical models.

6.2 DTNB Model Performance Analysis

This dataset was also modeled with a DTNB classifier in WEKA software. Same as well, the 2010 dataset was used for DTNB model training and decision rule learning, and the 2011 dataset was used as the test dataset for model validation and performance assessment. The detailed model performance measurements are shown in Tables.

Table 6-1 DTNB Model Classification Accuracy.

	Training		Test	
Correctly Classified Instances	Number	Percentage	Number	Percentage
	8506	74.06%	7494	62.73%
Incorrectly Classified Instances	Number	Percentage	Number	Percentage
	2980	25.94%	4453	37.27%
Total Number of Instances	11486		11947	

Table 6-2 DTNB Classification Performance by Driver Injury Severity.

Driver Injury Severity	TP Rate		FP Rate		Precision		F-Measure		ROC Area: AUC	
	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test
NO INJURY	0.825	0.788	0.325	0.621	0.807	0.68	0.816	0.73	0.804	0.631
INJURY	0.6	0.36	0.13	0.206	0.735	0.508	0.66	0.421	0.798	0.621
FATALITY	0.879	0.121	0.055	0.011	0.044	0.03	0.083	0.048	0.975	0.736
Weighted Average	0.741	0.627	0.251	0.465	0.778	0.614	0.755	0.613	0.802	0.627

As is shown in Table 6-1, the overall classification accuracies for the training and test datasets are 74.06% and 62.73%, showing reasonable classification performance. The relatively large variance (11.33%) between the classification accuracies for training and testing datasets indicates that the learned classifier is more specific to the training dataset, and a more comprehensive training dataset including sufficient records for each injury severity is desirable to produce a more compatible classifier.

Similarly to Section 4.2, the DTNB classifier also produced performance statistics regarding the popularly used classification performance measurements: TP rate, FP rate, precision, F-measure, and AUC values. The TP rates range from 0.121 for FATALTY to 0.788 for NO INJURY with a weighted average of 0.627 for the testing dataset, as illustrated in Table 6-2. These results demonstrate that the DTNB classifier is able to classify 78.8% of instances with no injuries correctly, while its capability of classifying injury and fatal instances is relatively inferior. This implies that this classifier performs

better on no injuries and injuries than fatal cases since the majority of the training dataset are no injury and injury records, with which more representative decision rules could be extracted for injury severity prediction. It is also shown in Table 6-2 that the DTNB hybrid model has the best performance in predicting no injury instances in the testing dataset, and its F-measure is equal to 0.73. For instances with fatalities, the trained classifier performs inferiorly due to the limited sample size, with its F-measure equal to 0.048. Overall, the average F-Measure is 0.613 for the entire test dataset, implying an acceptable model performance of the trained classifier.

The ROC curves and corresponding AUC values are also trained to support the promising performance of this DTNB classifier, as are shown in Figures. It is revealed that the DTNB model achieves the best performance for fatal records, with an AUC of 0.736. This is followed by that for no injury instances and injury instances, with AUCs of 0.631 and 0.621, respectively. The overall AUC for the test dataset is 0.627, suggesting that the learned DTNB classifier is able to effectively extract the injury severity patterns and produce an acceptable performance.

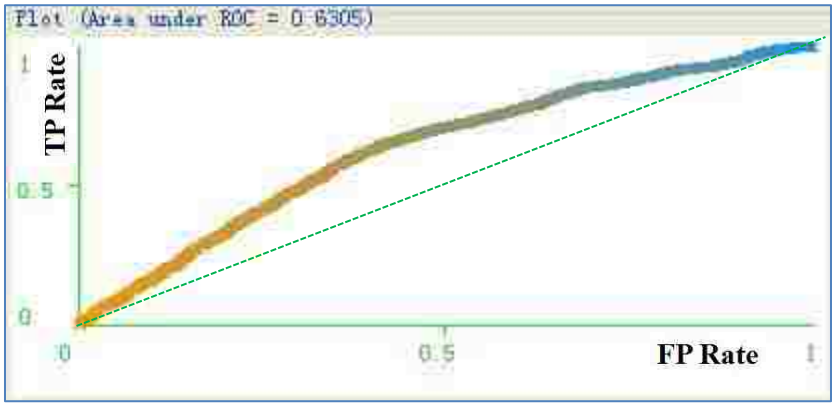

Figure 6-1 ROC Curve for the Category of NO INJURY.

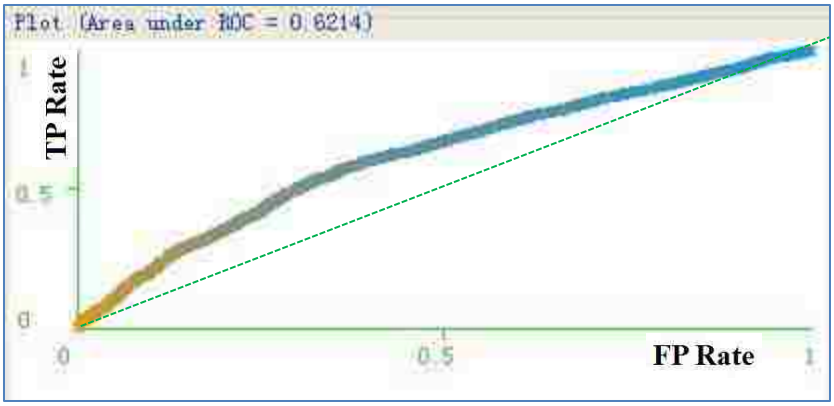

Figure 6-2 ROC Curve for the Category of INJURY.

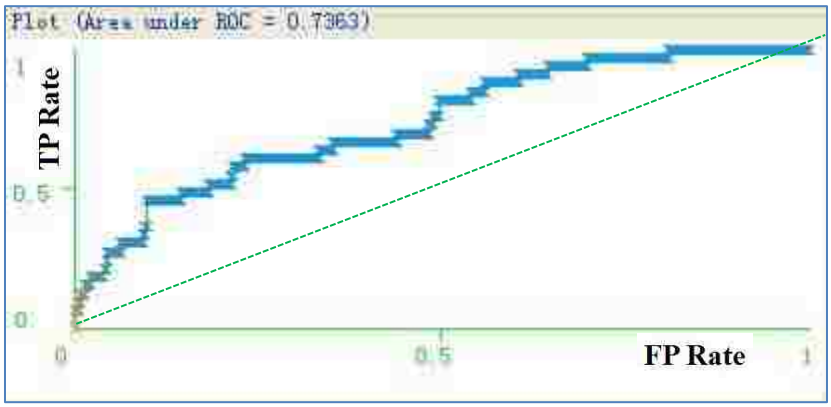

Figure 6-3 ROC Curve for the Category of FATALITY.

The DTNB model also produces a confusion matrix to illustrate misclassifications between each pair of injury severity levels, shown in Table 6-3**.** As is illustrated in Table 6-3**,** 1,531 instances of no injuries are misclassified as injury cases, 52 no injuries are misclassified as fatal instances, 2,764 injuries are misclassified as no injuries, 77 injuries are misclassified as fatalities, 12 fatalities are misclassified as no injury cases, and 17 fatalities are misclassified as injury cases. The overall match rate (accuracy) is 62.73%, illustrating an acceptable model performance was produced by the DTNB hybrid model.

Table 6-3 DTNB Classification Confusion Matrix for the Test Dataset.

		Predicted Instances Classified by Severity		
		NO INJURY (8670)	INJURY (3144)	FATALITY (133)
Observed Instances Classified by Severity	NO INJURY (7477)	5894	1531	52
	INJURY (4437)	2764	1596	77
	FATALITY (33)	12	17	4

6.3 Contributing Variable Selection and Decision Rule Extraction

In this study, 15 attributes are selected as the decisive feature set by the hybrid classifier as follows: *DAY*, *RDREL*, *LIGHT*, *WEATHER*, *RDGRD*, *NVEH*, *RDFUNC*, *MCINV*, *HEVINV*, *DTINC*, *RDPV*, *NLANE*, *DBELT*, *DALC*, and *MAXDAM*. These variables cover the information regarding weather, lighting condition, road geometry characteristics, driver behavior information, etc. Note that the attribute set is selected for formulating decision rules for all three injury severities based on the entire dataset, not only for a particular injury outcome. 2,865 decision rules are trained by the DTNB classifier based on the selected attributes, in which 1,366 rules are used for predicting no injury cases, 1,488 for injury prediction, and 11 for fatality prediction. As shown in Table 6-3, there are 8,670 instances in the test dataset predicted as no injuries, 3,144 as injuries, and 133 as fatalities. On average, a decision rule for no injury prediction is used to classify 6.3 instances in the testing dataset, a decision rule for injury prediction is used to classify 2.1 instances, and a decision rule for fatality prediction is used to classify 12.1 instances. However, if only the correctly classified instances are considered, the average numbers of correct predictions are 4.3 instances for a no injury decision rule, 1.1 instances for an injury decision rule, and 0.4 instances for a fatality decision rule. Based on these results, the learned decision rules for no injury are the most efficient in severity

outcome prediction, followed by those for injury. The learned decision rules for fatality are the least efficient in correct classification, which explains the lowest TP rate and F-measure of FATALITY for the testing dataset in Table 6-2.

6.4 Variable Influence Analysis

A trained DT lists all the decision rules for predicting the most probable driver injury severity in rear-end crashes under a set of specific conditions of these selected variables. The learned DT is fundamentally a matrix of if-then rules working with condition states and action state: if a specific set of conditions for the selected attributes is satisfied, a particular injury severity level that a driver is most likely to suffer in a rear-end crash would be returned. Table 6-4 shows the decision rules for fatality prediction. For example,

If *DAY*=SUN, and *RDREL*=ONWAY, and *LIGHT*=DARK, and *WEATHER*=CLEAR, and *RDGRD*=ONGRADE, and *NVEH*=TWO, and *RDFUNC*=RINT, and *MCINV*=N, and *HEVINV*=Y, and *DTINC*=NEAR, and *RDPV*=PAVED, and *NLANE*=TWO, and *DBELT*=Y, and *DALC*=N, and *MAXDAM*=DSABL, **Then** *SEV*=FATALITY.

Although there are not statistical summaries in the results, the significant effects of some condition-states on driver fatal injuries could be detected in Table 6-4.

RDREL has a unanimous condition state for all the 11 decision rules, *RDREL*=ONWAY, indicating that it is highly likely to result in fatalities if the first harmful event of a serial rear-end crash happens on the roadway. This is probably because in serial rear-end crashes with multiple vehicles, the first event on a roadway

segment would block traffic and result in consecutive collisions due to limited response time for following vehicle drivers. There are 5 out of the 11 decision rules with the presence of dark lighting conditions (*LIGHT*=DARK), indicating that insufficient light condition is an important factor in inducing driver fatal injuries in rear-end crashes. The other 6 decision rules for driver fatality prediction are associated with daylight conditions, which seem contradictory to commonsense. Similarly contradictory findings are also concluded for *WEATHER*, *RDGRD,* and *RDPV*, where clear weather, level road grade, and paved road surface are the most frequent conditions in predicting driver fatal injuries. These contradictions are explainable because drivers tend to be more aware when driving in adverse conditions, such as extreme weather, inferior environment lighting conditions, mountainous terrain, wet or icy pavement surfaces (associated with extreme weather), granular pavement, etc., while crash risk and severity might induce potential speeding or careless driving in comfortable driving environments. This finding receives support from multiple studies (Haque et al., 2012; Savolainen and Mannering, 2007; Shaheed et al., 2013; Yu and Abdel-Aty, 2014a). For instance, Yu and Abdel-Aty (2014) discovered that snowy weather conditions tend to reduce the likelihood of serious crashes. Savolainen and Mannering (2007) found that crashes occurring on wet road surfaces also tend to be less severe.

