ANALYZING CRASH FREQUENCY AND SEVERITY DATA USING NOVEL TECHNIQUES

by

GAURAV S. MEHTA

STEVEN JONES JR., COMMITTEE CHAIR
YINGYAN LOU
JAY LINDLY
MICHAEL ANDERSON
ROCKY DURRANS

A DISSERTATION

Submitted in the partial fulfillment of the requirements for the
degree of Doctor of Philosophy in the Department of
Civil, Construction and Environmental Engineering
in the Graduate School of
The University of Alabama

TUSCALOOSA, ALABAMA

2014

ABSTRACT

Providing safe travel from one point to another is the main objective of any public transportation agency. The recent publication of the Highway Safety Manual (HSM) has resulted in an increasing emphasis on the safety performance of specific roadway facilities. The HSM provides tools such as crash prediction models that can be used to make informed decisions. The manual is a good starting point for transportation agencies interested in improving roadway safety in their states. However, the models published in the manual need calibration to account for the local driver behavior and jurisdictional changes. The method provided in the HSM for calibrating crash prediction models is not scientific and has been proved inefficient by several studies. To overcome this limitation this study proposes two alternatives. Firstly, a new method is proposed for calibrating the crash prediction models using negative binomial regression. Secondly, this study investigates new forms of state-specific Safety Performance Function SPFs using negative binomial techniques.

The HSM's 1st edition provides a multiplier applied to the univariate crash prediction models to estimate the expected number of crashes for different crash severities. It does not consider the distinct effect unobserved heterogeneity might have on crash severities. To address this limitation, this study developed a multivariate extension of the Conway Maxwell Poisson distribution for predicting crashes. This study gives the statistical properties and the parameter estimation algorithm for the distribution.

The last part of this dissertation extends the use of Highway Safety Manual by developing a multivariate crash prediction model for the bridge section of the roads. Since the proposed model in the second article is a complex hierarchical model, this article demonstrates the method using a real life application. The study then compares the performance of the newly proposed multivariate Conway Maxwell Poisson (MVCMP) model with the multivariate Poisson Lognormal, univariate Conway Maxwell Poisson (UCMP) and univariate Poisson Lognormal model for different crash severities. This example will help transportation researchers in applying the model correctly.

DEDICATION

To my mother Charu Mehta, my father Satish Mehta, my sisters Khyati and Grishma

LIST OF ABBREVIATION AND SYMBOLS

AADT	Average Annual Daily Traffic
AIC	Akaike Information Criteria
CMF	Crash Modification Factor
FLDH	Four lane divided highway
HSM	Highway Safety Manual
LL	Log likelihood
MAD	Mean absolute deviance
MCMC	Markov Chain Monte Carlo simulation
MH	Metropolis Hastings algorithm
MLE	Maximum likelihood estimation
MPB	Mean prediction bias
MSPE	Mean square prediction error
MVP	Multivariate Poison distribution

MVPLN	Multivariate Poisson lognormal distribution
NB	Negative binomial distribution
PDF	Probability density function
SPF	Safety Performance Function
TLTWRR	Two-lane two way rural road
UCMP	Univariate Conway Maxwell Poisson distribution
Y	Random variable
y	Realization of random variable Y
λ	Mean parameter of Poisson distribution
ϵ	Error term
μ	Expected value of a distribution
ν	Dispersion parameter for Conway Maxwell Poisson distribution
x	Explanatory or observed variables
β	Coefficients to be estimated

α	Dispersion parameter
θ	Random parameter to be estimated
θ^*	Proposed value in the MCMC simulation
$\pi(\theta\|x)$	Prior distribution of the random parameter θ
$\pi(y\|\theta, x)$	Conditional likelihood of observed response given θ and x
$\pi(\theta\|y, x)$	Density of the posterior distribution of θ
$\pi(y\|x)$	Constant term dropped from the Bayes' Theorem
$\alpha(\theta_0, \theta^*)$	Probability of moving from θ to θ^* in MCMC simulation

ACKNOWLEDGEMENTS

I was very fortunate to meet great individuals during the course of my doctoral program at The University of Alabama.

First and Foremost, I would like to express my deepest and sincere gratitude to my advisor and co-chair Dr. Yingyan Lou. Her patience, motivation, enthusiasm, immense knowledge and work ethics made my journey of Ph.D. a privilege. Her passion and dedication to research has and will continue to be a source of inspiration throughout my life. I would like to thank Dr. Steven Jones Jr. for being chair of my committee. He has been extremely helpful and supportive mentor and it has been an honor to work with him on several projects over the past few years.

Thanks to Dr. Lindly for their very helpful inputs, insightful comments and support throughout my study. The two classes I took with him were some of the best classes I ever took. I am very thankful to Dr. Michael Anderson for being a wonderful advisor during my Master's study and becoming a part of my Ph.D. committee. His encouragement during my graduate study encouraged me to pursue Ph.D. Thanks are also due to Dr. Rocky Durrans for serving on my committee and for other helpful interactions.

Special thanks to Ms. Connie Harris for being very helpful and taking care of all the administrative work and travel reimbursements. I have to thank several fellow graduate students who worked with me during my graduate study here. Firstly, Mohammad Miralinaghi

and Francees Green for stimulating discussions, several lunch trips and sleepless nights working together before deadlines. Thanks to Abhay, Peiheng and Samwel for being very motivating and helpful office mates. Thanks to Tyler and Jing for being excellent colleagues and collaborating on several successful projects. Thanks are also due to Majeed and Kofi for endless discussions on events happening around the world and on religion.

Last, but not the least, I need to extend my thanks to my family member, my mother, father, and sisters. Without their unending support, trust and confidence I would not be where I am today.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1. INTRODUCTION

Motor vehicle travel has become an integral part of our life, with around 3 billion vehicle miles driven each year in U.S. (FHWA, 2012). This increasing automobile travel has taken toll on the economy in terms of fatalities, injuries and property damages. In 2010, 32,885 people died in car crashes along with 2.24 million people injured in crashes (NHTSA, 2012). The cost of these crashes to the society, in terms of lost wages, loss of life, reduced productivity etc., is estimated to exceed $230 billion annually (FHWA, 2012). To mitigate these conditions, the transportation agencies at all levels of government emphasize the importance of safety with respect to their responsibilities in providing and managing transportation infrastructure.

Crashes are random events and crash frequencies fluctuate over time and regress towards the mean. Regression towards mean is a phenomenon in which, a variable at its extreme on its first measurement will tend to be closer to the average in the second measurement. For example, a site, which has observed an extremely high number of crashes in one year, will have lower number of crashes the next year irrespective of site being treated with any traffic control measures. The use of short-term crash frequencies in such situations can lead to overestimation of the treatment effect, leading to inaccurate conclusions. To avoid the regression to the mean bias several researchers have used crash rate as a measure. The crash rate is the number of crashes that occur at a given site during a certain period. Intrinsically, the method assumes a linear relationship between the crash frequency and the exposure in the same period. This might not be a case for example crash rate per 50,000 vehicles will not necessarily double when the

volume increases to 100,000 vehicles. Studies have found that the use of crash rate for crash analysis can be misleading and result in the wasteful use of the very limited government resources (Highway Safety Manual, 2010). These limitations created a need for scientific tool that can help in quantitatively evaluating safety alongside other transportation performance measures. Therefore, after 10 years of cooperative research a manual called Highway Safety Manual (HSM) is developed that facilitates consideration of safety consequences while roadway planning, design, operations and maintenance stages.

1.1 Highway Safety Manual (HSM)

American Association of State Highway and Transportation Officials (AASHTO) published the Highway safety Manual (HSM) in 2010. It is the first national resource, which integrates new scientific techniques and knowledge to help transportation officials make informed decisions throughout the project development process, including: planning, design, operations, maintenance, and the roadway safety management process. For example, it can help in screening potential locations for improvement and choosing alternative roadway designs. Traditionally, safety analysis consisted of using descriptive statistical techniques like observed crash frequency and crash rate to quantify the roadway safety (Wade, Hammond, & Kim, 2004; HSM, 2010). These techniques can lead to misleading conclusions if no consideration is provided for regression to the mean bias and the variations in the environmental and roadway conditions.

Its benefits justify the development of HSM as a guidebook to help transportation officials conduct quantitative safety analysis, allowing safety to be quantitatively evaluated alongside transportation performance measures like level of service, environmental impacts and

construction cost. The HSM is organized in to four parts and three volumes. The Part A is and introductory part where it explains the purpose and scope of the HSM. The Part B presents topics, which help transportation planners and managers in identifying improvement sites, diagnosis, countermeasure selection, economic appraisal, project prioritization effectiveness evaluation (HSM, 2010). This research does not go into details of these parts, because they are useful, from the perspective of the policy makes. This research will deals with the Part C and part D of the manual. The HSM Part C provides with predictive methods, which estimate expected average crash frequency for a network.

The methods suggested in HSM provide reliable estimates using long-term crash frequency, often referred to as expected average crash frequency. A wide variety of statistical methods has been applied to estimate the expected average crash frequency. All these methods involve certain statistical regression models with different assumptions on their parameters. These regression models are titled differently in the literature, including accident prediction models (Sawalha and Sayed (2006), Greibe (2003)), crash prediction models (Caliendo, Maurizio and Parisi (2007), and Ackaah and Salifu (2011)), and Safety Performance Function (SPF) in HSM.

The HSM predictive method gives predictive models for different facility types. The key component of these models is the SPFs. These SPFs need to be calibrated to reflect the local or regional conditions. The following equation demonstrates how the base expected crash frequency value could be modified for unique site characteristics using CMFs and calibration factor.

$$N_{Pred} = N_{spfx} \times \left(CMF's_{yx}\right) \times C_x$$

where,

$N_{spf\,x} =$ Safety Performance Function for the base condition and site type x

$CMF's_{yx} =$ Crash Modification factors y specific to site type x

$C_x =$ Calibration factor.

Safety performance functions are the regression models developed for a range of facility types to predict average crash frequencies of road sections (or sites) over a certain time period (usually several years). For practical reasons, SPFs are often in a very concise form and include only limited numbers of variables. Each SPF adopts a set of underlying assumptions on the default values of certain facility characteristics, called base conditions. When some of these base conditions are violated, an adjustment to the prediction is required. Crash Modification factors are used to account for specific site conditions, which vary from the base conditions. The default SPFs and CMFs in HSM are developed based on data from selected states and may not apply universally. The following subsection describes each component of the predictive method in detail.

1.1.1 Safety Performance Functions (SPFs)

HSM identifies crash prediction models as safety performance function and hence this study uses crash prediction models and safety performance function (SPF) interchangeably. A SPF is a regression model, which develops a relationship between the number of crashes and road and traffic characteristics such as AADT, lane width, segment length etc. SPF's are developed for a range of facility types to predict expected average crash frequencies on roadway sections. These SPFs have three important applications listed in HSM. 1) They can be used to determine the safety impact of design changes at the project level. 2) They can be used to identify the sections of the roads that have highest potential for improvements commonly known

as network screening. 3) They can be used to evaluate the safety effects of engineering treatments as part of an empirical Bayes before-after study. For practical reasons, SPFs are often in a very concise form and include only limited numbers of variables. Each SPF adopts a set of underlying assumptions on the default values of certain facility characteristics, called base conditions. When some of these base conditions are violated, an adjustment to the prediction is required. HSM recommends using negative binomial regression for developing SPFs.

1.1.2 Crash Modification Factors (CMFs)

A CMF is a multiplicative factor applied to the base SPF when the conditions on the roadway segment differ from the base condition used for developing SPF. Crash Modification factors are used to account for change in specific site conditions, which vary from the base conditions. Transportation professionals can use CMFs in several ways. They can be used to estimate the impact of various roadway safety countermeasures on roadway safety, compare safety benefits among various alternatives and evaluate cost benefit analysis of the counter measures. The CMF equal to 1 indicates no effect of the changing conditions on the roadway safety, while CMF greater than 1 indicates increase in the crashes and vice-versa. The CMF clearinghouse is a web based database which maintains a list of CMFs developed by different studies and rates them based on the study size and rigor.

1.1.3 Calibration Factor

The SPFs and CMFs published in HSM are developed using data from selected states. Since several factors such as driving behavior, weather, traffic conditions etc. vary from state to state, HSM recommends calibrating these models, or even better, developingnew models specific

to local jurisdictions. The HSM defines calibration factor as the ratio of total number of crashes observed to the total number of crashes predicted using SPFs.

1.1.4 HSM Limitations

The HSM suggested calibration method is taking the ratio of the total observed crashes with the predicted crashes.However, this might not be a very accurate method always. Consider a hypothetical case of a facility with 5 sites as shown inTable 1. Let us assume that the expected crash frequency for five sites is already known and is as show in column 2. Let the second column be the predicted number of crashes obtained using the HSM base SPFs. If calibration factor is now computed using the HSM approach it can be seen that the calibration factor will be one. This implies that the base SPFs predicted by HSM are correct and it does not require any calibration. However, it is clearly seen that in this hypothetical case, the method will lead to erroneous results. This is one limitation of the current method of calibration, which can be avoided using the method proposed in this research.

Table 1 Hypothetical Scenario for Calibration Factor

	ECF	Predicted
Site1	1	5
Site2	2	4
Site3	3	3
Site4	4	2
Site5	5	1

Several states have initiated the implementation of HSM methods. During the processes, calibrating HSM base SPFs and developing new state specific SPFs have been studied extensively (Hauer, 1997; Hauer, 2004; Lord, Washington, & Ivan, 2005; Garber, Haas, &

Gosse, 2010; Tegge, Jo, & Ouyang, 2010; Sun, Margri, Shiraji, Gillella, & Li, 2011; Saito, Brimley, & Schultz, 2011). States like Utah and North Carolina have tried to calibrate the HSM base SPF using method recommended by HSM and compared its performance with state specific SPFs. They found that the calibrated models were inferior when compared to the newly developed state-specific SPFs. All of the studies in the literature that have estimated calibration factors have adopted the HSM-recommend method. The HSM-recommended method is very straightforward and easy to apply. However, the approach may not be exactly consistent with the recommended assumptions on the distribution of crash data, and may lead to suboptimal calibration factors.

HSM currently uses a univariate negative binomial (NB) regression model for predicting crashes. The NB regression performs well with the over-dispersed data sets; however, it can give misleading standard errors forunder-dispersed data(Oh, Washington, & Nam, 2006; Lord, Guikema, & Geedipally, 2008). This can result in falsely classifying a variable as having a significant effect of crashes. If these models are used for evaluating the effect of a particular treatment, the results can lead to no improvement of the roadway safety. To overcome these limitations, there is a need to identify a method that can accommodate under-dispersion. The univariate models published in HSM can be used to estimate the expected number of crashes by severities. However, these models cannot account for the interdependencies that might exist between crashes of different levels of severity for a specific roadway segment.

Finally, the current literature lacks the crash prediction models for bridges. There have been few studies that developed SPFs for freeways and other facilities not included in the 1st edition of HSM. However, there have been no efforts to identify the effects of bridges on roadway safety. Bridges are an integral part of the roadway infrastructure, but the majority of

bridge research efforts have related to bridge structural safety. There are very few to no studies on the traffic safety performance on the bridges. This study therefore develops a Multivariate Conway Maxwell Poisson (MVCMP) model for different crash severities on bridges. The study then compares the performance of proposed model with several univariate and multivariate Poisson lognormal models.

1.2 Research Objectives

The primary focus of this research is to create robust scientific models that help transportation agencies make well-informed decisions while planning new transportation facilities or improving the existing ones. The procedures provided in this research are in addition to the HSM and will result in the successful implementation of an HSM framework in the State of Alabama. The research is also useful to other states considering crash severity studies. Each of the specific objectives of this dissertation comprises an individual journal article. The specific objectives are discussed in the following sections and the three journal articles are presented as Chapter 2 – 4 of the dissertation. Chapter 5 then summarizes the interrelationships among the research objectives and the overall contribution of the dissertation research as a single body of work.

The first objective, described in Chapter 2, aims to contribute substantively to the growing stream of literature on application and implementation of the HSM. Specifically, the method of calibrating the base SPF models given in the HSM are examined, and a novel scientific method based on parametric distribution is proposed for calibrating the base SPFs given in HSM. The performance of the proposed new calibration method is evaluated by

comparing its prediction abilities with state-specific SPFs on two lane rural roads and four-lane divided highway crash data sets.

The second objective (Chapter 3) formulates MVCMP distribution to model traffic crashes at different level of severities simultaneously to account for possible correlations. The study presents statistical properties of the distribution along with an algorithm to estimate the parameter of the model using the Bayesian paradigm. The proposed method is very flexible and can be used with both over- and under-dispersed count data sets.

The third objective of the dissertation, presented in Chapter 4, is twofold. Firstly, it extends the scope of the HSM by developing SPFs for crash severities on bridge sections. This study is the first of its kind in considering bridges as separate entities from regular roadway facilities. As bridges have different physical and operational characteristics, this is considered an important addition to the literature. Secondly, it applies the MVCMP distribution in an actual field context and compares its performance with other widely used univariate and multivariate techniques such as univariate Conway Maxwell-Poisson distribution, univariate Poisson distribution and multivariate Poisson-lognormal distribution. The objective is achieved by applying all the aforementioned models for the crash severity analysis of the two lane rural roads and bridges in Alabama.

Chapter 2.

CALIBRATION OF SAFETY PERFORMANCE FUNCTIONS AND DEVELOPMENT OF NEW MODELS FOR TWO-LANE RURAL ROADS AND FOUR-LANE DIVIDED HIGHWAYS

2.1 Introduction

A detailed literature review indicates that the implementation of the HSM have been started by several states. The Highway Safety Manual (HSM) crash prediction modeling consists of three main components; safety performance functions, crash modification factors and calibration factor. There has been extensive research on the techniques used for developing Safety Performance Function (SPF): negative binomial, Poisson, zero truncated models, multivariate count models etc. The AASHTO has created the Crash Modification Factor (CMF) clearing house along with star rating for different crash modification factors. This helps in knowing the quality of the CMF being used. However, there is no effort in literature to check the validity of the calibration method. All the studies in literature either calibrate the models as per the HSM suggested method or have developed new SPFs (Saito, Brimley, & Schultz, 2011; Sun, Margri, Shiraji, Gillella, & Li, 2011; Garber, Haas, & Gosse, 2010; Srinivasan & Carter, 2011; Srinivasan, Haas, Dhakar, Hormel, Torbic, & Harwood, 2011). Since there is no research done to bridge the gap between calibrating the SPF and developing state specific SPFs, this research is proposing a new method of calibration based on maximizing the likelihood. The benefit of this approach is in the reduced resources required compared to developing the new SPFs and can provide a better estimation than using the existing calibration method.

2.2 Current Practice of Calibrating Safety Performance Function

The first edition includes three facilities types: Two-lane Two Way Rural Roads (TLTWR), Multilane Rural Highways and Urban and Suburban arterials. This study will include two-lane rural roads and four-lane divided highways. This research is laying the groundwork for implementing the HSM procedure in the state of Alabama. This research calibrates and develops new SPFs for two facilities types and the same methodology can be extended for other roadway types.

The base SPFs given in HSM are obtained using data from few states. This data set might not represent the entire country. The driver behavior between different states in North and South could be very different. Such demographic and jurisdictional differences between different regions need to be accommodated with the base SPF. It is therefore recommended by HSM to calibrate HSM models through calibration factors or develop new SPFs specific to local jurisdictions before actual implementation(HSM, 2010). Several states have calibrated base SPF given in HSM or developed state specific SPFs. A brief description of those studies is given below.

Xie et al.(2011) calibrated the base SPF for total crashes for the state of Oregon. The researchers estimated the calibration factors for different kinds of road segments as well as intersections. The sites were selected randomly such that the number of crashes for each site exceeded the minimum requirement of HSM. For every facility type, the homogeneous sites were created by dividing road segments into approximately 2 miles long sites. The researchers used wide variety of resources such as crash reports, video logs; Google maps, ODOT databases etc to obtain most of the required information. Information such as minor street AADT not

available from any data source was estimated using AADT prediction models. Two different calibration factors were computed. Firstly, a calibration factor for every year is estimated and the average of it is used as final calibration factor. Secondly, the calibration factor is computed using three years of data together. It was concluded that the values obtained from both the methods were almost similar. The report recommended that the HSM calibration procedure for estimating total crashes for Oregon State worked well, however, for estimating crash severities, local calibration factors specific to the state was needed. The state also calibrated the base SPFs for intersections. The intersections were first classified into different categories as defined by HSM and then a random selection was carried out to select the sites. It was observed that the calibration factors for both the segments and intersections were lower than one, indicating an over estimation by the HSM models. The probable reason for this over estimation was attributed to the crash reporting system in the state. The state of Oregon relies on self-reporting for PDO crashes under $1,500 resulting in lower number of crashes reported compare to other states.

Saito, Brimley, & Schultz (2011) conducted a study to calibrate the two-lane two way rural roads for the Utah Department of transportation (UDOT) using the crash data from 2005 – 2007. They calibrated the base SPF model for the state along with developing state specific models using Negative binomial and heirarchial bayesian model. The data required for the analysis was obtained from various sources such as Roadview, Google Earth, UDOT traffic tables, the UDOT crash database, and a UDOT construction project database for selecting road segments and obtaining the necessary facility and crash data. The final data set used for the analysis consisted of 157 sites averaging around 0.9 mi. All the segments used in the analysis were selected randomly and were on state or federal highway. The best model consisted of AADT, segment length, speed limit and percentage of multiple unit trucks as significant factors.

The comparison of the calibrated models with their state specific SPFs showed that there exisits some differences between the expected crashes in Utah and those predicted by HSM. The study recommended that the calibrated SPF performed better for the state than the proposed state specific SPFs, however, the state specific models required less data and the results were comparable.

The researchers at University of Florida undertook the calibration of the base SPFs for the Florida Department of Transportation. The research conducted by Srinivasan, et.al., (2011) calibrated the base SPFs for two-lane two way rural roads, rural multilane highways and urban and suburban arterial roads. Four years of data from 2005-08 was used in the analysis. The data used in the analysis was only from state highway system, since the data quality was not good for city and county roads. The geometric data used in the analysis was obtained from Florida Roadway Characteristics Inventory (RCI) and the crash data was obtained from crash reports. The data set consisted of 21 attributes of which 15 were used for creating the homogenous segments. The homogeneous sites were created such that certain attributes did not change in those sites. The shortest site for rural roads were 0.1 miles long while for the urban and suburban roads were 0.04 miles long. For the missing values of the road side hazard rating and the number of driveways, the researchers assumed certain values based on a their judgement. Sensitivity analysis was then conducted to examine the effect of the assumtion on the HSM crash estimation procedure. They concluded that assuming a default value for this variables would have not made any significant changes in their estimation. The state wide calibration factors were found to be under predicting the crashes and thus district specific calibration factors were recommended for rural two-lane two-way road segments. The calibration of Intersections was also conducted, however, it was critically impacted by the data issues.

Another leading state in the implementation of the HSM is North Carolina. They calibrated the base SPF for most of the road facility types and intersection given in HSM. Along with calibration, they also developed their own state specific SPFs for the different facilities with AADT as the only variable. Only two lane rural road segments had added variables such as shoulder width, type and terrain. The data used in the analysis was obtained from aerial photographs, GIS files, roadway inventories and other databases. Homogeneous segments were not created as recommended by HSM, however, the segments were created whenever the road had an intersection. 250 feet of the road segment was removed from the analysis as it was considered as the intersection influence area. Negative Binomial regression was then used to develop the state specis SPFs for different facility types.

2.3 Existing Methods forDeveloping State-Specific SPFs

There has been vast amount of research done in past two decades to gain a better understanding of the factors that can explain the relation to the vehicle crashes (Lord & Mannering, 2010). The factors considered in the literature include roadway geometry, environmental and traffic conditions, highway user attributes and crash related information. Several approaches have been utilized in developing the relation between the crashes and the above-mentioned factors. The methods found in the literature can be divided into two major streams: univariate count data models and multivariate count models. A brief review of the existing literature is given below.

2.3.1 Univariate Count Data Models

Univariate models are the simplest form of statistical models in which the analysis is carried out in terms of single variable. The relationship is developed between a single response variable and one or many predictor variables. The univariate models can be used for predicting the total crashes, however it might not be a very good method for analyzing crash severities, since there can be correlation between them, which cannot be estimated using these models. A brief overview of some of the most widely used models for crash predictions is given below.

2.3.2 Poisson Regression

Poisson regression is the most widely used regression for modeling the count data (Jovanis & Chang, 1986; Joshua & Garber, 1990; Jones, Janssen, & Mannering, 1991; Miaou & Lum, 1993). Since the crash data can never be negative, linear regression (requires continuous response variable) is not an option. Thus, Poisson regression has become the starting point for modeling the count data. As mentioned in the previous section, Poisson distribution has a strong assumption of mean equal to variance, which is generally not exhibited by the crash data. Hence there have been several extensions of the simple Poisson regression to accommodate the dispersion present in the crash data.

2.3.3 Negative Binomial Regression

Negative binomial regression or Poisson-gamma regression is the simple extension of the regular Poisson regression. The equi-dispersion assumption of the Poisson distribution is mostly inappropriate for crash data sets as they are found to be over or under-dispersed. Therefore, an error structure is added to the Poisson mean parameter assumed to follow gamma distribution with mean 1 and dispersion parameterα. The addition of the gamma error allows the variance of

the Negative binomial distribution to differ from its mean. As the parameter α approaches zero, the NB becomes a Poisson regression model. The Negative binomial regression has been one of the most widely used methods for predicting crashes (Hauer & Hakkert, 1988; Persaud, 1994; Mountain, Fawaz, & Jarrett, 1996; Johansson, 1996; Vogt & Bared, 1998; Vogt, 1999; Miaou, 2001). One main advantage of this method is that it is easy to account for over-dispersion. However, this regression method performs poorly when the data is under-dispersed or it has low sample mean and sample size (Lord & Mannering, 2010).

2.3.4 Poisson-Lognormal Model

This model is created as an alternative to the NB model. In this model, the error structure of the Poisson parameter is assumed to follow lognormal distribution instead of gamma in NB. This assumption gives more flexibility to the model. It is more flexible in accounting for over-dispersion, however, it over-dispersion parameter cannot be estimated, the method cannot handle under-dispersion and it performs poorly when the sample size is small and data has low mean. Some of the recent studies have used this method for analyzing crash data (Miaou, Song, & Mallick, 2003; Lord & Miranda-Moren0, 2008; Aguero-Valverde & Jovanis, 2008).

2.3.5 Zero Inflated Regression Model

Zero inflated models are used when the data has missing values or has far too many zeros than one would expect in the regular processes like Poisson. Zero inflated models are used for both the Poisson and NB regression models. These are called as the dual-state models, since it is assumed that the zero in the data sets come from two different processes. For example, consider a data set created by taking a survey of the class who went to a fishing trip recently. The question of interest here is how many fishes did individual student catch. As one would expect

there will be lot of zeros, however, this zeros will be from two different groups: a group of student who could not catch a fish and a group of student who did not go for the fishing trip. Lord Washington, & Ivan, (2005) argued that this would mean that there are some sites which can never have a crash and some sites where no crashes were recorded during the study period. This assumtion does not reflect the true crash data generating process. Some of the studies that have used this method are (Shankar, Milton, & Mannering, 1997; Lee & Mannering, 2002; Kumara & Chin, 2003)

2.3.6 Gamma Model

Gamma model is very rarely used in the transportation safety research. Oh, Washington, & Nam, (2006) used this model to predict the number of crashes occuring at the railway-highway at grade crossing. The big advantage of this model is that it can handle under-dispersed data. However, this model is also a dual state model and has same limitation as Zero Inflated Models discussed above.

2.3.7 Random-Parameter Model

In Random parameter regression models, the coefficients are allowed to vary across each observation. The Random parameter models account for the unobserved heterogeneity from one observation to another (Milton, shankar, & Mannering, 2008). The random parameter model have been used with both the Poisson regression and the NB regression recently (Anastasopoulos & Mannering, 2009; El-Basyouny & sayed, Accident Prediction Models with Random Corridor Parameter, 2009). The final model will provide a very good statistical fit; however, these models should be used cautiously, since the results may not be transferable to the other data sets because the results are observation specific.

2.3.8 Conway Maxwell Poisson Model

Conway Maxwell Poisson distribution was proposed by Conway and Maxwell in 1962 for modeling the queue and service rates. This distribution is an generalization of the Poisson distribution. The distribution was rarely used in the literature, after its introduction by Conway and Maxwell. In 2005, Shmueli et al. derived the statistical properties of this distribution. It is a unique distribution with Geometric, Bernoulli and Poisson distribution as its special cases. The main advantage of this distribution over other models studied above is its ability to handle both the over-dispersed and under-dispersed data. Lord et al. (2008) applied this distribution to the highway safety research and found that it performs comparable to the NB regression for the over-dispersed data and provides better fit for the under-dispersed data. It was also noted that the model does not perform well when the mean of the sample is low or the sample size is very small.

This section will describe the method used to calibrate the SPFs, the proposed method of computing the calibration factor and then the theory of negative binomial regression used to predict the state specific SPFs. The following section describes the data source and the method of creating homogeneous sites for the analysis. The same data source and technique of creating homogeneous sites will be used for both of the facility types to be studied.

2.4 Methodologies

The following section describes the different methods used for calibration the HSM base SPF, the new proposed calibration method and the negative binomial regression used for developing state specific SPFs.

2.4.1 Theoretical Foundations

2.4.1.1 Negative Binomial

Several statistical regression methods have been used for estimating the safety performance of a roadway system in terms of crash frequencies. Since crashes are rare and random events, it is assumed that the response variable in the regression models, the crash frequency of a particular road section, is a random variable following a certain distribution. The mean of the random response variable is assumed to be a function of characteristics of the roadway as well as the traffic. The goal of developing SPFs is to quantify such a relationship between the expected crash frequency and various explanatory variables. The variables in the model and their coefficients will be determined statistically such that the observed crash counts would be the most likely to realize or in other words, the probability of the observed crash counts to occur is the highest. This concept is called maximum likelihood estimation.

The most common methods used by transportation safety researchers are Poisson and Poisson-Gamma regression models (Lord, Washington, & Ivan, 2005; Lord & Bonneson, 2006; Lord & Park, 2008). Poisson regression assumes that the response variable follows a Poisson distribution with a mean of λ, which is a function to be determined from the data. This assumption on the random response variable implies that its mean is equal to its variance. However, the crash counts are usually over-dispersed with the mean smaller than the variance (Mitra & Washington, 2007; Lord, Guikema, & Geedipally, 2008; Lord & Mannering, 2010). To accommodate over-dispersed data, Poisson-Gamma regression is commonly adopted. Poisson-Gamma regression is an extension of the simple Poisson regression, and is a special type of NB models (Hilbe, 2011). By introducing a random term ε into the Poisson mean λ, the approach essentially assumes the crash counts follow a NB distribution. While other types of

NB models can be used for modeling count data, Poisson-Gamma parameterization is recommended by the HSM and is widely used in the transportation literature. This study will adopt Poisson-Gamma-type NB regression to estimate SPFs in terms of crash frequencies. From this point on, we will use Poisson-Gamma regression and NB regression interchangeably. Note that Poisson-Gamma parameterization can be further specified in multiple ways. NB1 and NB2 parameterizations (Hilbe, 2011) would both lead to a Poisson-Gamma-type distribution, with the only difference being the variance of the random response variable. This study will apply both NB1 and NB2 parameterizations,

Mathematically, for each homogeneous road section i, the crash counts at site i, Y_i, is first described as a Poisson random variable with a mean of λ_i in a Poisson-Gamma model. The probability density function (PDF) of Y_i defines the probability of the random variable Y_i equal to a given realized value y_i, and is given as follows.

$$f_{Y_i}(y_i;\lambda_i) = \frac{\exp(-\lambda_i)\,(\lambda_i)^{y_i}}{y_i!}$$

The Poisson-Gamma regression further assumes the Poisson mean λ_i takes the following form,

$$\lambda_i = \exp\left(\beta_0 + \sum\nolimits_j x_{ij}\beta_j + \varepsilon_i\right) = \exp\left(\beta_0 + \sum\nolimits_j x_{ij}\beta_j\right)\exp(\varepsilon_i)$$

where x_{ij} is the j^{th} explanatory variable for the mean of crash counts at site i, β_j is the corresponding coefficients, β_0 is a constant to capture other unknown or unobserved factors, and ε is a random term following Gamma distribution with a mean of one and a variance of $1/\alpha$. In this paper, we will call α the dispersion parameter. The βs and the α are parameters to be estimated. Let $\mu_i = \exp\left(\beta_0 + \sum_j x_{ij}\beta_j\right)$and $\nu_i = \exp(\varepsilon_i)$, the Poisson mean with Gamma heterogeneity can be written as $\lambda_i = \mu_i \nu_i$. With these assumptions, it can be shown that the

random response variable Y_i follows a negative binomial distribution with a mean of μ_i and a variance of $\mu_i + \alpha\mu_i^2$. The PDF of this distribution is given below.

$$f_{Y_i}(y_i; \mu_i, \alpha) = \begin{pmatrix} y_i + \frac{1}{\alpha} - 1 \\ \frac{1}{\alpha} - 1 \end{pmatrix} \left(\frac{1}{1+\alpha\mu_i}\right)^{1/\alpha} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}$$

With the PDF of crash counts established as a function of a set of parameters α and β, the goal of model estimation is to determine the form of μ_i and the values of the parameters so that the observed crash counts are the most likely to realize. Also note that from the equation of μ_i, it is clear that the mean of the negative binomial random variable Y_i follows a log-linear relationship with the explanatory variables: $\ln(\mu_i) = \beta_0 + \sum_j x_{ij}\beta_j$.

