A Full Bayes Approach to Road Safety: Hierarchical Poisson Mixture Models, Variance Function Characterization, and Prior Specification

Mohammad Heydari

A Thesis
in
The Department
of
Building, Civil and Environmental Engineering

Presented in Partial Fulfillment of the Requirements
for the Degree of Master of Applied Science (Civil Engineering) at
Concordia University

Montreal, Quebec, Canada

April 2012

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Mohammad Heydari**

Entitled: **A Full Bayes Approach to Road Safety: Hierarchical Poisson Mixture Models, Variance Function Characterization, and Prior Specification**

and submitted in partial fulfillment of the requirements for the degree of

MSc in Civil Engineering

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. Adel M. Hanna **Chair**

Dr. Luis Fernando Miranda-Moreno **Examiner**

Dr. Zachary Patterson **Examiner**

Dr. Ciprian Alecsandru **Examiner**

Dr. Luis Amador **Supervisor**

Approved by ----------
Chair of Department or Graduate Program Director

Dean of Faculty

Date ----------

ABSTRACT

A Full Bayes Approach to Road Safety: Hierarchical Poisson Mixture Models, Variance Function Characterization, and Prior Specification

Mohammad Heydari

Road safety is a major concern of every department of transportation. Allocating resources to improve safety requires the identification of hazardous sites (hotspots) and the assessment of safety countermeasures. For such tasks, reliable safety performance functions are noteworthy to predict collisions, prioritize improvements, and capture countermeasures effectiveness for overall road network management.

In this thesis, a case study from New Brunswick was used. Bayesian statistics were mainly applied in analyses by introducing Poisson mixture models in a hierarchical fashion. Poisson and Poisson mixture models were compared. Different characterizations of variance functions were verified. In a novel approach, the inverse of variance in Poisson-Lognormal models was examined to vary across sites as a function of site characteristics. In addition, accidents were analyzed by severity. Hierarchical Poisson-Gamma models presented the best fit. Traffic flow was the most influential factor in variance functions. Models with random variance structure provided the best fit, followed by those varying as a function of site characteristics. The interaction between precipitation and density of horizontal curves was statistically significant only for injury-

fatality accidents - these contributing factors weren't significant when considered separately.

Additionally, the effect of prior specifications in hierarchical Poisson-Gamma models was examined adopting a case study and a data simulation framework. Results showed that informative priors, especially for the inverse dispersion parameter, improve the accuracy of parameter estimates. Data with low sample mean and small sample size were dramatically affected by prior specification. However, hotspot identification and goodness-of-fit were not very sensitive to prior choice.

Acknowledgements

My sincere thanks are extended to my thesis supervisor, Professor Luis Amador for his encouragement and guidance. I would also like to thank him for the financial support provided during my studies.

My special thanks go to Professor Luis Fernando Miranda-Moreno for the insightful advice and very helpful comments.

Finally, I would like to thank my family, especially my mother, for their understanding, encouragement, and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

This chapter is divided into 4 sections: (1) background, (2) objectives, (3) scope and limitations, and (4) organization of the thesis, which are described as follows.

1.1 BACKGROUND

Traditionally, less consideration has been dedicated to road collisions from a management perspective. It was mainly over the past decade that road safety has widely attracted the attention of researchers, governments, and in particular, transportation authorities. Despite this attention and effort, there is still considerable potential for safety improvements in road networks in many countries; including developed ones. A study conducted in Canada estimates that accidents cost Canada $62.7 billion each year that is 4.9% of Canada's 2004 Gross Domestic Product (Transport Canada, 2007). The accident cost is mainly derived from two sources: economic and non-economic losses. Economic losses are related to physical damages to vehicles and infrastructures, injury recovery, administrative procedures, etc. And non-economic losses are those not directly measured in monetary terms like psychological consequences, pain, etc. Additionally, as reported by Transport Canada (2011), there were 2,209 fatalities, 11,451 serious injuries, and 172,883 injuries in Canada just in 2009. Currently 90% of the world's 1.2 million road fatalities per annum are in low and middle income countries, and by 2020 the number of road fatalities in these countries is expected to grow by 50% (International Road

Assessment Program, 2011). These numbers are compelling and clear evidence to urge for safety improvement programs.