The number of vehicles (*NVEH*) involved in a crash, as has been shown in Chapter 4 for multiple times, is significant in predicting driver fatal injuries in rear-end crashes, and two-vehicle rear-end crashes are the most common type resulting in fatalities, indicated in Table 6-4. Further analyses also found that the number of vehicles in a crash has significant influence on the mechanisms of inducing crash occurrences and casualties.

Venkataraman et al. (2013) discovered that the significant attribute sets affecting crash potentials vary for distinctive crash groups aggregated by the number of vehicles involved. Therefore, the intriguing effect of number of vehicles involved on driver injury outcomes should be further examined in future research with separate modeling of different crash groups by the number of vehicles involved.

Heavy vehicle involvement (*HEVINV*) is significant in predicting driver fatalities in rear-end crashes, indicated in Table 6-4. Heavy vehicle involvement was present in 8 of the 11 rules for fatality prediction, which is consistent with the statistical findings in Sections 2 and 4.1. Heavy vehicle type (*VTYPE*=HEV) is not found to be significant in predicting driver injury severity in rear-end crashes, which is probably because heavy vehicles make up only a slight portion of all the studied vehicles and its influence is not as significant as *HEVINV*. The number of driving lanes (*NLANE*) is significant in predicting driver injury severities in rear-end crashes, and rear-end crash fatalities are most likely to happen on two-lane roadways, as shown in Table 6-4. The influence of the number of lanes on crash severity has been assessed by a previous study. Jung et al. (2014) discovered that an increase in the number of lanes tends to increase the likelihood of incapable injury and fatalities in crashes occurring in rainy weather, and this study examined its interactive effects across other crash-related factors.

Road function (*RDFUNC*) is a significant factor contributing to driver fatal injury in rear-end crashes. As is shown in Table 6-4, fatal rear-end crashes are more likely to happen on rural roadways, including rural interstate (RINT) and rural non-interstate (RNINT) roadways. This finding is verified by the fact that 55% of the overall fatalities in traffic accidents occur on rural roads (NHTSA, 2013). This is explainable because

traffic in rural areas normally travels at high speeds, which may result in significant deformation of vehicles and, therefore, severe injuries on drivers in rear-end crashes. The safety performance of rural roads is generally discussed jointly with lane numbers. In New Mexico, 65% of crash-related fatalities occurred on rural highways. More than 80% of rural highways are two-lane highways (NMDOT, 2010), which explains the highest frequency of two-lane condition in Table 5 among all categories of lane numbers. Significant research has been done to address rural crash severities, including rural two-lane highways (Chen and Chen, 2011; de Oña et al., 2013; Farah et al., 2009; Karlaftis and Golias, 2002; Kashani and Mohaymany, 2011; Khorashadi et al., 2005b; Lord et al., 2005; Pardillo-Mayora et al., 2010; Siskind et al., 2011). For example, Farah et al. (2009) investigated drivers' overtaking strategies on rural-two-lane highways through driving simulations. Siskind et al. (2011) discovered that speeding, alcohol involvement, and traffic rule violations are major factors of fatal crashes on rural roadways. Table 6-4 also indicates that the condition state of rural roadways (RNINT and RINT) is closely associated with heavy vehicle involvement (*HEVINV*). This could be because a considerable portion of traffic on rural roadways is heavy vehicles travelling at high speeds due to light traffic, which increases the potential of severe injuries and fatalities in rear-end crashes.

Crash location (*DTINC*) is also intimately associated with crash injury severities, which was also found significant in Chapter 4. Table 6-4 shows that most of the fatal crashes occur within 0.1 mile of the nearest intersection (*DTINC*=NEAR). A reasonable explanation is that vehicles decelerate intensively from a high velocity and the headway between vehicles varies dramatically when approaching intersections, leading to

insufficient response time and severe rear-end collisions. Therefore, fatal rear-end crashes are most likely to happen when vehicles are approaching intersections with high speeds and inadequate acceleration, insufficient deceleration, short driver perception and reaction time, etc. may dramatically contribute to severe crash occurrences. Significant studies have been conducted to examine the characteristics of intersection-related crashes, including rear-end crashes. Kim et al. (2007) modeled crash risks for different severities at rural intersections via binomial hierarchical multilevel models. Xie et al. (2013) investigated the safety performance of signalized intersections taking corridor-level correlations into account. Huang et al. (2008) studied the driver injury and vehicle damage patterns in traffic crashes in urban intersections through hierarchical Bayesian models. Therefore, special attention should be paid at intersections, especially for rural intersections where vehicles approach at higher speeds.

Driver alcohol involvement (*DALC*) is also selected as a necessary factor to formulate driver injury severity prediction rules, as listed in Table 6-4, though driver alcohol involvement is only present in 1 of the 11 rules. This is likely because alcohol has influencing effects in impairing drivers' visibility and judgments, and the limited presence of driver alcohol involvement (*DALC*=Y) is due to the insufficient amount of fatality records. Consistent conclusions are also summarized by Hels et al. (2013), Poulsen et al. (2014) and Weiss et al. (2014). The most serious vehicle damage in a crash (*MAXDAM*) is found to be significant in predicting driver injury severities, and vehicle disabled damage (*MAXDAM*=DSABL) appears unanimously in all decision rules for driver fatality prediction. This indicates the significant association between vehicle disabled damage and driver fatality. A rational interpretation is that vehicle damage is a

reflection of the impact generated in a rear-end crash, which is transferrable from vehicle bodies to drivers, and severe vehicle damage is generally associated with high casualties. Comparing *DALC* with *MAXDAM*, it is discovered that disabled vehicle damage is shown in most of the fatality decision rules while driver alcohol or drug involvement is rarely present, which indicates that the variable *MAXDAM* has a higher weight in resulting in driver fatal injuries in rear-end crashes. However, the most serious vehicle damage is an aftermath of rear-end crashes while drivers' alcohol or drug involvement occurs before a crash happens. Therefore, the importance of drunken driving prohibition should not be understated and corresponding law enforcement should be enhanced. Similar to *DALC*, other variables, such as motorcycle involvement (*MCINV*), crash day (*DAY*), and seatbelt usage (*DBELT*), also illustrate unique patterns in predicting driver injury outcomes. Overall, the selected features and their conditions-states are consistent with the statistical analysis findings in Section 2, demonstrating the reasonableness of the results produced by the hybrid classifier.

Table 6-4 Decision Rules for Fatal Injury Classifications from the DTNB Hybrid Classifier.

DAY	RDREL	LIGHT	WEATHER	RDGRD	NVEH	RDFUNC	MCINV	HEVINV	DTINC	RDPV	NLANE	DBELT	DALC	MAXDAM	SEV
SUN	ONWAY	DARK	CLEAR	ONGRADE	TWO	RINT	N	Y	NEAR	PAVED	TWO	Y	N	DISABLE	FATALITY
TUE	ONWAY	DARK	SNOW	LEVEL	THREE	RNINT	N	Y	NEAR	PAVED	TWO	Y	N	DISABLE	FATALITY
WED	ONWAY	DARK	CLEAR	LEVEL	TWO	RINT	N	Y	NEAR	PAVED	TWO	Y	N	DISABLE	FATALITY
SAT	ONWAY	DARK	CLEAR	ONGRADE	TWO	RINT	N	Y	NEAR	PAVED	TWO	Y	N	DISABLE	FATALITY
MON	ONWAY	DAYLIGHT	CLEAR	LEVEL	TWO	RNINT	N	N	NEAR	PAVED	TWO	N	N	DISABLE	FATALITY
FRI	ONWAY	DAYLIGHT	CLEAR	LEVEL	TWO	RINT	N	Y	NEAR	PAVED	TWO	N	N	DISABLE	FATALITY
TUE	ONWAY	DAYLIGHT	SNOW	LEVEL	TWO	URBAN	N	Y	NEAR	PAVED	TWO	Y	N	DISABLE	FATALITY
TUE	ONWAY	DAYLIGHT	CLEAR	ONGRADE	THREE	RNINT	N	N	NEAR	PAVED	ONE	Y	N	DISABLE	FATALITY
FRI	ONWAY	DARK	RAIN	LEVEL	TWO	RNINT	N	N	NEAR	PAVED	TWO	Y	N	DISABLE	FATALITY
SAT	ONWAY	DAYLIGHT	CLEAR	LEVEL	MORE	URBAN	N	Y	FAR	PAVED	ONE	Y	Y	DISABLE	FATALITY
SAT	ONWAY	DAYLIGHT	CLEAR	LEVEL	MORE	URBAN	N	Y	FAR	PAVED	TWO	Y	N	DISABLE	FATALITY

6.5 Performance Comparison with Other Models

Model Performance comparison of this DTNB method with other methods consists of two parts: model performance comparison with proposed MNL-BN model, and model performance with a generalized MNL model (also used in Section 5.6), since they all use the same rear-end crash dataset for model training and calibration. As indicated before, there are several common measurements indicating model performance in both the semi-statistical machine-learning method and the proposed MNL-BN method: prediction accuracy, F-measure, ROC curve and AUC. Therefore, this comparison would be made regarding these measurements. For the comparison with the generalized multinomial logit model, the prediction accuracy would be used as the major measurement indicating model performance.

As shown in the above Table 5-3 and Table 6-1, in terms of prediction accuracy, the DTNB classifier outperforms the proposed MNL-BN model on the training dataset, but performs inferior on the testing dataset. Besides, the variance of estimation accuracies for the proposed MNL-BN model on training and testing datasets is around 1%, and that for the DTNB is around 12% indicating that the trained BN is more transferable and able to explain and model the testing data fairly well. As for the other measurements such as F-measure and AUC shown in Table 5-4 and Table 6-2, it shows the same pattern that the DTNB classifier performs better on training dataset and the proposed MNL-BN model performs better on the testing dataset, which indicates that the machine-learning method is more specific to learning scheme and training dataset and applicable to exploratory analysis, but the proposed MNL-BN model produces more reliable and less biased results for independent datasets once the model structure is trained.

As discussed before in Section 5.6, the total number of correctly predicted observations by the MNL model is 6664 (6664= 4881+1782+1), and the overall correction rate of this MNL prediction is 55.78%, so the DTNB model is more effective and accurate in predicting driver injury severities in rear-end crashes.

6.6 Conclusions

Based on a two-year rear-end crash dataset in New Mexico, this paper applies a DTNB hybrid classifier to select the attributable feature set regarding crash features, vehicle information and driver demographic and behavior characteristics for driver injury severities in rear-end crashes and extract the decision rules for driver injury severity prediction. The DTNB hybrid classifier produces a reasonable classification result, indicated by several performance measurements, such as F-measure, ROC curve, and AUC.