2.4.1.2 Goodness-of-Fit Measures

The goodness-of-fit of a statistical model describes how well the model fits the data. It describes the discrepancies between the observed value and the predicted or fitted value. In this study we have used log likelihood (LL) value and Akailke's information criterion (AIC) to evaluate the suitability of the models. This is because the models are estimated based on the concept of likelihood maximization, and LL and AIC are consistent with this approach. Deviance statistic is one of the goodness-of-fit measures used in the literature, but is not used in this study, since it has been observed that the deviance statistic fails to identify ill fitted models (Hilbe, 2011). As mentioned earlier, the likelihood is the occurrence probability that the data observed will actually be realized under the given parameter estimates. The higher the LL value is, the better the model. AIC on the other hand describes the tradeoff between bias and variance. AIC is related to the LL, and is computed based on the equation given below

$$AIC = -2 \times LL + 2 \times (Numberof Parameters)$$

A lower value of AIC indicates a better model. Because the number of parameters is a factor affecting the AIC, it effectively discourages over-fitting of data by penalizing addition of parameters.

2.4.1.3 Model Validation

To evaluate the prediction capability of the developed calibration factors and new state-specific SPFs, the models are run on a validation data set sampled from the original population. The validation data set is different from the data set used to estimate the models. Besides LL and AIC, three additional measures are also considered. They are mean absolute deviation (MAD), mean square prediction error (MSPE), and mean prediction bias (MPB). They have been commonly used in the literature (Lord, Guikema, & Geedipally, 2008; Washington, Persaud, Lyon, & Oh, 2005). One can argue that these measures are not comparing apples to apples, as they are essentially comparing the estimated mean of crash counts to the realization (observed) of random crash events. However, they might not be as statistically meaningful as LL and AIC, they are easier to compute. Moreover, they measure how far away the realizations are from the mean, which is a reasonable alternative to probabilistic measures such as LL and AIC.

Mean Absolute Deviation (MAD)

The MAD is suggested by (Washington, Persaud, Lyon, & Oh, 2005) as a goodness-of-fit measure for SPF. MAD is the ratio of the sum of absoulute difference between observed crash counts and predicted mean values to the number of sites n.

$$\boldsymbol{MAD} = \frac{(\sum_{\boldsymbol{i}=\boldsymbol{1}}^{\boldsymbol{n}} |\widehat{\boldsymbol{\mu}}_{\boldsymbol{i}} - \boldsymbol{y}_{\boldsymbol{i}}|)}{\boldsymbol{n}}$$

MAD gives the average magnitude of variability of prediction. Smaller values are preferred to larger values.

Mean Squared Prediction Error (MSPE)

MSPE is defined as the sum of the square of the difference between observed crash counts and predicted mean values divided by the number of sites. This statistic is used to assess the error associated with a validation or external data set. A lower value for MSPE implies a better model.

$$\boldsymbol{MPSE} = \frac{(\sum_{i=1}^{n}(\hat{\mu}_i - y_i)^2)}{n}$$

Mean Prediction Bias (MPB)

As MAD, MPB is also suggested by Washington et. al. (2005). It is defined as the sum of predicted mean values minus the observed crash counts, divided by the total number of sites considered. This statistic provides a measure of the magnitude and direction of the average model bias. Unlike MAD, MPB can be positive or negative and it is given by the following equation.

$$\boldsymbol{MPB} = \frac{(\sum_{i=1}^{n}(\hat{\mu}_i - y_i))}{n}$$

A positive value of MPB indicates that the SPF is overestimating the number of crashes, whereas negative value implies concluding a site to be safer than they actually are.

Log Likelihood (LL) and Akailke's Information Criterion (AIC)

When estimating the NB regression models using off-the-shelf software packages, the LL value is automatically reported along with the dispersion parameter α. When it comes to model validation, the LL value needs to be calculated manually. For each data point in the validation set, the probability that the random response variable (number of crashes) equals the observed value will be calculated following equation 3, where the estimated values of βs and α will be applied. LL value will then be computed as the natural log of the product of all these probability values. Once the LL value is calculated, the AIC value can be computed following equation 3.

2.4.2 Calibration Methods

In this study, two approaches are used to estimate the calibration factor. The first approach treats calibration factor estimation as a special case of SPF development, while the second approach is the HSM recommended method. A brief explanation of both methods is given below.

2.4.2.1 HSM Recommended Method

The HSM-recommended method for estimating the calibration factor is very straightforward. After the expected crash frequencies are calculated from the HSM base SPFs, the calibration factor is simply computed using the following formula.

$$C_r = \frac{\sum_i \text{Observed number of crashes for site i}}{\sum_i \text{Predicted expected number of crashes for site i}}$$

After computing the calibration factor this study proposes to develop new state specific SPFs and compare the performance of the state specific SPFs with the calibrated SPFs. Negative

Binomial regression recommended by the HSM is used for the analysis. The calibrated SPF resulting from the HSM-recommended method will be denoted as Model C2.

2.4.2.2 Calibration Factor Estimation as a Special Case of SPF Estimation

Calibration factor is a multiplicative factor used to adjust the base SPF in order to better fit local data. When a base SPF is adjusted by a calibration factorC_r, the new regression model becomes

$$\mu_i = \exp\left(\beta_{i0} + \sum_j x_{ij}\beta_{ij}\right) \cdot C_r \quad (1)$$

Assuming that C_r is independent of x_{ij} this equation can be rearranged as

$$\mu_i = \exp\left(\beta_{i0} + \sum_j x_{ij}\beta_{ij} + \ln(C_r)\right) \quad (2)$$

Note that the first two terms in the parenthesis in the above equation are both known, as they are the natural log of the base SPF. This model can be further simplified as

$$\mu_i = \exp(\ln(C_r) + \ln(\text{Base SPFfor site } i))$$

Equation (1) is in fact a NB regression model with one constant ($\ln(C_r)$) and one explanatory variable (log of the base SPF for site i) whose coefficient is fixed to one. To estimate the calibration factor C_r is essentially to estimate the intercept $\ln(C_r)$ in the NB regression model. Therefore, this method can be viewed as a special case of SPF estimation. The calibrated SPF resulting from this approach will be denoted as Model C1.

2.4.3 Safety Performance Function Development

It is recommended by in the HSM to develop state-specific SPFs, if data is available. After reviewing the efforts of SPF development made by researchers all over the world, four

different model specifications will be investigated for the Alabama data. All of the four models are applied to both TLTWRR and FLDH data.

2.4.3.1 Model 1

The first model keeps the form of HSM base SPF in view of its simplicity and its minimal requirements on data availability:

$$\hat{\mu}_i = \beta_0 AADT_i^{\beta_1} SL_i$$

In the above model, $\hat{\mu}_i$is the estimated expected number of crashes per year for site i, $AADT_i$ and SL_i are the annual average daily traffic and segment length for site i respectively, and β_0 and β_1 are parameters to be estimated. Note that the coefficient for segment length is fixed in the model. This is because AADT is always found to have a significant effect on crash prediction models, but segment length often has less impact (Council & Stewart, 1999).

2.4.3.2 Model 2

Model 2 takes the form of a SPF developed in UK (Mountain, Fawaz, & Jarrett, 1996) for TLTWRR. The functional form is shown below.

$$\hat{\mu}_i = \beta_0 \, AADT_i^{\beta_1} SL_i^{\beta_2} \, exp\left(\frac{bn_i}{SL_i}\right)$$

In the model, n_i is the number of minor junctions within site i, and β_0, β_1, β_2, and b are parameters to be estimated. In our study, n_i is considered as zero, because we have considered homogeneous roadway segments with no intersections. Although there might be driveways present within the roadway segments, they are not considered in this study.

2.4.3.3 Model 3

Model 3 takes the form developed by Connecticut transportation Institute (Milton & Mannering, The Relationship Between Highway Geometrics, Traffic Related Elements and Motor Vehicle Accidents, 1996)

$$\hat{\mu}_i = exp(\beta_0 + \beta_1 DY_i + \beta_2 lnAADT_i + \beta_3 \ln SL_i + \beta_4 LW_i + \beta_5 S_i)$$

In the above model, $DY_i$is the dummy variable to account for effects of the year on the intercept, LW_i is the lane width for site i, and S_i the speed limit (mph).

2.4.3.4 Model 4

Model 4 was originally developed by Council & Stewart (Council & Stewart, 1999) based on data from North Carolina, Washington, Minnesota and California. In this model, shoulder width SW_i is further introduced. The functional form of the model is given below.

$$\hat{\mu}_i = exp(\beta_0 + \beta_1 SW_i + \beta_2 LW_i) AADT_i^{\beta_3} SL_i$$

2.5 Case Studies

2.5.1 Data Source

The data required for the analysis is obtained from the Critical Analysis Reporting Environment (CARE) maintained by the Center for Advanced Public Safety at the University of Alabama. CARE receives the roadway inventory data from Alabama Department of Transportation and the crash data from the Alabama Department of Public Safety. The original data is in the form of individual crash record. This data is converted into a roadway-oriented data set where each record represents a 0.01-mile of roadway segment. The final data set consists of a total of 1,125,378 records or 1,126 miles of roads of Alabama. 91 variables are generated for every record based on the HSM and CMF Clearinghouse documentations (Federal

Highway Administration, 2008). These entries can be classified into 7 categories, including basic information, basic geometry and design, horizontal curve, vertical curve, intersection, traffic, and crash history. Note that each record in the dataset includes multiple traffic- and crash-related variables for years 2006 – 2009. While geometry- and design- related variables do not change, AADT and crash records for the same 0.01-mile road segment may vary from year to year.

2.5.2 Creating Homogeneous Sites

To correctly apply HSM methods, roadways need to be divided into homogeneous sections or sites. Within each site, the major geometric and traffic characteristics of the road should remain the same. The major characteristics considered in this study include facility type, number of lanes, lane width, shoulder width, median width, speed limit, and AADT. A new segment begins when any of the above characteristics changes. A brief graphical representation of creating homogeneous segments is shown in Figure 1**Error! Reference source not found.**.

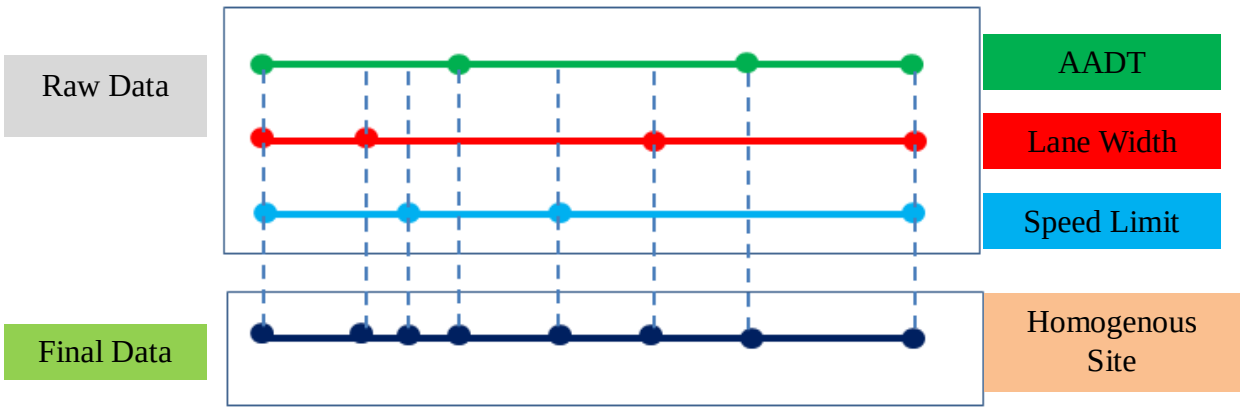


Figure 1 Creating Homogeneous sites

Consider two lane rural road raw data with 10 mile long segment. Each line in the Figure 1 indicates the length of the road. Every dot on the line indicates when a particular characteristic of road changes. The homogeneous sites are developed such that the first change in any variable from a selected set of variables begins a new site. From the above figure, it is evident that the lane width is the variable that changes first and hence a new site is started from that point onwards until the first change in any other variable. In the give case, the third site is created when the speed limit changes. This process is automated using Microsoft Excel Macros.

Horizontal and vertical curve related variables (curve present, radius, and grade) are not considered in generating homogeneous sites in this study due to data accuracy issues. Since the AADT for the same 0.01-mile road segment may vary from year to year, it is possible that the homogeneous sites created may also differ from one year to another. In view of this, this study considers a homogeneous site for a certain year as a data point, instead of using the same site mapping for all four years in the data set. This also increases the number of usable samples in the analysis and is likely to increase the reliability of the estimated SPFs and calibration factors. More specifically, the homogeneous sites are created for each year based on the geometric and traffic data of that year.

2.5.3 Site Selection Methods

Due to the large amount of data, a sample of the identified roadway sites needs to be selected for the analysis. These sample sites should be selected carefully to represent a variety of properties in terms of geography, socioeconomics, socio-demographics and jurisdictions. The selected sites should also represent most of the possible values for all the variables. Based on these considerations, a staged sampling approach is adopted in this task. In the first stage, a

random sampling is performed. The sampling results are then be compared with the original dataset in terms of value distributions of selected variables. If the distributions differ from those of the population significantly, a stratified sampling method (Richardson, Ampt, & Meyburg, 1995)is applied in the second stage. Subsets of the entire population (or strata) are created based on the value groups of selected variables. A random sample is then drawn from those groups that are underrepresented after the first stage sampling. This method ensures that some relatively small subgroups with rare characteristics are represented in the sample.

The following section gives the performance of the calibrated models and the state specific SPFs. This section compares calibration factors obtained from both the calibration methods. The next part of the section gives the coefficients of the four models considered for the analysis. All the results obtained are validated on the data that is not used for developing the model.

2.5.4 Calibration Factor Estimation

The two approaches to calibration factor estimation (discussed in section 2.7.1.1 and 2.7.1.2) are applied to the same prediction data set for each type of road facility. Two software packages are used for this analysis. Software Nlogit is used to implement the first approach, where the calibration factor is estimated as a special case of NB regression model. The resulting model is denoted as C1. Software SPSS is used to implement the simple HSM-recommended method. The resulting model is denoted as C2. To check the prediction capability of the calculated calibration factors, both models C1 and C2 are applied to the validation data set and are evaluated according to the proposed performance measures.

Using the first approach (C1), the calibration factor for TLTWRR is estimated as 1.522. The HSM-recommended approach (C2) leads to a calibration factor of 1.392. The performances of these two calibration factors will be reported later in the "Validation and Comparison" subsection. For FLDH, the first approach (C1) results in a calibration factor of 1.863, and the HSM-recommended method (C2) leads to a calibration factor of 1.103.

2.5.5 Safety Performance Function Estimation

This section presents the modeling results of recalibrated HSM model and three other models studied. The parameters of the model are estimated using SPSS. All the parameters included in the model are statistically significant.

2.5.5.1 Two-Lane Two-Way Rural Roads

Table 2 summarizes the parameter estimates together with the standard deviation of these estimates (presented in parenthesis) and the goodness-of-fit measures for all of the four models discussed earlier for TLTWRR.

All of the models make intuitive sense, since all of the coefficients have the same signs across all four models as they are expected. The log likelihood value for Model 3 is the highest comparing to other models, indicating Model 3 fits the data better. When we take a closer look at Model 3, it can be observed that the model has 6 parameters while other models only have 2 – 4 parameters. This explains the superior performance of Model 3. However, to avoid over-fitting, the model is further compared with others based on the AIC value, as AIC statistic accounts for the difference in number of parameters. Again, Model 3 has the lowest AIC value among all the models, which is consistent with the findings based on LL values. In view of this,

it can be concluded that the model 3 is best model among all the choices. It has lowest values for both the goodness of fit measure.

Table 2 Estimated Parameters for Two Lane Rural Roads

Variables	Model 1	Model 2	Model 3	Model 4
Intercept	-7.135(0.2711)	-6.295 (0.2727)	-7.971 (0.3610)	-7.020 (0.2828)
AADT	0.916(0.0332)	0.786 (0.0345)	0.775 (0.0351)	0.9279 (0.0345)
Segment Length	1	0.747(0.0261)	0.694 (0.0268)	1.00
Lane Width		-	0.0552 (0.0415)	0.0507 (0.0422)
Speed Limit		-	0.1641 (0.0229)	-
Dummy Year		-	0.0388 (0.0213)	-
Shoulder Width		-	-	-0.0961 (0.01465)
Dispersion Parameter	1.09	1.0189	0.9814	1.0593
Log Likelihood	-5315.058	-5092.591	-5069.970	-5116.921
AIC	1.715	1.7014	1.694	1.709

The values of the coefficients reveal the nature of the correlation between the variables. The positive coefficients for the AADT and lane width variables indicate that an increase in traffic volume and lane width would lead to increased mean crash counts, which is intuitive. The dummy variable for year has a negative coefficient, indicating there is a declining trend of mean crash frequencies over the years. One plausible reason for this trend could be the economic recession during 2008 and 2009, which resulted in reduced number of miles driven per capita and lower crash frequencies.

2.5.5.2 Four-Lane Divided Highways

The results for SPF development for FLDH are presented in Table 3. The positive coefficients for all the variables except lane width indicate that the expected crash frequencies

would increase as the values of these variables increase. Most of the other observations are similar to those of the TLTWRR case. However, it is worth mentioning that the signs of the lane width variable in Model 3 and Model 4 are different. This is an expected result. A few untested hypotheses for these observations are offered here: 1) There might be some correlations among the variables. Comparing the functional form of Model 3 and Model 4, it is likely that the correlations are among the lane width, the speed limit, and the year variable. In fact, correlations between lane width and speed limit are quite intuitive as the speed limit is usually higher for roads with better conditions. 2) Note that the estimated coefficient is very close to zero in both Model 3 and Model 4. This implies that although the signs are contrasting, the predicted mean crash frequencies are not significantly affected by the lane width variable.

Comparison of the LL and the AIC values indicates that Model 3 outperforms all the other models in predicting the expected crash frequencies. Moreover, Model 3 also has a smaller dispersion parameter comparing to the others.

Table 3 Estimated Parameters for Four Lane Divided Rural Highways

Variables	Model 1	Model 2	Model 3	Model 4
Intercept	-7.706 (0.4927)	-6.166 (0.6862)	-7.784 (0.6019)	-8.022 (0.5543)
AADT	0.974 (0.052)	0.748 (0.0530)	0.759 (0.0562)	0.991 (0.0524)
Segment Length	1	0.3562(0.0191)	0.354 (0.0213)	1
Lane Width		-	0.099 (0.0637)	-0.0528 (0.0559)
Speed Limit		-	0.049 (0.0165)	-
Dummy Year		-	0.081 (0.0280)	-
Shoulder Width		-	-	0.0520 (0.0158)
Dispersion Parameter	2.854	2.085	2.077	2.844
Log-Likelihood	-5285.61	-4977.22	-4966.408	-5282.298
AIC	2.6423	2.488	2.484	2.641

2.5.6 Validation and Comparison

After estimating various SPF models based on the Alabama data, the prediction ability of each model is tested using the five performance measures discussed earlier: LL, AIC, MAD, MPB, and MSPE.

2.5.6.1 Two-Lane Two-Way Rural Roads

The data set used for model validation consists of 750 sites from each year, or 3000 sites in total. The performance of each model on the validation data set is shown in Table 4.

Table 4 Performance Measures for All Six SPF Models for TLTWRR

Performance Measure	Model 1	Model 2	Model 3	Model 4	Model C1	Model C2
MAD	0.538	0.540	0.525	0.534	0.538	0.523
MPB	0.049	0.021	0.008	0.053	0.048	0.011
MSPE	0.781	0.721	0.702	0.765	0.793	0.768
Log Likelihood	-2398.525	-2368.353	-2343.780	-2390.953	-2404.33	-2428.337
AIC	1.601	1.582	1.567	1.598	1.605	1.621

The values of MPB and MSPE for Model 3 are smaller comparing to all the other models, including models C1 and C2 (adjusted base SPFs with calibration factors). This indicates that Model 3 leads to lower prediction error and less bias in the prediction. However, its variation in the prediction accounted by MAD is slightly higher comparing to Model C2, which is the calibrated HSM SPF using HSM-recommended method. Overall, Model 3 is the best choice among the four newly developed state-specific SPFs for TLTWRR. When we compare the calibrated models C1 and C2, it can be observed that Model C1 outperforms Model C2 in LL and AIC values (two probabilistic measures), while Model C2 performs better in terms of other criterion (mean-related measures). This is expected, since the proposed calibration method (C1) can be viewed as a special case of NB regression, and aims to maximize the

likelihood value. On the other hand, the HSM-recommended method (C2) essentially tries to bring the estimated expected number of crashes closer to the observed values.

2.5.6.2 Four-Lane Divided Highways

The data set used for the validation consists of 500 sites from each year or 2000 sites in total. The performance of all six models on this data set is shown in Table 5.

Table 5 Performance Measures for All Six SPF Models for FLDH

Performance Measure	Model 1	Model 2	Model 3	Model 4	Model C1	Model C2
MAD	1.4194	1.3359	1.1120	1.6497	1.4336	1.0927
MPB	0.5378	0.4644	-0.0373	0.8681	0.5587	-0.0547
MSPE	6.2003	3.6749	3.4352	8.5468	6.3445	4.0633
Log Likelihood	-2661.815	-2531.789	-2475.594	-2658.836	-2735.772	-2901.719
AIC	1.777	1.691	1.655	1.77	1.826	1.936

From Table 5, it can be observed that unlike other models, the MPB for Model 3 and Model C2 is negative. This indicates that Model 3 and Model C2 have negative bias, though the magnitude of the bias is rather small. In terms of MPB, Model 3 is better than all the other models, since the absolute value of the bias is the smallest. The MSPE for Model 3 is also the smallest, which indicates smaller prediction error. The MAD for Model C2 is slightly better than Model 3 and much better than all the other models. Moreover, Model 3 is significantly better than other models in terms of LL value, and it also has the lowest AIC value. Overall, Model 3 seems to perform better than all the other models except in MAD. Similar to the TLTWRR case, Model C1 performs better than Model C2 in terms of LL and AIC, while Model C2 outperforms Model C1 in terms of MAD, MPB, and MSPE.

2.6 Conclusion

This study has evaluated the applicability of the HSM predictive methods for two facility types, namely two-lane two-way rural roads and four-lane divided highways, in the State of Alabama. It has also estimated calibration factors to adjust the HSM base SPFs and has developed new state-specific SPFs that quantify the relationship between the expected crash frequencies and various road and traffic characteristics using Alabama data. Three tasks are performed in this study. The first task is to calibrate the HSM base SPF using HSM-recommended method for Alabama conditions. In the second task, a new calibration method is proposed, which treats the estimation of calibration factors as a special case of NB regression. Finally, four different new SPFs are investigated and parameterized using NB regression techniques using Alabama data. The prediction capabilities of these models are tested using a validation data set that is different from the original estimation data set. Multiple performance measures, such as MAD, MPB, MSPE, LL and AIC, are considered.

The calibration factors derived from both approaches and for both types of facilities are greater than one, implying that the HSM base SPFs are underestimating the mean crash frequencies on TLTWRR and FLDH in Alabama. The mean crash frequencies predicted by the proposed new calibration method is higher than that of the HSM-recommended method for both facility types. The HSM-recommended calibration method seems to outperform the proposed new calibration method for these two types of facilities. However, one particular state-specific SPF is found to outperform all other models, including both calibrated models.

The best state-specific model includes a few variables that are not part of the HSM base SPF. Lane width, speed limit, and year are all found to have statistically significant (does not

necessarily mean large magnitude though) impact on the mean crash frequencies. This indicates that the relationship between the crash and road characteristics in Alabama is different from what the HSM base SPF describes. The study has identified some inconsistency in the resulting models. For example, the coefficient of the lane width variable is positive in Model 3 but is negative in Model 4 for FLDH. However, a more comprehensive study is required to understand the effect among different variables. Overall, this study lays an important foundation towards the implementation of HSM methods in the state of Alabama. It can help the transportation officials in Alabama to make informed decisions regarding road safety programs. Though this study uses Alabama data, the framework provided can be used by other states for over-dispersed crash data.

2.7 References

Ackaah, Williams. and Salifu, Mohammad. (2011). Crash Prediction Models for Two Lane Rural Highways in the Ashanti Region of Ghana. IATSS Research, 35, 34-40.

Aguero-Valverde, J., & Jovanis, P. (2008). Analysis of Road Crash Frequency with Spatial Models. Transportation Research Record, 55-63.

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

Caliendo, C., Guida, M. and Parisi, A. (2007). A Crash Prediction Models for Multilane Roads. Accident Analysis and Prevention, 39, 657-670.

Council, F., & Stewart, R. (1999). Safety Effects of the Conversion of Rural Two-Lane to Four-Lane Roadways Based on Cross-Sectional Models. Transportation Research Record: Journal of the Transportation Research Board(1665), 35-43.

El-Basyouny, K., & sayed, T. (2009). Accident Prediction Models with Random Corridor Parameter. Accident Analysis and Prevention, 41(5), 1118-1123.

Federal Highway Administration. (2008). Crash Modification Factor Clearing House. Retrieved July 23, 2011, from http://www.cmfclearinghouse.org

FHWA. (2012). Office of Safety. Retrieved February 2012, from FHWA Safety: http://safety.fhwa.dot.gov/facts_stats/

Garber, N., Haas, P., & Gosse, C. (2010). Development of Safety Performance Function for Two Lane Roads Maintained by Virginia Department of Transportation. Virginia Department of Transportation.

Greibe, B. (2003). Accident Prediction Models for Urban Roads. Accident Analysis and Prevention, 273-285.

Hauer, E. (1997). Observational Before-After Studies in Road Safety.

Hauer, E. (2004). Satistical Road safety Modelling. Transportation Research Record, 1897, 81-87.

Hauer, E., & Hakkert, A. (1988). Extent and some Implications of Incomplete Accident Reporting. Transportation Research Record(1185), 1-10.

Hilbe, J. (2011). Negative Binomial Regression. Cambridge, United Kingdom: Cambridge University Press.

HSM. (2010). Highway Safety Manual.

Johansson, P. (1996). Speed Limitation and Motorway Casualties: A time-series Count Data Regression Approach. Accident Analysis and Prevention, 28(1), 73-87.

Jones, B., Janssen, L., & Mannering, F. (1991). Analysis of the Frequency and Duration of Freeway Accidents in Seattle. Accident Analysis and Prevention, 23(2), 239-255.

Joshua, S., & Garber, N. (1990). Estimating truck Accident Rate and Involvement Using Linear and Poisson Regression Models. Transportation Planning and Technology, 15(1), 41-58.

Jovanis, P., & Chang, H. (1986). Modeling the Relationship of Accidents to Miles Travelled. Transportation Research Record, 1068, 42-51.

Kumara, S., & Chin, H. (2003). Modeling Accident Occurrence at Signalized Tee Intersections with Special Emphasis on Excess Zeros. Traffic Injury Prevention, 3(4), 53-57.

Lee, J., & Mannering, F. (2002). Impact of Roadside Features on the Frequency and Severity of Run-off Roadway Accidents: an Empirical Analysis. Accident Analysis and Prevention, 34(2), 149-161.

Lord, & Mannering. (2010). The Statistical Analysis of Crash Frequency Data: A Review and Assessment of Methodological Alternatives. Transportation Research Part A, 44(5), 291-305.

Lord, D., & Bonneson, J. (2006). Development of Accident Modification Factors for Rural Frontage Road Segments in Texas. College Station, TX.: Texas A&M University.

Lord, D., & Miranda-Moren0, L. (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(5), 751-770.

Lord, D., & Park, P. (2008). Investigating the Effects of the Fixed and Varying Dispersion Parameters of Poisson-Gamma Models on Empirical Bayes Estimates. Accident Analysis and Prevention, 40, 1441-1457.

Lord, D., Guikema, S., & Geedipally, S. (2008). Application of the Conway-Maxwell-Poisson Generalized Linear Model for Analyzing Motor Vehicle Crashes. Accident Analysis and Prevention, 40(3), 1123-1134.

Lord, D., Washington, S., & Ivan, J. (2005). Poisson, Poisson-Gamma and Zero Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory. Accident Analysis and Prevention, 37(1), 35-46.

Lord, Guikema, & Geedipally. (2008). Application of the Conway-Maxwell Poisson Generalized Linear Model for Analyzing Motor Vehicle Crashes. Accident Analysis and Prevention, 40(3), 1123-1134.

Miaou, S. (2001). Estimating Roadside Encroachment rates with The Combined Strengths of Accident-and Encroachment-Based Approaches. Federal Highway Administration.

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

Miaou, S., Song, J., & Mallick, B. (2003). Roadway Traffic Crash Mapping: A Space Time Modeling Approach. Journal of Transportation and Statistics, 6(1), 33-57.

Milton, J., & Mannering, F. (1996). The Relationship Between Highway Geometrics, Traffic Related Elements and Motor Vehicle Accidents. Seattle: Washington State Transportation Center.

Milton, J., shankar, V., & Mannering, F. (2008). Highway Accident Severities and the Mixed Logit Model: An Exploratory Empirical Analysis. Accident Analysis and Prevention, 40(1), 260-266.

Mitra, S., & Washington, S. (2007). On the Nature of Over-Dispersion in Motor Vehicle Crash Prediction Models. Accident Analysis and Prevention, 39(3), 459-468.

Mountain, L., Fawaz, B., & Jarrett, D. (1996). Accident Prediction Models for Roads with Minor Junctions. Accident Analysis and Prevention, 28(6), 695-707.

NHTSA. (2012, February). Traffic Safety Facts. Retrieved February 25, 2012, from National Highway Traffic Safety Administration: http://www-nrd.nhtsa.dot.gov/Pubs/811552.pdf

Oh, J., Washington, S., & Nam, D. (2006). Accident Prediction Model for Railway-Highway Interfaces. Accident Analysis and Prevention, 38, 346-356.

Persaud, B. (1994). Accident Prediction Models for Rural Roads. Canadian Journal of Civil Engineering, 21(4), 547-554.

Richardson, A., Ampt, E., & Meyburg, A. (1995). Survey Methods for Transportation Planning. (Eucalyptus Press) Retrieved 05 25, 2012, from http://www.transportsurveymethods.com.au/

Saito, M., Brimley, B., & Schultz, G. (2011). Transportation Safety Data and Analysis Volume 2: Calibration of Highway Safety Manual and Development of New Safety Performance Functions. Utah Department of Transportation.

Sawalha, Z. and Sayed, T. (2006). Transferability of Accident Prediction Models. Journal of Safety Science, 209-219.

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

Srinivasan, R., & Carter, D. (2011). Development of Safety Prerformance Function for North Carolina. University of North Carolina Highway Safety Research Center.

Srinivasan, S., Haas, P., Dhakar, N., Hormel, R., Torbic, D., & Harwood, D. (2011). Development and Calibration of Highway Safety Manual Equations for Florida Conditions. Florida Department of Transportation.

Sun, X., Margri, D., Shiraji, H., Gillella, S., & Li, L. (2011). Application of Highway safety Manual: Louisiana Experience with Rural Multilane Highways. Washington D.C.: 90th Annual Meeting of The Transportation Research Board.

Tegge, R., Jo, J., & Ouyang, Y. (2010). Development and Application of Safety Performance Functions for Illinois. Illinois Department of Transportation.