In fact, road safety has lately become one of the major concerns in the transportation engineering community (e.g., in the USA and European Countries). This is first because of the global awareness of the problem. Second, there is a broad accordance on the fact that prevention is more desirable than post-crash medical care especially when one considers the social and economic cost of fatalities. Third, economic costs of countermeasures (treatments) for safety improvements are usually rapidly compensated by reduction on the number of observed accidents. At this stage, evaluating treatment effects and prioritizing sites where countermeasures should be destined become remarkable. In the literature, these prioritized sites that suffer from unsuitable safety conditions are called hotspots. It should be taken into account that road safety as part of an overall road network management requires reliable estimation of the safety of each entity (e.g., road segment). This estimation is then used to guide the decision making process in allocation of funds related to safety improvements that result in a safer network for road users. Furthermore, transportation engineering decisions and projects usually cause variations in road network characteristics (e.g., change in geometric design) that in turn could affect the safety of the network (Hauer, 1997). The latter also requires dependable evaluation of the safety before the implementation of such decisions and projects. There are different methods for both hotspot identification and countermeasure assessment; the most important ones rely on safety performance functions (SPF). In other words, the evaluation of the safety mainly depends on the quality and reliability of SPFs.

SPF is a mathematical equation that provides a relationship between accident frequencies and a series of site characteristics.

1.2 OBJECTIVES

The major goal of this work is to achieve reliable SPFs for accident modeling in traffic safety so that the management processes can be more effective from a cost and safety perspective. Specifically, the aim is to investigate some of the most important modeling approaches in road safety and to try to improve them theoretically and practically. Therefore, the first specific objective is to compare the basic Poisson model and two of the most common Poisson mixture hierarchical models (Poisson-Gamma and Poisson-Lognormal). The second specific objective is to explore various characterizations of the variance function in Poisson mixture models. The reason for this objective is the fact that characterization of the inverse dispersion parameter as a function of site characteristics in Poisson-Gamma models has been shown, by some researchers, to be able to improve goodness-of-fit and parameter estimation precision. The third specific objective is to analyze accidents by type and severity; property damage only, injury-fatality, and total accidents. Hence, it is possible to explore the effect of various site characteristics on the mean and the variance functions. Finally, since providing the prior distribution for model parameters in the Bayesian approach is necessary, another objective is to focus on the prior specification issues and to investigate on the model outcomes through various prior choices. Specific objectives are summarized in more detail as follows:

a) To compare Poisson and hierarchical Poisson mixture models

For this objective, in a fully Bayesian framework, three regression approaches (1) the Poisson model, (2) the hierarchical Poisson-Gamma model, and (3) the hierarchical Poisson-Lognormal model will be applied to a case study to compare each of the model outcomes that are parameter estimations and model fitting using a Bayesian goodness-of-fit measure.

b) To examine the effect of different variance function characterizations in hierarchical Poisson mixture models

To follow this objective, first, different characterization of the inverse dispersion parameter as a function of site characteristics will be investigated in the hierarchical Poisson-Gamma model. Second, in a novel approach this characterization methodology will be extended to the hierarchical Poisson-Lognormal model; this time for the inverse of variance. Third, a randomly varying approach will be adopted based on which the inverse dispersion parameter and the inverse of variance, instead of being fixed or variable as a function of site characteristics, will be allowed to vary randomly across sites.

c) To analyze accidents by type and severity

In this research study, accidents will be divided into three categories: property damage only, injury-fatality, and total accidents. Each category will be investigated separately, and the effect of different site characteristics on the mean and the variance functions will be estimated.

d) To examine the effect of the prior specification for model parameters on estimation accuracy, hotspot identification, and goodness-of-fit

Almost all studies in road safety that apply Bayesian statistics adopt a non-informative prior for model parameters. Since some problems have been reported to merge from the non-informative prior specification regarding accuracy of estimates especially when working with limited datasets, here, the impact of the informative prior specification on results will be examined. For this purpose, model outcomes obtained from different prior specifications will be compared in terms of parameter estimates, hotspot identification, and goodness-of-fit.

1.3 SCOPE AND LIMITATIONS

The scope of this thesis is limited to investigation on some of the most important accident modeling aspects in road safety. In other words, the aim is not only to verify the reliability of SPFs developed based on current methodologies in road safety, but also to suggest some practical improvements for these methodologies. The final outcome can then be used for the road safety management process in hotspot identification and in evaluation of countermeasure effectiveness. The most important limitation was related to data availability and the case study – provided by the New Brunswick Department of Transportation. In fact, not many site characteristics were available. Nevertheless, to overcome this limitation, firstly, a macro-level approach has been employed to develop SPFs. Secondly, to increase the size of data (observations) three different types (severities) of accidents were considered.