The DTNB hybrid classifier outputs the selected feature set for driver injury severity prediction, accompanied by a decision table with learned decision rules based on the applied dataset. 15 attributes were selected as significant in predicting driver injury fatalities, including crash day (*DAY*), first harmful event location (*RDREL*), lighting condition (*LIGHT*), weather condition (*WEATHER*), road grade (*RDGRD*), number of vehicles involved (*NVEH*), road function (*RDFUNC*), motorcycle involvement (*MCINV*), heavy vehicle involvement (*HEVINV*), distance from crash location to the nearest intersection (*DTINC*), road pavement condition (*RDPV*), number of driving lanes (*NLANE*), seatbelt use (*DBELT*), driver alcohol involvement (*DALC*), and maximum

vehicle damage (*MAXDAM*). Decision rules for fatality prediction reveal that the involvement of heavy vehicles in rear-end crashes increases the probability of driver fatalities, and motorcycle involvement is also significant in predicting driver injury and fatalities. Driver fatalities are more likely to occur in a comfortable traffic environment, such as clear weather, level road grade, and paved road surface, whereas drivers would be more aware of potential risk under adverse driving conditions. Driver fatal injuries are most likely to happen on rural roads, especially on rural two-lane highways. Maximum vehicle damage in rear-end crashes is positively associated with driver injury severities, and drivers are most likely to suffer fatal injuries when vehicles involved in rear-end crashes are disabled. The number of vehicles in a rear-end crash significantly affects driver injury outcomes, and two-vehicle rear-end crash is the most frequent type resulting in driver fatalities. Fatal rear-end crashes are more likely to happen near intersections, where vehicles accelerate and decelerate dramatically, resulting in limited time for proper responses. The effectiveness of seatbelt use and drunk driving prohibition in reducing driver injury severities are verified in the extracted decision rules.

Chapter 7 Conclusions and Future Research

7.1 Conclusions of This Study

Traffic crashes induce significant life and property loss and have imposed heavy economic and emotion burden on social welfare. Examination of crash injury severity patterns and the major contributing factors to crash injury severity is of practical necessity and importance. Transportation researchers have applied multiple types of analysis models to examine crash injury severity and the causal mechanisms in the past decades. Traditional crash data analysis techniques summarize crash severity patterns and investigate contributing factors and their influence solely based on the studied dataset, from which the estimation results might be biased due to limited data size. A Bayesian method is able to provide more accurate posterior estimation results by incorporating parameter prior distribution information and evidence from the studied dataset, and therefore has been increased in traffic safety studies in recent years. Using driver injury severity as a representative, this study is proposed to systematically evaluate the applicability and effectiveness of Bayesian method in traffic crash injury severity analysis, and three major types of Bayesian models are included in this study: hierarchical Bayesian regression models, Bayesian non-regression model and knowledge-based Bayesian non-parametric method, and a Bayesian model selection framework is developed based on discrete research purpose, crash data availability and data structure.

Regression models are the mostly applied research models in traffic crash injury severity analysis, and it is found that hierarchical regression modeling is more robust and produces less biased results due to the hierarchical structure of crash data, such as

national-region-roadway-crash-vehicle/driver/occupant hierarchy. With Bayesian inference method utilized for posterior parameter estimation, three hierarchical Bayesian regression models are considered in this research: hierarchical Bayesian binary logit model, hierarchical Bayesian ordered logit model, and hierarchical random intercept model with cross-level interactions based on the difference in driver injury categorization and model development, and three rural crash datasets are selected respectively for model applicability and performance evaluation. In the calibration procedure of these three models, parameter non-informative prior and the Gibbs Sampler are used in for model simulation in model simulation. Model performances are compared with the control models without considering unobserved heterogeneity based on the DIC criteria. The statistical significant of parameters of interest are evaluated by 95% BCI and variable influence are assessed by the estimated odds ratio or average pseudo-elasticity. Research results indicate that the three proposed Bayesian models outperform their respective counterparts, or provide competitive performance after penalized by the high model complexity, in model fit and estimation effectiveness (Table 7-1). Significant variables contributing to driver injury severities from crash and vehicle/driver levels, such as crash location, road surface condition, lighting condition, vehicle type, driver age, driver sobriety level, seatbelt use, are identified, and their corresponding influence are evaluated.

Table 7-1 Hierarchical Bayesian Regression Model Performance Comparison Summary.

	Model Type	Model Name	DIC Value
Hierarchical Bayesian Regression Models	Regression with Binary Reponses	Hierarchical Bayesian binary logit model	2522.69
		Ordinary binary logit model (control model)	2928.65
	Regression with Ordered Responses	Hierarchical Bayesian ordered logit model	15708.30
		Ordinary ordered logit model (control model)	16716.50
	Regression with Multinomial Reponses	Hierarchical random intercept model with cross-level interactions	6092.45
		Hierarchical random intercept model without cross-level interactions (control model)	6005.42

Regression analyses on crash data are based on certain assumptions on model development and crash data, but non-regression causal relationship may exist between driver injury severity and contributing factors, and violation of these assumptions may lead to biased estimation results. In this study, a MNL-BN hybrid model is utilized as a non-regression machine-learning method in this study by relaxing certain hierarchical model in model assumptions to predict driver injury severities, where the multinomial logit model is utilized to select significant variables for driver injury prediction and the BN model is used to train an optimal classifier. A two-year rear-end crash dataset is used for a case study to evaluate model applicability and performance. The test results demonstrate that the proposed hybrid approach performs reasonably well and outperforms traditional multinomial logit model in prediction accuracy. The Bayesian network reference analyses indicate that the factors, such as truck-involvement, inferior lighting conditions, windy weather conditions, the number of vehicles involved, etc. could significantly increase driver injury severities in rear-end crashes.

In this study, a DTNB hybrid classifier, which is an incorporation of a decision table and a naïve Bayes classifier and has never been used in traffic safety analysis before, is utilized as a representative model of the knowledge-based non-parametric Bayesian machine-learning models to identify the deterministic attribute set that best predicts driver injury severities and extract the corresponding decision rules based on these attributes. A same rear-end crash dataset with the MNL-BN model is also used for case study analysis. The test results show that the hybrid classifier performs fairly well, but is less transferrable to independent testing datasets comparing with the propose MNL-BN model, showing that the machine-learning method is more specific to learning scheme and training dataset. It is also superior to traditional MNL model in injury severity prediction accuracy. Fifteen significant attributes were found to be significant in predicting driver injury severities, including weather, lighting conditions, road geometry characteristics, driver behavior information, etc. The extracted decision rules demonstrate that heavy vehicle involvement, a comfortable traffic environment, inferior lighting conditions, two-lane rural roadways, vehicle disabled damage, and two-vehicle crashes would increase the likelihood of drivers sustaining fatal injuries.

In summary, a framework of selecting the most appropriate Bayesian approaches for traffic crash driver injury severity analyses, and five representative models are developed and utilized to evaluate the applicability of Bayesian methods in data-driven based traffic crash driver injury severity studies. Analysis results from all of these models indicate promising performance of Bayesian methods in predicting driver injury outcome in traffic crashes, capturing the causal relationship between injury outcome and crash, environment, vehicle, driver characteristics, and assessing the heterogeneous influence of

the identified contributing factors, among which some are found significant through more than one model, such as crash type (number of vehicle in a crash), driver age and gender, driver drug or alcohol involvement, seatbelt use, etc. The proposed methods are of theoretical and practical importance for transportation researchers and engineers to better understand crash mechanisms, develop effective crash severity reduction countermeasures and improve traffic system safety performance.

7.2 Future Work Recommendation

Although the proposed three types of Bayesian models illustrate promising model performance on driver injury severity pattern discovery and variable influence assessment, further research is still needed regarding model structure development and model calibration specification.

For the three hierarchical Bayesian regression models, the hierarchical Bayesian binary logit model is simplified from the random intercept model and serves as the basis in model development. The hierarchical Bayesian ordered logit model is developed based on the binary logit model by assuming driver injury severity is an ordinal variable with multiple injury severity levels, which is able to capture more accurate variable impact on different severity levels. Both of these models account for unobserved heterogeneity in crash data only using a random error term representing crash-level variance. The hierarchical Bayesian random intercept model with cross-level interactions overcomes the limitations of the above two models and systematically examines the interaction effects between crash-level and vehicle/driver-level variables. More detailed interactive

effects among variables within a same hierarchical level should be investigated to supplement this study in the future. Due to the model complexity issue in hierarchical random intercept model, variables with too many values were simplified with fewer categories for model simplicity purpose, where to some extent loss of information is inevitable. Therefore, it is desirable to reduce model complexity as well as minimize loss of information through model structure design and specification in further studies. In these three models, due to the limited availability of historical crash data, non-informative prior is used for estimations of all parameters of interest. Comprehensive historical crash data information is desired for informative prior development for more reliable estimation results.

As is shown in the case study, the proposed MNL-BN model is effective in predicting driver injury severity and explicitly formulating statistical associations between driver injury severity outcomes and explanatory attributes. In the development of this model, the input variables are selected based on MNL modeling, which might not be comprehensive. Further research is recommended to introduce more variable importance ranking and selection procedures and focus on the influence of these procedures on BN classifier performance, which is a necessary step to improve the applicability and effectiveness of the BN models. Also, historical crash knowledge is desired in Bayesian probability inference procedure (Equation 3-27) discussed in Section 3.3.3 for more accurate probability inference results.

The DTNB classifier shows it effectiveness in causal relationship examination using decision rules. As indicated before, the attribute set for decision rule learning were selected for all three injury severities based on the entire dataset, where an attribute that is

critical in predicting a specific injury level may not be significant in predicting the others. Therefore, discriminative analysis should be conducted and a unique feature set for each injury outcome should be examined in future research. Besides, the DTNB classifier extracts a total of 2,865 decision rules for three severity levels, which is a complicated presentation even with a succinct and understandable tabular format. Hence, additional effort should be made to elaborate these rules in a clustered and ordered way.

REFERENCES

Abdalla, I.M., 2005. Effectiveness of safety belts and Hierarchical Bayesian analysis of their relative use. Safety Science 43, 91–103. doi:10.1016/j.ssci.2005.02.003

Abdel-Aty, M., Keller, J., 2005. Exploring the overall and specific crash severity levels at signalized intersections. Accident; analysis and prevention 37, 417–25. doi:10.1016/j.aap.2004.11.002

Abdel-Aty, M.A., Abdelwahab, H.T., 2004. Predicting injury severity levels in traffic crashes: a modeling comparison. Journal of Transportation Engineering 130, 204–210.

Abdel-Aty, M.A., Chen, C.L., Schott, J.R., 1998. An assessment of the effect of driver age on traffic accident involvement using log-linear models. Accident Analysis & Prevention 30, 851–861. doi:10.1016/S0001-4575(98)00038-4

Abdelwahab, H.T., Abdel-Aty, M.A., 2001. Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections. Transportation Research Record 1746, 6–13.

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 & Prevention 43, 1581–1589. doi:10.1016/j.aap.2011.03.021

Altmann, A., Tolosi, L., Sander, O., Lengauer, T., 2010. Permutation importance: a corrected feature importance measure. Bioinformatics 26, 1340–1347.

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

Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident; analysis and prevention 41, 153–9. doi:10.1016/j.aap.2008.10.005

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, 628–33. 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, 110–9. 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, 768–75. doi:10.1016/j.aap.2007.09.006

Bedard, M., Guyatt, G.., Stone, M.J., Hireds, J.P., 2002. The Independent Contribution of Driver, Crash, and Vehicle Characteristics to Driver Fatalities. Accident Analysis & Prevention 34, 717–727.

Bedeley, R.T., Lee, E., Attoh-Okine, N.O., 2013. Modelling pedestrian crossing behaviour using Bayesian networks. Proceedings of the Institution of Civil Engineers-Transport 166, 282–288.

Borg, A., Bjelland, H., Njå, O., 2014. Reflections on Bayesian Network models for road tunnel safety design: A case study from Norway. Tunnelling and Underground Space Technology 43, 300–314. doi:10.1016/j.tust.2014.05.004

Bouckaert, R.R., 2008. Bayesian Network Classifiers in Weka for Version 3-5-7. The University of Waikato, Hamilton, New Zealand.

Brooks, S.P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7, 434–455. doi:10.1080/10618600.1998.10474787

Buntine, W., 1991. Theory refinement on Bayesian networks, in: Proceedings of Seventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers, Inc., San Francisco, California, pp. 52–60.