Vogt, A. (1999). Crash Models for Rural Intersections: Four-Lane by Two-Lane Stop-controlled and Two-lane by Two-lane Signalized. Federal Highway Adminstration.

Vogt, A., & Bared, J. (1998). Accident Models for Two Lane Rura Roads: Segments and Intersections. Federal Highway Adminstration.

Wade, M., Hammond, C., & Kim, C. (2004). ACCIDENT ANALYSIS OF SIGNIFICANT CRASH RATES FOR LOW TO VERY LOW VOLUME ROADWAYS IN 10 MINNESOTA COUNTIES. Minnesota Department of Transportation.

Washington, S., Persaud, B., Lyon, C., & Oh, H. (2005). Validation of Accident Models for Intersections. Federal Highway Administration.

Xie, F., Gladhill, K., Dixon, K., & Monsere, C. (2011). Calibrating the Highway Safety Manual Predictive Models for Oregon state Highways. Transportation Research Record.

Chapter 3.

A BAYESIAN ANALYSIS OF CRASH SEVERITIES WITH MULTIVARIATE CONWAY-MAXWELL POISSON DISTRIBUTION

3.1 Introduction

There has been considerable amount of research done in the past two decades on statistical modeling and analysis of crash data (see Lord & Mannering (2010) for a recent review). The studies involve developing regression equations to estimate expected number of crashes for a given roadway segment (or a site). The factors considered in the literature include roadway geometry, environmental and traffic conditions, highway user attributes and crash related information. Several approaches have been utilized in developing the relation between the crashes and the above-mentioned factors. The methods found in the literature can be divided into two major streams: univariate count data models and multivariate count models.

Traditionally, univariate models have been used for analyzing total crash frequencies or crash frequencies of a certain severity, where there is only one response variable. Since crash counts are discrete positive numbers, the Poisson distribution and its generalizations are widely used for modeling crash data sets (Shmueli, Minka, Kadane, Borle, & Boatwright, 2005; Lord, Washington, & Ivan, 2005; Tegge, Jo, & Ouyang, 2010; Zou, Geedipally, & Lord, 2013; Aitchison & Ho, 1989; Castillo & Perez-Casany, 2005; Jovanis & Chang, 1986; Joshua & Garber, 1990). One key property of Poisson distribution is that its mean is equal to its variance. However, crash counts have often been found to have extra Poisson dispersion, i.e. the mean

crash frequency is not equal to its variance (Mitra & Washington, 2007). When mean is smaller than the variance, the data is over-dispersed; otherwise it is under-dispersed. To address this issue, several hierarchical distributions have been used such as Poisson-Gamma or Negative Binomial (NB) (Mehta & Lou, 2013; Kim & Washington, 2006; Miaou, 1994), Poisson-lognormal (Valverde & Jovanis, 2008), zero-inflated Poisson (Miaou, 1994; Shankar, Milton, & Mannering, 1997), zero-inflated NB (Lord, Washington, & Ivan, 2007) etc. for over-dispersed datasets (Lord & Mannering, 2010); and double Poisson (DP) (Zou, Geedipally, & Lord, 2013), weighted Poisson (Consul, 1989) generalized Poisson (Famoye, 1993), Gamma (Winklemann, 1995) and Conway-Maxwell Poisson (CMP) (Lord, Guikema, & Geedipally, 2008) distributions for under-dispersed datasets (Lord, Geedipally, & Guikema, 2010). The NB and zero-inflated models are preferred when the crash data is over-dispersed (Lord & Miranda-Moreno, 2008). On the other hand, CMP and DP distributions are even more flexible and can be used for any kind of extra Poisson dispersion. They are able to accommodate both over- and under-dispersed data (Zou, Geedipally, & Lord, 2013; Lord, Guikema, & Geedipally, 2008), while most of the other distributions suffer from theoretical or logical soundness when it comes to under-dispersion (Famoye, 1993; Winklemann, 2008). Zou et al. (Zou, Geedipally, & Lord, 2013) compared the performance of CMP and DP models on different kinds of crash data and found that CMP outperforms DP for modeling under-dispersed datasets and performs on par with DP for over-dispersed or equal-dispersed datasets.

Although easier to estimate and apply, univariate models cannot account for correlation between crash frequencies of different severities, which is more likely than not to exist. For example, consider a particular roadway segment having unobserved characteristics such as poor pavement surface, adjacent land use, grade, sight distance, lighting etc. All these factors might

influence the crash counts at all severity levels for this site. Knowing such correlation can improve the estimation efficiency (Ma, Kockelman, & Damien, 2008).

Several existing studies have applied bi-variate and multivariate models to crash severity analysis, where regression equations for two or more response variables are estimated concurrently with consideration of correlated error structures. Some of the multivariate models used in crash severity analysis include multivariate Poisson (MVP) (Ma & Kockleman, 2006), multivariate NB (MVNB) (Subrahmaniam & Subrahmaniam, 1973; Maher, 1990), multivariate zero-inflated Poisson (Dong, Richards, Clarke, Xuemei, & Ma, 2014) and multivariate Poisson-lognormal (MVPLN) (Ma, Kockelman, & Damien, 2008). MVP regression models have the same limitation as their univariate counterpart that the mean and the variance are assumed equal. MVP and MVNB are restrictive since they cannot account for negative correlations between crash severities (Aitchison & Ho, 1989; Park & Lord, 2007). This implies that there cannot be any site with one type of crash severity increasing and other type decreasing at the same time. Aitchison and Ho (1989) relaxed this restrictive correlation structure by proposing an MVPLN model. Since their work, MVPLN has been adopted by several highway safety studies (Ma, Kockelman, & Damien, 2008; Park & Lord, 2007; Chib & Winklemann, 2001; El-Basyouny & sayed, 2009). Shoukri(Shoukri, 1982) generalized a bivariate extension of DP model to incorporate both over- and under-dispersion. There are, however, no studies or applications of this model in the highway safety literature. Neither is there extension of CMP for modeling correlated datasets, despite its desired properties such as flexibility of spanning over three special distributions and capability to handle any kind of extra Poisson dispersion.

This study aims to fill the gap in the literature of multivariate regression models that can accommodate both over- and under-dispersion for crash severity analysis by proposing a

multivariate extension of the univariate CMP. Adopting Bayesian inference paradigm, a Markov Chain Monte Carlo (MCMC) method is used to estimate the parameters of the regression model. A component (block)-wise Metropolis-Hastings (MH) algorithm is developed and implemented. The algorithm is coded in MATLAB. The proposed formulation and algorithm are first validated using a simulated dataset. Finally, a case study on crash severity analysis of two-lane rural roads is presented as an application of the proposed multivariate Conway-Maxwell Poisson (MVCMP) formulation and the component-wise MH algorithm.

3.2 Existing Count Data Models

This section first presents the history of the Conway-Maxwell distribution along with studies involving use of CMP in the crash analysis followed by multivariate count data models mainly concentrating on multivariate Poisson lognormal model.

3.2.1 Conway-Maxwell Poisson Distribution

Conway-Maxwell Poisson distribution was proposed by Conway and Maxwell in 1962 for modeling the queue and service rates. This distribution is a generalization of the Poisson distribution. The distribution was rarely used in the literature, after its introduction by Conway and Maxwell in 1962. In 2005, Shmueli et al. derived the statistical properties of this distribution. It is a unique distribution with Geometric, Bernoulli and Poisson as its special cases. The main advantage of this distribution over other models studied above is its ability to handle both the over-dispersed and under-dispersed data.

Lord et al. (2008) applied this distribution to the highway safety research and found that it performs comparable to the NB regression for the over-dispersed data and provides better fit

for the under-dispersed data. It was also noted that the model does not perform well when the mean of the sample is low or the sample size is very small.

3.2.2 Multivariate Count Data Models

Several Bivariate and Multivariate models are applied to crash severity analyses. Some of the multivariate models used in crash severity analysis include multivariate Poisson, multivariate NB regression, multivariate zero inflated Poisson and Poisson-lognormal regression models. Subhramaniam and Subhramaniam(1973) derived the estimation procedure for the bivariate negative binomial distribution. Maher (1990) used the bivariate NB regression to explain the crash migration effect in purely probabilistic terms, without recourse to the concept of physical migration. Most of these studies had very restrictive or no covariance structure estimated. Park and Lord (2007) introduced Poisson-lognormal model to jointly model the crash counts by severities. This new specification can handle over-dispersion as well as full general correlation structure.

The multivariate Poisson regression model, one of the most widely used model for count data do not allow for over dispersion. Most of the specifications found in literature are very restrictive with covariance structure being strictly positive. Aitchison and Ho (1989) therefore proposed the Multivariate Poisson-Lognormal model specification, which provides very flexible covariance structure. In this specification, the crashes occurring at the disjoint time intervals are considered as mixture of independent Poisson distribution and the normally distributed errors. The normal mixing distribution adds the flexibility of having negatively correlated crash counts. There have been several applications of the MVPLN in the crash analysis literature (Chib &

Winklemann, 2001; Park & Lord, 2007; Ma, Kockelman, & Damien, 2008; El-Basyouny & sayed, 2009).

Chib and Greenberg (1996) studied several Markov chain Monte Carlo simulation methods that can help in simulating large sets of multivariate density function. They use Bayesian paradigm to solve the intractable integrals and explained how it works in theory and practice. The Gibbs and Metropolis Hastings algorithms are explained, and they show how the simulated data obtained using these algorithms converge to the posterior distribution. The application of the above methods on different models namely seemingly unrelated regression model, Tobit, Probit, Random coefficient panel model and regression model with autoregressive errors are shown.

Chib and Winklemann(2001) applied the above-mentioned MCMC technique to two different problems. The first problem jointly modeled the six measures of the medical-care demand by elderly and the second problem dealt with airline incidents recorded for 16 US passenger air carriers between 1957 and 1986. The counts were assumed independent Poisson variate conditioned on a vector of correlated unobserved heterogeneities. The multivariate normal and multivariate t-distributions were used as the distribution of the unobserved heterogeneity. They simulated 6000 iterations following an initial burn-in of 500 simulations. They observed that the coefficients obtained were robust to the starting values of the simulation.

Park and Lord (2007) applied the same specification used by Chib and Winklemann to the intersection data from California. They compared the crash severities obtained from the MVPLN model and compared it with univariate Poisson and NB regression models. The data used in the analysis consisted of 451 intersections and 10 years of crash records. Matlab was

used for estimating the parameters of MVPLN model. It was observed that the standard errors obtained using univariate models were underestimated resulting in the incorrect model specification. Although, this model specification performed better for the over-dispersed crash counts it cannot handle under-dispersed crash counts.

Ma et al. (2008) applied the MVPLN specification to estimate the crash severities on two lane rural roads. The data consisted of 7,773 sites of two lane highway totaling to 510 miles of road. The parameter of the model were estimated using the MCMC framework and was coded in R. A total of 8000 simulations are performed and first 1000 simulations were discarded as the burn-in time. The burn-in time helps in eliminating the effects of the starting values. To check the convergence of the chain they used different starting values and compared the posterior distributions. It was observed that the chain converged to the same posterior distribution of parameters. A statistically significant positive correlation between the crash counts was observed.

3.2.3 Limitation of the Existing Count Data Models

The univariate count data models can be used for estimating the crash severities individually, however, there may exist some interdependence between the crashes of different severities for a specific segment of roadway. For example, sites with high fatality rate might also have very high injury rate and lower PDO crashes. There can also be some correlation between different environmental conditions, roadway characteristics and specific roadway segment, which cannot be realized if univariate models are used. There are several count data regression model as described above, however all these models can at best accommodate the equi-dispersed or over-dispersed data. Unfortunately, crash data can sometimes be under-dispersed and all the

above models will be misleading when the data is heavily under-dispersed. As a solution to this problem, this dissertation proposes a multivariate extension of the Conway Maxwell Poisson distribution, which is proved to perform better for both under-dispersed and over-dispersed data.

3.3 Bayesian Paradigm Parameter Estimation Background

3.3.1 Parameter Estimation Background

This section starts with the debate on Frequentist vs Bayesian Inference along with a list of the differences in computation and interpretation of the two methods. The next subsection talks about the Bayes' theorem followed by the computation of the Bayes' theorem. The Bayesian analysis mostly involves computation of higher dimensional integrals that require simulation techniques. The Markov Chain Monte Carlo simulation technique is used in this study and it is described in detail along with the explanation on the Markov chains and the conditions that need to be satisfied for successful implementation of this simulation technique. Finally, a detailed explanation of Gibbs sampler and Metropolis Hastings algorithm that is used in this study is presented.

3.3.1.1 Bayesian Inference vs Frequentist Inference

Frequentist and the Bayesian techniques (Gelman, Carlin, Stern, & Rubin, 2014) are two main approaches for statistical inference. The frequentist techniques include Maximum Likelihood Estimation (MLE), Method of Moments (Hogg, Craig, & McKean, 2006) etc., whereas, Bayes' theorem is the anchor of Bayesian inference. The frequentist methods treat data as random and parameters as fixed and unknown. On the other hand, Bayesian inference assumes that the parameters are random variables with associated probabilities rather than a fixed value. This allows for more flexibility in modeling uncertainties related to parameter

values. Bayesian analysis incorporates prior knowledge or results from a previous model into the model building procedure, which is not feasible with the frequentist methods. When the likelihood L does not have a closed form, traditional methods like MLE, which optimizes L cannot be implemented directly. A simulation method like Iteratively Reweighted Least Squares (Wolke & Schwetlick, 1988) or Expectation Maximization (Dempster, Laird, & Rubin, 1977) algorithms will be required to estimate the parameters. However, these methods become inefficient as the number of parameters to be estimated increases (Bolstad, 2010), especially when numerical methods for evaluating higher dimensional integrals are involved. Moreover, the quality of the approximation of L resulted from these numerical methods largely governs the quality of parameter estimation as L is the objective function to be maximized. Instead of numerical evaluation of integrals, Bayesian inference allows for drawing random samples directly from the posterior distribution. Finally, the two approaches differ in the interpretation of the credible and confidence intervals. Credible interval gives the actual probability of a parameter lying between two points; whereas confidence interval concludes with some level of confidence that with probability 0 or 1 a parameter is between two points (Gelman, Carlin, Stern, & Rubin, 2014). Moreover, Bayesian techniques can handle multivariate data much more easily compared to frequentist techniques. In this study, Bayesian paradigm is adopted to estimate the parameters. Bayesian inference combines both prior information and evidence (data) to arrive at parameter estimates that are expressed in terms of posterior probabilities based on Bayes' theorem.

3.3.1.2 Bayes' Theorem

The Bayes' theorem is a theorem of probability named after the Rev. Thomas Bayes (Stigler, 1983). The essence of the theorem is to provide a mathematical rule explaining how one

should change their existing belief based on the light of new evidence. To understand Bayes' theorem let's consider a dataset, where x is a matrix of explanatory variables and y is a vector of dependent variable. If θ is assumed to be a vector of parameters to be estimated, then the Bayes' Theorem can be written as

$$\pi(\theta|y,x) = \frac{\pi(\theta|x)\pi(y|\theta,x)}{\pi(y|x)} \propto \pi(\theta|x)\,\pi(y|\theta,x)$$

where,

$\pi(\theta|x)$is the prior distribution of the random parameters θ. It represents the best estimate of the probability of the parameter prior to any consideration of the data or evidence.

$\pi(y|\theta,x)$is the conditional likelihood of the observed response given a particular set of values for θ and x.

$\pi(y|x)$term is usually dropped from the equality part in the equation, since it does not have any parameters in it and can be considered as constant.

$\pi(\theta|y,x)$is the density of the posterior distribution of θ which is obtained by updating the prior belief about the parameters after incorporating the new information contained in the data.

3.3.1.3 Bayesian Computation

Bayesian inference mainly consists of computation of posterior distribution, which is conditional distribution known at least up to a constant. The conditional distribution for most of the higher dimensional models does not follow any standard distributions and neither have a closed form solution. In such instances, simulations using Markov Chain Monte Carlo technique is very popular and widely used (Ma, Kockelman, & Damien, 2008; Chib & Greenberg, 1996; Park & Lord, 2007). Markov chain simulation is a technique of generating sample of the

required posterior distribution from an approximate distribution, which corrects itself sequentially by drawing large number of correlated samples depending on the previous draw. The chain process in which the sample is obtained is called Markov Chain. The sample obtained after the Markov Chain reaches the stationary distribution is believed to resemble the target distribution. A brief explanation of the background of Markov Chains is given below.

3.3.2 Markov Chains

Markov process is a stochastic process with discrete or continuous states governed by the transition probability. It is a sequence of random variables satisfying the Markov property, which states that the current state of the variable or the system is dependent only on the most recent previous state and is independent of all the other past states. This is called the memory less property of Markov process. The Markov chain considered in this study is defined as a discrete time stochastic process having finite or countable possible states.

The Markov chain is defined as follows. Let $X_t$be the values that the chain takes at time t (or state). Let S be a finite and countable state space and $S = (s_1, s_2, \dots\dots\dots, s_s)$ be the possible values of X. There can be any number of possible finite state space, but for our application it is considered to be part of R^p, where p represents the dimension of X. The Markov chain starts at one of the S possible states and moves from one state to another. This process of moving from one state to another is called a step. The new state is selected randomly from all the possible states based on a certain probability of a move. The new state can be the same state or any new state. This new state is the transition state and the probability associated with it is the transition probability. The set of these transition states and associated probabilities characterizes a Markov chain. According to the Markov property the transition probability depends only on the most

recent draw X_{t-1}, implying that the probability of moving to state *t* only depends on the previous state *t-1* and is independent of all other states. This can be written as follows.

$$P(X_t = s_j | X_{t-1} = s_i,\ i = 1,2,\ldots . t) = P(X_t = s_j | X_{t-1} = s_i\ ).$$

Some of the properties of the Markov Chains, which are important for its successful application in this research are described below.

3.3.2.1 Time Homogeneity

A chain is called as time homogeneous, if the transition matrix does not change for any step. For such chains the k^{th} step transition probability can be calculated, by raising the transition matrix to power *k*. Mathematically, a chain is time homogeneous when $P(X_{h+t} = s_j | X_h = s_i\ )$ is independent of h , implying when $h = 0$ the following equation is true

$$P(X_{h+t} = s_j | X_h = s_i\ ) = P(X_t = s_j | X_0 = s_i\ )$$

3.3.2.2 Irreducible

A Markov Chain is irreducible, if the chain can jump from any state to any state, alternatively, it cannot get stuck between few states. A Markov chain is irreducible if

$$P(X_t = s_i | X_0 = s_j) > 0 \qquad \forall\ s_i, s_j\ \in S\ \ni\ t \geq 0$$

3.3.2.3 Aperiodic

A Markov Chain is said to be periodic, if a chain can return to a particular state from some other state only after a fixed number of periods. Any chain that is not periodic is called aperiodic. Thus, a process that can be in state s_i at time t or $t + 1$ having started at state s_i is aperiodic.

3.3.2.4 Stationary distribution

Let us denote π^* as the stationary distribution of a Markov chain. An irreducible and aperiodic Markov Chain is said to have converged to a stationary distribution π^* when the following sufficient condition holds.

$P(s_i|s_j)\pi_i^* = P(s_j|s_i)\pi_j^*$

Where

$P(s_i|s_j)$is the probability of moving to state s_i from state s_j

$\pi_i^*$is the probability distribution at state s_i

The above equation is also called as the reversibility condition. The stationary distribution implies $\pi = \pi P$. This is the very important property of the Markov chain, which will be useful in our analysis. According to this property, one can start a chain with any initial distribution $\pi^{(0)}$, the irreducible and aperiodic chain will converge to stationary distribution as we run the chain for a sufficiently long period of time.

$$\lim_{n\to\infty} \pi^{(n)} = \pi$$

Thus, once the distribution converges to the stationary distribution the marginal distribution of the state at all future times is again given by the stationary distribution, so these values are an identically distributed sample from this distribution.

The distribution of the state of the chain at time t, given the state at time $t-1$ is given by a transition kernel $P(x_t|x_{t-1})$, taking values in the arbitrary state space S and having the property the conditional distribution given the past, depends only on the present stateX_t. Let x_t

be the values that the state of the chain takes at time (or state). The probability of the current state given all the past states is gives below

$$P(x_t|x_{t-1}, x_{t-2}, \dots\dots\dots x_0) = P((x_t|x_{t-1})$$

Two widely used methods for constructing a Markov Chain are the Metropolis-Hastings Algorithm and the Gibbs Sampler. A brief review of both the methods is presented below.

3.3.3 Metropolis Algorithm

Metropolis algorithm proposed by Metropolis et al. (1953) is one of the methods of obtaining sample from a complex probability distribution generally encountered in Monte Carlo integration. The idea behind the algorithm is to draw samples from some proposal or jumping distribution and accepting the values, which increases the density of the posterior distribution. Given a target density $\pi(\theta)$ from which the sample is to be drawn, and $\pi(\theta) = f(\theta)/C$, where the normalizing constant C may not be known or very difficult to compute, then the Metropolis algorithm can be applied in following sequence to obtain a sample from the target density as follows.

Step1) Initialize θ_0 with any value such that $f(\theta_0) > 0$

Step2) For m=1,2 …..M, where M denotes the number of iterations, sample a candidate value θ^* from a symmetric proposal distribution i.e. $q(\theta_0, \theta^*) = q(\theta^*, \theta_0)$, such that the state space of the proposal distribution is the same as the sample space of the target density $\pi(\theta)$. It is important to select a proposal distribution that can be easily sampled.

Step3) The acceptance ratio is then calculated by taking the ratio of the posterior densities at the starting point θ_0 and the new candidate point θ^*as $\alpha(\theta_0, \theta^*) = \{\min\left\{\frac{\pi(\theta^*)}{\pi(\theta_0)}\right\}, 1\}$. Given the current state of the chain at θ_0, the value of next step in chain θ_1 is set at $\theta_1 = \theta^*$ with

probability $\alpha(\theta_0, \theta^*)$ and at $\theta_1 = \theta_0$ with probability $1 - \alpha(\theta_0, \theta^*)$. If the ratio $\frac{\pi(\theta_0)}{\pi(\theta^*)}$ is greater than 1, then the candidate value is definitely accepted, however, if the ratio is less than 1 than the candidate is accepted if a random draw of a standard uniformly distributed $U(0,1)$ variable is small than the ratio.

Step4) This concludes first iteration. The procedure is repeated for M times to obtain a sample which after a suitable burn-in approximates the target distribution $\pi(\theta)$. The proof that the sequence of iterations converges to the target distribution is very easy to prove and is a two-step process. Firstly, it is important that the sequence simulated be a Markov Chain, which holds if the chain is irreducible, aperiodic and not transient. According to Gelman et al (2014) the aperiodicity and non-transiency holds for a random walk chain on any proper distribution. The condition of irreducibility holds, if the chain has positive probability of moving to any state from any other state. Secondly, to prove that the stationary distribution converges to the target distribution, it is sufficient to show that the detail balance equation holds given by

$$(\theta^*)q(\theta^*, \theta_0)\alpha(\theta^*, \theta_0) = \pi(\theta_0)q(\theta_0, \theta^*)\alpha(\theta_0, \theta^*). \quad (1)$$

If $\theta_0 \neq \theta^*$ and it is such that $\frac{\pi(\theta_0)q(\theta_0,\theta^*)}{\pi(\theta^*)q(\theta^*,\theta_0)} > 1$.

Then, $\alpha(\theta_0, \theta^*) = \frac{\pi(\theta^*)q(\theta^*,\theta_0)}{\pi(\theta_0)q(\theta_0,\theta^*)}$and $\alpha(\theta^*, \theta_0) = 1$.

Based on the assumption where the acceptance probability is 1, the unconditional probability density of transition from $\theta^*$to θ_0 is

$$\pi(\theta^*)q(\theta^*, \theta_0) = \pi(\theta_0)q(\theta_0, \theta^*)\alpha(\theta_0, \theta^*)$$

$$\pi(\theta^*)q(\theta^*, \theta_0) = \pi(\theta_0)q(\theta_0, \theta^*)\frac{\pi(\theta^*)q(\theta^*, \theta_0)}{\pi(\theta_0)q(\theta_0, \theta^*)}$$

$$= \pi(\theta^*)q(\theta^*, \theta_0)\alpha(\theta^*, \theta_0)$$

Thus the detailed balance equation holds as shown in equation (1).

3.3.4 Metropolis Hastings Algorithm

Hastings (1970) generalized the Metropolis algorithm by removing the restriction of symmetric proposal distribution. The Hastings algorithm can be implemented by modifying the acceptance ratio in the Metropolis algorithm as follows. In the above setting, the acceptance ratio for the new candidate value is given by

$$\alpha(\theta_0, \theta^*) = \left\{ \min \frac{\pi(\theta^*) q(\theta^*, \theta_0)}{\pi(\theta_0) q(\theta_0, \theta^*)}, 1 \right\}$$

where, $q(\theta^*, \theta_0) \neq q(\theta_0, \theta^*)$. The proof of the convergence to a unique stationary distribution is similar to the proof of Metropolis algorithm.

3.3.5 Gibbs Sampler

The Gibbs sampler or alternating conditional sampling is a second method of developing a Markov Chain proposed by Geman and Geman(1984). It is a special case of MH algorithm, which requires knowledge of the conditional distribution of the unknown parameters. To understand Gibbs algorithm, let's assume a parameter space Θ with two parameters (θ_1, θ_2). Sometimes, it is easy to obtain a marginal distribution by simulating from conditional distributions $\theta_1|\theta_2$ and $\theta_2|\theta_1$ rather than integrating the joint density $\pi(\theta_1) = \int p(\theta_1, \theta_2)\, d\theta_2$. In such instances, Gibbs sampler can be used by first initializing one of the parameters say θ_1^0. Then, a sample is generated for parameter θ_2 using the conditional distribution $p(\theta_2|\theta_1 = \theta_1^0)$ and denoted by θ_2^1. The next sample for θ_1 is obtained using the conditional distribution $p(\theta_1|\theta_2 = \theta_2^0)$. The process is repeated M times to obtain a sequence of sample from the full conditional distribution. After a suitable burn-in the sample obtained from each full conditional

distribution is converges to a unique stationary distribution, which is the required target or posterior distribution.

3.3.6 Convergence Tests

MCMC is a simulation process that can be run for any number of iterations with any starting value. While theoretically MCMC converges to the desired posterior distribution under certain conditions (Gelman, Carlin, Stern, & Rubin, 2014), in implementation, testing the convergence of the chain generated could be a challenge. There is no test or diagnostic in the literature that promises to correctly identify convergence every time. Nonetheless, there are several different techniques with their own advantages and disadvantages. Some of the tests perform well with Gibbs sampler while some perform well with MH algorithm (Cowles & Carlin, 1996). The convergence diagnostics can be classified into two categories: visual and test diagnostics. The visual diagnostics commonly check how the chain is mixing or moving around the parameter space, including trace, density, running mean, and autocorrelation plots (Cowles & Carlin, 1996). The test diagnostics generally involve computing some statistics to identify convergence, including Gelman and Rubin (Gelman & Rubin, 1992), Raftery and Lewis (Raftery & Lewis, 1992), Geweke (Geweke, 1992), Roberts (Roberts, 1992), Ritter and Tanner (Ritter & Tanner, 1992) etc. The readers are referred to Cowles and Carlin (Cowles & Carlin, 1996) for a more comprehensive review.

Two visual diagnostics, the trace and the running mean plot, are used in this research. Trace plots display the sample values drawn from each iteration of MCMC simulation against the iteration number. It is mainly used to see the pattern of mixing. An ideal plot should not have any lumps or blocks of iterations indicating improper mixing. Running mean plots draw

the mean of the sample points up to each iteration against the iteration number. Ideal plots should look like an increasing or decreasing function that is converging. The chain is considered not well-mixed if the plots have peaks and lows.

3.4 Methodologies

3.4.1 Multivariate Conway-Maxwell Poisson (MVCMP) Formulation

In Poisson distribution, the probability of observing y_i crashes at roadway segment $i$is given by

$$p(y_i) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!},$$

whereλ_i is the mean of the Poisson distribution for site i. Comparing to Poisson distribution, CMP replaces the exponential term with an infinite sum and introduces an additional scale parameter ν. Under CMP, the probability of observing y_i crashes at roadway segment i is

$$p(y_i) = \frac{\lambda_i^{y_i}}{(y_i!)^{\nu}\, Z(\lambda_i, \nu)}$$

where

$$Z(\lambda_i, \nu) = \sum_{j=0}^{\infty} \frac{\lambda_i^j}{(j!)^{\nu}}. \quad (2)$$

CMP distribution has Geometric ($\nu = 0$), Poisson ($\nu = 1$) and Bernoulli ($\nu = \infty$) as its special cases. Shmueli et.al.(Shmueli, Minka, Kadane, Borle, & Boatwright, 2005)further approximated $Z(\lambda_i, \nu)$ using asymptotic theory by

$$\tilde{Z}(\lambda_i, \nu) = \frac{\exp\left(\nu\lambda_i^{\frac{1}{\nu}}\right)}{\lambda_i^{\frac{\nu-1}{2\nu}}(2\pi)^{\frac{\nu-1}{2}}\sqrt{\nu}}\left\{1 + O\left(\lambda_i^{-\frac{1}{\nu}}\right)\right\} \quad (3)$$

This approximation works well when $\nu > 1$ or $\lambda_i^{\frac{1}{\nu}}$>10. The approximation will be adopted in this study, because the computation time almost halved when compared to the time required summing the infinite series until its convergence.

To derive a formulation of the MVCMP distribution, let Y denote an n-by-Smatrix of multivariate crash severity counts, where n is the total number of sites under consideration and S is the number of severity levels. Note that Y is a matrix of random variables. Each element of the matrix, denoted as Y_{is}, is the random crash count for site i and severity s, for $i = 1,2 \ldots . n$, and $s = 1,2, \ldots S$. And y_{is} is a realization of Y_{is} (observed crash counts). Further, let $Y_i = (Y_{i1}, Y_{i2}, \ldots . Y_{iS})'$and $y_i = (y_{i1}, y_{i2}, \ldots . y_{iS})'$. It is assumed that there is no spatial correlation among the roadway segments, but the crash counts of different severity levels for site i are correlated. Thus the variance-covariance matrix of Y can be expressed as

$$cov(Y) = \begin{bmatrix} \alpha_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \alpha_n \end{bmatrix},$$

where each diagonal block of $cov(Y)$, α_i, is an S-by-S matrix representing the variance-covariance matrix of the multivariate random variable Y_i. Since there is no spatial correlation, the off-diagonal blocks of $cov(Y)$ are all zeros. On the other hand, $cov(Y_i) = \alpha_i$ is not a diagonal matrix due to correlation among crash counts of different severity levels at site i.

Let $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \ldots . \epsilon_{iS})'$ denote the unobserved heterogeneity specific to crash severity for site i. The unobserved heterogeneity ϵ_i is assumed to follow an S-dimensional multivariate normal distribution with zero mean and a variance-covariance matrix Σ, i.e. $\epsilon_i \sim N_S[0, \Sigma]$. The S-by-S matrix Σ is assumed to be constant across the sites.

Let K be the number of covariates (explanatory variables) considered in the analysis and x denote an n-by-K matrix of covariates. Each row of this matrix is denoted by $x_i =$

$(x_{i1}, x_{i2}, \dots, x_{iK})$, which represents a K-dimensional row vector corresponding to the characteristics of the i^{th} roadway segment. Let $\beta = (\beta_1, \beta_2, \dots, \beta_S)$ denote a K-by-S matrix of the regression coefficients whose columns $\beta_s = (\beta_{1s}, \beta_{2s}, \dots, \beta_{Ks})'$ are K-dimensional column vectors consisting of the regression coefficients for the crash count of the s^{th} severity level.

It is assumed that the marginal distribution of Y_{is} conditioned on the heterogeneity ϵ_i follows a univariate CMP distribution. The dispersion parameter $\nu = [\nu_1, \dots, \nu_S]$ denotes an S-dimensional row vector, which remains constant across the sites and varies across severities. This provides flexibility in modeling crash counts of different severities at the same site as both over- and under-dispersed. The Poisson mean λ_{is} is determined by the attribute vectorx_i, the coefficient vector β_s, and the site-specific unobserved heterogeneity ϵ_{is}: $\lambda_{is} = \exp(x_i'\beta_s + \epsilon_{is})$. Now, the conditional probability of y_{is} can be written as

$$p_{CMP}(y_{is}|\epsilon_i, \beta_s, x_i, \nu_s) = \frac{\lambda_{is}^{y_{is}}}{(y_{is}!)^{\nu_s} Z(\lambda_{is}, \nu_s)} \tag{4}$$

This can be further approximated as

$$\tilde{p}(y_{is}|\epsilon_i, \beta_s, x_i, \nu_s) = \frac{\lambda_{is}^{y_{is}}}{(y_{is}!)^{\nu_s} \tilde{Z}(\lambda_{is}, \nu_s)}$$

In the above formulation, β, ν, ϵ and Σ are the parameters to be estimated with observations $x$and y.