1.4 ORGANIZATION OF THE THESIS

This thesis consists of 5 chapters as explained in the following:

Chapter 1 provides an introduction by presenting the background, research objectives, and the scope and limitations.

Chapter 2 reviews literature to provide the reader with adequate background knowledge. This chapter has three parts: (i) review of road safety management, (ii) review of Safety Performance Functions, and (iii) review of Bayesian statistics.

The work described in Chapters 3 and 4 have been written as self contained papers and as such, each of these chapters has its own abstract, introduction, and methodology.

Chapter 3 compares some of the most important regression approaches in road safety: Poisson and hierarchical Poisson mixture models. Subsequently, this chapter examines different variance function characterizations in hierarchical Poisson mixture models. Lastly, it discusses the accident analysis by different severities.

Chapter 4 investigates the effect of various prior specifications on the analysis results in terms of parameter estimation accuracy, hotspot identification, and goodness-of-fit using a case study and a data simulation framework.

Chapter 5 provides a summary and conclusions.

CHAPTER 2

LITERATURE REVIEW

In this chapter some of the most important and basic aspects of road safety are reviewed. As briefly explained in the previous chapter, road safety management comprises two phases that are prioritizing sites for safety improvements and estimating countermeasure effects. The first section of this chapter is dedicated to the managerial road safety issues and the second section's focus is on the accident modeling or development of SPFs. Finally, the third section represents a concise review of Bayesian statistics.

2.1 ROAD SAFETY MANAGEMENT

The safety of a site is defined as the number accidents, or accident consequences, by kind and severity, expected to occur on that site during a specified time period (Hauer, 1997). For road safety management purposes, there are two periods of before and after treatment that are studied to evaluate the impact of a treatment (countermeasure). Indeed, these studies that take into account before and after periods are called before-after road safety studies. In the literature the widely accepted duration for each of the before and the after period is three years. Consequently, two concepts are typically used: (1) prediction, and (2) estimation. As explained by Hauer (1997), prediction is what would have been the safety of a site in the after period had treatment not been implemented, and estimation is what the safety of a treated site in the after period was with treatment in place.

The effectiveness of a countermeasure can be traced with different levels of accuracy. For example, this effectiveness can be estimated by using a raw approach that is just based on the comparison between accident counts from before and after the period in which countermeasures have been implemented. This approach is known as naïve before-after study (Hauer, 1997; Miaou and Song, 2005). The naïve approach has two main drawbacks; firstly, it ignores the fact that every change in the safety of a site after a treatment takes place may be due to other factors like variation in some site characteristics (e.g., traffic flow). Secondly, it may be characterized by the regression-to-the-mean phenomenon (Hauer and Persaud, 1984). This phenomenon, basically, comes from the fact that an abnormally high accident frequency on a site may decrease whether or not a countermeasure was implemented.

As an alternative, the accident rate has been also used to verify the effectiveness of a treatment and to prioritize sites. Accident rate, for instance for a road segment, is usually presented as the number of accidents in a specific period of time divided by vehicle kilometers traveled in that period (Hauer, 1997). However, this method also is not reliable enough since it assumes a linear relationship between the accident frequencies and the traffic flow. Moreover, the accident rate does not account for the importance of a transportation facility. This means that, for example, a road segment with very low accident frequency and very limited traffic flow may be given the priority over a more important site with higher accident frequency and much higher traffic flow.

To overcome the above mentioned problems and inaccuracies, the empirical Bayes (EB) method has been used being, in fact, the most common approach in road safety (Persaud et al., 1999; Heydecker and Wu, 2001; Hauer et al., 2002). The EB method can be

applied to prioritize sites for the safety improvements and also to estimate the impact of countermeasures on the safety of a site. This method combines the accident history of sites under investigation with their expected accident frequencies - estimations of the mean obtained from a reference group of sites with similar characteristics - to evaluate the safety effect of a treatment. The EB approach is a model based method in which an SPF, which is developed for an untreated reference group of sites, is used to predict the expected accident frequency for a series of treated sites under examination. In other words, the accident frequencies related to the before treatment period for the treated sites are known from the observed data. Subsequently, it's possible to verify how a particular countermeasure affected the occurrence of accidents on the treated sites by using the expected accident frequencies obtained from the reference sites, which have not been subject to any treatment.