Cafiso, S., Di Graziano, A., Di Silvestro, G., La Cava, G., Persaud, B., 2010. Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. Accident Analysis & Prevention 42, 1072–9. doi:10.1016/j.aap.2009.12.015

Caliendo, C., Guida, M., Parisi, A., 2007. A crash-prediction model for multilane roads. Accident Analysis & Prevention 39, 657–70. doi:10.1016/j.aap.2006.10.012

Carriquiry, A.L., Pawlovich, M., 2004. From empirical Bayes to full Bayes:methods for analyzing traffic safety data. Ames, Iowa.

Castillo, E., Menéndez, J.M., Sánchez-Cambronero, S., 2008. Predicting traffic flow using Bayesian networks. Transportation Research Part B: Methodological 42, 482–509. doi:10.1016/j.trb.2007.10.003

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, 365–75. doi:10.1016/j.jsr.2005.06.013

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

Chang, L.-Y., Wang, H.-W., 2006. Analysis of traffic injury severity: an application of non-parametric classification tree techniques. Accident Analysis & Prevention 38, 1019–27. doi:10.1016/j.aap.2006.04.009

Chen, E., Tarko, A.P., 2014. Modeling safety of highway work zones with random parameters and random effects models. Analytic Methods in Accident Research 1, 86–95. doi:10.1016/j.amar.2013.10.003

Chen, F., Chen, S., 2011. Injury severities of truck drivers in single- and multi-vehicle accidents on rural highways. Accident Analysis & 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, 323–32. doi:10.1016/j.aap.2014.09.025

Chen, W.-H., Jovanis, P.P., 2000. Method for identifying factors contributing to driverinjury severity in traffic crashes. Transportation Research Record: Journal of the Transportation Research Board 1717, 1–9.

Chiang, V.X.Y., Cheng, J.Y.X., Zhang, Z.C., Teo, L.-T., 2014. Comparison of severity and pattern of injuries between motorcycle riders and their pillions: a matched study. Injury 45, 333–7. doi:10.1016/j.injury.2013.01.040

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 & Prevention 35, 253–259. doi:10.1016/S0001-4575(02)00003-9

Chliaoutakis, J.E., Gnardellis, C., Drakou, I., Darviri, C., Sboukis, V., 2000. Modelling the factors related to the seatbelt use by the young drivers of Athens. Accident Analysis & Prevention 32, 815–825. doi:10.1016/S0001-4575(00)00006-3

Christoforou, Z., Cohen, S., Karlaftis, M.G., 2010. Vehicle occupant injury severity on highways: an empirical investigation. Accident; analysis and prevention 42, 1606–

20. doi:10.1016/j.aap.2010.03.019

Chung, Y., 2010. Development of an accident duration prediction model on the Korean Freeway Systems. Accident; analysis and prevention 42, 282–9. doi:10.1016/j.aap.2009.08.005

Chung, Y., Song, T.-J., Yoon, B.-J., 2014. Injury severity in delivery-motorcycle to vehicle crashes in the Seoul metropolitan area. Accident; analysis and prevention 62, 79–86. doi:10.1016/j.aap.2013.08.024

Congdon, P., 2005. Bayesian Models for Categorical Data, 1st ed. Wiley.

Conroy, C., Hoyt, D.B., Eastman, A.B., Erwin, S., Pacyna, S., Holbrook, T.L., Vaughan, T., Sise, M., Kennedy, F., Velky, T., 2006. Rollover crashes: predicting serious injury based on occupant, vehicle, and crash characteristics. Accident; analysis and prevention 38, 835–42. doi:10.1016/j.aap.2006.02.002

Cooper, G., Herskovits, E., 1991. A Bayesian method for the induction of probabilistic network from data.

Cooper, G., Herskovits, E., 1992. A Bayesian method for the induction of probabilistic networks from data. Machine Learning 9, 309–347.

Cowles, K., 2003. Practical Considerations for WinBUGS Users.

Czech, S., Shakeshaft, A.P., Byrnes, J.M., Doran, C.M., 2010. Comparing the cost of alcohol-related traffic crashes in rural and urban environments. Accident Analysis & Prevention 42, 1195–8. doi:10.1016/j.aap.2010.01.010

Das, A., Abdel-Aty, M., 2011. A combined frequency–severity approach for the analysis of rear-end crashes on urban arterials. Safety Science 49, 1156–1163.

Davis, G.A., Swenson, T., 2006. Collective responsibility for freeway rear-ending accidents? An application of probabilistic causal models. Accident Analysis & Prevention 38, 728–736.

de Lapparent, M., 2006. Empirical Bayesian analysis of accident severity for motorcyclists in large French urban areas. Accident 38, 260–268. doi:10.1016/j.aap.2005.09.001

de Lapparent, M., 2008. Willingness to use safety belt and levels of injury in car accidents. Accident; analysis and prevention 40, 1023–32. doi:10.1016/j.aap.2007.11.005

de Oña, J., López, G., Mujalli, R., Calvo, F.J., 2013. Analysis of traffic accidents on rural highways using Latent Class Clustering and Bayesian Networks. Accident Analysis

& Prevention 51, 1–10. doi:10.1016/j.aap.2012.10.016

de Oña, J., Mujalli, R.O., Calvo, F.J., 2011. Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks. Accident Analysis & Prevention 43, 402–11. doi:10.1016/j.aap.2010.09.010

Deublein, M., Schubert, M., Adey, B.T., Köhler, J., Faber, M.H., 2013. Prediction of road accidents: A Bayesian hierarchical approach. Accident Analysis & Prevention 51, 274–291. doi:10.1016/j.aap.2012.11.019

Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3.

Dissanayake, S., Lu, J.J., 2002. Factors influential in making an injury severity difference to older drivers involved in fixed object–passenger car crashes. Accident Analysis & Prevention 34, 609–618. doi:10.1016/S0001-4575(01)00060-4

Dobbertin, K.M., Freeman, M.D., Lambert, W.E., Lasarev, M.R., Kohles, S.S., 2013. The relationship between vehicle roof crush and head, neck and spine injury in rollover crashes. Accident; analysis and prevention 58, 46–52. doi:10.1016/j.aap.2013.04.020

Domingos, P., Plazzani, M., 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 29, 103–130.

Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2014. Multivariate random-parameters zero-inflated negative binomial regression model: an application to estimate crash frequencies at intersections. Accident; analysis and prevention 70, 320–9. doi:10.1016/j.aap.2014.04.018

Dougherty, J., Kohavi, R., Sahami, M., 1995. Supervised and unsupervised discretization of continuous features, in: Proceedings of the 12th International Conference on Machine Learning. San Francisco, California, pp. 194–202.

Duan, J., Li, Z., Salvendy, G., 2013. Risk illusions in car following: Is a smaller headway always perceived as more dangerous? Safety Science 53, 25–33.

Eksler, V., 2010. Measuring and understanding road safety performance at local territorial level. Accident Analysis & Prevention 48, 1197–1202.

El-Basyouny, K., Sayed, T., 2010. Application of generalized link functions in developing accident prediction models. Safety Science 48, 410–416.

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, 119–27.

doi:10.1016/j.aap.2012.01.027

Eluru, N., Bhat, C.R., 2007. A joint econometric analysis of seat belt use and crash-related injury severity. Accident Analysis & Prevention 39, 1037–1049.

Eluru, N., Bhat, C.R., Hensher, D.A., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accident Analysis & Prevention 40, 1033–1054.

Elvik, R., 2008. The predictive validity of empirical Bayes estimates of road safety. Accident Analysis & Prevention 40, 1964–1969. doi:10.1016/j.aap.2008.07.007

Elvik, R., 2013. A before–after study of the effects on safety of environmental speed limits in the city of Oslo, Norway. Safety Science 55, 10–16.

Farah, H., Bekhor, S., Polus, A., 2009. Risk evaluation by modeling of passing behavior on two-lane rural highways. Accident Analysis & Prevention 41, 887–894. doi:10.1016/j.aap.2009.05.006

Farmer, C.M., Lund, A.K., 2002. Rollover Risk of Cars and Light Trucks after Accounting for Driver and Environmental Factors. Accident Analysis & Prevention 34, 163–173.

Feng, T., Timmermans, H.J.P., 2013. Transportation mode recognition using GPS and accelerometer data. Transportation Research Part C: Emerging Technologies 37, 118–130. doi:10.1016/j.trc.2013.09.014

Flask, T., Schneider, W., 2013. A Bayesian analysis of multi-level spatial correlation in single vehicle motorcycle crashes in Ohio. Safety Science 53, 1–10. doi:10.1016/j.ssci.2012.08.008

Flask, T., Schneider, W.H., Lord, D., 2014. A segment level analysis of multi-vehicle motorcycle crashes in Ohio using Bayesian multi-level mixed effects models. Safety Science 66, 47–53. doi:10.1016/j.ssci.2013.12.006

Frith, W.J., 1984. Adoption of right turn on red—Effects on injury accidents at signalized intersections. Accident Analysis & Prevention 16, 75–76. doi:10.1016/0001-4575(84)90032-0

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

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis, Third. ed. Chapman and Hall/CRC, London,UK.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1995. Markov Chain Monte Carlo in Practice. Chapman & Hall/CRC.

Gkritza, K., Mannering, F.L., 2008. Mixed logit analysis of safety-belt use in single- and multi-occupant vehicles. Accident Analysis & Prevention 40, 443–51. doi:10.1016/j.aap.2007.07.013

Goodheart, B., 2013. Identification of Causal Paths and Prediction of Runway Incursion Risk by Means of Bayesian Belief Networks. Transportation Research Record: Journal of the Transportation Research Board 2400, 9–20. doi:10.3141/2400-02

Gregoriades, A., 2007. Towards a user-centred road safety management method based on road traffic simulation, in: Proceedings of the 39th Conference on Winter Simulation: 40 Years! The Best Is Yet to Come. Washington, D.C., pp. 1905–1914.

Gregoriades, A., Mouskos, K.C., 2013. Black spots identification through a Bayesian Networks quantification of accident risk index. Transportation Research Part C: Emerging Technologies 28, 28–43. doi:10.1016/j.trc.2012.12.008

Gross, E.A., Axberg, A., Mathieson, K., 2007. Predictors of seatbelt use in American Indian motor vehicle crash trauma victims on and off the reservation. Accident; analysis and prevention 39, 1001–5. doi:10.1016/j.aap.2007.01.008

Haleem, K., Abdel-Aty, M., 2010. Examining traffic crash injury severity at unsignalized intersections. Journal of safety research 41, 347–57. doi:10.1016/j.jsr.2010.04.006

Haleem, K., Gan, A., 2013. Effect of driver's age and side of impact on crash severity along urban freeways: a mixed logit approach. Journal of safety research 46, 67–76. doi:10.1016/j.jsr.2013.04.002

Haleem, K., Gan, A., 2015. Contributing factors of crash injury severity at public highway-railroad grade crossings in the U.S. Journal of safety research 53, 23–9. doi:10.1016/j.jsr.2015.03.005

Hall, M., Frank, E., 2008. Combining Naive Bayes and Decision Tables, in: 21st Florida Artificial Intelligence Research Society Conference. AAAI Press, Miami, Florida, pp. 318–319.

Haque, M.M., Chin, H.C., Debnath, A.K., 2012. An investigation on multi-vehicle motorcycle crashes using log-linear models. Safety Science 50, 352–362. doi:10.1016/j.ssci.2011.09.015

Haque, M.M., Chin, H.C., Huang, H., 2010. Applying Bayesian hierarchical models to examine motorcycle crashes at signalized intersections. Accident Analysis & Prevention 42, 203–212. doi:10.1016/j.aap.2009.07.022

Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Botstein, D., Brown, P., 2000. "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns. Genome Biology. Genome Biology 1.