This study proposes the use of Bayesian paradigm to estimate the parameters of the model. The justification of the use of Bayesian approach over Frequentist is a big ongoing debate among statisticians, which is outside the scope of this dissertation. However, this paper prefers Bayesian analysis because of following advantages. Firstly, frequentist methods consider the

parameter as fixed and unknown unlike in Bayesian approach, where it is considered random and thus having a distribution. Secondly, Bayesian approach can easily handle non-homogenous variances and multivariate data, which sometimes can be difficult for the traditional methods. The Bayesian statistics takes into account the prior belief about the parameter before looking at the sample data. The Bayes theorem is a tool for combining the information from the prior belief and the evidence (data) to get information about the parameters or the hypothesis of interest.

3.4.2 Parameter Estimation

3.4.2.1 Bayesian Inference

Frequentist and the Bayesian techniques (Gelman, Carlin, Stern, & Rubin, 2014) are two main approaches for statistical inference. The frequentist techniques include Maximum Likelihood Estimation (MLE), Method of Moments (Hogg, Craig, & McKean, 2006) etc., whereas Bayes' theorem is the anchor of Bayesian inference. The frequentist methods treat data as random and parameters as fixed and unknown. On the other hand, Bayesian inference assumes that the parameters are random variables with associated probabilities rather than a fixed value. This allows for more flexibility in modeling uncertainties related to parameter values. Bayesian analysis incorporates prior knowledge or results from a previous model into the model building procedure, which is not feasible with the frequentist methods. When the likelihood L does not have a closed form, traditional methods like MLE, which optimizes L cannot be implemented directly. A simulation method like Iteratively Reweighted Least Squares (Wolke & Schwetlick, 1988) or Expectation Maximization (Dempster, Laird, & Rubin, 1977) algorithms will be required to estimate the parameters. However, these methods become inefficient as the number of parameters to be estimated increases (Bolstad, 2010), especially when numerical methods for evaluating higher dimensional integrals are involved. Moreover,

the quality of the approximation of L resulted from these numerical methods largely governs the quality of parameter estimation as L is the objective function to be maximized. Instead of numerical evaluation of integrals, Bayesian inference allows for drawing random samples directly from the posterior distribution. Finally, the two approaches differ in the interpretation of the credible and confidence intervals. Credible interval gives the actual probability of a parameter lying between two points; whereas confidence interval concludes with some level of confidence that with probability 0 or 1 a parameter is between two points (Gelman, Carlin, Stern, & Rubin, 2014). Moreover, Bayesian techniques can handle multivariate data much more easily compared to frequentist techniques. In this study, Bayesian paradigm is adopted to estimate the parameters.

Bayesian inference combines both prior information and evidence (data) to arrive at parameter estimates that are expressed in terms of posterior probabilities based on Bayes' theorem. Consider a dataset with vector *Y* as the random dependent variable, *X* as a matrix of covariates and θ as the vector of parameters. Let $Prior(\theta|x)$ denote the prior belief of the parameter with observed explanatory variables x, $p(y|\theta,x)$ the joint probability density of observation y conditioned on θ and x, and $\pi(\theta|y,x)$ the posterior distribution of parameter θ. According to Bayes' theorem, $\pi(\theta|y,x) \propto Prior(\theta|x)p(y|\theta,x)$.

For the crash severity analysis in this study, let's consider the joint distribution of $y_i = (y_{i1}, y_{i2}, \dots, y_{iS})$. Since the response variables are assumed conditionally independent across different crash severities, we have

$$p(y_i|\beta,\nu,\epsilon_i,x_i) = \prod_{s=1}^{S} p(y_{is}|x_i,\beta_s,\nu_s,\epsilon_{is})$$

(5)

Recall that $\epsilon_i \sim N_S[0, \Sigma]$. It is further assumed that y and ϵ are independent. Denote ϕ_S as the probability density function of an S-variate normal distribution. Therefore, the joint conditional distribution of y_i and ϵ_i can be written as follows.

$$p(y_i, \epsilon_i|\beta, x_i, \nu, \Sigma) = p(y_i|\epsilon_i, \beta, x_i, \nu)\phi_s(\epsilon_i|0, \Sigma) \tag{6}$$

The likelihood of parameters given y, assuming independence across sites can be written as

$$L(\beta, \nu, \Sigma|y, x) = \prod_{i=1}^{n} \int p(y_i, \epsilon_i|\beta, x_i, \nu, \Sigma)\, d\epsilon_i = \prod_{i=1}^{n} L_i(\beta, \nu, \Sigma|y_i, x_i)$$

Each component of L, L_i, is an S-variate integral to be computed from equations () and (), which does not have a closed-form solution.

Applying Bayes' theorem, the joint posterior density is given by

$$\pi(\beta, \nu, \Sigma, \epsilon|y, x) \propto Prior(\beta, \nu, \Sigma, \epsilon) \cdot p(y|\beta, \nu, x, \Sigma, \epsilon)$$

It is further assumed that all the parameters are independent of each other. Thus, the prior term in the above equation can be written as product of four probability distributions as shown below.

$$Prior(\beta, \nu, \Sigma, \epsilon) = Prior(\beta) \cdot Prior(\nu) \cdot Prior(\Sigma) \cdot \phi(\epsilon|0, \Sigma) \tag{7}$$

Assume $\beta_s \sim N_K(\mu_{\beta_s}, V_{\beta_s})$, a K-variate normal distribution with mean vector μ_{β_s} and variance-covariance matrix V_{β_s}; $\nu_s \sim \gamma(a_{\nu_s}, b_{\nu_s})$, a Gamma distribution with scale parameter a_{ν_s} and shape parameter b_{ν_s}; $\Sigma \sim f_w(\tau_\Sigma, V_\Sigma)$, an inverse Wishart distribution (Hogg, Craig, & McKean, 2006) with scale matrix V_Σ and degrees of freedom τ_Σ, which is a conjugate prior (Press, 1982; Gelman, Carlin, Stern, & Rubin, 2014). These parameters defining the prior

distributions of the parameters to be estimated are called hyper parameters. With these assumptions, equation () can then be written as

$$Prior(\beta, \nu, \Sigma, \epsilon) = \left[\prod_{s=1}^{S} \phi_K(\mu_{\beta_s}, V_{\beta_s})\gamma(a_{\nu_s}, b_{\nu_s})\right] \cdot f_w(\tau_\Sigma, V_\Sigma) \cdot \left[\prod_{i=1}^{n} \phi_S(\epsilon_i|0, \Sigma)\right]$$

Therefore, the following holds for the joint posterior distribution $\pi(\beta, \nu, \Sigma, \epsilon|y, x)$.

$$\pi(\beta, \nu, \Sigma, \epsilon|y, x)$$

$$\propto Prior(\beta, \nu, \Sigma, \epsilon) \cdot p(y|\beta, \nu, x, \Sigma, \epsilon)$$

$$= \left[\prod_{s=1}^{S} \phi_K(\mu_{\beta_s}, V_{\beta_s})\right] \cdot \left[\prod_{s=1}^{S} \gamma(a_{\nu_s}, b_{\nu_s})\right] \cdot f_w(\tau_\Sigma, V_\Sigma) \cdot \left[\prod_{i=1}^{n} \phi_S(\epsilon_i|0, \Sigma)\right]$$

$$\cdot \left[\prod_{i=1}^{n}\prod_{s=1}^{S} p_{CMP}(y_{is}|\epsilon_i, \beta_s, x_i, \nu_s)\right]$$

$$= f_w(\tau_\Sigma, V_\Sigma) \cdot \left[\prod_{s=1}^{S} \phi_K(\mu_{\beta_s}, V_{\beta_s})\gamma(a_{\nu_s}, b_{\nu_s})\right] \cdot \left[\prod_{i=1}^{n}\left(\phi_S(\epsilon_i|0, \Sigma)\prod_{s=1}^{S} p_{CMP}(y_{is}|\epsilon_i, \beta_s, x_i, \nu_s)\right)\right] \quad (8)$$

MCMC simulation is a widely used method to draw samples from the posterior distribution of interest. MCMC starts with an approximate distribution, and corrects itself over its run to better approximate the target distribution. A sequence of samples are drawn in such a way that the new sample only depends on the previous sampled distribution, forming a Markov Chain (Gelman, Carlin, Stern, & Rubin, 2014). The main advantage of this method is it only requires the conditional distribution up to proportionality. After a certain burn-in, the algorithm will result in a sample from the posterior distribution of interest (Gelman, Carlin, Stern, & Rubin, 2014). There are several methods for creating a functional Markov Chain to estimate the

parameters from equation (8). Some common methods in the literature include Metropolis (Metropolis & Ulam, The Monte Carlo Method, 1949), MH (Hastings, 1970), and Gibbs (Geman & Geman, 1984) algorithms. As an enhancement to the original Metropolis algorithm, the MH algorithm has three steps: 1) Initialize the parameters with some starting value; 2) Choose a proposal distribution to generate candidate values; 3) Accept or reject the candidate values based on the calculated acceptance ratio, which evaluates if the newly generated value increases the probability of the conditional distribution or not (Hastings, 1970). Gibbs sampler is a special case of MH algorithm, and is generally used when the posterior distribution is some known distribution. In this study, a component-wise MH algorithm (Chib & Winklemann, 2001) is implemented.

3.4.2.2 Component-wise Metropolis-Hastings Algorithm

In a component-wise MH algorithm, instead of updating all the variables at a time, a component or a block of variables are updated using the full conditional distribution while other variables are held constant. Since the samples are drawn from the posterior distribution up to proportionality, only the terms pertaining to the parameters being updated are relevant in the full conditional distribution. This is because other terms are considered constant and can be dropped without affecting proportionality. The process is implemented iteratively until all parameter estimates converge. Applying this method, the parameter components to be estimated are $\epsilon|(y, x, \beta, \nu, \Sigma)$, $\Sigma|(\epsilon)$, $\beta|(y, x, \nu, \epsilon, \Sigma)$, and $\nu|(y, x, \beta, \epsilon, \Sigma)$. The iterative process is described below.

Step 1). Initialize the hyper-parameters for the priors and initialize the variables. Let $m = 0$ be the current iteration count. Denote the current values of the variables as $\beta[m], \nu[m], \Sigma[m], \epsilon[m]$.

Step 2). Sample $\Sigma[m+1]|\epsilon[m]$ using Gibbs sampler.

Step 3). Sample $\epsilon[m+1]|y, x, \beta[m], \nu[m], \Sigma[m+1]$ using MH algorithm

Step 4). Sample $\beta[m+1]|y, x, \nu[m], \epsilon[m+1], \Sigma[m+1]$ using MH algorithm

Step 5). Sample $\nu[m+1]|y, x, \beta[m+1], \epsilon[m+1], \Sigma[m+1]$ using MH algorithm

Step 6). Set $m = m + 1$. If m is less than a predetermined maximum number of iterations, stop; otherwise go to Step 2).

Steps 2 – 6 are usually repeated thousands of times to ensure the final sample resembles the posterior distribution of interest. The details of each sampling step are given below.

Step 2). Sampling Σ

Inverse Wishart distribution, $f_w(\tau_\Sigma, V_\Sigma)$, is used as a prior for the variance-covariance matrix parameter Σ, and the terms that pertain to Σfrom equation () are given below. Since inverse Wishart distribution is a conjugate prior (Press, 1982; Gelman, Carlin, Stern, & Rubin, 2014), a Gibbs sampler is implemented.

$$\pi(\Sigma[m+1]|\epsilon[m]) \propto f_w(\Sigma|\tau_\Sigma, V_\Sigma) \prod_{i=1}^{n} \phi_S(\epsilon_i[m]|0, \Sigma[m])$$

$$\pi(\Sigma[m+1]|\epsilon[m]) \propto f_w\left(n + \tau_\Sigma, \left[V_\Sigma + \sum_{i=1}^{n} (\epsilon_i'[m] \cdot \epsilon_i[m])\right]^{-1}\right)$$

Step 3). Sampling ϵ

The terms pertaining to ϵ in equation (8) are

$$\prod_{i=1}^{n}\left(\phi_S(\epsilon_i|0,\Sigma)\prod_{s=1}^{S}p_{CMP}(y_{is}|\epsilon_i,\beta_s,x_i,\nu_s)\right)$$

This conditional probability density is not given by any known density function, hence a MH algorithm is setup to generate a sequence of samples from the posterior distribution. Since ϵ_i is considered independent across the sites, each ϵ_i is sampled separately where values for ϵ_{-i} (any $\epsilon_j, j \neq i$) are held constant at their values from last iteration. In other words,

$$\pi(\epsilon_i|y_i,x_i,\beta,\nu,\Sigma,\epsilon_{-i}) \propto \phi_S(\epsilon_i|0,\Sigma)\prod_{s=1}^{S}p_{CMP}(y_{is}|\epsilon_i,\beta_s,x_i,\nu_s)$$

We further use $\tilde{Z}$ (equation (4)) to replace Z (equation (5)) in the CMP probability density function p_{CMP} (equation (6)). Also note that the term $(y_{is}!)^{\nu}$ in p_{CMP} does not pertain to ϵ and thus can be dropped without affecting proportionality. Therefore,

$$\pi(\epsilon_i[m+1]|y_i,x_i,\beta[m],\nu[m],\Sigma[m+1],\epsilon_{-i}[m]) \propto \phi_S(\epsilon_i|0,\Sigma[m])\prod_{s=1}^{S}\frac{(\lambda_{is})^{y_{is}}}{\tilde{Z}(\lambda_{is},\nu_s[m])}$$

where,$\lambda_{is} = \exp(x_i'\beta_s[m]+\epsilon_{is}[m])$.

A multivariate t-distribution with degrees of freedom d and location parameter D is used as a proposal density to generate a proposal vector ϵ_i^*. The degrees of freedom is used as tuning parameter to ensure a satisfactory acceptance rate. A proposal vector is accepted with probability

$$\min\left\{\frac{\pi(\epsilon_i^*|y_i,x_i,\beta,\nu,\Sigma)\;p(\epsilon_i|d,D)}{\pi(\epsilon_i|y_i,x_i,\beta,\nu,\Sigma)\;p(\epsilon_i^*|d,D)},1\right\}$$

where$p(\cdot\,|d,D)$ is the probability density of the proposal t-distribution.

Step 4). Sampling β

The terms pertaining to β in equation 8) are

$$\left[\prod_{s=1}^{S}\phi_K(\mu_{\beta_s}, V_{\beta_s})\right]\cdot\left[\prod_{i=1}^{n}\prod_{s=1}^{S}p_{CMP}(y_{is}|\epsilon_i, \beta_s, x_i, \nu_s)\right]$$

$$=\prod_{s=1}^{S}\left(\phi_K(\mu_{\beta_s}, V_{\beta_s})\prod_{i=1}^{n}p_{CMP}(y_{is}|\epsilon_i, \beta_s, x_i, \nu_s)\right)$$

Just as ϵ, this conditional probability density of β is not given by any known density function either. Therefore, a MH algorithm is implemented. Since β_s is considered independent across severities, each β_s is sampled separately where values for β_{-s} (any $\beta_j, j \neq s$) are held constant at their values from last iteration. In other words,

$$\pi(\beta_s|y, x, \beta_{-s}, \nu_s, \Sigma, \epsilon) \propto \phi_K(\beta_s|\mu_{\beta_s}, V_{\beta_s})\prod_{i=1}^{n}p_{CMP}(y_{is}|\epsilon_i, \beta_s, x_i, \nu_s)$$

Similarly, we use $\tilde{Z}$ to replace Z in the CMP probability density function p_{CMP}. Again, the term $(y_{is}!)^{\nu}$ in p_{CMP} does not pertain to the parameter of interest β, and can thus be dropped. Therefore,

$$\pi(\beta_s[m+1]|y_i, x_i, \epsilon_i[m+1], \nu_s[m], \Sigma[m+1], \beta_{-s}[m]) \propto \phi_K(\beta_s|\mu_{\beta_s}, V_{\beta_s})\prod_{i=1}^{n}\frac{(\lambda_{is})^{y_{is}}}{\tilde{Z}(\lambda_{is}, \nu_s[m])}$$

where $\lambda_{is} = \exp(x_i'\beta_s[m] + \epsilon_{is}[m+1])$.

A multivariate normal distribution with mean vector u_{β_s} and variance B_{β_s} is used as a proposal density to generate a proposal vector β_s^*. The mean vector u_{β_s} is used as tuning parameter to ensure a satisfactory acceptance rate. A proposal is accepted with the probability

$$\min\left\{\frac{\pi(\beta_s^*|y_i,x_i,\epsilon_s\ ,\nu,\sum)\ p(\beta_s|u_{\beta_s},B_{\beta_s})}{\pi(\beta_s|y_i,x_i,\epsilon_s,\nu,\sum)\ p(\beta_s^*|u_{\beta_s},B_{\beta_s})},1\right\}$$

Step 5) Sampling ν_s

The terms pertaining to ν in equation () are

$$\left[\prod_{s=1}^{S}\gamma(a_{\nu_s},b_{\nu_s})\right]\cdot\left[\prod_{i=1}^{n}\prod_{s=1}^{S}p_{CMP}(y_{is}|\epsilon_i,\beta_s,x_i,\nu_s)\right]$$

$$=\prod_{s=1}^{S}\left(\gamma(a_{\nu_s},b_{\nu_s})\prod_{i=1}^{n}p_{CMP}(y_{is}|\epsilon_i,\beta_s,x_i,\nu_s)\right)$$

Similar to sampling for $\epsilon$and β, a MH algorithm is setup and each ν_s is sampled separately where values for ν_{-s} (any $\nu_j, j\neq s$) are held constant at their values from last iteration. In other words,

$$\pi(\nu_s|y,x,\nu_{-s},\beta_s,\Sigma,\epsilon)\ \propto\gamma(a_{\nu_s},b_{\nu_s})\prod_{i=1}^{n}p_{CMP}(y_{is}|\epsilon_i,\beta_s,x_i,\nu_s)$$

Approximating Z by $\tilde{Z}$ and dropping the constant term $\lambda_{is}^{y_{is}}$in p_{CMP}, we have

$$\pi(\nu_s[m+1]|y,x,\epsilon[m+1],\beta_s[m+1],\Sigma[m+1],\nu_{-s}[m])$$

$$\propto\gamma(a_{\nu_s},b_{\nu_s})\prod_{i=1}^{n}\frac{1}{\tilde{Z}(\lambda_{is},\nu_s[m])(y_{is}!)^{\nu_s[m]}}$$

where$\lambda_{is}=\exp(x_i'\beta_s[m+1]+\epsilon_{is}[m+1])$

A univariate gamma distribution with shape parameter c and scale parameter h is used as a proposal density to generate a proposal value ν_s^*.c and $h$are considered as tuning parameters to ensure a satisfactory acceptance rate. A proposal is accepted with the probability

$$\min\left\{\frac{\pi(\nu_s^*|y,x,\epsilon\ ,\beta_s,\Sigma)\ p(\nu_s|c,h)}{\pi(\nu_s|y,x,\epsilon,\beta_s,\Sigma)\ p(\nu_s^*|c,h)},1\right\}$$

3.5 Numerical Example

3.5.1 Setting

To illustrate, validate and evaluate the performance of the proposed approach, a numerical example is first carried out using simulated data. A multivariate dataset is simulated from the proposed MVCMP distribution such that the frequencies of some severities are over-dispersed and others are u nder-dispersed. The simulated data consists of 2000 observations and four covariates, resulting in a 2000-by-4 matrix of x's; and five crash severities, resulting in a 4-by-5 matrix of β's and five ν's. The inversed cumulative distribution function (CDF) method is used for generating the random variables and a detailed procedure is given below.

First, the values of all four covariates are simulated from a uniform distribution between 0 and 1 for all 2000 data points. Fixed arbitrary values are assigned to the β matrix. As ν's represents the dispersion in the data, they are assigned values such that the datasets represent both under- and over- dispersed data conditions. In this study, three ν's are assigned values greater than one (under-dispersion) and the remaining two are assigned values less than one (over-dispersion). A positive definite matrix is assigned as Σ. Based on the assigned Σ, a multivariate vector ϵ is generated for each of 2000 observations. Using the covariates, the parameter values and ϵ, the MVCMP distribution parameter λ is computed for each of the 2000 data points. The function $\tilde{Z}(\lambda, \nu)$ is then computed for each data point following equation (4). Finally, the observation y's are computed using the inversed-CDF method. Since the conditional marginal MVCMP is a CMP distribution, which corresponds to a discrete random variable, the cumulative distribution can be easily calculated. A detailed procedure for generating the MVCMP random variable is given below.

Let's assume the univariate random variable Y follows CMP distribution with parameters (λ, ν) and probability $p(Y = i) = \frac{\lambda^i}{(i!)^\nu \tilde{Z}(\lambda,\nu)}$, for $i = 0,1,2, \ldots$. The key to the use of inverse transform method is $p(Y = i + 1) = \frac{\lambda^{i+1}}{((i+1)!)^\nu \tilde{Z}(\lambda,\nu)} = \frac{\lambda}{(i+1)^\nu} p(Y = i)$. Using this recursive property, one can generate a random variable by following the 5 step procedure give below for each random observation.

Step1) Generate a random number U from [0, 1]

Step2) For $i = 0, p(Y = i) = \tilde{Z}(\lambda, \nu)$, and the cumulative distribution$F = p(Y = i)$

Step3) If $U < F$, set $y = i$ and stop otherwise go to step 4. Here, y implies the realization.

Step4) Compute $p(Y = i + 1) = \frac{\lambda}{(i+1)^\nu}$, $F = F + p(Y = i + 1)$ and $i = i + 1$

Step5) Go to Step 3

After simulating the dataset, the proposed component-wise MCMC algorithm is run to estimate the parameters β, ν, ϵ and Σ. The initial values are assigned randomly. Vague prior distributions are used to check the validity of model in case of no prior information about the variables. The algorithm is a success, if the estimated values of β's, ν's, ϵ's and Σ are close to the assigned true values.

3.5.2 Convergence Tests

Figure 2 shows the typical trace and running mean plots of the dispersion parameter ν for two selected severity levels: non-incapacitating crashes (ν_A) and property damage only (ν_0). Note that non-incapacitating crashes are under-dispersed and property damage only crashes are

over-dispersed. From the trace plots (Figure 2 a & b), it is observed that the chain has stabilized after about 5000 iteration and the randomness in the chain shows that the chain was not stuck at any point for a large number of iterations. Hence it can be concluded that based on the trace plots the chain has explored the parameter space well.

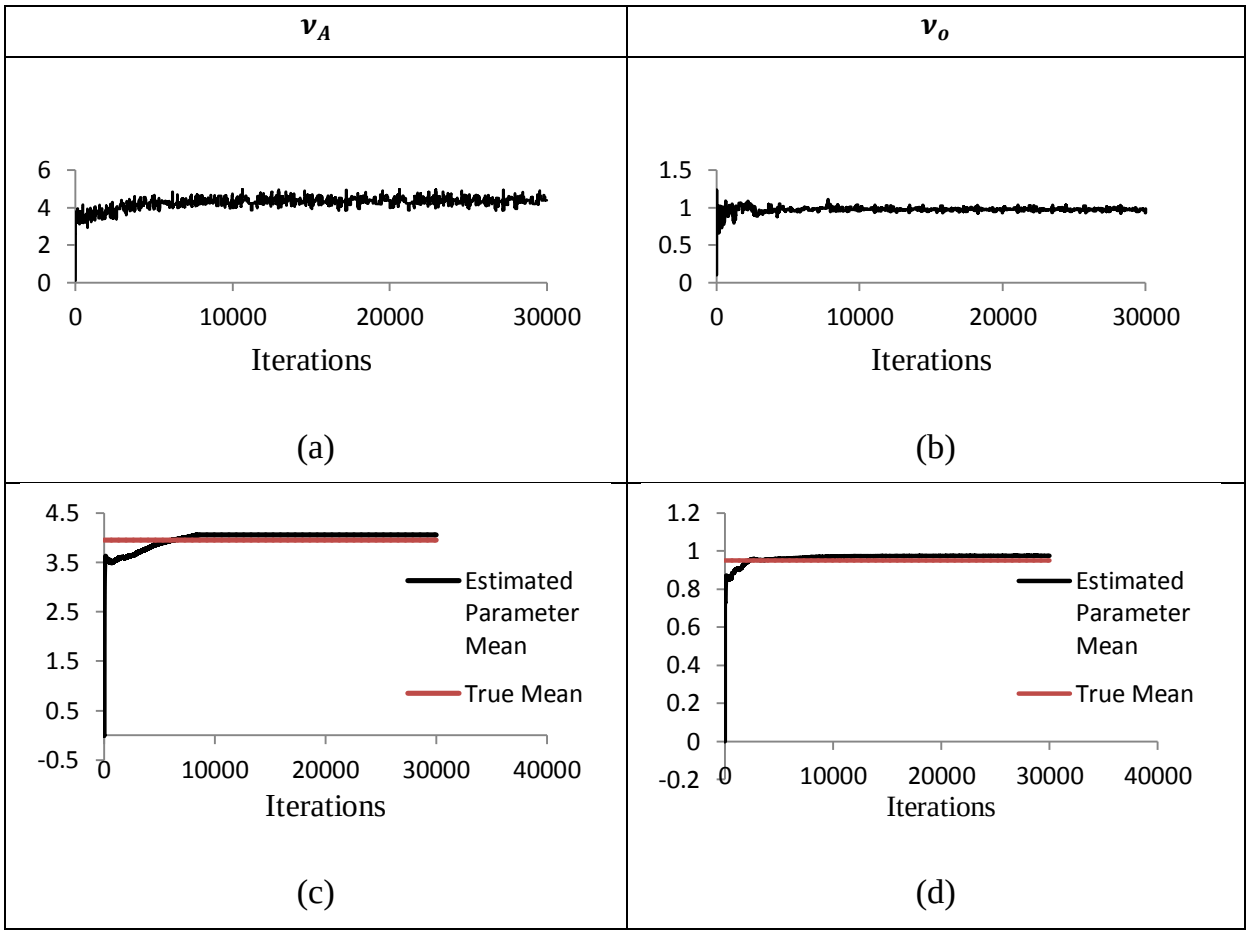


Figure 2 Trace Plot (a, b) and Running Mean Plot (c, d) for dispersion coefficients of non-incapacitating crashes and property damage only crashes.

The running means plot (Figure 2 c&d) shows that both parameter estimates are steadily converging to their true values as the number of iterations increase. Similar analysis was performed for each of the 20 β's, 5 ν's and 25 elements of Σ. It was observed that all the parameters had well-mixing chains all of which converged at about 15000 iterations. The

graphical analysis for ϵ's is not possible since there are 10,000 of them. However, the ϵ convergence was monitored based on the acceptance rate. It was ensured the acceptance rate was around 20%, which is a commonly accepted rate for higher dimensional problems (Gelman, Carlin, Stern, & Rubin, 2014). The tests proved that the algorithm is capable of estimating parameters with extra Poisson dispersion and is able to converge to the target distribution after a sufficiently large number of iterations.

3.5.3 Estimation Results

Table 6reports the parameter values obtained from two separate trials. The first trial consisted of 20,000 iterations and the second trial consisted of 40,000 iterations. The first 10,000 iterations of both the trials are discarded as burn-in. The table includes the assigned true coefficient values, estimated mean parameter value, standard error of the mean and the High Density Region (HDR) estimates given by the 2.50^{th}- and 97.5^{th}-percentile of the simulated sample values. It can be seen that both chains have converged to almost similar posterior distributions for all the parameters, and the estimated values are very comparable to the assigned true values. The two percentile values provide a 95% equal-tail credible interval for each set of parameter estimates. The probability of the estimates being statistically significant (non-zero) is 95%, if the 95% credible interval does not contain zero (Gelman, Carlin, Stern, & Rubin, 2014). It can be seen from the table that none of the 95% credible interval of the estimated parameters contain zero. For Trial 1 (20,000 iterations), most estimates are very close to the assigned true values except ν_K, ν_A, and β_{O_0}. For Trail 2 (40,000 iterations), all the parameter estimates converged very well to their assigned true values. This shows that the proposed component-wise MCMC algorithm is able to reasonably converge to the posterior distribution after 20,000 iterations. But more iterations may improve the quality of the estimates. Table 7 gives the

estimated value of variance-covariance matrix of the error term ϵ. The estimated matrix is close to the assigned values, showing that the model is efficient enough to catch even this small covariance's between different severities. All the elements of the matrix are found to be significantly different than 0.

Table 6 Parameter Estimates for Numerical Example

Variables	Assigned Coeffts	Trial 1				Trial 2			
		Mean	SE	2.50%	97.50%	Mean	SE	2.50%	97.50%
Fatal Crashes (K)									
β_{Ko}	-5.49	-5.94	0.0026	-5.96	-4.70	-5.59	0.0007	-5.85	-5.46
β_{K1}	-1.35	-0.98	0.0012	-1.58	-0.98	-1.45	0.0009	-1.75	-1.40
β_{K2}	1.61	1.36	0.0014	0.85	1.51	1.52	0.0004	0.50	0.68
β_{K3}	1.33	1.01	0.0017	0.86	1.63	1.29	0.0007	1.09	1.42
ν_K	1.65	2.39	0.0024	1.08	2.67	1.69	0.0004	1.21	1.89
Incapacitating Crashes (A)									
β_{Ao}	-5.96	-5.38	0.0013	-5.43	-4.66	-5.31	0.0013	-5.61	-4.97
β_{A1}	-2.68	-1.93	0.0013	-2.25	-1.56	-1.98	0.0014	-2.33	-1.64
β_{A2}	1.89	1.37	0.0015	0.89	1.81	1.79	0.0011	1.08	2.00
β_{A3}	1.31	1.19	0.0012	1.00	1.62	1.38	0.0014	1.10	1.63
ν_A	3.85	4.36	0.0029	3.22	4.94	4.03	0.0009	3.66	4.40
Non-Incapacitating Crashes (B)									
β_{Bo}	-5.20	-4.99	0.0023	-5.34	-3.78	-4.96	0.0034	-5.58	-4.66
β_{B1}	-1.93	-1.37	0.0014	-1.91	-1.06	-1.99	0.0001	-2.49	-1.70
β_{B2}	1.80	0.94	0.0014	0.94	1.58	1.91	0.0004	1.41	2.11
β_{B3}	2.30	1.69	0.0013	1.26	1.97	1.77	0.0005	1.37	1.98
ν_B	0.80	0.75	0.0001	0.72	0.85	0.85	0.0642	0.74	0.88
Possible Injury Crashes (C)									
β_{Co}	-5.29	-4.94	0.0022	-5.03	-4.20	-5.23	0.0004	-5.86	-5.14
β_{C1}	-1.88	-1.44	0.0013	-2.37	-1.34	-1.55	0.0017	-1.92	-1.43
β_{C2}	1.80	1.75	0.0007	1.54	1.79	1.56	0.0009	1.30	1.75
β_{C3}	1.44	1.93	0.0019	1.40	2.07	1.52	0.0001	1.31	1.63
ν_C	1.50	1.22	0.0008	0.89	1.43	1.38	0.0001	0.95	1.49
Property Damage only Crashes (O)									
β_{Oo}	-5.91	-4.53	0.0007	-4.59	-4.53	-5.81	0.0002	-6.21	-5.20
β_{O1}	-2.87	-2.25	0.0014	-2.74	-1.65	-2.78	0.0015	-2.91	-1.54
β_{O2}	1.91	1.65	0.0011	1.50	2.24	1.72	0.0003	1.69	1.78
β_{O3}	1.23	1.37	0.0005	1.22	1.40	1.48	0.0016	1.21	1.62
ν_O	0.90	0.95	0.0001	0.91	0.98	0.97	0.0001	0.91	1.00

Table 7 Assigned and Estimated Variance-covariance Matrix

Actual Variance-covariance Matrix					Estimated Variance-covariance Matrix				
.0176	.0014	.0011	.0180	.0040	.0358	.0028	.0027	.0195	.0035
.0014	.0648	.0063	.0099	.0018	.0028	.1630	.0117	.0096	.0080
.0011	.0063	.2360	.0250	.0708	.0027	.0117	.2018	.0119	.0540
.0185	.0099	.0250	.4051	.2526	.0195	.0096	.0119	.3007	.1062
.0040	.0018	.0708	.2523	.1246	.0035	.0080	.0540	.1062	.2580

3.6 Case Studies

3.6.1 Application on Two-Lane Rural Roads

For a real-world example, this study analyzes the two-lane rural road crash dataset compiled by the research team in a previous study (Mehta & Lou, 2013). Raw data was obtained from Critical Accident Reporting Environment system maintained by Center for Advance Public Safety at The University of Alabama. The processed dataset consists of homogeneous roadway sites ranging from 0.05 miles to 6 miles. A sample of 800 sites is selected randomly from the processed dataset used in (Mehta & Lou, 2013). Care is taken to ensure that the sample is representative of the population. The analysis includes five crash severities. Basic statistics of the crashes of each severity level are reported in Table 8. Six covariates are considered in this analysis, namely annual average daily traffic (AADT), segment length (SL), speed limit, shoulder width, shoulder type and highway class.