Recently, some studies have suggested the use of the full Bayes approach for before-after studies (Lan et al., 2009; Persaud et al., 2010). The full Bayes approach can provide some improvements and advantages with respect to the EB approach such as: the possibility to add more flexibility and complexity in the model, and to obtain a better and more interpretative uncertainty around the estimated values. Both approaches, the EB and the full Bayes, are model based methods (Miaou and Song, 2005) in which the use of an SPF is indispensable for the analysis. At this point, the importance of developing reliable SPFs becomes clear, which is reviewed in the next section.

2.2 ACCIDENT MODELLING & SAFETY PERFORMANCE FUNCTIONS

An accident model, or an SPF, is a mathematical equation that describes the accident frequencies (as dependent variable) based on a series of site characteristics (as independent variables). These independent variables are also known as contributing factors. Equation 2.1 represents a generic SPF, for road segments, where μ_i = accident frequency of site i; α = vector of SPF parameters, and x = vector of site characteristics of site i.

$$ln\ (\mu_i) = \alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_n x_{in} \quad [2.1]$$

Statistical methods are used to develop SPFs; these models mainly depend on the historical observation of accident counts. To develop SPFs the first step is the choice of a model function that may include different contributing factors; then, a regression approach should be applied to the determined model function for the parameter estimation purpose. Regarding regression approaches, accident occurrences are typically assumed to follow the Poisson distribution due to their random nature (Equations 2.2).

$$k_i \sim Poisson\ (\theta_i) \quad [2.2]$$

where

k_i = observed accident frequency for site i;

θ_i = Poisson parameter or expected accident frequency for site i.

Moreover, the Poisson probability density function is defined in Equation 2.3 where the probability of having k accidents for an specific site, $P(k)$, is calculated based on θ; expected accident frequency (mean) of that site.

$$P(k_i) = (e^{-\theta_i} \theta_i^{K_i})/k_i! \quad [2.3]$$

The basic regression approach in developing SPFs is the Poisson regression, and almost all other regression approaches (Poisson mixtures) are an extension of this regression; for instance, Poisson-Gamma and Poisson-Lognormal models. In these models the expected accident frequency (θ_i) is defined as $\mu_i r_i$ where μ_i is calculated from the SPF and r_i is the multiplicative random effect. In Poisson-Gamma models, r_i is assumed to follow a Gamma distribution while in Poisson-Lognormal models r_i is assumed to be log-normally distributed.

Since the Poisson regression is not completely appropriate for accident data in most of the cases, Poisson mixtures are usually used. The assumption of the simple Poisson model is that the mean and the variance are equal; that is, these two parameters are only described by the SPF. However, this assumption is not satisfied in many accident data where heterogeneity across sites is a usual fact (Mitra and Washington, 2006). The heterogeneity mainly shows itself in the form of over dispersion, which implies that the variance is greater than the mean. Therefore, to circumvent such a problem the Poisson-Gamma (Negative Binomial) model is often used in accident data analysis (Poch and Mannering, 1996; Hinde and Demetrio, 1998; Miaou and Lord, 2003; Anastasopoulos and Mannering, 2008). In the Poisson-Gamma model, the variance is greater than the mean so that this model accounts for the over dispersion phenomenon. Moreover, other

Poisson mixture models like Poisson-Lognormal have been used for accident data analysis in different studies (Lord and Miranda-Moreno, 2008; El-Basyouny and Sayed, 2010). Poisson-Lognormal can also overcome the over dispersion problem.

In road safety, Poisson-Gamma models with fixed dispersion parameter have been commonly used; however, recently some researchers have proposed a methodology in which the dispersion parameter varies across sites as a function of some site characteristics such as traffic flow and segment length (Hauer, 2001; Miaou and Lord, 2003; Geedipally et al, 2009). This approach was first examined by Hauer (2001) where the dispersion parameter was a function of the segment length. Moreover, Geedipally and Lord (2008) have studied the effect of the varying dispersion parameter as a function of site characteristic on confidence intervals of SPFs estimations. This study was focused on intersection data, and the dispersion parameter was a function of the minor and the major traffic flows. Results showed that in general a varying dispersion parameter approach provides more precise results.