Hauer, E., 1992. Empirical bayes approach to the estimation of "unsafety": The multivariate regression method. Accident Analysis & Prevention 24, 457–477. doi:10.1016/0001-4575(92)90056-O

Hauer, E., 2001. Overdispersion in modelling accidents on road sections and in Empirical Bayes estimation. Accident Analysis & Prevention 33, 799–808. doi:10.1016/S0001-4575(00)00094-4

Heckerman, D., Geiger, D., Chickering, D.M., 2013. Learning Bayesian networks: the combination of knowledge and statistical data. Machine Learning 20, 197–243.

Hefny, A.F., Barss, P., Eid, H.O., Abu-Zidan, F.M., 2012. Motorcycle-related injuries in the United Arab Emirates. Accident Analysis & Prevention 49, 245–8. doi:10.1016/j.aap.2011.05.003

Hels, T., Lyckegaard, A., Simonsen, K.W., Steentoft, A., Bernhoft, I.M., 2013. Risk of severe driver injury by driving with psychoactive substances. Accident Analysis & Prevention 59, 346–356. doi:10.1016/j.aap.2013.06.003

Hensher, D. a., Greene, W., 2003. The mixed logit models: The state of practice. Transportation 30, 133–176.

Hilakivi, I., Veilahti, J., Asplund, P., Sinivuo, J., Laitinen, L., Koskenvuo, K., 1989. A sixteen-factor personality test for predicting automobile driving accidents of young drivers. Accident Analysis & Prevention 21, 413–8.

Hossain, M., Muromachi, Y., 2012. A Bayesian network based framework for real-time crash prediction on the basic freeway segments of urban expressways. Accident Analysis & Prevention 45, 373–381. doi:10.1016/j.aap.2011.08.004

Hu, W., Donnell, E.T., 2011. Severity models of cross-median and rollover crashes on rural divided highways in Pennsylvania. Journal of safety research 42, 375–82. doi:10.1016/j.jsr.2011.07.004

Huang, H., Abdel-Aty, M., 2010. Multilevel data and bayesian analysis in traffic safety. Accident Analysis & Prevention 42, 1556–1565. doi:10.1016/j.aap.2010.03.013

Huang, H., Chin, H.C., Haque, M.M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis & Prevention 40, 45–54. doi:10.1016/j.aap.2007.04.002

Huang, H., Hu, S., Abdel-Aty, M., 2014. Indexing crash worthiness and crash

aggressivity by major car brands. Safety Science 62, 339–347. doi:10.1016/j.ssci.2013.09.002

Huang, H., Siddiqui, C., Abdel-Aty, M., 2011. Indexing crash worthiness and crash aggressivity by vehicle type. Accident Analysis & Prevention 43, 1364–1370. doi:10.1016/j.aap.2011.02.010

Islam, S., Mannering, F., 2006. Driver aging and its effect on male and female single-vehicle accident injuries: some additional evidence. Journal of safety research 37, 267–76. doi:10.1016/j.jsr.2006.04.003

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.

Jiang, H., Deng, Y., Chen, H.-S., Tao, L., Sha, Q., Chen, J., Tsai, C.-J., Zhang, S., 2004. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics 5, 81.

Jin, S., Wang, D., Qi, H., 2010. Bayesian Network Method of Speed Estimation from Single-Loop Outputs. Journal of Transportation Systems Engineering and Information Technology 10, 54–58. doi:10.1016/S1570-6672(09)60022-2

Jones, A.P., Jørgensen, S.H., 2003. The use of multilevel models for the prediction of road accident outcomes. Accident Analysis & Prevention 35, 59–69. doi:10.1016/S0001-4575(01)00086-0

Jonsson, T., Ivan, J., 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.

Jonsson, T., Lyon, C., Ivan, J., Washington, S.P., Van Schalkwyk, I., Lord, D., 2009. Differences in the performance of safety performance functions estimated for total crash count and for crash count by crash type. Transportation Research Record: Journal of the Transportation Research Board 2102, 115–123.

Jung, S., Jang, K., Yoon, Y., Kang, S., 2014. Contributing factors to vehicle to vehicle crash frequency and severity under rainfall. Journal of Safety Research 50, 1–10. doi:10.1016/j.jsr.2014.01.001

Jung, S., Qin, X., Noyce, D.A., 2010. Rainfall effect on single-vehicle crash severities using polychotomous response models. Accident; analysis and prevention 42, 213–24. doi:10.1016/j.aap.2009.07.020

Karlaftis, M.G., Golias, I., 2002. Effects of road geometry and traffic volumes on rural

roadway accident rates. Accident Analysis & Prevention 34, 357–365. doi:10.1016/S0001-4575(01)00033-1

Kashani, A.T., Mohaymany, A.S., 2011. Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models. Safety Science 49, 1314–1320. doi:http://dx.doi.org/10.1016/j.ssci.2011.04.019

Kelly, P., Sanson, T., Strange, G., Orsay, E., 1991. A prospective study of the impact of helmet usage on motorcycle trauma. Annals of Emergency Medicine 20, 852–856. doi:10.1016/S0196-0644(05)81426-X

Khorashadi, A., Niemeier, D., Shankar, V., Mannering, F., 2005a. Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. Accident Analysis & Prevention 37, 910–21. doi:10.1016/j.aap.2005.04.009

Khorashadi, A., Niemeier, D., Shankar, V., Mannering, F., 2005b. Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis. Accident Analysis & Prevention 37, 910–921. doi:10.1016/j.aap.2005.04.009

Kim, D.-G., Lee, Y., Washington, S., Choi, K., 2007. Modeling crash outcome probabilities at rural intersections: application of hierarchical binomial logistic models. Accident Analysis & Prevention 39, 125–34. doi:10.1016/j.aap.2006.06.011

Kim, J.-K., Kim, S., Ulfarsson, G.F., Porrello, L.A., 2007. Bicyclist injury severities in bicycle-motor vehicle accidents. Accident; analysis and prevention 39, 238–51. doi:10.1016/j.aap.2006.07.002

Kim, J.-K., Ulfarsson, G.F., Kim, S., Shankar, V.N., 2013. Driver-injury severity in single-vehicle crashes in California: A mixed logit analysis of heterogeneity due to age and gender. Accident Analysis & Prevention 50, 1073–81. doi:10.1016/j.aap.2012.08.011

Kim, J.-K., Ulfarsson, G.F., Shankar, V.N., Kim, S., 2008. Age and pedestrian injury severity in motor-vehicle crashes: a heteroskedastic logit analysis. Accident Analysis & Prevention 40, 1695–1702. doi:10.1016/j.aap.2008.06.005

Kim, J.-K., Ulfarsson, G.F., Shankar, V.N., Mannering, F.L., 2010. A note on modeling pedestrian-injury severity in motor-vehicle crashes with the mixed logit model. Accident; analysis and prevention 42, 1751–8. doi:10.1016/j.aap.2010.04.016

Kjaerulff, U.B., Madsen, A.L., 2008. Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis. Springer, New York, NY.

Kockelman, K.M., Kweon, Y.-J., 2002. Driver injury severity: an application of ordered probit models. Accident Analysis & Prevention 34, 313–321. doi:10.1016/S0001-4575(01)00028-8

Kohavi, R., 1995. "A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence 2, 1137–1143.

Kononov, J., Lyon, C., Allery, B., 2011. Relating flow, speed and density of urban freeways to functional form of an spf, in: Compendium of the 2011 TRB Annual Meeting. Washington, D.C.

Kuhnert, P.M., Do, K.-A., McClure, R., 2000. Combining non-parametric models with logistic regression: an application to motor vehicle injury data. Computational Statistics & Data Analysis 34, 371–386. doi:10.1016/S0167-9473(99)00099-7

Kutner, M.H., Neter, J., Nachtsheim, C.J., Li, W., 2004. Applied Linear Statistical Models, 5th Intern. ed. McGraw-Hill Education, Columbus, Ohio.

Lambert-Bélanger, A., Dubois, S., Weaver, B., Mullen, N., Bédard, M., 2012. Aggressive driving behaviour in young drivers (aged 16 through 25) involved in fatal crashes. Journal of Safety Research 43, 333–8. doi:10.1016/j.jsr.2012.10.011

Lee, C., Abdel-Aty, M., 2008. Presence of passengers: does it increase or reduce driver's crash potential? Accident; analysis and prevention 40, 1703–12. doi:10.1016/j.aap.2008.06.006

Lemp, J.D., Kockelman, K.M., Unnikrishnan, A., 2011. Analysis of large truck crash severity using heteroskedastic ordered probit models. Accident; analysis and prevention 43, 370–80. doi:10.1016/j.aap.2010.09.006

Levine, E.M., Bedard, M., Molloy, D.W., Basilevsky, A., 1999. Determinants of Driver Fatality Risk in Front Impact Fixed Object Collisions. Mature Medicine, Toronto, Canada.

Lew, A., 1991. Fuzzy decision tables for expert systems. Computers Math Application 21, 111–116.

Li, X., Lord, D., Zhang, Y., Xie, Y., 2008. Predicting motor vehicle crashes using Support Vector Machine models. Accident; analysis and prevention 40, 1611–8. doi:10.1016/j.aap.2008.04.010

Li, Y., Bai, Y., 2008. Comparison of characteristics between fatal and injury accidents in the highway construction zones. Safety Science 46, 646–660.

Liang, Y., Lee, J.D., 2014. A hybrid Bayesian Network approach to detect driver

cognitive distraction. Transportation Research Part C: Emerging Technologies 38, 146–155. doi:10.1016/j.trc.2013.10.004

Liu, Y., Feng, X., Wang, Q., Zhang, H., Wang, X., 2014. Prediction of Urban Road Congestion Using a Bayesian Network Approach. Procedia - Social and Behavioral Sciences 138, 671–678. doi:10.1016/j.sbspro.2014.07.259

Lord, D., 2006. Modeling motor vehicle crashes using poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accident Analysis & Prevention 38, 751–766.

Lord, D., Geedipally, S.R., Guikema, S.D., 2010. Extension of the application of conway-maxwell-poisson models: analyzing traffic crash data exhibiting underdispersion. Risk Analysis 30, 1268–1276.

Lord, D., Manar, A., Vizioli, A., 2005. Modeling crash-flow-density and crash-flow-V/C ratio relationships for rural and urban freeway segments. Accident Analysis & Prevention 37, 185–199. doi:10.1016/j.aap.2004.07.003

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., Miranda-Moreno, L.F., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: A Bayesian perspective. Safety Science 46, 751–770. doi:10.1016/j.ssci.2007.03.005

Lord, D., Park, P.Y.-J., 2008. Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates. Accident; analysis and prevention 40, 1441–57. doi:10.1016/j.aap.2008.03.014

Lord, D., Washington, S., Ivan, J.N., 2007. Further notes on the application of zero-inflated models in highway safety. Accident; analysis and prevention 39, 53–7. doi:10.1016/j.aap.2006.06.004

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 & Prevention 40, 964–975. doi:10.1016/j.aap.2007.11.002

MacNab, Y.C., 2003. A Bayesian hierarchical model for accident and injury surveillance. Accident Analysis & Prevention 35, 91–102.