Table 8 Summary Statistics for Two-lane Rural Road Data Set

Crashes	Min	Max	Average (SD*)	Total Crashes
Fatal	0	2	0.026 (0.0076)	83
Incapacitating	0	4	0.24 (0.0246)	140
Non-Incapacitating	0	2	0.038 (0.0092)	46
Possible Injury	0	1	0.042 (0.0089)	53
Property Damage Only	0	6	0.5580 (0.0426)	356

* Denotes standard deviation

3.6.2 Model Estimation

The MVCMP model is estimated using the proposed Bayesian formulation and the component-wise MH algorithm. The initial values for both β and ν are set as results from MLE of individual univariate CMP distribution for each crash severity. The starting value of Σ is set as an identity matrix. The prior distributions used for each parameters to be estimated are $\beta_s \sim N_s(0,\ 0.1I)$, $\nu_s \sim \gamma(0.03,0.1)$, $\Sigma \sim f_w(20, 10I)$, and $\epsilon = [0]$. The component-wise MH algorithm is run for 40,000 iterations with the first 10,000 iterations discarded as burn-in. The trace and running mean plots are produced for each parameter to check their convergence. A typical trace plot and running mean plot for the coefficient of AADT and SL for fatal crash severity is shown here. The trace plots for both the parameters demonstrate good mixing, implying that the parameter space is well explored. From the running mean plots, it can be observed that the coefficient for AADT took almost 20,000 iterations to converge to a stationary distribution. On the other hand, the coefficient for SL converged in just about 10,000 iterations. These plots show that 40,000 iterations seem to be a reasonable balance between better estimation quality and computation time. However, this might not be the case for other studies using different data sets. Again, visual tests are not conducted for the parameter ϵ's. Table 9 reports the parameter estimates and

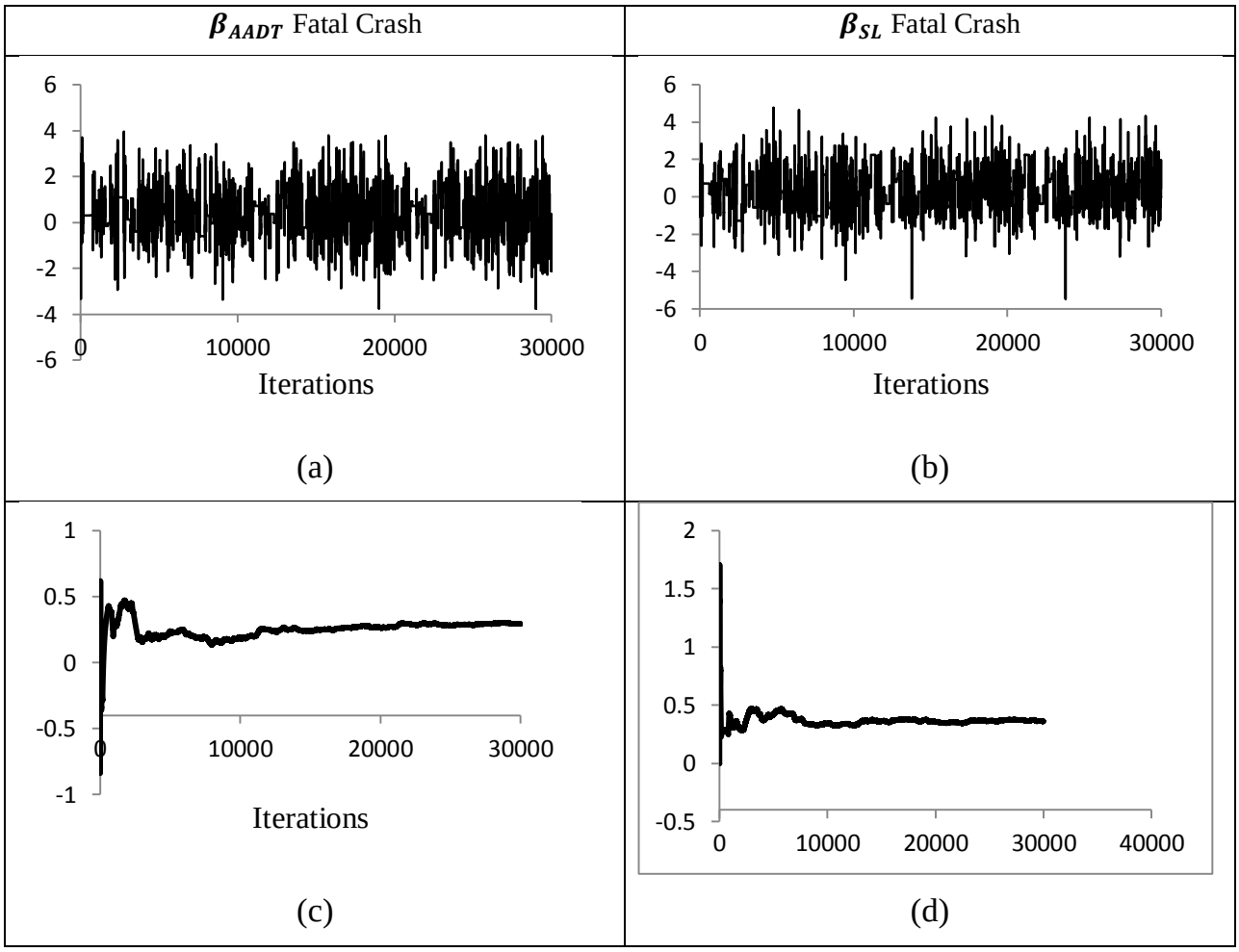


Figure 3 Trace plots and running mean plots for the coefficient of fatal crash parameters AADT and segment length(SL)

Table 9 Parameter Estimates of MVCMP

Severities	Variables	MVCMP		HDR	
		Mean	SE	2.50 %	97.5%
Fatal Crashes	Const	**-7.8500***	0.0128	-10.8484	-4.8224
	AADT	**0.3327**	0.0080	0.1399	0.5940
	SL	**0.2767**	0.0087	0.0591	2.2456
	SpeedLimit	**0.2847**	0.0072	0.04142	0.4045
	Shold Width	0.3374	0.0092	-1.8460	2.5239
	Shold Type	0.0744	0.0060	-1.4230	1.5371
	Highway Class	0.2217	0.0065	-1.2776	1.8696
	Nu	1.9169	0.0042	1.0993	3.9466
Incapacitating Crashes	Const	**-8.0079**	0.0115	-10.6817	-5.2552
	AADT	**0.3837**	0.0096	0.0866	0.7944
	SL	**0.3996**	0.0104	0.0759	0.9323
	SpeedLimit	0.3434	0.0093	-1.8319	2.6578
	Shold Width	0.4336	0.0105	-1.9868	3.1546
	Shold Type	0.2937	0.0094	-1.8456	2.6503
	Highway Class	0.4396	0.0093	-1.7547	2.7228
	Nu	0.2946	0.0013	0.0001	0.9084
Non Incapacitating Crashes	Const	**-8.0203**	0.0106	-10.2373	-5.7039
	AADT	**0.3030**	0.0073	0.2468	0.4616
	SL	**0.2660**	0.0077	0.1912	0.4128
	SpeedLimit	0.2552	0.0070	-1.4046	1.7981
	Shold Width	0.1355	0.0073	-1.6322	2.0514
	Shold Type	0.0529	0.0050	-1.2135	1.2217
	Highway Class	0.1521	0.0052	-0.9429	1.6442
	Nu	2.7812	0.0051	1.2817	4.5633
Possible Injury	Const	**-8.0105**	0.0115	-10.6417	-5.2325
	AADT	**0.3813**	0.0094	0.0844	0.5608
	SL	**0.3544**	0.0100	0.0975	0.4330
	SpeedLimit	0.3164	0.0089	-1.7661	2.4490
	Shold Width	0.4270	0.0107	-2.1305	2.9153
	Shold Type	0.2595	0.0090	-1.8332	2.5280
	Highway Class	0.3942	0.0092	-1.7941	2.6264
	Nu	0.6736	0.0028	0.0003	0.8994
Property Damage Only	Const	**-8.0324**	0.0119	-10.9440	-7.1553
	AADT	**0.4121**	0.0092	0.0711	0.6735
	SL	**0.4122**	0.0102	0.2890	0.5439
	SpeedLimit	0.3317	0.0092	-1.7883	2.5528
	Shold Width	**0.3683**	0.0108	0.1404	0.6710
	Shold Type	0.2678	0.0091	-1.8477	2.4989
	Highway Class	0.3321	0.0094	-1.8377	2.5914
	Nu	0.6340	0.0026	0.0003	1.9146

*Note: Variables in bold indicate statistically significant variables.

Table 10 Variance-Covariance matrix Σ

1.1135	0.4132	0.7502	0.1841	0.1896
0.4132	1.2131	0.3533	0.5643	0.7028
0.7502	0.3533	1.1542	0.5640	0.0945
0.1841	0.5643	0.5640	1.0756	0.7463
0.1896	0.7028	0.0945	0.7463	1.0420

3.6.3 Results Interpretation

The parameter estimates obtained from the MVCMP specifications should not be directly compared with the univariate CMP or other distributions. This is because MVCMP specifications accounts for correlation across crash severities, which is not considered in univariate models. However, comparing the signs of the parameters could be justified.

The nature or the parameter estimates across all crash severities is consistent and logical. It is observed that AADT and SL are statistically significant across all five severities. The coefficient of AADT is positive for all five severity levels. This implies that the increase in traffic volume leads to increase in the expected number of crashes for every severity. However, the magnitude of the impact of AADT varies across the severities. Similarly, SL is also positive across all the severities, implying an increase in expected number of crashes as the length of the roadway segment increases. The speed limit variable is significant only for the fatal crash severity. This is very intuitive, because as speed increases the possibility of damage during collision increases and hence the fatality increases. This observation also implies that there is no correlation between speed limit and crashes of other severity types.

One would expect that as the shoulder width increases the expected number of crashes on that road should decrease. However, in this case study, shoulder width has significant impact only on the property damage only crashes. The positive sign of the parameter estimate is

counter-intuitive. One probable reason could be that confounding factors, such as locations where wider shoulders are provided may explain this outcome. In other words, the sections of roads with high probability of property damage only crashes (for example, running out of the road) could be a reason for providing wider shoulders in the first place. This finding is worth of further investigation.

Finally, the variance-covariance matrix (

Table 10) is not a diagonal matrix, indicating that the crashes of different severities taking place on a particular roadway segment are not independent. All the off-diagonal elements of the variance co-variance matrix are positive, suggesting positive dependence among different crash severities. This finding is intuitive because certain properties of a site (such as sharp curve or poor drainage) can lead to increase in all types of crash severities. It is also consistent with the findings from Ma et al. (2008). Additional factors for such correlation may include pavement quality, sight distance, demography etc.(El-Basyouny & sayed, 2009). These unobserved factors and the correlation they cause cannot be modeled by univariate models, making the multivariate models superior in this sense.

3.7 Discussion

There are many implementation issues associated with Bayesian analysis as described by Gelman et al. (2014). Some implementation issues of the proposed component-wise MCMC approach particularly faced in this study are discussed here. Firstly, it is observed from both our artificial numerical example and the real case study that the use of unreasonably high or low starting values can result in the chains being stuck. Take the real case study for example, when the starting value of constant coefficient for any severity is set at -25 or higher, the acceptance rate for that parameter was 0. This is because the likelihood of the current point became negative infinity, resulting in no movement of the chain. The same issue also exists with prior and proposal distributions. Very unrealistic prior and proposal distributions results in very slow chain movement with acceptance rate of less than 1%. Secondly, it is also advised to use log transformation while computing likelihood. This is because most of the times the simulated values can go to extreme; with the exponential function of parameter λ, the likelihood value would approach infinity. Finally, identification of the proposal distribution is one of the big challenges in this method. A good discussion on developing different kinds of proposal distribution is given by Chib et al.(Chib & Winklemann, 2001). This study has adopted Random Walk proposal distribution, which requires identifying the hyper-parameters through trial runs to obtain suitable candidates. With recent advancements in the MCMC simulation techniques, several adaptive MCMC techniques are proposed, which can tune the proposal distribution itself, based on certain criteria (Roberts & Rosenthal, 2009). The efficiency of such methods with this formulation could be an interesting topic for future research.

3.8 Conclusion

This study presents a new multivariate approach for modeling crash counts by severity. More specifically, a formulation of Multivariate Conway-Maxwell Poisson distribution is developed for the first time with the assumption that the conditional distributions are univariate Conway-Maxwell Poisson. A heterogeneous error term is introduced to capture the correlation among crash frequencies of different severities caused by unobserved common factors, such as lighting or presence of vertical curve. The dispersion parameters are allowed to vary across the severities. This setting gives flexibility to the model for accommodating crash counts severities displaying both over- and under-dispersion. The parameter estimated with consideration of correlation could be more precise. This feature of the model can help avoid serious errors in analysis, which can be caused by misspecification of the distributional assumptions while using Poisson or Negative Binomial regression models.

A component-wise Markov Chain Monte Carlo algorithm with both Gibbs and Metropolis Hastings samplers is proposed and coded in Matlab to estimate the parameters following the Bayesian paradigm. Bayesian paradigm considers parameters as random variables rather than fixed values. This helps in associating uncertainties with the parameter rather than data. Another notable positive of this approach is that it allows the use of prior information. If new data is available in the future, the results from this study can be used as priors to refine the model. The Markov Chain Monte Carlo approach draws samples directly from the posterior distribution, avoiding expensive numerical evaluation of high-dimensional integral that does not have closed-form functions. The algorithm requires starting values for all the parameters and hyper-parameters for the prior distributions. The priors could be informative or flat. This study

used flat priors to study the behavior of the algorithm with no prior information. The results showed that the algorithm performs well despite the lack of prior knowledge.

The proposed MVCMP assumes no spatial correlation across different roadway sites. This is a very strong assumption as driving conditions and road users' driving behaviors vary from region to region. There can be some common unobserved variables specific to a particular neighborhood or jurisdictions. Additionally, the proposed distribution is established based on parametric assumptions of crash severities conditioned on lognormal errors following CMP distribution. Misspecifications of the distributional assumptions can lead to incorrect analysis. An alternative to this problem could be non-parametric specification, whichis a good topic for future research. The comparison of the performance of this model with the other multivariate models like multivariate Poisson-lognormal model and Double Poisson distribution will be an interesting topic. Moreover, Lord and Miranda-Moreno (Lord & Miranda-Moreno, 2008) found that the parameter estimates with univariate CMP can be affected when the sample mean is low. So another extension of this study could be an investigation of the efficiency of the proposed MVCMP with different levels of sample means.

3.9 References

Aitchison, J., & Ho, C. (1989). The Multivariate Poisson-Lognormal Distribution. Biometrika, 76(4), 643-653.

Bolstad, W. (2010). Understandting Computational Bayesian Statistics. Hoboken: John Wiley & Sons Inc.

Castillo, J., & Perez-Casany, M. (2005). Over-Dispersed and Under-Dispersed Poisson Generalizations. Journal of Statistical Planning and Inference, 134(2), 496-500.

Chib, S., & Greenberg, E. (1996). Markov Chain Monte Carlo Simulation Methods in Econometrics. Econometric Theory, 409-431.

Chib, S., & Winklemann, R. (2001). Markov Chain Monte Carlo Analysis of Correlated Count Data. Journal of Business and Economic Statistics, 19(4), 428-435.

Consul, P. (1989). Generalized Poisson Distributions: Properties and Applications. New York: Marcel Dekker.

Cowles, M., & Carlin, B. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparitive Review. American Statistical Association, 91(434), 883-904.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, 39(1), 1-38.

Dong, C., Richards, S., Clarke, D., Xuemei, Z., & Ma, Z. (2014). Examining Signalized Intersection Crash Frequency Using Multivariate Zero-Inflated Poisson Regression. Safety Science, 70, 63-69.

El-Basyouny, K., & sayed, T. (2009). Collision Prediction Models using Multivariate Poisson-Lognormal Regression. Accident Analysis and Prevention, 41(4), 820-829.

Famoye, F. (1993). Restricted Generalized Poisson Regression Model. . Communications in Statistics-Theory and Methods, 22(5), 1335-1354.

Gelman, A., & Rubin, D. (1992). Inference from Iterative Simulations Using Multiple Sequences . Statistical Science, 7, 457-511.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2014). Bayesian Data Analysis. Boca Raton, Florida: Chapman & Hall/CRC.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 12, 609-628.

Geweke, J. (1992). Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. Oxford: Oxford University Press.

Hastings, W. (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika, 57, 97-109.

Hogg, R., Craig, M., & McKean, J. (2006). Introduction to Mathematical Statistics. Pearson Prentice Hall.

Joshua, S., & Garber, N. (1990). Estimating Truck Accident Rate and Involvements Using Linear and Poisson Regression Models. Transportation Planning and Technology, 15(1), 41-58.

Jovanis, P., & Chang, H. (1986). Modeling the Relationship of Accidents to Miles Travelled. Transportation Research Record, 1068, 42-51.

Kim, D., & Washington, S. (2006). The Significance of Endogeneity Problems in Crash Models: An Examination of Left-Turn Lanes in Intersection Crash Models. Accident Analysis and Prevention, 38(6), 1094-1100.

Lord, & Mannering. (2010). The Statistical Analysis of Crash Frequency Data: A Review and Assessment of Methodological Alternatives. Transportation Research Part A, 44(5), 291-305.

Lord, D., & Mannering, F. (2010). The statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. Transportation Research Part A: Policy and Practice, 44(5), 291-305.

Lord, D., & Miranda-Moreno, L. (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(5), 751-770.

Lord, D., Geedipally, S., & Guikema, S. (2010). Extension of The Application of Conway Maxwell Poisson Models: Analyzing traffic Crash Data Exhibiting Underdispersion. Risk Analysis, 30(8), 1268-1276.

Lord, D., Guikema, S., & Geedipally, S. (2008). Application of the Conway-Maxwell-Poisson Generalized Linear Model for Analyzing Motor Vehicle Crashes. Accident Analysis and Prevention, 40(3), 1123-1134.

Lord, D., Washington, S., & Ivan, J. (2005). Poisson, Poisson-Gamma and Zero Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory. Accident Analysis and Prevention, 37(1), 35-46.

Lord, D., Washington, S., & Ivan, J. (2007). Further Notes on The Application of Zero Inflated Models in Highway Safety. Accident Analysis and Prevention, 39(1), 53-57.

Ma, J., & Kockleman, K. (2006). Bayesian Multivariate Poisson Regression for Models of Injury Count, By Severity. Transportation Research Record, 1950, 24-34.

Ma, J., Kockelman, K., & Damien, P. (2008). A Multivariate Poisson-Lognormal Regression Model for Prediction of Crash Counts by Severity, Using Bayesian Methods. Accident Analysis and Prevention, 40(3), 964-975.

Maher, M. (1990). A Bivariate Negative Binomial Model to Explain Traffic Accident Migration. Accident Analysis and Prevention, 22(5), 487-498.

Mehta, G., & Lou, Y. (2013). Safety Performance Function Calibration and Development for the State of Alabama: Two-Lane Two-Way Rural Roads and Four-Lane Divided Highways. Transportation Research Record(In Press).

Metropolis, N., & Ulam, S. (1949). The Monte Carlo Method. American statistical Association, 44, 335-341.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., & Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. The Journal of Chemical Physics, 1087.

Miaou, S. (1994). The Relationship Between Truck Accidents and Geometric Design of Road Sections: Poisson versus Negative Binomial Regressions. Transportation Research Record, 26(4), 471-482.

Mitra, S., & Washington, S. (2007). On the Nature of Over-dispersion in Motor Vehicle Crash Prediction Models. Accident Analysis and Prevention, 39(3), 459-468.

Park, E., & Lord, D. (2007). Multivariate Poisson-lognormal Models for Joint Modeling of Crash Frequency by Severity. Transportation Research Record, 1-6.

Press, S. (1982). Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference (2nd ed.). Malabar, Florida: Robert E. Krieger Publishing Company.

Raftery, A., & Lewis, S. (1992). How Many Iterations in the Gibbs Sampler? Oxford: Oxford University Press.

Ritter, C., & Tanner, M. (1992). Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy Gibbs Sampler. American Statistical Association, 87, 861-868.

Roberts, G. (1992). Convergence Diagnostics of the Gibbs Sampler. Oxford: Oxford University Press.

Roberts, G., & Rosenthal, J. (2009). Examples of Adaptive MCMC. Journal of Computational and Graphical Statistics, 18(2), 349-367.

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

Shmueli, G., Minka, T., Kadane, J., Borle, S., & Boatwright, P. (2005). A Useful Distribution for Fitting Discrete Data: A Revival of Conway Maxwell Poisson Distribution. Journal of the Royal Statistical Society, 54(1), 127-142.

Shoukri, M. (1982). On the Generalization and Estimation for The Double Poisson Distribution. Trabajos de Estadistica Y de Investigacion Operativa, 33(2), 97-109.

Stigler, S. (1983). Who Discovered Bayes's Theorem. american Statistician, 290-296.

Subrahmaniam, K., & Subrahmaniam, K. (1973). On the Estimation of The Parameters in the Bivariate Negative Binomial Distribution. Journal of Royal Statistical Society, 131-146.

Tegge, R., Jo, J., & Ouyang, Y. (2010). Development and Application of Safety Performance Functions for Illinois. Illinois Department of Transportation.

Valverde, J., & Jovanis, P. (2008). Analysis of Road Crash Frequency with Spatial Models. Transportation Research Record(2061), 55-63.

Winklemann, R. (1995). Duration Dependence and Dispersion in Count Data Models. Journal of Business and Economic Statistics, 13, 467-474.

Winklemann, R. (2008). Econometric Analysis of Count Data (Fifth Eddition). Berlin: Springer-Verlag.

Wolke, R., & Schwetlick, H. (1988). Iteratively Reweighted Least Squares: Algorithms, Convergence Analysis, and Numerical Comparisons. Journal of Scientific and Statistical Computing, 9(5), 907-921.

Zou, Y., Geedipally, S., & Lord, D. (2013). Evaluating The Double Poisson Generalized Linear Model. Accident Analysis and Prevention, 59, 497-505.

Chapter 4.

EVALUATING THE PERFORMANCE OF MULTIVARIATE CONWAY-MAXWELL POISSON DISTRIBUTION FOR TWO-LANE RURAL ROADS AND BRIDGES

4.1 Introduction

The recent publication of the Highway Safety Manual (HSM) has resulted in an increasing emphasis on the safety performance of different roadway facilities (AASHTO, 2010). In particular, the HSM sets out the development of facility-specific safety performance functions (SPFs) for predicting the expected number of crashes on various components of the roadway system. The first edition of the HSM contains SPFs for three facility types. Additional SPFs have been developed for two-lane roads, freeways, intersections, and other facilities since then(e.g., Quin et al., 2005; Parajuli, et al., 2006; Garber, 2010; Sun et al., 2011; Manan, et al., 2013).The HSM currently provides univariate models for estimating the total crash frequency on road segments for two-lane rural roads(HSM, 2010). The HSM provides factors, which, when multiplied with the expected average total crash frequency, estimates the expected crash frequencies for different severities. However, this method of predicting frequency of crash severities suffer from the same limitation of common unobserved heterogeneities or missing information across severities as described in Section 3.3 of Chapter 3. This study therefore advocates the use of the Multivariate Conway Maxwell Poisson (MVCMP) distribution for jointly estimating the expected number of crashes across different severities. The study

evaluates the performance of the proposed MVCMP formulation on the two-lane rural road dataset described in Chapter 2.

Further, this study extends the proposed MVCMP to bridges by developing a joint crash severity model for bridge sections. This study is the first of its kind to develop a safety performance function for different crash severities on bridges.

Bridges are an important component of the transportation infrastructure in the U.S, with roughly 17,000 bridges just in the state of Alabama (FHWA 2013). However, none of the studies so far have considered bridges as a different identity from the regular roadway segments. The physical properties and operational characteristics of bridges differ substantially from roads and highways (Retting, et al., 2000). For example, bridges are usually narrower than approach roads and often lack shoulders that can safely accommodate emergency maneuvers. Bridges consist of various components (e.g., bridge railings, piers, headwalls, abutments and guardrails) that may themselves pose safety hazards to vehicles travelling on or under them (Gates &Noyce, 2005). Bridge decks are designed to accommodate certain load-bearing capacities and other specific properties that often render safety improvements such as resurfacing (e.g., asphalt overlays) impractical (Retting, et al., 2000). Such bridge-specific characteristics, then, affect the safety of traffic traveling along and near bridges.

All the previous bridge safety studies have focused on structural properties (Crespo-Minguillión et al., 1997; Rocha &Calcada, 2012; Zhao & Uddin, 2013; Ju, 2013) and the performance of bridge components such as railings (Michie, 1981; Thanh&Itoh, 2013; Hirai, et al., 2006; Soltani, et al., 2013). Prior to the HSM and its emphasis on Empirical Bayes and accounting for regression-to-the-mean biases, Turner (1984) proposed a regression-based method for predicting crashes on bridges. As one of the seminal efforts towards predicting bridge-

related crashes, it accounted for traffic levels and physical characteristics such as the width of the bridge and the approaching roadway.

Bridges are subjected to periodic and consistent bridge inspections that typically cover the condition of the deck, superstructure, substructure and functional obsolescence (FHWA, 1968, Retting, et al., 2000). Identifying probable safety implications in terms of vehicle crashes is not a part of these routine inspections. In recognition of the importance of bridges as part of the roadway system and in keeping with the HSM, this study develops SPFs for different crash severities and evaluates its performance with other univariate techniques.

The discrete choice models have been widely used for analyzing crash severities. Therefore, the following section provides a review of literature involving use of multivariate econometric models used for analyzing crash severities. Section 4.3 describes the methodology used in this chapter for analyzing the performance of the proposed method. It reviews Univariate Conway Maxwell Poisson and Multivariate Poisson lognormal formulation. Section 4.4 then compares the predictive capabilities of the formulations describe in section 4.3 with the proposed MVCMP formulation. Finally, the conclusion section summarizes the advantages and limitations of the proposed method.

4.2 Existing Multivariate Econometric Models

The crash severity data is generally represented by five severities such as fatal, incapacitating, non-incapacitating, possible injury and property damage only crashes. There have been myriad studies applying different modeling techniques to understand the relationship between crash severities, roadway geometry and traffic characteristics. A detailed review of the statistical techniques used for estimating crash severities can be seen in a review paper by

Savolaien et al. (2011). The techniques found in the literature can be broadly categorized into probabilistic models, data mining techniques and econometric models.

The application of the data mining techniques has been very scarce as they are comparatively newer techniques. Some of the data mining techniques used include artificial neural networks (Delen, Sharda, & Bessonov, 2006; Chimba & Sando, 2009), classification and decision trees (Chang & Wang, 2006).

The probabilistic model assumes certain parametric distribution over different crash severities. Multivariate Poisson distribution, multivariate Poisson lognormal distribution and multivariate negative binomial distribution are some of the examples of the probabilistic models used in the transportation safety literature. A review of these probabilistic models is provided in section 3.2 of Chapter 3. The econometric models are one of the most widely used models in the literature (Savolainen, Mannering, Lord, & Quddus, 2011).

The econometric models are further divided in to two categories based on the type of response variable used in the analysis. If the response variable is a binary or nominal variable then an unordered model such as multinomial logit choice model (Khorashadi, Niemeier, Shankar, & Mannering, 2005; Islam & Mannering, 2006; Schneider & Savolainen, 2009; Kim, Ulfarsson, Kim, & Shankar, 2013), logistic regression (Eluru, Bhat, & Henser, 2008), nested logit (Holdridge, Shankar, & Ulfarsson, 2005; Savolainen & Mannering, 2007) etc. are used for the analysis. On the other hand, if the response variable is ordinal, then ordered models such as ordered logit (Gardner, 2006; Siddiqui, Chu, & Guttenplan, 2006; Haleem & Abdel-Aty, 2010), ordered probit (Lee & Abdel-Aty, 2005; Pai, 2009; Ye & Lord, 2011), generalized ordered logit (Eluru, Bhat, & Henser, 2008), etc. are used for the analysis.

The Ordered models are some of the most widely used models for crash severity analysis (Savolainen, Mannering, Lord, & Quddus, 2011). They assume that the level of severity are inherently related to one another and contain a single unobserved heterogeneity term for all the crash severities. The resulting probabilities for different severities are determined by partitioning the single heterogeneity terms in parts for each severity. This can be seen as one of the major draw backs of the ordered response model. On the contrary, the unordered, MVCMP and MVPLN models allow different heterogeneity terms for different crash severities. The ordered response models do not allow the independent variables to vary across the crash severities. For example if AADT is considered in the model, its coefficient will remain the same across all severities (Yasmin & Eluru, 2013). This limitation hinders the explanatory power of the independent variables.

The unordered response model do not suffer from the same limitations. However, they are still a two stage model like ordered response models (El-Basyouny & sayed, 2009) and does not include correlations. The first stage of the model is the total number of collision and the second stage of the model is to predict the severity of the crash.

Considering the above limitations of these most popular methods, the multivariate extensions of the Poisson lognormal and the Conway-Maxwell Poisson distributions are more flexible models. They allow for the extra Poisson variation often observed in the collision data as well as a full general correlation structure. Considering all drawbacks of widely used econometric models in crash severity analysis, and the benefits of the probabilistic models describe in Chapter 2 and 3, this study first applies four probabilistic models, namely MVCMP, MVPLN, univariate CMP and univariate Poisson lognormal on two-lane rural road and bridge

datasets. The performances of the four models are then compared to identify which model fits the data well. The objective of this comparison is to check the predictive capabilities of all the different models. The expected number of crashes is used by transportation professional in ranking sites for treatments. If the expected number of crashes computed is incorrect, it can result in less efficient use of money.

4.3 Methodologies

This study compares the performance of MVCMP model with both univariate Conway Maxwell Poisson, Univariate Poisson lognormal and Multivariate Poisson Lognormal (MVPLN) specifications. The MVCMP formulation developed in Chapter 3 is applied to the two lane rural road and bridge data sets. The details of the specification can be seen in the methodology section of Chapter 3. The parameterization for the univariate CMP and MVPLN formulation along with parameter estimation techniques is presented in this section.

4.3.1 Conway Maxwell Poisson Distribution

Conway-Maxwell Poisson distribution is the generalization of the Poisson distribution with an additional parameter, which governs the rate of successive probabilities. Conway and Maxwell introduced it in 1962 to handle queuing system. Shmueli et al. later derived the distributional properties in 2005.

To derive the mathematical formulation, let's assume Y to be a CMP random variable representing crash frequency for a particular severity. If a homogeneous road section is represented by i, then y_i represents the discrete crash count observed at that road section. The probability mass function (PMF) of CMP random variable Y_i equal to an observed crash count y_i can be written as follows.

$$f_{Y_i}(y_i; \lambda_i, \nu) = \frac{\lambda_i^{y_i}}{y_i!\, Z(\lambda_i, \nu)}$$

where,

$$Z(\lambda_i, \nu) = \sum_{j=0}^{\infty} \frac{\lambda_i^j}{(j!)^\nu}$$

$i = 1,2 \ldots\ldots, n$represents the total number of roadway segments considered in the analysis

$\lambda_i > 0$ and $\nu \geq 0$are the two parameters that are to be estimated.

The CMP distribution extends over three distribution and it has Geometric ($\nu = 0$), Poisson ($\nu = 1$) and Bernoulli ($\nu = \infty$) as its special cases. When $\nu < 1$, the data set is called as over-dispersed data set, while when $\nu > 1$ it is called as under-dispersed. This flexibility makes CMP an attractive choice for different variety of problems involving count data.

Let X be the matrix of the observed traffic and roadway conditions (predictor variables), then a log link function is used to associate the predictor variables with the parameter λ and it can be written as follows

$$\lambda_i = exp(\beta_0 + \sum\nolimits_j x_{ij}\beta_j)$$

where,x_{ij} is the j^{th} explanatory variable for the mean of crash counts at site i, β_j is the corresponding coefficients, β_0 is a constant to capture other unknown or unobserved factors. The parameters are estimated using the Maximum likelihood estimation and the log-likelihood function for a road segment $i$is given below.

$$logL_i(\lambda_i, \nu | y_i, x) = y_i \log(\lambda_i) - \nu \log(y_i!) - \log Z(\lambda_i, \nu)$$

For all the n road segments the above equation can be rewritten as

$$\log L = \sum_{i=1}^{n} y_i \log(\lambda_i) - \nu \sum_{i=1}^{n} \log(y_i!) - \sum_{i=1}^{n} \log Z(\lambda_i, \nu)$$

The R routine is used to estimate the parameters obtained by maximizing the above log-likelihood equation.

4.3.2 Multivariate Poisson Lognormal Distribution.

The derivation of the MVPLN formulation starts using the similar notation as in MVCMP formulation in Chapter 3. Let Y denote a n-by-S matrix of crash counts, and y_{is} denote the crash count observed for roadway segment i and severity s, for $i = 1,2 \ldots\ldots n$, and $s = 1,2, \ldots S$, where n is the total number of sites under consideration and S is the number of severity levels. Let $y_i = (y_{i1}, y_{i2}, \ldots. y_{is})'$ denote the vector of crash counts for site i and S is the number of severity levels. It is assume that y_i are independently distributed and the probability mass function of the Poisson distribution for s^{th} element is given by the following equation.

$$f_{Y_{is}}(y_{is}|\lambda_{is}) = \frac{\lambda_{is}^{y_{is}} exp(-y_{is})}{(y_{is}!)}$$

where, λ_{is} is the mean and variance of the Poisson distribution and $\lambda_i = (\lambda_{i1}, \lambda_{i2}, \ldots\ldots \lambda_{iS})'$. This property of Poisson distribution does not accommodate dispersion in the data. Let x denote an n-by-K matrix of covariates with K number of covariates. Each row of this matrix is denoted by $x_i = (x_{i1}, x_{i2}, \ldots, x_{iK})$, which represents a K-dimensional row vector corresponding to the characteristics of the i^{th} roadway segment. Let $\beta = (\beta_1, \beta_2, \ldots, \beta_S)$ denote a K-by-S matrix of the regression coefficients whose columns $\beta_s = (\beta_{1s}, \beta_{2s}, \ldots, \beta_{Ks})'$ are K-dimensional column vectors consisting of the regression coefficients for the crash count of the s^{th} severity level. To model the extra Poisson variation it is further assumed that $\lambda_{is} = exp(\beta_0 +$

$\sum_j x_{ij}\beta_j + \epsilon_{is})$ where, $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, \ldots\ldots \epsilon_{iS})'$ denote the unobserved heterogeneity specific to crash severity for site i. The unobserved heterogeneity ϵ_i is assumed to follow an S-dimensional multivariate normal distribution with zero mean and a variance-covariance matrix Σ, i.e. $\epsilon_i \sim N_S[0, \Sigma]$. The S-by-S matrix Σ is assumed to be constant across the sites.

Thus given all the parameters and data (β, X, Σ), the joint distribution of y_i can be written as

$$p(y_i|\beta, \epsilon_i, x_i) = \prod_{s=1}^{S} f_{Y_{is}}(y_{is}|\lambda_{is}) = \prod_{s=1}^{S} p(y_{is}|x_i, \beta_s, \epsilon_{is}) \quad (1)$$

4.3.3 Parameter Estimation

As explained in the case of MVCMP formulation the marginal distribution computations from equation 1 requires evaluation of S-variate integral which does not have a closed form solution. Hence, Bayesian Inference via Markov Chain Monte Carlo simulation is used to estimate the parameters of the model. The algorithm is coded in Matlab and a brief description of it is given below.

The joint conditional distribution of y_i and ϵ_i can bewritten as follows

$$p(y_i, \epsilon_i|\beta, x_i, \nu, \Sigma) = p(y_i|\epsilon_i, \beta, x_i)\phi_s(\epsilon_i|0, \Sigma)$$

The likelihood of parameters given y, assuming independence across sites can be written as $L(\beta, \Sigma|y, x) = \prod_{i=1}^{n} \int p(y_i, \epsilon_i|\beta, x_i, \Sigma)\, d\epsilon_i = \prod_{i=1}^{n} L_i(\beta, \Sigma|y_i, x_i)$

Applying Bayes' theorem, the joint posterior density is given by

$\pi(\beta, \Sigma, \epsilon|y, x) \propto Prior(\beta, \Sigma, \epsilon) \cdot p(y|\beta, x, \Sigma, \epsilon)$

It is further assumed that all the parameters are independent of each other. Thus, the prior term in the above equation can be written as product of four probability distributions as shown below.

$$Prior(\beta, \Sigma, \epsilon) = Prior(\beta) \cdot Prior(\Sigma) \cdot \phi(\epsilon|0, \Sigma) \quad (2)$$

Assume $\beta_s \sim N_K(\mu_{\beta_s}, V_{\beta_s})$, a *K*-variate normal distribution with mean vector μ_{β_s} and variance-covariance matrix V_{β_s}; $\Sigma^{-1} \sim f_w(\tau_\Sigma, V_\Sigma)$, an Wishart distribution(Hogg, Craig, & McKean, 2006) with scale matrix V_Σ and degrees of freedom τ_Σ, which is a conjugate prior (Press, 1982; Gelman, Carlin, Stern, & Rubin, 2014). These parameters defining the prior distributions of the parameters to be estimated are called hyper parameters. With these assumptions, equation () can then be written as

$$Prior(\beta, \Sigma, \epsilon) = \left[\prod_{s=1}^{S} \phi_K(\mu_{\beta_s}, V_{\beta_s})\right] \cdot f_w(\tau_\Sigma, V_\Sigma) \cdot \left[\prod_{i=1}^{n} \phi_S(\epsilon_i|0, \Sigma)\right]$$

Therefore, the following holds for the joint posterior distribution $\pi(\beta, \Sigma, \epsilon|y, x)$.

$$\pi(\beta, \Sigma, \epsilon|y, x)$$

$$\propto Prior(\beta, \Sigma, \epsilon) \cdot p(y|\beta, x, \Sigma, \epsilon)$$

$$= \left[\prod_{s=1}^{S} \phi_K(\mu_{\beta_s}, V_{\beta_s})\right] f_w(\tau_\Sigma, V_\Sigma) \cdot \left[\prod_{i=1}^{n} \phi_S(\epsilon_i|0, \Sigma)\right] \cdot \left[\prod_{i=1}^{n} \prod_{s=1}^{S} p_{Poi}(y_{is}|\epsilon_i, \beta_s, x_i)\right]$$

$$= f_w(\tau_\Sigma, V_\Sigma) \cdot \left[\prod_{s=1}^{S} \phi_K(\mu_{\beta_s}, V_{\beta_s})\right] \cdot \left[\prod_{i=1}^{n} \left(\phi_S(\epsilon_i|0, \Sigma) \prod_{s=1}^{S} p_{Poi}(y_{is}|\epsilon_i, \beta_s, x_i)\right)\right] \quad (3)$$

The joint posterior distribution is simulated by component-wise Metropolis-Hastings algorithm. The process is implemented iteratively by sampling $\epsilon|(y,x,\beta,\Sigma)$, $\Sigma^{-1}|(\epsilon)$, $\beta|(y,x,\epsilon,\Sigma)$. The details of the estimation procedure are presented in Chib and Winklemann (1999).

The sampling process can be divided into four parts and it is as shown below.

Step 1) The first step is to initialize all the hyper parameters for the prior and proposal distribution.

Step 2). Sampling Σ^{-1}

Inverse Wishart distribution, $f_w(\tau_\Sigma, V_\Sigma)$, is used as a prior for the variance-covariance matrix parameter Σ, and the terms that pertain to Σfrom equation (3) are given below. Note that all the studies (Chib & Winklemann, Markov Chain Monte Carlo Analysis of Correlated Count Data, 2001; Ma, Kockelman, & Damien, A Multivariate Poisson-Lognormal Regression Model for Prediction of Crash Counts by Severity, Using Bayesian Methods, 2008; Park & Lord, 2007) so far have simulated Σ^{-1} rather than Σ using Wishart distribution. However, in this study inverse Wishart distribution is used which is also a conjugate prior(Press, 1982; Gelman, Carlin, Stern, & Rubin, 2014). A Gibbs sampler is implemented as follows.

$$\pi(\Sigma[m+1]|\epsilon[m]) \propto f_w(\Sigma|\tau_\Sigma, V_\Sigma) \prod_{i=1}^{n} \phi_S(\epsilon_i[m]|0,\Sigma[m])\, \pi(\Sigma[m+1]|\epsilon[m])$$

$$\propto f_w\left(n+\tau_\Sigma, \left[V_\Sigma + \sum_{i=1}^{n} (\epsilon_i'[m]\cdot\epsilon_i[m])\right]^{-1}\right)$$

Step 3). Sampling ϵ

The terms pertaining to ϵ in equation (3) are

$$\prod_{i=1}^{n}\left(\phi_S(\epsilon_i|0,\Sigma)\prod_{s=1}^{S} p_{Poi}(y_{is}|\epsilon_i,\beta_s,x_i)\right)$$

This conditional probability density is not given by any known density function, hence a MH algorithm is setup to generate a sequence of samples from the posterior distribution. Since ϵ_i is considered independent across the sites, each ϵ_i is sampled separately where values for ϵ_{-i} (any $\epsilon_j, j \neq i$) are held constant at their values from last iteration. In other words,

$$\pi(\epsilon_i|y_i, x_i, \beta, \Sigma, \epsilon_{-i}) \propto \phi_S(\epsilon_i|0, \Sigma) \prod_{s=1}^{S} p_{Poi}(y_{is}|\epsilon_i, \beta_s, x_i)$$

On further solving,

$$\pi(\epsilon_i[m+1]|y_i, x_i, \beta[m], \Sigma[m+1], \epsilon_{-i}[m]) \propto \phi_S(\epsilon_i|0, \Sigma[m]) \prod_{s=1}^{S} (\lambda_{is})^{y_{is}} \exp(-\lambda_{is})$$

where $\lambda_{is} = \exp(x_i'\beta_s[m] + \epsilon_{is}[m])$.

A multivariate t-distribution with degrees of freedom d and location parameter D is used as a proposal density to generate a proposal vector ϵ_i^*. The degrees of freedom is used as tuning parameter to ensure a satisfactory acceptance rate. A proposal vector is accepted with probability

$$\min\left\{\frac{\pi(\epsilon_i^*|y_i, x_i, \beta, , \Sigma)\ p(\epsilon_i|d, D)}{\pi(\epsilon_i|y_i, x_i, \beta, \Sigma)\ p(\epsilon_i^*|d, D)}, 1\right\}$$

where $p(\cdot\,|d, D)$ is the probability density of the proposal t-distribution.

Step 4). Sampling β

The terms pertaining to β in equation (3) are

$$\left[\prod_{s=1}^{S} \phi_K(\mu_{\beta_s}, V_{\beta_s})\right] \cdot \left[\prod_{i=1}^{n} \prod_{s=1}^{S} p_{Poi}(y_{is}|\epsilon_i, \beta_s, x_i)\right]$$

$$= \prod_{s=1}^{S} \left(\phi_K(\mu_{\beta_s}, V_{\beta_s}) \prod_{i=1}^{n} p_{Poi}(y_{is}|\epsilon_i, \beta_s, x_i) \right)$$

Just as ϵ, this conditional probability density of β is not given by any known density function either. Therefore, a MH algorithm is implemented. Since β_s is considered independent across severities, each β_s is sampled separately where values for β_{-s} (any $\beta_j, j \neq s$) are held constant at their values from last iteration. In other words,

$$\pi(\beta_s|y, x, \beta_{-s}, \Sigma, \epsilon) \propto \phi_K(\beta_s|\mu_{\beta_s}, V_{\beta_s}) \prod_{i=1}^{n} p_{Poi}(y_{is}|\epsilon_i, \beta_s, x_i,)$$

Therefore,

$$\pi(\beta_s[m+1]|y_i, x_i, \epsilon_i[m+1], \Sigma[m+1], \beta_{-s}[m]) \propto \phi_K(\beta_s|\mu_{\beta_s}, V_{\beta_s}) \prod_{i=1}^{n} (\lambda_{is})^{y_{is}} \exp(-\lambda_{is})$$

where $\lambda_{is} = \exp(x_i'\beta_s[m] + \epsilon_{is}[m+1])$.

A multivariate normal distribution with mean vector u_{β_s} and variance B_{β_s} is used as a proposal density to generate a proposal vector β_s^*. The mean vector u_{β_s} is used as tuning parameter to ensure a satisfactory acceptance rate. A proposal is accepted with the probability

$$\min\left\{ \frac{\pi(\beta_s^*|y_i, x_i, \epsilon_s, \sum) \, p(\beta_s|u_{\beta_s}, B_{\beta_s})}{\pi(\beta_s|y_i, x_i, \epsilon_s, \sum) \, p(\beta_s^*|u_{\beta_s}, B_{\beta_s})}, 1 \right\}$$

4.4 Case Studies

4.4.1 Data Preparation

The following section describes the different data sources from where data sets are obtained and the processing done on those data sets to bring it into a format that can be used for analysis. The software's used in the data cleaning are also listed in this section. The first subsection describes the data used for two-lane rural road analysis and the second subsection describes the data used for bridge section analysis.

4.4.1.1 Two Lane Rural Roads

The dataset used for the analysis of two-lane rural road is the same as the one used in Chapter 2. The dataset consists of 5991 homogenous segments created such that the geometric and the traffic characteristics remain constant within each segment. The variables considered in the analysis include AADT, segment length, speed limit, shoulder width, shoulder type and highway class. All the observations are used for training the model. The detailed statistics of the dataset is given inTable 2.1

4.4.1.2 Bridges

Bridge Data

The bridge data used in the analysis is the same as the dataset used to develop bridge SPFs for bridges by Mehta et al. (2014). The data set consists of only the bridges that carry state or interstate highways (including ramps) in Alabama. Initially, the bridge data is obtained from the National Bridge Inventory (NBI). The NBI contains raw inventory data for 17,513 bridges in Alabama. After screening, 1,122 bridges are coded into a point shape files in ArcGIS 10 and converted to bridge vectors onto road shape files obtained from the U.S. Census Bureau's Master

Address File/Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER). Chase et al. (1999) reported that some 98% of bridge coordinates obtained from NBI are within two miles of actual locations. As such, the NBI-specified point locations of the 1,122 Alabama bridges are moved to the nearest roads and the road names in the shape files are checked to ensure they matched the road names in the NBI data. Zhang et al. (2012) showed that NBI coordinates may represent any other point along a bridge (i.e., at an end or along the span). For identifying bridge-related crashes by spatial relations, the NBI point locations for the Alabama bridges are assumed to be midpoints to create bridge vectors.

Crash Data

Crash data from 2009 through 2012 are collected from the Critical Analysis Reporting Environment (CARE) database maintained by the Center for Advanced Public Safety at the University of Alabama. The crash data contain GPS coordinates, which are used to associate crashes onto bridge vectors using ArcGIS. As previously explained, the original bridge coordinates are taken as the mid-point of the bridge for creating a vector. Due to lack of direct bridge related information in CARE, a crash is designated bridge-related, if its GPS coordinates places it directly onto a bridge or within one half of the length of each individual bridge on both sides of the bridge vector. This method of associating crashes with the bridge is adopted because of two reasons. Firstly, reported crash coordinates might not be accurate as the crashed vehicles are usually moved on the side of the road before filing a crash report. Secondly, increasing search areas on either ends of a bridge can account for crashes in the bridge influence area. Crash data are further examined to ensure that the coded location referred to a crash actually

occurred on a bridge and not on a surface road below the bridge (i.e., underpass). As a result, 9,958 crashes are used to develop the SPF for total crash counts.

Data Integration

For model development, the crash count for an individual bridge in each study year served as a single observation. A total of 4,488 overall crash observations were available, and a total of 3,366 single-vehicle crash observations were available. It should be noted that the annual average daily traffic (AADT) information given in the NBI is not consistent for all the bridges and ranged from 2003 to 2011. There are several methods published in the literature for forecasting AADT such as exponential smoothing (Holt, 2004), support vector regression (Castro-Neto et al., 2009), fuzzy approach (Gastaldi et al., 2014) etc. However, the observed AADT in the NBI data set is only for one year between 2003 and 2011, and the forecasted AADT for 2025. It is impossible to use any statistical methods with only two data points for a particular bridge. As the majority of AADT data in the NBI data set are from 2009, a simple linear interpolation/extrapolation from a 2009 base year was conducted to predict AADT for SPF development.

A sample set of bridge-related crashes was split into a training set and a validation set. The training set was used to develop the SPF models, while the other set was used to validate the best-fit model(s). The results are presented in the following sections.

4.4.2 Results

The following section describes model estimation for MVCMP and MVPLN formulation for both two-lane rural roads and Bridges. As describe in the introduction section four different models are estimated for each data set. The MVCMP and MVPLN models are estimated using the component-wise MH algorithm. The Univariate CMP model is estimated using "COMPoisson" routine written in R by Sellers(Sellers & Shmueli, 2010). The Univariate Poisson lognormal model is estimated using SPSS. Results obtained from each model are compared to check the behavior of different model specifications on the same data set.

4.4.2.1 Bridges

The MVCMP and MVPLN model for bridges are estimated using the proposed Bayesian formulation and the component-wise MH algorithm. The initial values of β for both the models and ν for MVCMP model is set to the results obtained from MLE of individual univariate CMP and Univariate Poisson lognormal distribution for each crash severity. The starting value of Σ is set as an identity matrix. The prior distributions used for each parameters to be estimated are $\beta_s \sim N_s(0,\ 0.1I)$, $\nu_s \sim \gamma(0.03,0.1)$, $\Sigma \sim f_w(20, 10I)$, and $\epsilon = [0]$. The component-wise MH algorithm is run for 200,000 iterations with the first 100,000 iterations discarded as burn-in. The trace and running mean plots are produced for each parameter to check their convergence. The Table 11 reports the parameter estimates for the bridges in Alabama. The estimates in the bold face are the variables whose 95% credible interval does not contain 0.

Table 11Parameter Estimates for Bridges

Severities	Variables	MVCMP	UCMP	MVPLN	UP
Fatal Crashes	Const	**-3.1042**	**-15.2518**	**-5.1954**	**-13.7033**
	ln(AADT)	0.1478	**1.2102**	**0.1161**	**1.0402**
	Bridge Length	**0.3136**	**0.4940**	**-0.4743**	**0.3137**
	Percentage Truck	**-0.0804**	**0.0215**	**0.2477**	0.0233
	Traffic Lanes	-0.0039	-0.1537	-0.0200	-0.1010
	Approach Width	0.0018	**-0.0241**	-0.0104	-0.0219
	Highway Class	0.1452	**1.5498**	0.7447	**1.3620**
	Pearson Chi sq	-	-		7.043
	Nu	2.327	3.217	-	-
Incapacitating Crashes	Const	**-4.9315**	**-13.3666**	**-8.6303**	**-18.1120**
	ln(AADT)	**0.3023**	**1.1007**	**0.1143**	**1.6080**
	Bridge Length	**0.4171**	**0.1560**	**-0.3097**	**0.3440**
	Percentage Truck	-0.0736	0.0047	**0.1788**	-0.0020
	Traffic Lanes	-0.0054	-0.0418	-0.0437	-0.0020
	Approach Width	-0.005	**-0.0269**	**-0.0438**	**-0.0590**
	Highway Class	0.3113	**1.0611**	0.8895	**1.5270**
	Pearson Chi sq	-	-		129.327
	Nu	2.098	0.040	-	-
Non Incapacitating Crashes	Const	**-3.8181**	**-18.5430**	**-6.2169**	**-18.4680**
	ln(AADT)	**0.1669**	**1.5681**	**0.323**	**1.5600**
	Bridge Length	**0.3089**	**0.5300**	**-0.3156**	**0.5260**
	Percentage Truck	-0.1033	0.0219	0.1379	0.0220
	Traffic Lanes	-0.0054	**-0.3094**	-0.0240	**-0.3070**
	Approach Width	0.005	**-0.0289**	-0.0319	-0.0290
	Highway Class	0.3113	**1.6732**	**0.8351**	**1.6680**
	Pearson Chi sq	-	-		0.61
	Nu	2.184	1.013		
Possible Injury	Const	**-3.2484**	**-15.6293**	**-6.5048**	**-23.3570**
	ln(AADT)	**0.6557**	**1.3575**	**0.2021**	**2.1560**
	Bridge Length	**0.3666**	**0.1618**	**-0.3504**	**0.4080**
	Percentage Truck	-0.0171	-0.0110	**0.1953**	-0.0200
	Traffic Lanes	-0.0028	-0.0995	-0.0295	-0.1100
	Approach Width	-0.0068	**-0.0262**	**-0.0754**	**-0.0570**
	Highway Class	0.2417	**0.9866**	0.6429	**1.5980**
	Pearson Chi sq	-	-		1.467
	Nu	1.943	0.030		

Property Damage Only	Const	**-3.5415**	**-10.8049**	**-5.6961**	**-22.0070**
	ln(AADT)	**0.878**	**1.0135**	**0.4038**	**2.1820**
	Bridge Length	**0.9387**	**0.2469**	**-0.2118**	**0.4300**
	Percentage Truck	-0.1104	**0.0060**	**0.1316**	**0.0180**
	Traffic Lanes	-0.0049	-0.2031	-0.0238	**-0.1380**
	Approach Width	-0.0049	**-0.0129**	**-0.0184**	**-0.0520**
	Highway Class	0.9700	0.6949	**1.4450**	**1.8220**
	Pearson Chi sq	-	-		275.808
	Nu	1.296	0.000		

Note: Variables in bold indicate variables whose 95% credible intervals do not contain 0.

From the above table it can be noted that the MVCMP model has the least number of variables that are significant for all the crash severity types. All the other models have similar or higher number of parameters that are significant.

The constant term, AADT and segment length are the three variables that are significant in all the models and for all the crash severities except for Fatal Crashes. The sign of the coefficients of constant term and AADT are same across all the severities. The positive sign of the coefficient implies that with the increase in AADT, the expected number of crashes for all the severities increase. It is important to note that the AADT has no significance on the fatal crashes. This is very uncommon observation however; it is in accord with the finding of the Ma et al (2008). This finding suggests that change in traffic volume does not associate with the frequency of fatal crashes.

The bridge length term is significant in all four models and is positive for MVCMP, UCMP and UPLN models. This implies that as the length of the bridge increases the expected number of crashes on the bridge also increases. This observation is in sync with the 2LRR observation for total crashes presented in Chapter 2. It is very intuitive because, as the length of the bridge increases the amount of time travelled on the section will increase leading to higher

probability of being involved in a crash on that section. The MVPLN model on the other hand contradicts the observation withits negative coefficient for bridge length across all the severities. This finding needs further investigation as travelling on longer bridges should not be any safer than on shorter bridges.

Apart from AADT and bridge length, the percentage of trucks on the bridge also affects the fatal crash severity. However, this observation is counterintuitive for MVCMP model. The negative sign of the coefficient indicates that as the as the percentage of truck increases the expected number of fatal crashes reduce. This finding is consistent with the study by Mehta et al. (2014 under review) on the same data set but for total crash frequency. One possible explanation for such behavior could be that the drivers are more cautious while driving around large number of trucks. A behavior study of drivers around trucks could probably answer this anomaly. The percentage of truck variable has the most variation across all the crash severities. The variable is not significant for any other severity type according to the MVCMP model. For all the other models the variable has both positive coefficient whenever it is significantly impacting the crash frequency. This implies that as the percentage of truck increases the expected number of crashes also increases.

The number of traffic lanes variable is insignificant in all the MVCMP and UCMP and MVPLN models. The variable is significant for UPLN model for non-incapacitating and property damage only crashes. The coefficient is negative for both severities indicating that increasing the number of lanes reduces the rate of expected crashes on the bridges.

The approach width variable is significant for different crash severities in all the models except MVCMP. The coefficient of the variable is negative indicating reduction in the expected

number of crashes for bridges on the wider roads. This behavior could be because of the confounding effect of the deck and bridge width. As the approach width to the bridge increases, the bridge width might be increasing leading to safer driving conditions.

The highway class variable is positive across all the crash severities and all the models implying that as the highway class increases (arterial to freeways) the expected number of crashes also increases. The probable reason behind this observation could be the confounding effect of other factors such as speed limit, which is not included in this study, because of lack of data. As the class of the highway increases certain unobserved factors such as speed limits, shoulder widths etc. change, which can have some impact on crash severities.

Another important observation is the estimated ν values. The univariate CMP model suggests that the fatal crash severity is under-dispersed and all the other crash severities are over-dispersed. However, the multivariate model suggests that all the crash severities are under-dispersed. This difference can be explained by the mean of the two models. The characteristic of the dispersion is conditioned on the mean, therefore different mean values will lead to different degrees of dispersion. Examining the predictive capability of the two models will reveal the better model as seen in the validation section. It should also be noted that a bootstrap analysis is done on UCMP to check the parameter estimates for ν. It was observed that for some instances the parameter estimates were very unstable with some cases had parameter estimates range between 2 and 22 just in 5 trials. This is not the case for MVCMP formulation. Several different proposal distribution and prior distributions are set to test the variability in the estimates. The variability was found to be very low compare to the UCMP estimation. The two models cannot be compared for their efficiency as they are estimated using different techniques.

Finally, it should be noted for PDO crashes that all the variables considered in the analysis are significant from the univariate Poisson lognormal model perspective. However, it can be observed that the value of person chi-square divided by degrees of freedom is 275 implying an extreme over-dispersion. As described in chapter 2 the over-dispersion leads to serious underestimation of the standard errors resulting in incorrect interval estimation and inability to capture the true parameter values. This problem does not exist in other models as can be seen from the number of variables significant in those models.

The Table 12and Table 13contain the MCMC estimates of the covariance matrix and correlation matrix of the MVCMP and MVPLN models.

Table 12Correlation Matrix for Crash Severities MVCMP (Bridges)

1	0.9371	0.9275	0.8797	0.974
0.9371	1	0.9507	0.9207	0.9231
0.9275	0.9507	1	0.9786	0.9754
0.8797	0.9207	0.9786	1	0.8978
0.974	0.9231	0.9754	0.8978	1

Table 13Correlation Matrix for Crash Severities MVPLN (Bridges)

1	0.95	0.9497	0.9585	0.9623
0.95	1	0.9536	0.9619	0.9691
0.9497	0.9536	1	0.9639	0.9683
0.9585	0.9619	0.9639	1	0.9787
0.9623	0.9691	0.9683	0.9787	1

All the elements of the above matrix are significant and their 95% credible interval does not contain 0. All the non-diagonal elements of the matrix are positive, implying that the positive correlation between the severities. Thus using a MVCMP and MVPLN formulation can

help in incorporating this correlation in estimating the crash severities that cannot be done by the univariate models.

4.4.2.2 Two-Lane Rural Roads

The MVCMP and MVPLN model for two lane rural roads are estimated using the proposed Bayesian formulation and the component-wise MH algorithm. The initial values for both β for both the models and ν for MVCMP model is set to the results obtained from MLE of individual univariate CMP and Univariate Poisson distribution for each crash severity. The starting value of Σ is set as an identity matrix. The prior distributions used for each parameters to be estimated are $\beta_s \sim N_s(0,\ 0.1I)$, $\nu_s \sim \gamma(0.03,0.1)$, $\Sigma \sim f_w(20, 10I)$, and $\epsilon = [0]$. The component-wise MH algorithm is run for 200,000 iterations with the first 100,000 iterations discarded as burn-in. The trace and running mean plots are produced for each parameter to check their convergence. The Table 14 reports the parameter estimates for the bridges in Alabama. The estimates in the bold face are the variables whose 95% credible interval does not contain 0.

Table 14 Parameter Estimates for Two-Lane Rural Roads

Severities	Variables	MVCMP	UCMP	MVPLN	UP
Fatal Crashes	Const	**-9.6292***	**-3.2169**	-0.0564	**-7.0860**
	AADT	**0.3259**	**0.4554**	0.2250	**0.6550**
	SL	-0.1238	**-2.2461**	**-4.2533**	**-2.2710**
	SpeedLimit	**0.3314**	0.0495	-0.1821	0.0540
	Lane Width	0.0752	-0.0287	-0.0721	-0.0030
	Shold Width	-0.0606	**-0.1565**	-0.1365	-0.1310
	Shold Type	**0.1486**	-0.0798	-0.0084	0.0590
	Pearson Chi sq	-	-	-	1.836
	Nu	3.003	2.564	-	-
Incapacitating Crashes	Const	**-9.5627**	**-6.0499**	-0.2158	**-6.4920**
	AADT	**0.5087**	**0.5415**	**0.0510**	**0.6060**
	SL	**0.2940**	**0.2956**	**-1.6458**	**0.3370**
	SpeedLimit	**0.3703**	-0.0382	-0.3432	-0.0440
	Lane Width	0.1541	**0.1311**	0.0765	0.1450
	Shold Width	-0.0806	**-0.0962**	-0.0722	-0.1070
	Shold Type	**0.1363**	**0.0555**	-0.0594	0.0610
	Pearson Chi sq	-	-	-	1.098
	Nu	2.098	0.523	-	-
Non Incapacitating Crashes	Const	**-10.3376**	**-8.0800**	-0.0854	**-8.0180**
	AADT	**0.2881**	**0.8572**	0.1893	**0.8320**
	SL	0.1422	**0.4145**	0.1581	**0.4020**
	SpeedLimit	**0.4782**	-0.3012	**-0.4532**	-0.2850
	Lane Width	0.0787	-0.0419	-0.0056	-0.0400
	Shold Width	-0.045	-0.2459	-0.1764	-0.2400
	Shold Type	**0.1517**	**0.2314**	0.0240	0.2250
	Pearson Chi sq	-	-	-	0.992
	Nu	2.968	1.812	-	
Possible Injury	Const	**-10.9488**	**-13.1572**	-0.1107	**-9.313**
	AADT	**0.3679**	**0.7654**	0.0914	**0.9790**
	SL	0.1566	**0.4835**	0.2541	**0.4940**
	SpeedLimit	**0.5123**	0.3585	**-0.3282**	0.3640
	Lane Width	0.0950	0.1148	0.0069	0.1160
	Shold Width	-0.0455	-0.2007	-0.1452	**-0.2040**
	Shold Type	**0.1284**	**0.1288**	-0.0941	0.1310
	Pearson Chi sq	-	-	-	0.962
	Nu	2.929	0.723	-	

Property Damage Only	Const	**-8.6055**	**-5.3064**	-0.3863	**-6.696**
	AADT	**0.8430**	**0.5393**	**0.51**	**0.7460**
	SL	**0.5620**	**0.3096**	**0.3937**	**0.4650**
	SpeedLimit	**0.2354**	-0.0297	**-0.3741**	-0.0430
	Lane Width	0.1123	0.0597	0.0186	0.0740
	Shold Width	-0.0509	**-0.0329**	-0.0158	-0.0450
	Shold Type	0.0735	**0.0153**	**0.1058**	0.0160
	Pearson Chi sq	-	-	-	1.380
	Nu	2.578	0.336	-	

Note: Variables in bold indicate variables whose 95% credible intervals do not contain 0

From the above table it is observed that the constant term, AADT and segment length in both the univariate modelsfor all the severities have similar signs and are consistent with the literature. The positive sign of AADT and segment length are very intuitive. It implies that as the total volume of vehicles and the length of the road under consideration increases, the expected crash frequency for all the severities also increases. This observation is also consistent with the state specific SPFs developed using NB regression in chapter 2. The multivariate MVCMP model does show similar results for all the severities except for fatal severity. The segment length variable is not significant according to the MVCMP model. The MVPLN model is very different from all the models. The model suggests that for fatal crash severity, segment length is the only variable that has significant impact on the crash frequency. The sign of the coefficient is negative and is consistent with the MVCMP model.