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

Malyshkina, N. V, Mannering, F.L., 2009. Markov switching multinomial logit model: An application to accident-injury severities. Accident; analysis and prevention 41, 829–38. doi:10.1016/j.aap.2009.04.006

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

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, 217–26. doi:10.1016/j.aap.2008.11.001

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

Massie, D.L., Campbell, K.L., Williams, A.F., 1995. Traffic Accident involvement rates by driver age and gender. Accident Analysis & Prevention 27, 73–87. doi:10.1016/0001-4575(94)00050-V

Mbakwe, A.C., Saka, A.A., Choi, K., Lee, Y.-J., 2014. Modeling Highway Traffic Safety in Nigeria Using Delphi Technique and Bayesian Network, in: TRB 93rd Annual Meeting Compendium of Papers. Washington, D.C., p. 21p.

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

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, 260–6. doi:10.1016/j.aap.2007.06.006

Moore, D.N., Schneider, W.H., Savolainen, P.T., Farzaneh, M., 2011. Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and non-intersection locations. Accident Analysis & Prevention 43, 621–630. doi:10.1016/j.aap.2010.09.015

Morgan, A., Mannering, F.L., 2011. The effects of road-surface conditions, age, and

gender on driver-injury severities. Accident Analysis & Prevention 43, 1852–63. doi:10.1016/j.aap.2011.04.024

Mujalli, R.O., 2011. Application of Bayesian Networks for the Analysis of Traffic Accidents Injury Severity on Rural Highways. University of Granada.

Mujalli, R.O., de Oña, J., 2011. A method for simplifying the analysis of traffic accidents injury severity on two-lane highways using Bayesian networks. Journal of Safety Research 42, 317–326. doi:10.1016/j.jsr.2011.06.010

Mussone, L., Ferrari, A., Oneta, M., 1999. An analysis of urban collisions using an artificial intelligence model. Accident Analysis & Prevention 31, 705–718. doi:10.1016/S0001-4575(99)00031-7

National Health Council, 2013. Estimating the Costs of Unintentional Injuries [WWW Document]. URL http://www.nsc.org/news_resources/injury_and_death_statistics/Pages/EstimatingtheCostsofUnintentionalInjuries.aspx

National Highway Traffic Safety Adminisitration(NHTSA), 2011. Fatality analysis reporting system (FARS) encyclopedia [WWW Document].

National Highway Traffic Safety Adminisitration(NHTSA), 2013. Fatality Analysis Reporting System Encyclopedia [WWW Document].

National Highway Traffic Safety Administration, 2013a. Traffic Safety Facts 2011:A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System. Washington, D.C.

National Highway Traffic Safety Administration, 2013b. Traffic Safety Facts 2011 Data: Rural/Urban Comparison. Washington, D.C.

National Highway Traffic Safety Adminstriation, 2001. Traffic Safety in the New Millennium: Strategies for Law Enforcement. Washington, D.C.

National Highway Traffic Safety Adminstriation, 2012. Traffic Safety Facts 2010 Data. Washington, D.C.

National Highway Traffic Safety Adminstriation, 2013. Traffic Safety Facts 2012:A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System. Washington, D.C.

New Mexico Department of Transportation, 2012. New Mexico traffic crash annual report 2011.

NHTSA, 2006. Large-Truck Crash Causation Study: An Initial Overview. Washington,

D.C.

Noland, R.B., Quddus, M.A., 2004. A spatially disaggregate analysis of road casualties in England. Accident; analysis and prevention 36, 973–84. doi:10.1016/j.aap.2003.11.001

Oh, J., Washington, S.P., Nam, D., 2006. Accident prediction model for railway-highway interfaces. Accident; analysis and prevention 38, 346–56. doi:10.1016/j.aap.2005.10.004

Ouyang, Y., Shankar, V., Yamamoto, T., 2002. Modeling the Simultaneity in Injury Causation in Multivehicle Collisions. Transportation Research Record: Journal of the Transportation Research Board 1784, 143–152. doi:10.3141/1784-18

Ozbay, K., Noyan, N., 2006. Estimation of incident clearance times using Bayesian Networks approach. Accident Analysis & Prevention 38, 542–555. doi:10.1016/j.aap.2005.11.012

Pai, C.-W., Hwang, K.P., Saleh, W., 2009. A mixed logit analysis of motorists' right-of-way violation in motorcycle accidents at priority T-junctions. Accident; analysis and prevention 41, 565–73. doi:10.1016/j.aap.2009.02.007

Pardillo Mayora, J.M., Jurado Piña, R., 2009. An assessment of the skid resistance effect on traffic safety under wet-pavement conditions. Accident Analysis & Prevention 41, 881–6. doi:10.1016/j.aap.2009.05.004

Pardillo-Mayora, J.M., Domínguez-Lira, C. a, Jurado-Piña, R., 2010. Empirical calibration of a roadside hazardousness index for Spanish two-lane rural roads. Accident Analysis & Prevention 42, 2018–2023. doi:10.1016/j.aap.2010.06.012

Park, B.-J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accident; analysis and prevention 41, 683–91. doi:10.1016/j.aap.2009.03.007

Patil, S., Geedipally, S.R., Lord, D., 2012. Analysis of crash severities using nested logit model--accounting for the underreporting of crashes. Accident Analysis & Prevention 45, 646–53. doi:10.1016/j.aap.2011.09.034

Persaud, B., Lan, B., Lyon, C., Bhim, R., 2010. Comparison of empirical Bayes and full Bayes approaches for before-after road safety evaluations. Accident Analysis & Prevention 42, 38–43. doi:10.1016/j.aap.2009.06.028

Persaud, B., Lyon, C., 2007. Empirical Bayes before-after safety studies: lessons learned from two decades of experience and future directions. Accident; analysis and prevention 39, 546–55. doi:10.1016/j.aap.2006.09.009

Poulsen, H., Moar, R., Pirie, R., 2014. The culpability of drivers killed in New Zealand road crashes and their use of alcohol and other drugs. Accident Analysis & Prevention 67C, 119–128. doi:10.1016/j.aap.2014.02.019

Provost, F., Domingos, P., 2001. Well-trained PETs: Improving Probability Estimation Trees, CeDER Working Paper #IS-00-04. New York.

Pulugurtha, S.S., Otturu, R., 2014. Effectiveness of red light running camera enforcement program in reducing crashes: evaluation using "before the installation", "after the installation", and "after the termination" data. Accident; analysis and prevention 64, 9–17. doi:10.1016/j.aap.2013.10.035

Qin, X., Ivan, J.N., Ravishanker, N., Liu, J., Tepas, D., 2006. Bayesian estimation of hourly exposure functions by crash type and time of day. Accident Analysis & Prevention 38, 1071–1080.

Quddus, M., Wang, C., Ison, S.G., 2010. Road traffic congestion and crash severity: an econometric analysis using ordered response models. Journal of Transportation Engineering 136, 424–435.

Quigley, J., Hardman, G., Bedford, T., Walls, L., 2011. Merging expert and empirical data for rare event frequency estimation: Pool homogenisation for empirical Bayes models. Reliability Engineering & System Safety 96, 687–695. doi:10.1016/j.ress.2010.12.007

Refaeilzadeh, P., Tang, L., Liu, H., 2009. Cross-Validation. Encyclopedia of Database Systems.

Rifaat, S.M., Chin, H.C., 2005. Analysis of severity of single-vehicle crashes in Singapore, in: TRB 2005 Annual Meeting CD-ROM. Transportation Research Board, National Research Counci, Washington, D.C.

Riviere, C., Lauret, P., Ramsamy, J.F.M., Page, Y., 2006. A Bayesian Neural Network approach to estimating the Energy Equivalent Speed. Accident Analysis & Prevention 38, 248–259. doi:10.1016/j.aap.2005.08.008

Robinson, R.W., 1977. Counting unlabeled acyclic digraphs. Combinatorial Mathematics V Lecture Notes in Mathematics 622, 28–43.

Rodenburg, W., Heidema, A.G., Boer, J.M.A., Bovee-Oudenhoven, I.M.J., Feskens, E.J.M., Mariman, E.C.M., Keijer, J., 2008. A framework to identify physiological responses in microarray-based gene expression studies: selection and interpretation of biologically relevant genes. Physiological Genomics 33, 78–90.

Russo, B.J., Savolainen, P.T., Schneider, W.H., Anastasopoulos, P.C., 2014. Comparison

of factors affecting injury severity in angle collisions by fault status using a random parameters bivariate ordered probit model. Analytic Methods in Accident Research 2, 21–29. doi:10.1016/j.amar.2014.03.001

Savolainen, P., Mannering, F., 2007. Probabilistic models of motorcyclists' injury severities in single- and multi-vehicle crashes. Accident Analysis & 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 & Prevention 43, 1666–1576. doi:10.1016/j.aap.2011.03.025

Shaheed, M.S., Gkritza, K., 2014. A latent class analysis of single-vehicle motorcycle crash severity outcomes. Analytic Methods in Accident Research 2, 30–38. doi:10.1016/j.amar.2014.03.002

Shaheed, M.S.B., Gkritza, K., Zhang, W., Hans, Z., 2013. A mixed logit analysis of two-vehicle crash severities involving a motorcycle. Accident Analysis & Prevention 61, 119–28. doi:10.1016/j.aap.2013.05.028

Shankar, V., Mannering, F., 1996. An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity. Journal of Safety Research 27, 183–194. doi:10.1016/0022-4375(96)00010-2

Shankar, V., Mannering, F., Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accident Analysis & Prevention 27, 371–389. doi:http://dx.doi.org/10.1016/0001-4575(94)00078-Z

Shankar, V.N., Albin, R.B., Milton, J.C., Mannering, F.L., 1998. Evaluating Median Cross-Over 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.

Shively, T.S., Kockelman, K.M., Damien, P., 2010. A bayesian semi-parametric model to estimate relationships between crash counts and roadway characteristics. Transportation Research Part B: Methodological 44, 699–715.

Simoncic, M., 2004. A Bayesian network model of two-car accidents. Journal of transportation and Statistics 7, 13–25.

Simons-Morton, B., Lerner, N., Singer, J., 2005. The observed effects of teenage passengers on the risky driving behavior of teenage drivers. Accident Analysis & Prevention 37, 973–82. doi:10.1016/j.aap.2005.04.014

Siskind, V., Steinhardt, D., Sheehan, M., O'Connor, T., Hanks, H., 2011. Risk factors for fatal crashes in rural Australia. Accident Analysis & Prevention 43, 1082–1088. doi:10.1016/j.aap.2010.12.016

Snijders, A.B., Bosker, R.J., 2000. Multilevel Analysis, An Introduction to Basic and Advanced Multilevel Modeling, Second. ed. SAGE Publication Ltd, London,UK.

Sohn, S.Y., Shin, H., 2001. Pattern recognition for road traffic accident severity in Korea. Ergonomics 44, 107–117.

Spiegelhalter, D.J., Best, N.G., Carlin, B.., 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

Spiegelhalter, D.J., Thomas, A., Best, N.G., Lunn, D., 2003. WinBUGS User Manual Version 1.4. Cambridge, UK.

Srinivasan, K., 2002. Injury severity analysis with variable and correlated thresholds: ordered mixed logit formulation. Transportation Research Record: Journal of the Transportation Research Board 1784, 132–142.

Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. Bioinformatics 8, 25.

Svetnik, V., Liaw, A., Tong.C., WANG, T., 2004. Application of Breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. Multiple Classifier Systems Lecture Notes in Computer Science 3077, 334–343.

Tang, R., Sinnwell, J., Li, J., Ride, D., de Andrade, M., Biernacka, J., 2009. Identification of genes and haplotypes that predict rheumatoid arthritis using random forests, in: BMC Proceedings.