For fatal crash severity, all the models conclude the same thing. Apart from AADT and segment length, the only other variable significantly affecting fatal crashes is speed limit for MVCMP model and shoulder width for MVPLN model. The positive coefficient of the speed limit implies that as the speed limit on the road increases the expected number of fatal crash increases. The speed limit observation is a very important and is in line with the intuition. As

the speed limit increases, the damages resulting out of crashes at higher speed will obviously be higher.

The speed limit variable in MVCMP model is significant for all the crash severities with positive coefficient. The variable is significant for non-incapacitating, possible injury and property damage only crashes based on MVPLN formulation. However, the negative coefficient is counter intuitive. It is interesting to note that speed limit variable is not significant in any of the univariate models. This suggests that multivariate models provide statistically significant insights into crash count dependence on speed limit.

The lane width and shoulder width variables are insignificant for both the multivariate variables for all the severities. This behavior could be because of the apparent over-dispersion seen from the ratio of person chi-sq, degrees of freedom and ν values. It has been shown before that dispersion in the data can lead to inflated standard errors resulting into inaccurate results (Park & Lord, 2007). The shoulder width variable is significant for fatal crashes, incapacitating and property damage only crashes for the UCMP formulation. The coefficient is negative implying reducing expected crashes as the shoulder width increases. The lane width is significant for only one severity and it has a positive coefficient. The shoulder width variable is also significant for one severity of UPLN formulation, with negative coefficient as expected.

The shoulder type variable is significant for all the non-property damage crash severities based on the MVCMP formulation. The variable is also significant for all the non-fatal crash severities based on the UCMP formulation. The coefficient of shoulder type in all the models where it is significant is positive and counter-intuitive. The positive sign implies that as the shoulder type improves from SOD to paved the expected number of crashes for different

severities also increases. This observation could be the result of some confounding factors, which lead the authorities to have a particular kind of shoulder. A further investigation to this method would be a good future research. It should be noted that the MVPLN or UPLN model does not identify shoulder type as a significant variable for any crash severity.

The Table 15and Table 16 gives the correlation matrix obtained from the MVCMP and MVPLN formulation. All the off diagonal elements of the matrix are greater than 0.15 and their 95% credible interval does not contain the term 0. Hence, it can be concluded that there exists some correlation between different crash severities.

Table 15 Correlation Matrix for Crash Severities MVCMP (2LRR)

1	0.3555	0.6617	0.1682	0.1760
0.3555	1	0.2985	0.4940	0.6251
0.6617	0.2985	1	0.5061	0.0861
0.1682	0.4940	0.5069	1	0.7049
0.1760	0.6251	0.0861	0.7049	1

Table 16 Correlation Matrix for Crash Severities MVPLN (2LRR)

1	0.8133	0.6885	0.8143	0.8351
0.8133	1	0.7482	0.8813	0.8883
0.6885	0.7482	1	0.7495	0.7541
0.8143	0.8813	0.7495	1	0.896
0.8351	0.8883	0.7541	0.896	1

The univariate models would not be able to estimate these correlations and hence it will be missing on important information. The magnitude of correlation between MVPLN and MVCMP models is different for some of the severities; however, the overall conclusion of the study remains the same.

4.4.3 Validation

Model validation is conducted to verify the applicability of newly developed models on data set not used for training the models. There are different methods of validating trained models. However, in this study, the prediction capability of the models is compared using three performance measures, which are widely used in the literature (Washington, et al., 2005; Lord, et al., 2008). Three commonly used performance measures – mean absolute deviance (MAD), mean prediction bias (MPB), and mean square prediction error (MPSE) are used in this study (refer to Chapter 2 section 2.4.5).

4.4.3.1 Bridges

To check the prediction performance of the MVCMP model a validation data set not used in training the model is used. The dataset consists of 1793 observations with same six variables used in the training model. Note that the use of this multivariate model on external data is different from the regular univariate models. Firstly, using the variance covariance matrix estimated for the training data set , a multivariate random error term is generated from the normal distribution with mean vector **0** and variance covariance matrix. Then, the expected number of crashes conditioned on the covariance matrix is computed. The results obtained from the two-step process are given in Table 17.

Table 17 Expected Number of Crashes vs Observed Number of Crashes for Bridges

	Observed Crashes	MVCMP	UCMP	MVPLN	UPLN
Fatal	21	24	88	164	248
Incapacitating	300	361	407	1002	443
Non-Incapacitating	263	286	307	2053	313
Possible Injury	305	374	467	482	986
Property damage Only	3604	3403	3620	5212	5663

It can be seen from the table that MVCMP formulation predicts the total number of crashes very close to the total observed crashes for all the severities except PDO crashes. Although, the comparison is between the expected number of crashes and realization of crashes in a particular year, this exercise shows that the proposed model does not give out unreasonable crash numbers. The most important thing to note here is that the expected number of crashes for severe injuries is very close to that being observed. Underestimation of severe injuries could result in huge losses to the nation in terms of human life and productivity, while overestimation can result into unnecessary expenditure on certain projects that will not result into any benefit.

Table 18 Validation Results for Bridges

MSPE	MVCMP	UCMP	MVPLN	UPLN
Fatal	0.0299	0.1340	**0.0196**	0.1102
Incapacitating	**0.6172**	0.7280	1.9843	0.9676
Non-Incapacitating	**0.4121**	1.4651	0.5655	0.5982
Possible Injury	0.7031	**0.6693**	10708.9193	4.0251
Property damage Only	**114.9561**	927.7006	7605.6751	167.7774
MPB				
Fatal	**-0.0015**	-0.2732	-0.0339	-0.1141
Incapacitating	**-0.0305**	-0.3533	-0.0542	-0.0718
Non-Incapacitating	**-0.0116**	-0.9002	-0.0225	-0.0253
Possible Injury	**-0.0345**	-0.0894	-9.6928	-0.3426
Property damage Only	**0.1011**	-0.8097	-49.3727	-1.0355
MAD				
Fatal	0.1132	0.3314	**0.0537**	0.1290
Incapacitating	0.3166	0.5648	0.3330	**0.2786**
Non-Incapacitating	0.2732	0.9722	0.2548	**0.2211**
Possible Injury	**0.3339**	0.3626	9.9601	0.4751
Property damage Only	**2.4842**	3.3686	51.2697	2.7491

Table 18 gives the three validation tests used throughout the dissertation. The Mean Absolute Deviance, Mean Prediction Bias and Mean Square Prediction Error compare the

predicted value with the observed crash values. The closer the values are to zero the better is the model. For the Bridge data set, the MVCMP model performs better for 10 crash severities out of fifteen. This observation along with the observation from Table 17shows that MVCMP model performs better than the other univariate and MVPLN model.

4.4.3.2 Two-Lane Rural Roads

The data set used in validation of the 2LRR model consist of 5991 homogeneous sites. As describe in the previous section, a multivariate random number is generated using Matlab and the sum of expected number of crashes conditioned on the covariance matrix are given in Table 19.

Table 19 Expected Number of Crashes vs Observed Number of Crashes for (2LRR)

	Observed Crashes	MVCMP	UCMP	MVPLN	UPLN
Fatal	67	87	103	194	48
Incapacitating	731	876	946	1220	490
Non-Incapacitating	149	183	205	2526	697
Possible Injury	171	218	359	482	307
Property damage Only	1833	2180	2006	2244	1411

It is observed that all the severe crashes and crashes with low sample mean are very well predicted by MVCMP model. This observation is in accord with the finding of Geedipally et al. (2008), where they observed that the UCMP performs better for under-dispersed data with low sample mean. The MVCMP model addresses the issue of unobserved heterogeneity and allows for correlation among crash counts at different crash severities resulting in better performance in crash prediction. Finally, Table 20 gives the result of other validation tests on two-lane rural road data sets.

Table 20 Validation for Two-Lane Rural Roads

MPSE	MVCMP	UCMP	MVPLN	UPLN
Fatal	0.0362	0.1876	0.3595	**0.0229**
Incapacitating	0.3640	0.3793	2.5306	**0.1617**
Non-Incapacitating	**0.0605**	1.6714	0.2027	0.1523
Possible Injury	0.0682	0.0755	0.4028	**0.0506**
Property damage Only	1.3117	1.1250	22.8824	**0.6051**
MPB				
Fatal	**-0.0298**	-0.3851	-0.0731	-0.3409
Incapacitating	-0.2563	-0.9947	**-0.1080**	-0.4769
Non-Incapacitating	**-0.0502**	-0.1326	-0.3047	-1.1870
Possible Injury	**-0.0668**	-0.2066	-0.1151	-0.2139
Property damage Only	-0.6640	-1.5165	**-0.3252**	-0.7944
MAD				
Fatal	**0.0126**	0.4036	0.0943	0.3784
Incapacitating	0.4258	1.1206	**0.2944**	0.5864
Non-Incapacitating	**0.1577**	0.1774	0.3342	1.1899
Possible Injury	0.1721	0.2546	**0.1606**	0.2577
Property damage Only	0.8966	1.7954	**0.6165**	0.9600

It can be observed from Table 20 that of the three tests on five severities, the MVCMP model performs better for 6 severities while MVPLN performs better for 5 severities. However, based on the total number of crash predictions MVCMP performs much better than all the other model formulations. The validation exercise shows great promise for the application of the MVCMP formulation in state of Alabama for jointly modeling crash severities. An interesting future work will be to validate this MVCMP model on the data from other states to check the predictive capability of the proposed model.

4.5 Conclusion

The bridges are integral part of the infrastructure and they have very different physical properties from regular roadway sections. This study lays an important foundation towards developing the Safety Performance Function for bridges. It extends the applicability of the concepts promulgated in the Highway Safety Manual (HSM) by developing the SPFs for the bridges. The new bridge-specific SPFs can be used to support safety decision-making efforts in Alabama and other areas with similar physical and operation characteristics (traffic composition, driving behaviors, climate, bridge construction techniques, etc.). Additionally, as well form the basis for additional research into the development of similar relationships whether based on new location-specific SPFs for bridges or on calibration factors as set out in the HSM.

The study then demonstrates the use of the newly proposed MVCMP model to jointly analyze the crash counts classified by different severities for two-lane rural roads and four-lane divided highways. To illustrate the importance of the multivariate technique, it compares the coefficients with the univariate Poisson, univariate Conway Maxwell Poisson and Multivariate Poisson lognormal distribution. The parameters for the univariate Poisson lognormal model are estimated in SPSS, the parameter of CMP are estimated in R and Matlab is used to estimate the parameters for the other two models.

The comparison of all the four models on two different data set indicates that the both the MVCMP and MVPLN model perform better than the univariate models. The number of variables having significant impact on the expected crash severity is much higher for the univariate Poisson model. This is as expected, because the presences of apparent extra Poisson dispersion in the data set inflate the standard errors resulting in false identification of a variable

as having significant effect on the crash severity. The univariate CMP model does little better as the results obtained by this model are close to MVCMP and MVPLN model. It is also observed that for fatal crashes, the uncertainty estimates from MVCMP model are better than all the three models. As reducing severe crashes is an important objective of any transportation official, using MVCMP model will be a better choice.

Finally, there are some discrepancies observed that needs to be further investigated. The dispersion parameter estimates of MVCMP model for both the data sets do not completely match with the univariate models. Although, the bootstrapping procedure showed that for some of the univariate model, the dispersion parameter estimates are unstable. However, a further investigation is suggested to address this issue. The shoulder type variable sign for the possible injury and PDO crashes is counterintuitive. The coefficient suggests that as the type of shoulder improves from SOD to paved, the expected crashes also increase. This observation could be because of some confounding factors not included in the model, hence, it warrants a further investigation.

4.6 References

Chang, L., & Wang, H. (2006). Analysis of Traffic Injury Severity: An Application of Non-Parametric Classification Tree Techniques. Accident Analysis and Prevention, 1019-1027.

Chib, S., & Winklemann, R. (2001). Markov Chain Monte Carlo Analysis of Correlated Count Data. Journal of Business and Economic Statistics, 19(4), 428-435.

Chimba, D., & Sando, T. (2009). Neuromorphic Prediction of Highway Injury Severity. Advances in Transportation Studies, 19(1), 17-26.

Crespo-Minguillión, C., Casas, J. 1997. A Comprehensive Traffic Load Model for Bridge Safety Checking. Structural Safety, 19(4), 339-359.

Delen, D., Sharda, R., & Bessonov, M. (2006). Identifying Significant Predictors of Injury Severity in Traffic Accidents using a Series of Artificial Neural Networks. Accident Analysis and Prevention, 38(3), 434-444.

Eluru, N., Bhat, C., & Henser, D. (2008). A Mixed Generalized Ordered Response Model for Examining Pedestrian and Bicyclist Injury Severity Level in Traffic Crashes. Accident Analysis and Prevention, 40(3), 1033-1054.

Gardner, P. (2006). Segment Characteristics and Severity of Head-on Crashes on Two-Lane Rural Highways in Maine. Accident Analysis and Prevention, 38(4), 652-661.

Gates, T. & Noyce, D., 2005. The safety and Cost Effectiveness of Bridge-Approach Guardrail for County State-Aid Bridges in Minnesota, St. Paul : Minnesota Department of Transportation.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2014). Bayesian Data Analysis. Boca Raton, Florida: Chapman & Hall/CRC.

Haleem, K., & Abdel-Aty, A. (2010). Examining Traffic Crash Injury Severity at Unsignalized Intersections. Journal of Safety Research, 41(4), 347-357.

Hirai, T., Itoh, Y. & Liu, B., 2006. A Study on the Strain rate Effect of Vehicle Guard Fences Using Numerical Collision Analysis. 8th International Conference on Computational Structures Technology. pp. 527-542.

Hogg, R., Craig, M., & McKean, J. (2006). Introduction to Mathematical Statistics. Pearson Prentice Hall.

Holdridge, J., Shankar, V., & Ulfarsson, G. (2005). The Crash Severity Impacts of Fixed Roadside Objects. Journal of Safety Research, 36(2), 139-147.

HSM. (2010). Highway Safety Manual.

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(3), 267-276.

Khorashadi, A., Niemeier, D., Shankar, V., & Mannering, F. (2005). Differences in Rural and Urban Driver-Injury Severities in Accidents involving Large-Trucks: An Exploratory Analysis. Accident Analysis and Prevention, 37(5), 910-921.

Kim, J.-K., Ulfarsson, G., Kim, S., & Shankar, V. (2013). Drivere Injury Severity in Single Vehicle Crashes in California: A Mixed Logit Analysis of Heterogeneity due to Age and Gender. Accident Analysis and Prevention, 50, 1073-1081.

Lee, C., & Abdel-Aty, M. (2005). Comprehensive Analysis of Vehicle-Pedestrian Crashes at Intersections in Florida. Accident Analysis and Prevention, 37(4), 775-786.

Ma, J., Kockelman, K., & Damien, P. (2008). A Multivariate Poisson-Lognormal Regression Model for Prediction of Crash Counts by Severity, Using Bayesian Methods. Accident Analysis and Prevention, 40(3), 964-975.

Mehta, G., & Lou, Y. (2013). Safety Performance Function Calibration and Development for the State of Alabama: Two-Lane Two-Way Rural Roads and Four-Lane Divided Highways. Transportation Research Record 2398(1),pp.75-82.

Michie, J. 1981. Recommended Procedures for the Safety Performance Evaluation of Highway Appurtenances. NCHRP Report 230. National Cooperative Highway Research Program, Transportation Research Board, National Academy of Sciences. Washington, DC.

Pai, C. (2009). Motorcyclist Injury Severity in Angle Crashes at T-Junctions: Identifying Significant Factors and Analyzing What Made Motorists Fail to Yield to Motorcycles. Safety Science, 47(8), 1097-1106.

Park, E., & Lord, D. (2007). Multivariate Poisson-lognormal Models for Joint Modeling of Crash Frequency by Severity. Transportation Research Record, 1-6.

Press, S. (1982). Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference (2nd ed.). Malabar, Florida: Robert E. Krieger Publishing Company.

Retting, R., Williams, J. & Schwartz, S., 2000. Motor Vehicle Crashes on Bridges and Countermeasure Opportunities. Journal of Safety Research, 31(4), pp. 203-210.

Savolainen, P., & Mannering, F. (2007). Probabilistic Models of Motorcyclists' Injury Severities in Single- and Multi-Vehicle Crashes. Accident Analysis and Prevention, 955-963.

Savolainen, P., Mannering, F., Lord, D., & Quddus, M. (2011). The Statistical Analysis of Highway Crash-Injury Severities: A Review and Assessment of Methodological Alternatives. Accident Analysis and Prevention, 1666-1676.

Schneider, W., & Savolainen, P. (2009). Driver Injury Severity Resulting from Single Vehicle Crashes along Horizontal Curves on Rural Two Lane Highways. Transportation Research Record: Journal of Transportation Research Board, 2102, 85-92.

Siddiqui, N., Chu, X., & Guttenplan, M. (2006). Crossing Locations, Light Conditions, and Pedestrian Safety. Transportation Research Record: Journal of Transportation Research Board, 1982, 141-149.

Soltani, M., Moghaddam, T. & Karim, M., 2013. Analysis of Developed Transition Road Safety Barrier Systems. Accident Analysis and Prevention, Volume 59, pp. 240-252.

Thanh, L. & Itoh, Y., 2013. Performance of Curved Steel bridge Railings Subjected to Truck Collisions. Engineering Structures, Volume 54, pp. 34-46.

Turner, D. 1984. Prediction of Bridge Accident Rates. ASCE Journal of Transportation Engineering, 110(1), 45-54.

Yasmin, S., & Eluru, N. (2013). Evaluating Alternate Discrete Outcome Frameworks for Modeling CrashInjury Severity. Accident Analysis and Prevention, 506-521.

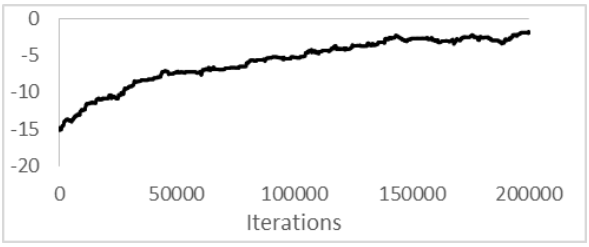
Ye, F., & Lord, D. (2011). Comparing Three Commonly Used Crash Severity Models on Sample Size Requirements: Multinomial Logit, Ordered Probit and Mixed Logit Models. Proceedings of the 90th Annual Meeting of the Transportation Research Board. Washington DC: Transportation Research Board.

Zhao, Z. & Uddin, N., 2013. Field Calibrated Simulation Model to Perform bridge Safety Against Emergency Breaking of Trucks. Engineering Structures, pp. 2253-2262.