Tape, T.G., 2001. Interpretation of diagnostic tests. Annal of Interval Medicine 135, 72. doi:doi:10.7326/0003-4819-135-1-200107030-00043

Tay, R., 2015. A random parameters probit model of urban and rural intersection crashes. Accident; analysis and prevention 84, 38–40. doi:10.1016/j.aap.2015.07.013

Ukkusuri, S. V., Hasan, S., Aziz, H., 2011. Random Parameter Model Used to Explain Effects of Built-Environment Characteristics on Pedestrian Crash Frequency. Transportation Research Record: Journal of the Transportation Research Board 2237, 98–106.

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 & Prevention 36, 135–147. doi:10.1016/S0001-4575(02)00135-5

Usman, T., Fu, L., Miranda-Moreno, L.F., 2010. Quantifying safety benefit of winter road maintenance: accident frequency modeling. Accident; analysis and prevention 42, 1878–87. doi:10.1016/j.aap.2010.05.008

van Petegem, J.W.H.J.H., Wegman, F., 2014. Analyzing road design risk factors for run-off-road crashes in The Netherlands with crash prediction models. Journal of safety research 49, 121–7. doi:10.1016/j.jsr.2014.03.003

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 & Prevention 59, 309–18. doi:10.1016/j.aap.2013.06.021

Wang, C., Quddus, M., Ison, S., 2009. The effects of area-wide road speed and curvature on traffic casualties in England. Journal of Transport Geography 17, 385–395. doi:10.1016/j.jtrangeo.2008.06.003

Wang, M., Chen, X., Zhang, H., 2010. Maximal conditional chi-square importance in random forests. Bioinformatics 26, 831–837.

Wang, X., Abdel-Aty, M., 2006. Temporal and spatial analyses of rear-end crashes at signalized intersections. Accident; analysis and prevention 38, 1137–50. doi:10.1016/j.aap.2006.04.022

Wang, X., Abdel-Aty, M., 2008. Modeling left-turn crash occurrence at signalized intersections by conflicting patterns. Accident Analysis & Prevention 40, 76–88. doi:10.1016/j.aap.2007.04.006

Wang, Z., Chen, H., Lu, J., 2009. Exploring Impacts of Factors Contributing to Injury Severity at Freeway Diverge Areas. Transportation Research Record: Journal of the Transportation Research Board 2102, 43–52. doi:10.3141/2102-06

Washington, S.P., Congdon, P., Karlaftis, M.G., Mannering, G., 2005. Bayesian multinomial logit models: exploratory assessment of transportation applications., in: TRB 2005 Annual Meeting CD-ROM. Transportation Research Board, National Research Counci, Washington, D.C.

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 & Prevention 65, 142–155. doi:10.1016/j.aap.2013.12.020

Weiss, S.J., Ellis, R., Ernst, A.A., Land, R.F., Garza, A., 2001. A comparison of rural and urban ambulance crashes. The American journal of emergency medicine 19, 52–6. doi:10.1053/ajem.2001.20001

Winston, C., Maheshri, V., Mannering, F.L., 2006. An exploration of the offset hypothesis using disaggregate data: the case of airbags and antilock brakes. Journal of Risk and Uncertainty 32, 89–99.

Witlox, F., Antrop, M., Bogaert, P., De Maeyer, P., Derudder, B., Neutens, T., Van Acker, V., Van de Weghe, N., 2009. Introducing functional classification theory to land use planning by means of decision tables. Decision Support Systems 46, 875–881. doi:10.1016/j.dss.2008.12.001

Witten, I.H., Frank, F., Hall, M.A., 2011. Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann Publishers, San Francisco.

World Health Organization, 2013. Global Status Report on Road Safety 2013 Supporting a Decade of Action. Geneva, Switzerland.

Wu, Q., Chen, F., Zhang, G., Liu, X.C., Wang, H., Bogus, S.M., 2014. Mixed logit model-based driver injury severity investigations in single- and multi-vehicle crashes on rural two-lane highways. Accident; analysis and prevention 72C, 105–115. doi:10.1016/j.aap.2014.06.014

Wu, Z., Sharma, A., Mannering, F.L., Wang, S., 2013. Safety impacts of signal-warning flashers and speed control at high-speed signalized intersections. Accident Analysis & Prevention 54, 90–8. doi:10.1016/j.aap.2013.01.016

Xie, K., Wang, X., Huang, H., Chen, X., 2013. Corridor-level signalized intersection safety analysis in Shanghai, China using Bayesian hierarchical models. Accident Analysis & Prevention 50, 25–33. doi:10.1016/j.aap.2012.10.003

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

Xie, Y., Zhang, Y., Liang, F., 2009. Crash injury severity analysis using Bayesian ordered probit models. Journal of Transportation Engineering 135, 18–25.

Xie, Y., Zhao, K., Huynh, N., 2012. Analysis of driver injury severity in rural single-vehicle crashes. Accident Analysis & 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

Xiong, Y., Tobias, J.L., Mannering, F.L., 2014. The analysis of vehicle crash injury-severity data: A Markov switching approach with road-segment heterogeneity. Transportation Research Part B: Methodological 67, 109–128. doi:10.1016/j.trb.2014.04.007

Xu, P., Huang, H., 2015. Modeling crash spatial heterogeneity: random parameter versus geographically weighting. Accident; analysis and prevention 75, 16–25. doi:10.1016/j.aap.2014.10.020

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., Hashiji, J., Shankar, V.N., 2008. Underreporting in traffic accident data, bias in parameters and the structure of injury severity models. Accident Analysis & Prevention 40, 1320–1329.

Yanmaz-Tuzel, O., Ozbay, K., 2010. A comparative Full Bayesian before-and-after analysis and application to urban road safety countermeasures in New Jersey. Accident Analysis & Prevention 42, 2099–2107.

Yasmin, S., Eluru, N., Pinjari, A.R., Tay, R., 2014. Examining driver injury severity in two vehicle crashes - A copula based approach. Accident Analysis & Prevention 66, 120–35. doi:10.1016/j.aap.2014.01.018

Yau, K.K.W., 2004. Risk factors affecting the severity of single vehicle traffic accidents in Hong Kong. Accident; analysis and prevention 36, 333–40. doi:10.1016/S0001-4575(03)00012-5

Yu, R., Abdel-Aty, M., 2013. Multi-level Bayesian analyses for single- and multi-vehicle freeway crashes. Accident Analysis & Prevention 58, 97–105. doi:10.1016/j.aap.2013.04.025

Yu, R., Abdel-Aty, M., 2014a. 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., 2014b. Using hierarchical Bayesian binary probit models to analyze crash injury severity on high speed facilities with real-time traffic data.

Accident Analysis & Prevention 62, 161–167. doi:10.1016/j.aap.2013.08.009

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

Zador, P., Moshman, J., Marcus, L., 1982. Adoption of right turn on red: Effects on crashes at signalized intersections. Accident Analysis & Prevention 14, 219–234. doi:10.1016/0001-4575(82)90033-1

Zhao, L., Wang, X., Qian, Y., 2012. Analysis of factors that influence hazardous material transportation accidents based on Bayesian networks: A case study in China. Safety Science 50, 1049–1055. doi:10.1016/j.ssci.2011.12.003

Zou, Y., Zhang, Y., Lord, D., 2013. Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis. Accident; analysis and prevention 50, 1042–51. doi:10.1016/j.aap.2012.08.004

Zou, Y., Zhang, Y., Lord, D., 2014. Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models. Analytic Methods in Accident Research 1, 39–52. doi:10.1016/j.amar.2013.11.001

CURRICULUM VITAE

1. PERSONAL DATA

RESEARCH INTERESTS

- Traffic Safety Analyses and Accident Modeling
- Traffic Congestion Pricing
- GIS-based Transportation Infrastructure Management
- Sustainable Transportation Infrastructure Design and Maintenance
- Human Factors in Transportation System
- Intelligent Transportation System
- Road Safety Design
- Construction Scheduling and Cost Estimate
- Construction Management

EDUCATION

- January 2012-Present Research Assistant, Department of Civil Engineering, University of New Mexico (UNM)
  Dissertation: Data-Driven Bayesian Method-based Traffic Crash Driver Injury Severity Formulation, Analysis, and Inference

- 2011 M.S. School of Transportation Engineering, Tongji University, China
  Thesis: Research of Evaluation Indexes of Highway Tunnel Safety Based on Visual Information

- 2008 B.S. School of Transportation Engineering, Tongji University, China
  Thesis: Discovering Greening Patterns of Highways in Shanghai Suburban Area

PROFESSIONAL WORK EXPERIENCE

May- August, 2014 New Mexico Governor's Fellowship Intern Program
-Summer Intern in **New Mexico Department of Transportation**, District 3
Work summary: signal warrant analysis; rehabilitation project field review; project plan review; field survey for drainage design; traffic safety permit review.

PROFESSIONAL LICENSE AND CERTIFICATE

State of New Mexico Engineering in Training (EIT) Certificate (FE Exam passed, to be issued after graduation), May 2015

MAJOR COURSES

- Mechanics of Materials
- Structural Mechanics
- Reinforced Concrete Structure
- Traffic Engineering
- Soil Mechanics
- Rail Transportation
- Subgrade Engineering
- Transportation Planning

- Highway Geometry Design
- Transport Economics
- Pavement Design
- Bridge Design
- Roadway Materials
- Road Safety Design
- Transportation Ergonomics
- Road and Airport Facility Management System
- CAD on Transportation
- Road Environment and Landscape Design

HONORS AND AWARDS

- Feb. 2015 **Doctoral Conference Presentation Award** ($1000), awarded by UNM Office of Graduate Studies
- Nov. 2014 **Research Travel Grant** ($500), awarded by UNM Civil Engineering Department
- Mar. 2014 **Student Conference Award Program** ($600), awarded by UNM Career Service
- Mar. 2014 **Graduate Research and Travel Grant** ($530), awarded by Office of Graduate Studies, UNM
- Mar. 2013 **Student Conference Award Program** ($600), awarded by UNM Career Service
- Mar. 2013 **Graduate Research and Travel Grant** ($700), awarded by Office of Graduate Studies, UNM
- Jan. 2010 **Level-A Graduate Scholarship**, awarded by Tongji University, China (total tuition waiver)

COMPUTER SKILLS

Proficient in using AutoCAD, ArcGIS, HCM Software, PTV VISSIM, Microsoft Office, Matlab, SAS, R.

2. PROJECT EXPERIENCE, RESEARCH PUBLICATIONS AND PRESENTATIONS

PROJECT EXPERINECE

- *Safety Performance Enhancement Analysis of Rumble Stripes with Elements: A Case Study on Rural Highway US 285 in New Mexico*, University of New Mexico University Transportation Center and New Mexico Department of Transportation (September 2014-Present)

Research assistant and leading researcher. Main work: Comprehensive Literature Review, Crash Data Collection, Data Analysis and Model Cross-validation, Results Conclusion and Final Report Composition.

- *New Mexico Department of Transportation Pavement Evaluation Program 2012* (January 2012-December 2012)

Research assistant. Main/Participated work: Applicants organization and Interview, Daily Operation and Management, Data Collection from Crew and Validation, Data Submission to NMDOT, Data Quality Control (QC) Analysis, Final Report Composition.

- *Research on Visual Stimulus and Visual Environment Improvement for Drivers in Freeway Tunnels*, China Natural Science Fund Project (2008-2011)
- *Study on Key Traffic Safety Technique and Its Application on Expressway Tunnels in Zhejiang Province'*, funded by Ministry of Transportation of Zhejiang Province, China (2008-2010)
- *Research on Accident-Prone Analysis and Countermeasures about Management Safety of Tunnel Group',* funded by Ministry of Transportation of Zhejiang Province, China (2008-2011)

Research assistant in the above three projects in China. Main work: On-site Driving Experiment Design, Experiment Calibration and Data Collection, Data Analysis and Model Cross-validation, Results Conclusion and Final Report Composition

- *Lines and Markings Design for Underground Parking Lots of Fuzhou Wanda Plaza, China*, (April 2010-August 2010)

Primary designer. Main work (with AutoCAD): On-site Investigation, Traffic Markings Design, Signs Dimension Design and Implementation, Signs Information Design and Implementation

RESEARCH REPORT

1. Susan Bogus Halter, Vanessa Valentin, Guohui Zhang, David Barboza, **Cong Chen** and Su Zhang. 2012 Pavement Evaluation Report Northern New Mexico. New Mexico Department of Transportation. State Maintenance Bureau SB-2. 2012.

JOURNAL PUBLICATIONS

1. **Cong Chen**, Su Zhang, Guohui Zhang, Susan M. Bogus, and Vanessa Valentin. Discovering Temporal and Spatial Patterns and Characteristics of Pavement Distress Condition Data on Major Corridors in New Mexico. *Journal of Transport Goegraphy*, Volume 38, 2014, pp. 148-158.

2. **Cong Chen**, Guohui Zhang, Rafiqul Tarefder, Jianming Ma, Heng Wei, and Hongzhi Guan. A Multinomial Logit Model-Bayesian Network Hybrid Approach for Driver Injury Severity Analyses in Rear-end Crashes. *Accident Analysis and Prevention.* Volume 80, 2015, pp. 76-88.

3. **Cong Chen**, Guohui Zhang, Zong Tian, Susan M. Bogus, and Yin Yang. Hierarchical Bayesian Random Intercept Model-based Cross-level Interaction Decomposition for Truck Driver Injury Severity Investigations. *Accident Analysis and Prevention*, Volume 85, 2015, pp. 186-198.

4. **Cong Chen**, Guohui Zhang, Hua Wang, Jinfu Yang, Peter J. Jin, and C. Michael Walton. Bayesian Network-based Formulation and Analysis for Toll Road Utilization Supported by Traffic Information Provision. *Transportation Research: Part C: Emerging*

Technologies, Volume 60, 2015, pp. 339-359.

5. Su Zhang, Susan M. Bogus, Chris Lippitt, Paul R. H. Neville, **Cong Chen**, Guohui Zhang, and Vanessa Valentin. Extracting Pavement Distress Condition Patterns based on High Spatial Resolution Multispectral Digital Aerial Photography. *Photogrammetric Engineering and Remote Sensing,* Volume 81, No. 9, 2015, pp. 709-720.

6. Qiong Wu, Xiaodong Pan, Hui Yang, **Cong Chen**. Research on Driving Safety Experiment of Tunnel Based on Sidewall Effect, *Highway Engineering*, Vol. 38, No. 5, 2013, pp 99-102 (Chinese Edition).

7. Yongchao Song, Xiaodong Pan, **Cong Chen**, and Zewen Yu. Study of Connectivity of Traffic Nodes on Mountainous Highway for Emergency Evacuation. *China Journal of Highway and Transport*, 23(8), 2010, pp. 102-106 (Chinese Edition).

8. Xiaodong Pan, **Cong Chen**, Tao Lin, and Yongchao Song. Research on the Dimension of Crosswalk-notice Mark on Highways. *Highway Engineering,* 34(6), 2009, pp.144-148 (Chinese Edition).

PAPERS UNDER PEER-REVIEW

1. **Cong Chen**, Guohui Zhang, Jinfu Yang, John C. Milton, and Adélamar "Dely" Alcántara**.** Rear-end Crash Severity Analysis Using a Decision Table-Naïve Bayes Hybrid Classifier. Under review. *Accident Analysis and Prevention.*

2. **Cong Chen**, Guohui Zhang, Helai Huang, Jianming Ma, Yanyan Chen, and Hongzhi Guan. Hierarchical Bayesian Modeling of Driver Injury Severity in Rural Interstate Freeway Crashes. Under review. *Journal of Safety Research.*

3. **Cong Chen**, Guohui Zhang, and Helai Huang, Jiangfeng Wang, and Rafiqul A. Tarefder. Examining Driver Injury Severity Outcomes in Rural Non-interstate Roadway Crashes Using a Hierarchical Ordered Logit Model. Under review. *Accident Analysis and Prevention.*

4. **Cong Chen**, Guohui Zhang, Zhen Qian, Rafiqul A. Tarefder, and Zong Tian. Investigating Driver Injury Severity Patterns in Rollover Crashes Using a Support Vector Machine Model. Under Review. *Accident Analysis and Prevention*.

5. **Cong Chen**, Yanyan Chen, Jianming Ma, Guohui Zhang, and C. Michael Walton. Driver Behavior Formulation in Intersection Dilemma Zones with Phone Use Distraction via a Logit-Bayesian Network Hybrid Approach. Under Review. *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*.

6. Qiong Wu, Guohui Zhang, **Cong Chen**, Haizhong Wang, and Adélamar "Dely" Alcántara. Heterogeneous Imapacts of Gender-Interpreted Contributing Factors on Driver Injury Severities in Single-Vehicle Rollover Crashes. Under review, *Accident Analysis and Prevention.*

7. Sikai Xie, **Cong Chen**, Qiong Wu, Qi Lu, Su Zhang, Guohui Zhang, Yin Yang, A Cost-Effective Kinect-Based Approach for 3D Pavement Surface Reconstruction and Cracking Recognition. Under Review. *IEEE Transactions on Intelligent Transportation Systems.*

CONFERENCE PRESENTATIONS

1. **Cong Chen**, Guohui Zhang, Zong Tian, Susan M. Bogus, and Yin Yang. Investigating Truck Driver Injury Severity Using a Hierarchical Bayesian Random Intercept Model with Cross-Level Interactions. Accepted for Presentation at the 95th Transportation Research Board Annual Meeting, Washington, D.C., Jan. 2016.

2. Sikai Xie, **Cong Chen**, Qiong Wu, Qi Lu, Su Zhang, Kelly R. Montoya, Guohui Zhang and Yin Yang. 3D Pavement Surface Reconstruction and Cracking Detection Based on Kinect Fusion Techniques. Accepted for Presentation at the 95th Transportation Research Board Annual Meeting, Washington, D.C., Jan. 2016.

3. Qiong Wu, Guohui Zhang, **Cong Chen**, Haizhong Wang, and Adélamar "Dely" Alcántara. Heterogeneous Analysis of Gender on Driver Injury Severities in Single-Vehicle Rollover Crashes. Accepted for Presentation at the 95th Transportation Research Board Annual Meeting, Washington, D.C., Jan. 2016.

4. Stephen Lujan, **Cong Chen**, Guohui Zhang, Rafiqul A. Tarefder, Timothy Parker, and Francisco Sanchez. Enhancing Safety Performance of Rumble Strips Through The Use of Reflective Striping: An Empirical Study on U.S. 285 in New Mexico. Accepted for Presentation at the 95th Transportation Research Board Annual Meeting, Washington, D.C., Jan. 2016.

5. **Cong Chen**, Guohui Zhang, Hua Wang, Peter J. Jin, and C. Michael Walton. Examining Toll Road Utilization Supported by Traffic Information Provision Using a Nest-logit-based Bayesian Network Approach. Presented at the 94th Transportation Research Board Annual Meeting, Washington, D.C., Jan. 2015.

6. **Cong Chen**, Guohui Zhang, Jinfu Yang, John C. Milton, Adélamar "Dely" Alcántara. Prediction of Driver Injury Severity in Rear-end Crashes: A Decision Table/Naïve Bayes (DTNB) Classification Approach. Presented at the 94th Transportation Research Board Annual Meeting, Washington, D.C., Jan. 2015.

7. **Cong Chen**, Guohui Zhang, Helai Huang, Jianming Ma, Yanyan Chen, and Hongzhi Guan. Examining Driver Injury Severity on Rural Interstate Highways Using a Hierarchical Bayesian Approach. Presented at the 94th Transportation Research Board Annual Meeting, Washington, D.C., Jan. 2015.

8. Qiong Wu, Cheng Wang, **Cong Chen**, and Guohui Zhang, 2015. Developing a VISSIM-Based Simulation Platform for Connected Autonomous Vehicle Control Optimization at Intersections. Accepted for presentation at the UTC Spotlight Conference (Nov 4-5, 2015)

9. **Cong Chen**, Su Zhang, Guohui Zhang, Susan M. Bogus, and Vanessa Valentin. Temporal-spatial Pattern Discovery of Pavement Distress on New Mexico Major Corridors. Presented at 2014 New Mexico Tech Fiesta Student Poster Competition, University of New Mexico, Sep. 2014.

10. **Cong Chen**, Qiong Wu, Guohui Zhang, Jianming Ma, Heng Wei, and Hongzhi Guan. Rear-end Crash Casualty Severity Analysis using Multinomial Logit Model and Bayesian Network. Presented at the 93rd Transportation Research Board Annual Meeting, Washington, D.C., Jan. 2014.

11. Qiong Wu, **Cong Chen** and Guohui Zhang. Formulating Alcohol-Impaired Driver Injury Severities in Intersection-Related Crashes in New Mexico. Presented at 51st Paving and Transportation Conference, Albuquerque, New Mexico, Jan. 2014.

12. **Cong Chen**, Su Zhang, Guohui Zhang, Susan M. Bogus, and Vanessa Valentin. Analysis of Pavement Surface Distress Condition on Major Corridors in New Mexico. Presented at the 92nd Annual Meeting of Transportation Research Board, Washington, D.C., Jan. 2013.

13. **Cong Chen**, David Barboza, Susan M. Bogus, Guohui Zhang, and Vanessa Valentin. Pavement Distress Condition Data Collection, Process, Analysis, and Interpretation on Major Corridors in New Mexico. Presented at 50th Paving and Transportation Conference, Albuquerque, New Mexico, Jan. 2013.

14. **Cong Chen** and Xiaodong Pan. Determining the Sight-Insufficient Locations in Tunnel Entrances: Based on a Driving Visibility Experimental Study in Zhejiang Province, China. Accepted for presentation at the 91st Transportation Research Board Annual Meeting, Washington, D.C., Jan. 2012.

3. SERVICE AND PROFESSIONAL ACTIVITIES

EXTRACURRICULAR ACTIVITIES

- January 2015-Present Vice President of New Mexico ITE Student Chapter
- 2015 Volunteer at 52st Paving and Transportation Conference, Albuquerque, New Mexico
- 2014 Volunteer at 51st Paving and Transportation Conference, Albuquerque, New Mexico
- 2013 Lecturer of UNM Civil Engineering Brycon Career Expo Day
- 2013 Lecturer at Transportation Session in UNM Civil Engineering Open House
- 2013 Volunteer at 50th Paving and Transportation Conference, Albuquerque, New Mexico
- 2012 Lecturer on UNM Civil Engineering Brycon Career Expo Day
- 2012 Lecturer at Transportation Session in UNM Civil Engineering Open House

ACTIVE MEMBERSHIP

- Student member of Institute of Transportation Engineers (ITE)
- Student member of American Society of Civil Engineers (ASCE)
- Student member of Engineers without Boarders (EWB)
- Student member of Chinese Overseas Transportation Association (COTA)

PEER-REVIEW EXPERIENCE

- The TRB Annual Meeting and Journal of the Transportation Research Board
- The International IEEE conference on Intelligent Transportation Systems
- COTA International Conference for Transportation Professionals (CICTP)
- Accident Analysis and Prevention