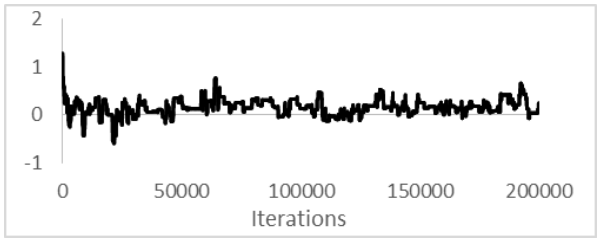
4.7 Appendix

Trace Plots for MVCMP Bridge Rail in Chapter 4

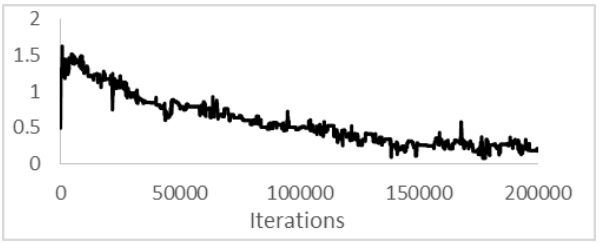
Trace Plot for Const



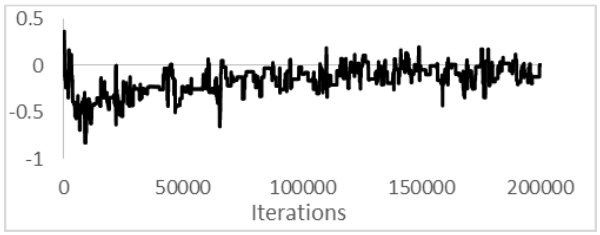
Trace Plot for β_1



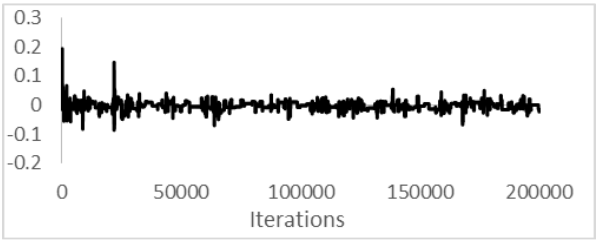
Trace Plot for β_2



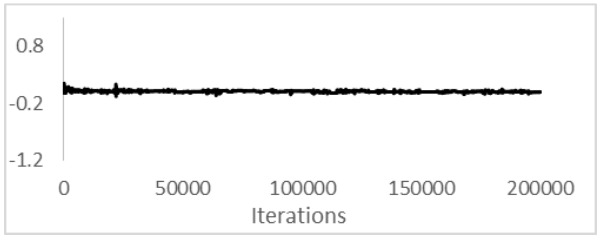
Trace Plot for β_3



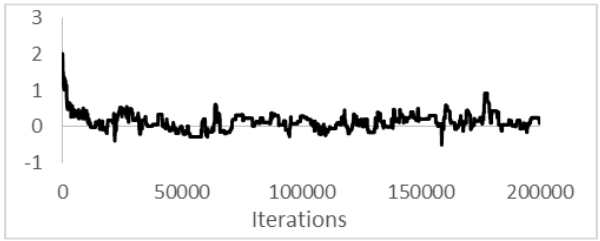
Trace Plot for β_4



Trace Plot for β_5

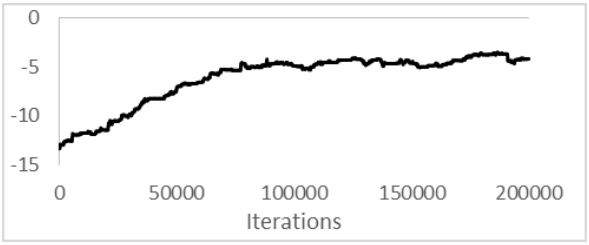


Trace Plot for β_6

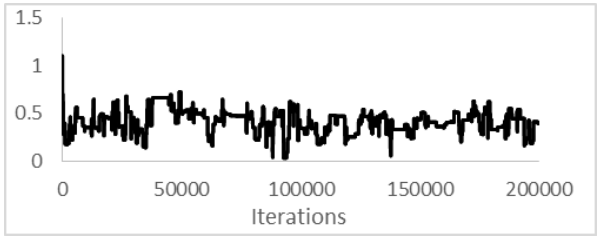


Trace Plots for Fatal Crashes

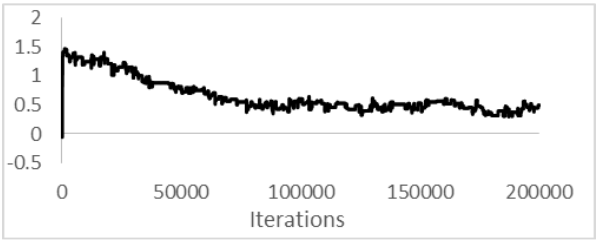
Trace Plot for Const



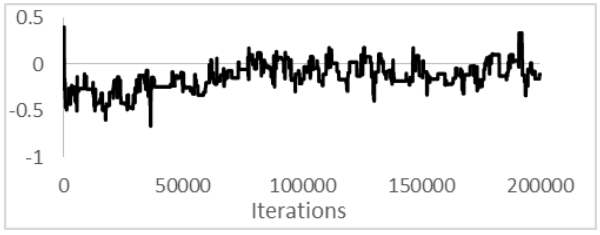
Trace Plot for β_1



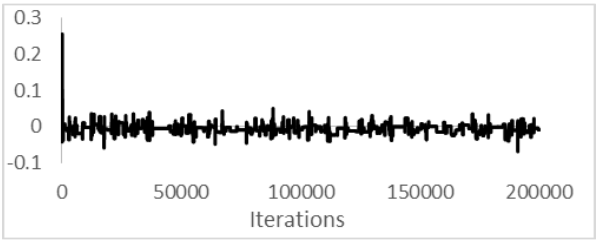
Trace Plot for β_2



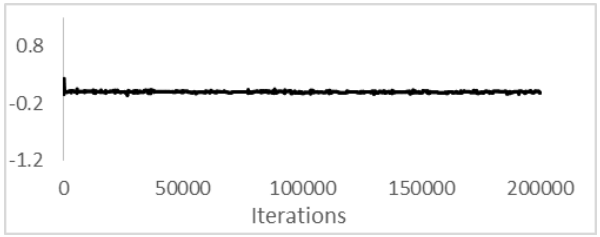
Trace Plot for β_3



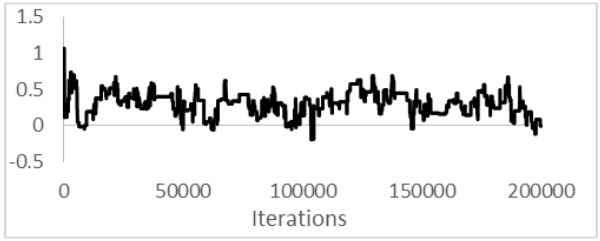
Trace Plot for β_4



Trace Plot for β_5

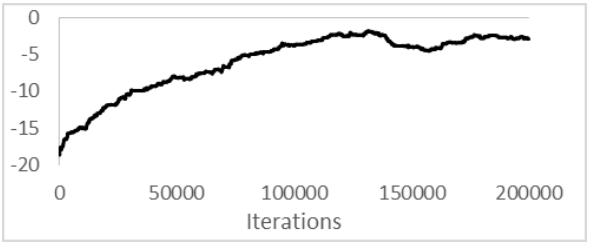


Trace Plot for β_6

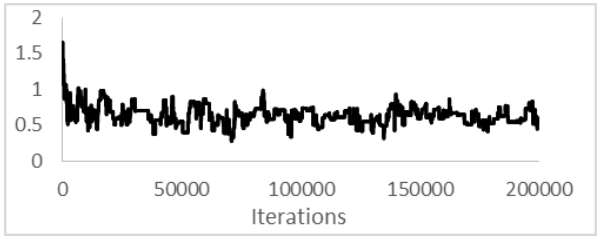


Trace Plots for Incapacitating Crashes

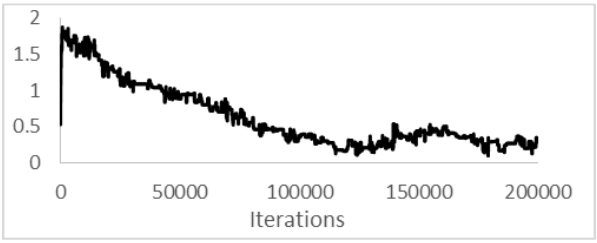
Trace Plot for Const



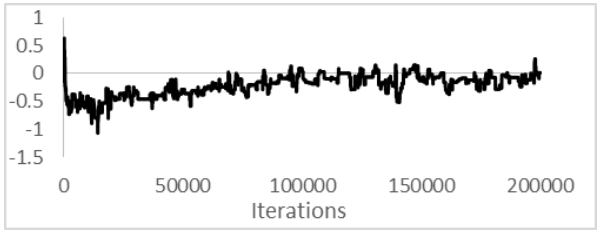
Trace Plot for β_1



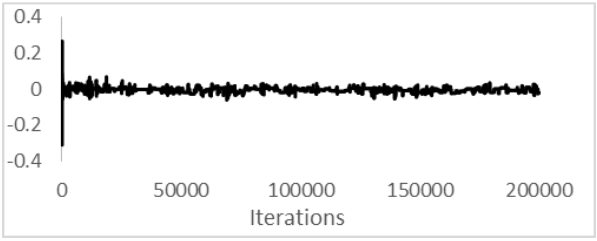
Trace Plot for β_2



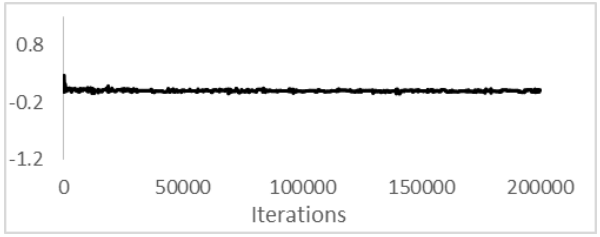
Trace Plot for β_3



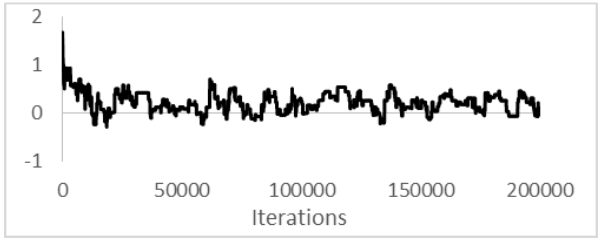
Trace Plot for β_4



Trace Plot for β_5

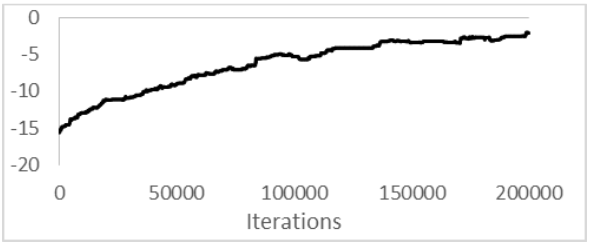


Trace Plot for β_6

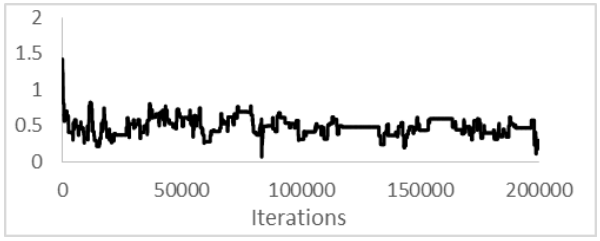


Trace Plots for Non-incapacitating Crashes

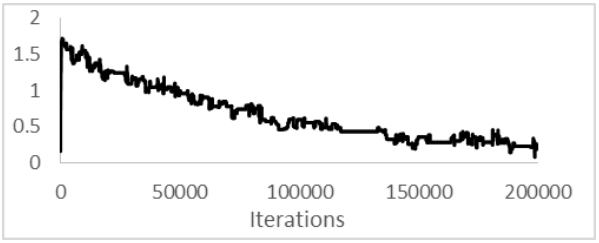
Trace Plot for Const



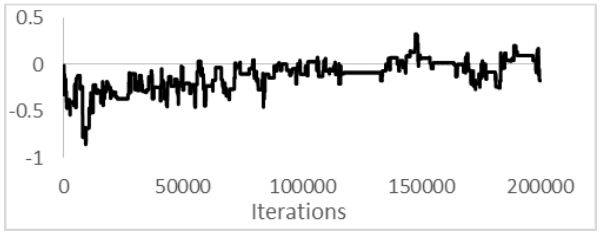
Trace Plot for β_1



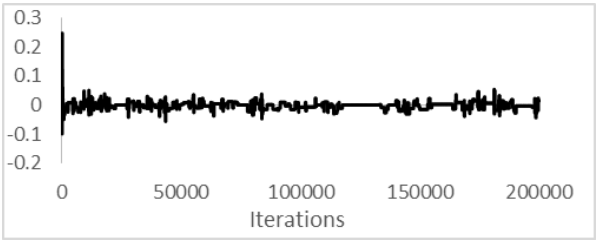
Trace Plot for β_2



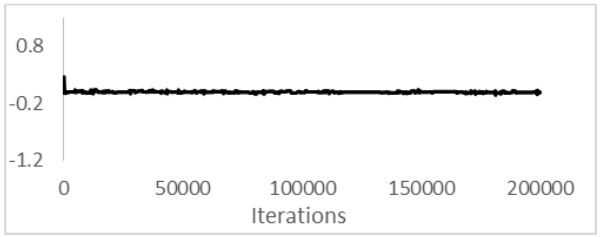
Trace Plot for β_3



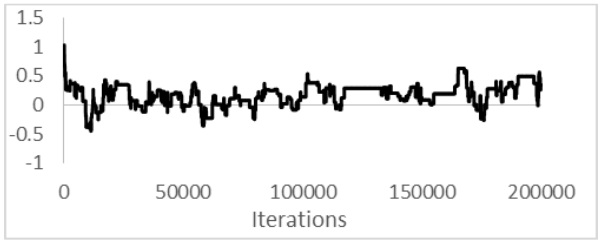
Trace Plot for β_4



Trace Plot for β_5

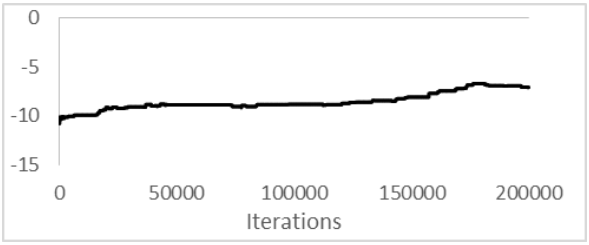


Trace Plot for β_6

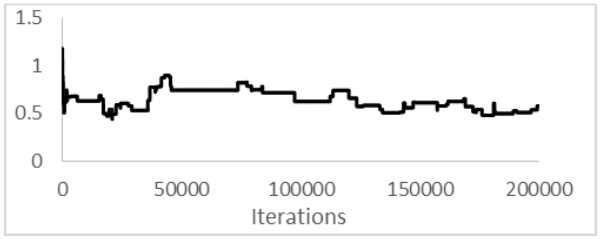


Trace Plots for Possible Injury Crashes

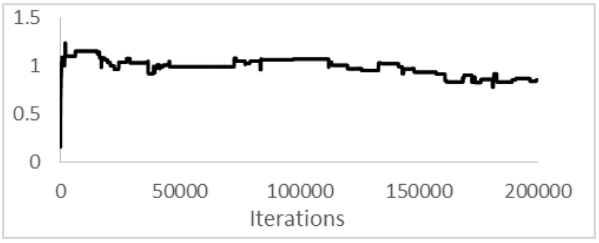
Trace Plot for Const



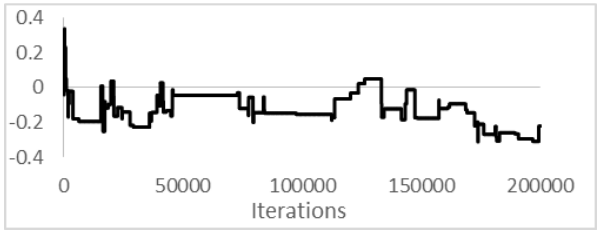
Trace Plot for β_1



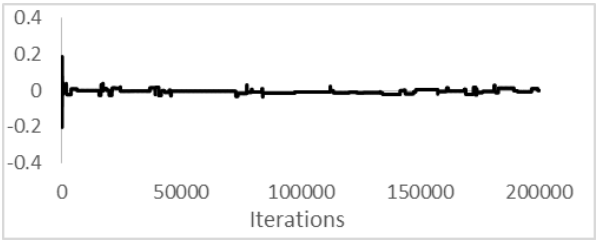
Trace Plot for β_2



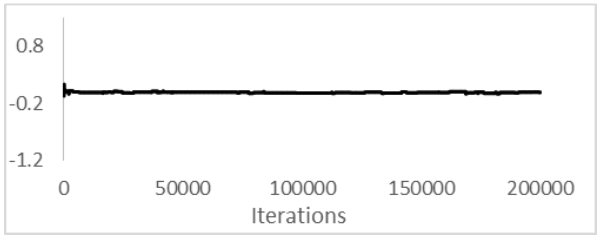
Trace Plot for β_3



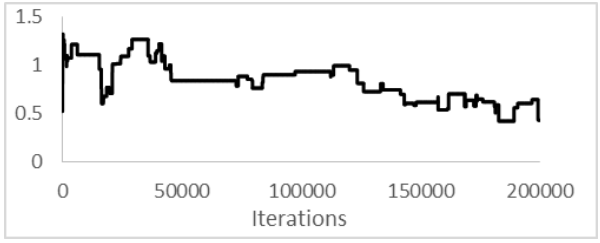
Trace Plot for β_4



Trace Plot for β_5

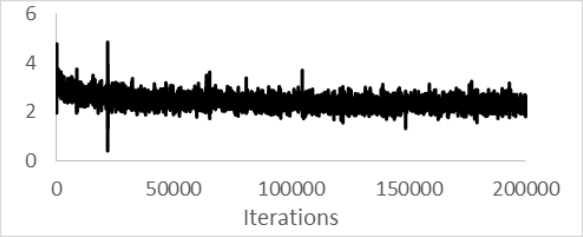


Trace Plot for β_6

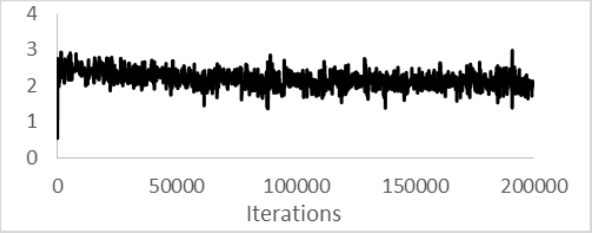


Trace Plots for Property Damage Only Crashes

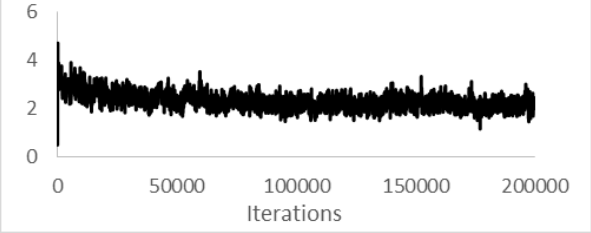
Trace Plot for nu Fatal Crashes



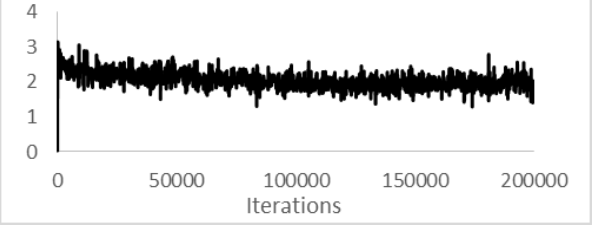
Trace Plot for nu Incapacitating Injury



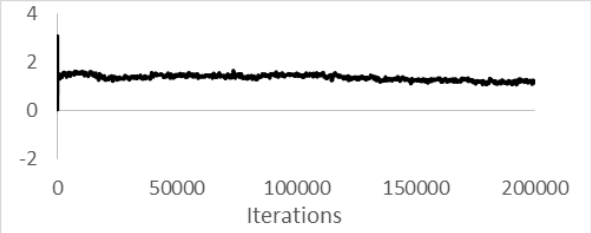
Trace Plot for nu Non-incapacitating Crashes



Trace Plot for nu Possible Injury Crashes



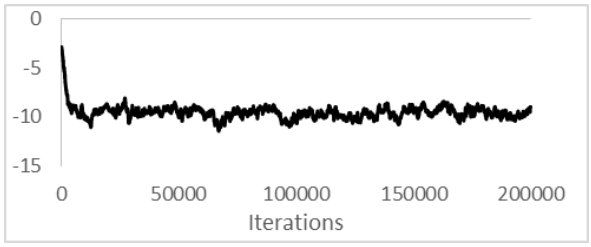
Trace Plot for nu Property Damage Only Crashes



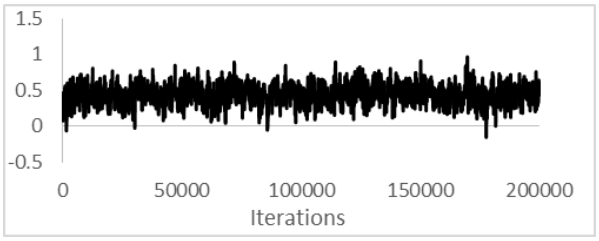
Trace Plots for Nu

Trace Plots for MVCMP Two Lane Rural Roads in Chapter 4

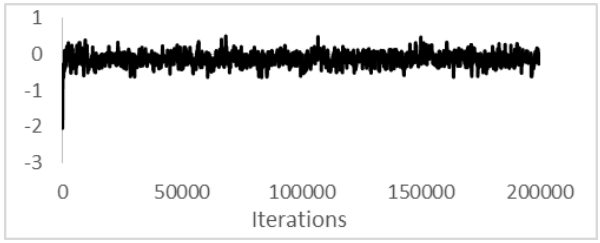
Trace Plot for Const



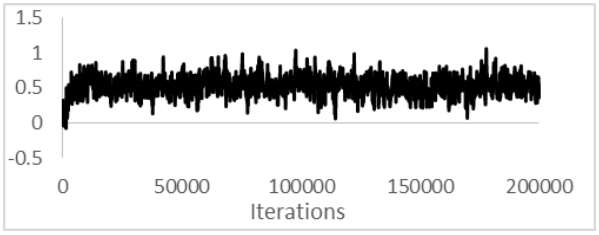
Trace Plot for β_1



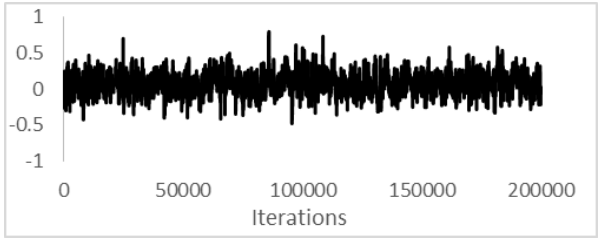
Trace Plot for β_2



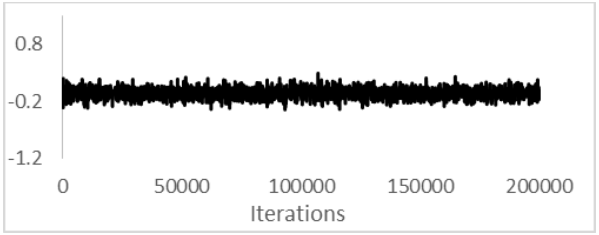
Trace Plot for β_3



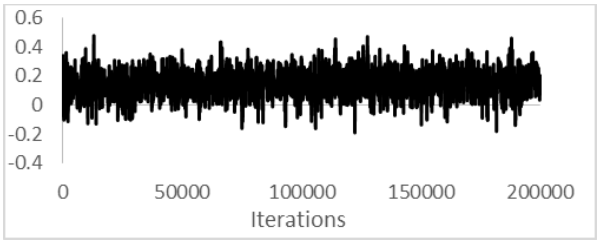
Trace Plot for β_4



Trace Plot for β_5

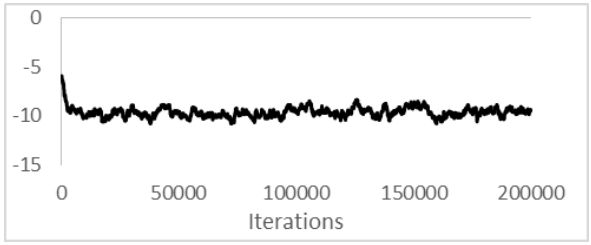


Trace Plot for β_6

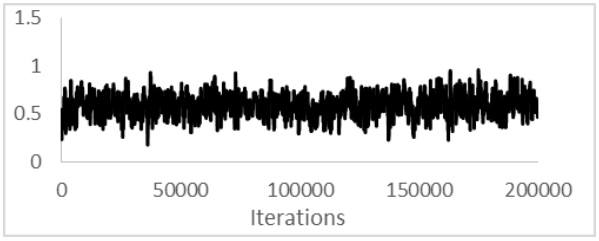


Trace Plots for Fatal Crashes

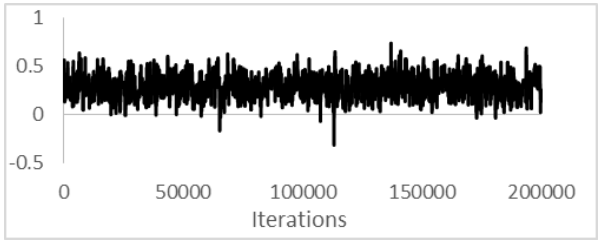
Trace Plot for Const



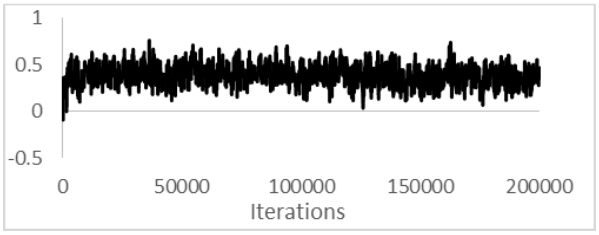
Trace Plot for β_1



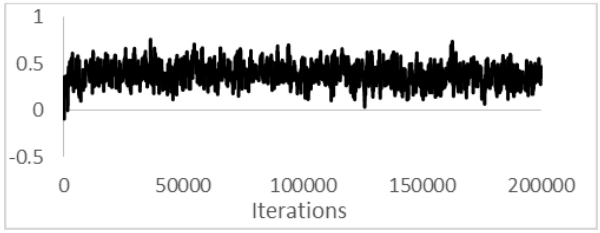
Trace Plot for β_2



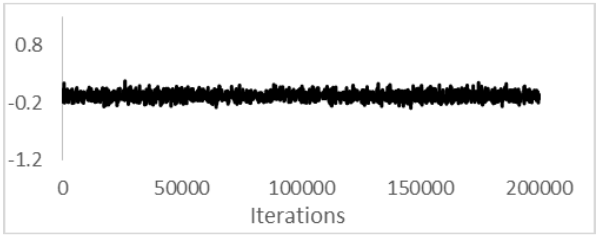
Trace Plot for β_3



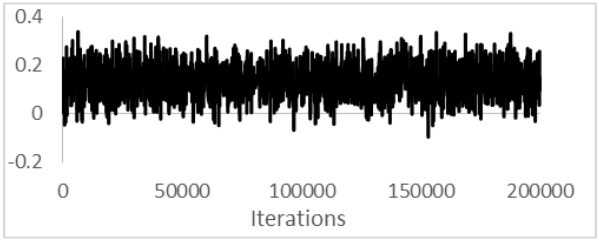
Trace Plot for β_4



Trace Plot for β_5

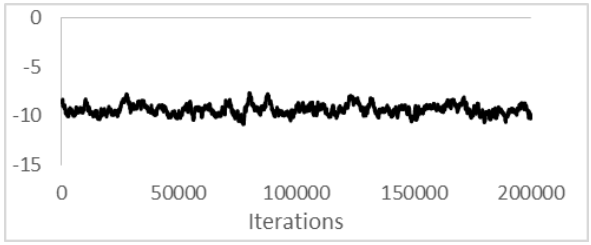


Trace Plot for β_6

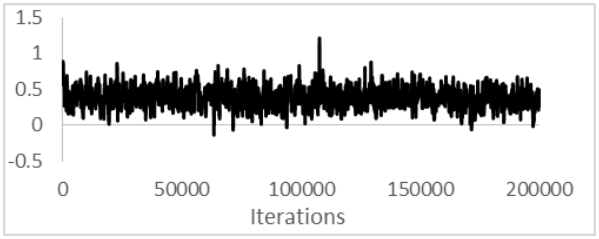


Trace Plots for Incapacitating Crashes

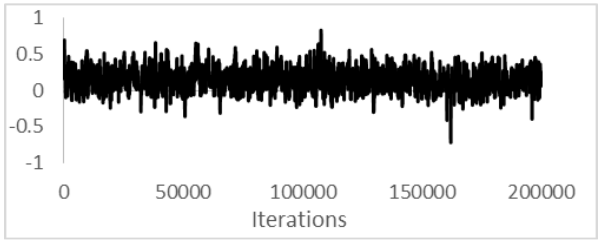
Trace Plot for Const



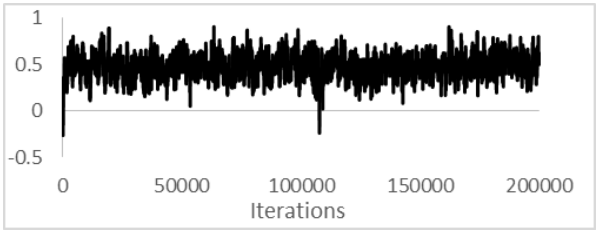
Trace Plot for β_1



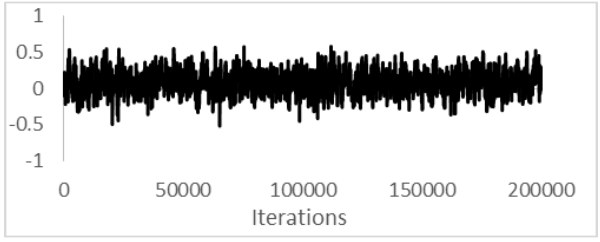
Trace Plot for β_2



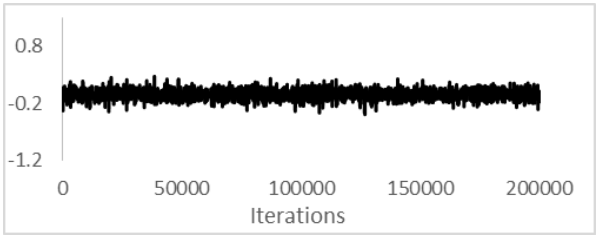
Trace Plot for β_3



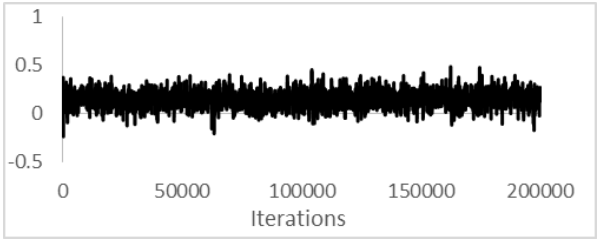
Trace Plot for β_4



Trace Plot for β_5

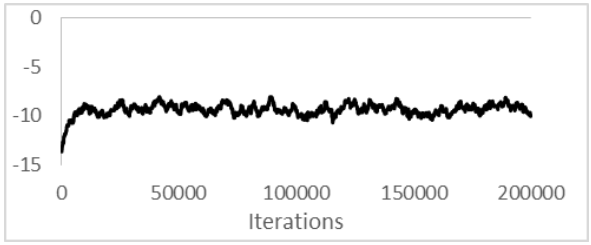


Trace Plot for β_6

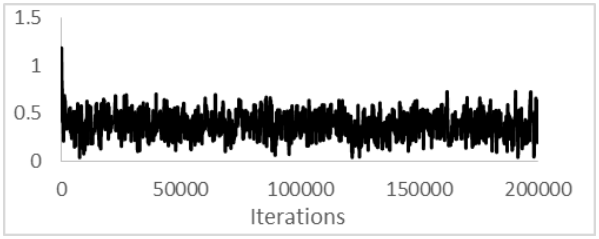


Trace Plots for Non-incapacitating Crashes

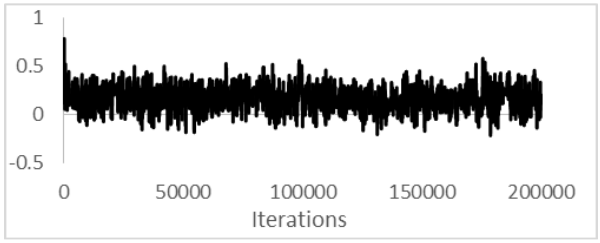
Trace Plot for Const



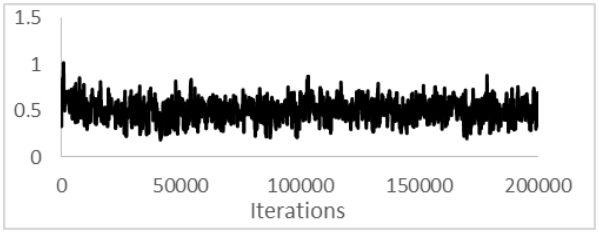
Trace Plot for β_1



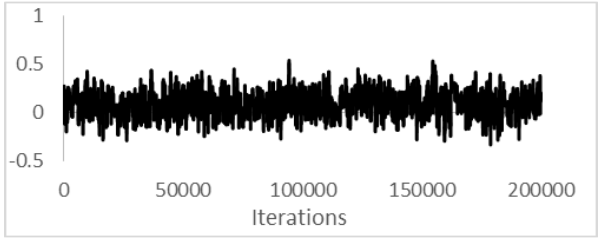
Trace Plot for β_2



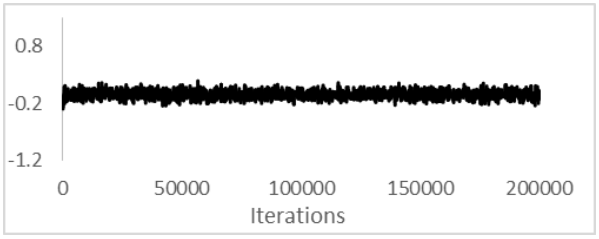
Trace Plot for β_3



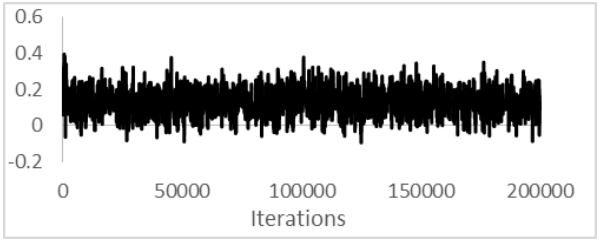
Trace Plot for β_4



Trace Plot for β_5

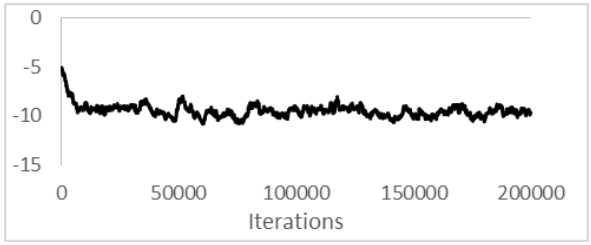


Trace Plot for β_6

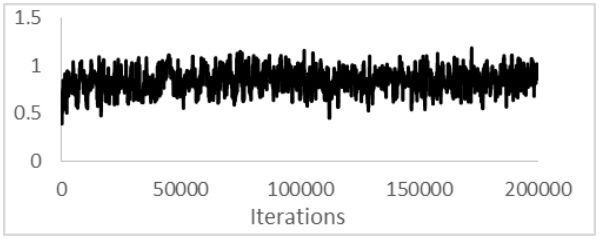


Trace Plots for Possible Injury Crashes

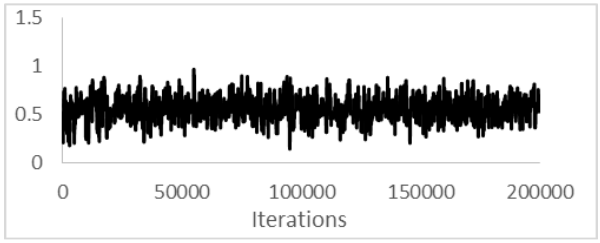
Trace Plot for Const



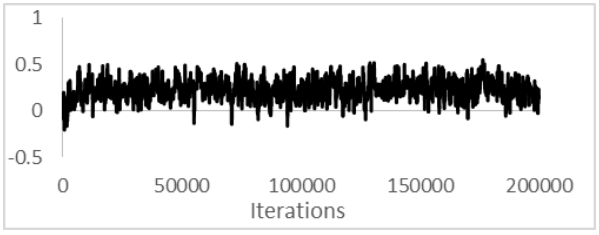
Trace Plot for β_1



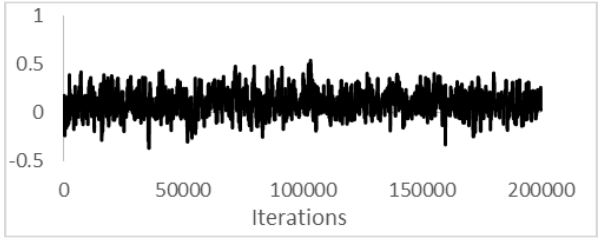
Trace Plot for β_2



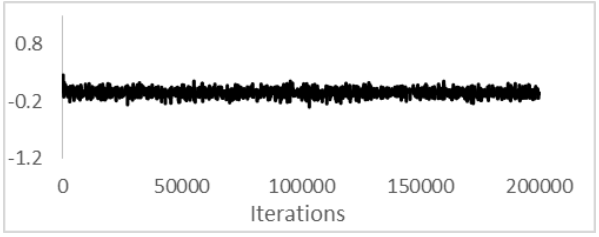
Trace Plot for β_3



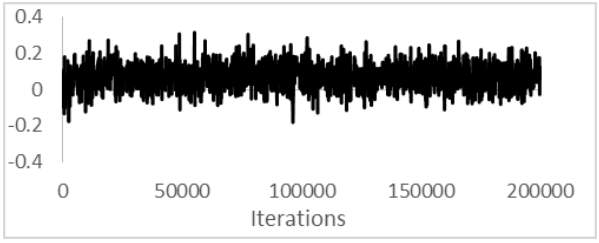
Trace Plot for β_4



Trace Plot for β_5

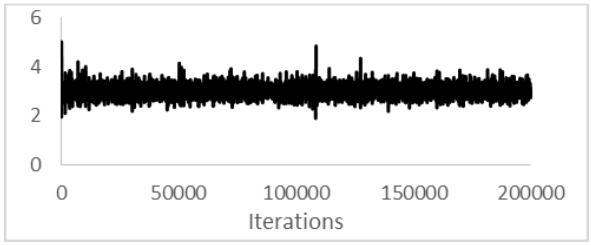


Trace Plot for β_6

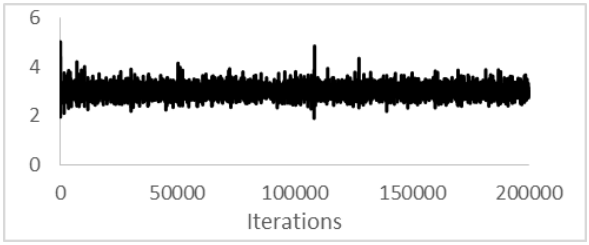


Trace Plots for Property damage Only Crashes

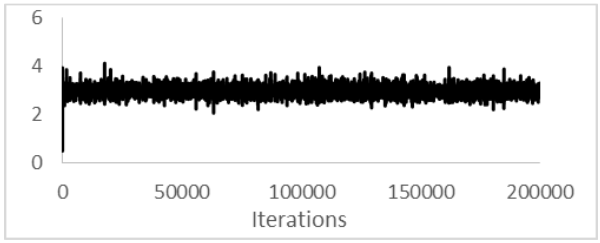
Trace Plot for nu Fatal Crashes



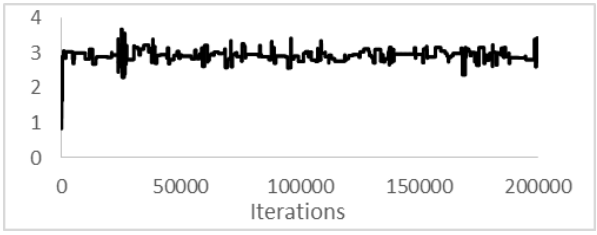
Trace Plot for nu Incapacitating Injury



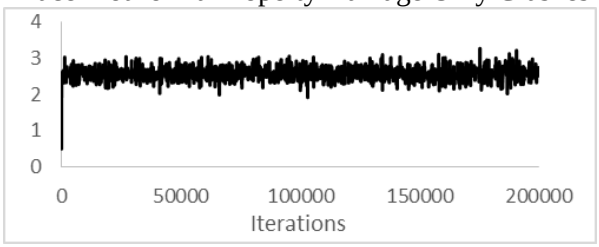
Trace Plot for nu Non-incapacitating Crashes



Trace Plot for nu Possible Injury Crashes



Trace Plot for nu Property Damage Only Crashes



Trace Plots for Nu

Chapter 5. CONCLUSION

Providing safe travel on roads is the main objective of any public transportation agency. The recent publication of the Highway Safety Manual has resulted in an increasing emphasis on the safety performance of specific roadway facilities. Crash prediction models are a tool provided by HSM that helps transportation agencies in making informed decisions concerning the improvement of roadway safety. Crash prediction models are used to estimate the expected number of crashes for a given type of roadway facility. This information can help in identifying sites for possible safety treatment and evaluating the effect of such treatments. However, only one edition of the HSM has been published, and there are some areas with room for extension. The first chapter of the dissertation introduces HSM along with different components of the prediction models. Chapters 2-4 are the core of this dissertation, and they extend the applicability of HSM.

The Chapter 2 focuses on proposing an alternative method of calibrating the base SPFs in HSM. The SPFs given in HSM are developed using data sets from a small number of states. The data is not guaranteed to be transferable to other states because of differences in driver behavior and jurisdictional differences. To account for these differences, the HSM recommends calibrating the model by using the ratio of observed and predicted crashes for a particular facility and taking it as a multiplier for estimating crashes on that facility type. However, this method is not scientific and can lead to inefficient results as shown in Chapter 2. Therefore, this study

proposes a probabilistic approach based on negative binomial regression to calibrate base SPF models given in HSM.

The proposed method is used on two lane rural road and four lane divided highway facilities. The results are compared with the calibration method suggested by HSM. State specific SPFs are developed as one of the other tasks in this chapter. The calibration factors derived from both approaches and for both types of facilities are greater than one, implying that the HSM base SPFs are underestimating the mean crash frequencies on TLTWRR and FLDH in Alabama. The proposed method performs better in terms of log likelihood; however, the mean crash frequencies predicted is higher than that of the HSM-recommended method for both facility types. The HSM-recommended calibration method seems to outperform the proposed new calibration method for these two types of facilities. However, the state-specific SPFsare found to outperform all other models, including both calibrated models. The calibration method suggested is a more scientific method, and it is definitely recommended to evaluate its performance in other states. The new calibration method can specially help transportation officials with not enough data or statistical expertise to develop their state specific SPFs.

The Chapter 3 extends the application of HSM to crash severity analysis. The HSM's 1st edition uses the univariate negative binomial regression model for predicting crashes. This models works well with the over-dispersed data; however, it cannot handle under-dispersed data sets. Lord, Geedipally, & Guikema (2010) showed that negative binomial models do not perform well, when the data is under-dispersed or characterized by low sample mean and small sample size. They recommend using the Conway Maxwell Poisson distribution to address the issue of modeling under-dispersed dataset. The univariate models also assume that there is no correlation

between the crash counts at different levels of severity for a particular road segment. However, they do not consider the distinct effect unobserved heterogeneity might have on crash severities.

To address this limitation this study proposed a multivariate extension of the Conway Maxwell Poisson distribution for predicting crashes. This study gives the statistical properties of the distribution. The study considers a general correlation structure of the counts at different level of severities along with both the over and under dispersion in the data. The parameters are estimated under the Bayesian paradigm using Gibbs sampler and Metropolis-Hastings algorithm. The proposed method is applied to the two lane rural road data, and a significant correlation matrix is obtained from the analysis. This indicates that considering a multivariate model for different severity types allows sharing of information across severities resulting in increasing predictive accuracy.

Finally, Chapter 4 further extends the applicability of the concepts promulgated in the Highway Safety Manual (HSM) by developing a safety performance function for bridge segments on roadway facilities. Bridges are integral part of the infrastructure and they have very different physical properties from regular roadway sections. This study lays an important foundation towards developing SPFs for bridges. The new bridge-specific SPFs can be used to support safety decision-making efforts in Alabama and other areas with similar physical and operation characteristics (traffic composition, driving behaviors, climate, bridge construction techniques, etc.). The study then demonstrates the use of the newly proposed MVCMP model to jointly analyze the crash counts classified by different severities for two-lane rural roads and four-lane divided highways. The comparison of all the four technique shows that the MVCMP and MVPLN model perform better than both the univariate models. It is also observed that for fatal crashes, the uncertainty estimates from MVCMP model are better than all the three models.

As reducing severe crashes is an important objective of any transportation official, using the MVCMP model should be considered for crash severity analysis.

There are certain limitations of this study that can be considered for future research.

1) The proposed Multivariate Conway Maxwell Poisson distribution model assumes no spatial correlation. This is a very strong assumption since driver characteristics in a particular region or some local jurisdiction could have some distinct effect on the crash frequency. The spatial correlation can account for these similarities.
2) The current model consists of parametric assumptions, such as normally distributed unobserved heterogeneities and CMP distributed conditional probabilities. Misspecification of parametric assumptions can result into erroneous results. Hence, a non-parametric extension of this model would be an interesting future study.
3) All the models developed in this study are based on Alabama crash data sets. The proposed model should be evaluated on data sets from other states to check their transferability.
4) Finally, the proposed MVCMP algorithm requires intense computation resources. The algorithm should be explored to identify if it can be parallelized to reduce the computation time.