2.3 BAYESIAN STATISTICS

Road safety, heavily, relies on statistical analysis in order to develop SPFs based on local observations. Maximum Likelihood Estimation (MLE) is perhaps the most common method used to estimate model parameters in statistical analyses (Hauer, 1997; Bedford and Cooke, 2001; Winkelmann, 2003). However, the use of Bayesian estimation that requires a large amount of computation (Gelman et al., 1995; Carlin and Louis, 2009) has became very popular especially in the last decade because of the computational capacities found in personal computers. Bayesian statistics have some advantages with respect to

MLE such as (i) interesting probabilistic interpretative properties, (ii) superiority in dealing with uncertainty and randomness, and (iii) the ability to analyze complex data and data comprising small number of observations (Mitra and Washington, 2006; Gelman and Hill, 2007; Amador and Mrawira, 2011). Additionally, in the Bayesian approach hierarchical models can be introduced in the analysis adding more flexibility in the model and some improvements in the analysis (Congdon, 2010). Hierarchical models are those in which one or more parameters of the model are in turn dependent on a series of other parameters (called hyper-parameters) based on certain probability density functions (hyper-priors). In this case, hyper-parameters follow a particular prior distribution too. So that different levels of hierarchy can be set up in the analysis. The Bayesian paradigm is widely used in some fields; for instance, reliability engineering and medicine, particularly epidemiology. In road safety, also, some researchers have applied Bayesian methods for hotspot identification, evaluation of countermeasure effectiveness, and parameter estimations in developing SPFs (Oh and Washington, 2006; Lord and Miranda-Moreno, 2008; El-Basyouny and Sayed, 2010).

Bayesian statistics include three elements: (1) prior distribution, (2) likelihood distribution, and (3) posterior distribution. In Bayesian statistics, it's necessary to provide a prior for each parameter. The prior consists of some sort of knowledge that exists for a certain parameter based on previous studies or expert criteria. The likelihood is obtained by the observed data, and consequently, the posterior inference can be made based on these two; the prior and the likelihood. In particular, the process of making posterior inferences takes advantage of Markov Chain Monte Carlo (MCMC) methods (Gamerman and Lopes, 2006) to overcome computational complexity and difficulties of the Bayesian

approach. Equations 2.4 and 2.5 show the estimation of the posterior in Bayesian statistics.

$$P(a|Data) \propto f(Data|a).P(a) \quad [2.4]$$

$$P(a|Data) = \frac{P(Data|a)P(a)}{\int P(Data|a)P(a)da} \quad [2.5]$$

where

P(a): prior distribution of parameter *a*;

P(Data|a): likelihood function; observed data given parameter *a*;

P(a|Data): posterior distribution of parameter *a* given observed data;

and the denominator represents the marginal likelihood.

One of the most important concerns in running MCMC simulations for the posterior inference is whether or not iterations are stable and convergence is reached. In fact, by running more than one chain the convergence can be verified graphically in conventional software such as OpenBUGS (used for running MCMC simulations for Bayesian inference). Finally, deviance information criterion (DIC) is used as a goodness-of-fit measure in Bayesian statistics. DIC is a generalization of the Akaike information criterion (AIC) and can be used to compare model-fitting of statistical models that are developed for the same data. Models with smaller DIC value are those that fit the data better. Usually, differences of smaller than 5 in DIC values are not considered significant, whereas differences that are greater than 10 are generally important and indicate the superiority of the model that has the smaller DIC value.

CHAPTER 3

Hierarchical Poisson Mixture Models, Variance Function Characterizations, and Severity Modeling

Abstract

The Objectives of this research are**:** (1) to examine differences in parameter estimation and goodness-of-fit between Poisson and hierarchical Poisson mixture models, (2) to investigate different variance function characterizations under hierarchical Poisson-Gamma models with a novel extension to hierarchical Poisson-Lognormal models, (3) to investigate a randomly varying approach in which the inverse dispersion parameter and the inverse of variance vary randomly across sites, (4) to verify the statistical significance of various site characteristics in variance functions, and (5) to examine the effect of different contributing factors in the specification of models by severity and accident type.

This study applied Poisson and hierarchical Poisson mixture models in the Bayesian context, to a case study. For Poisson-Gamma models, the inverse dispersion parameter was incorporated in the models as (a) fixed, (b) varying as a function of site characteristics, and (c) randomly varying. Similarly, for Poisson-Lognormal models same procedures were adopted. That is, the inverse of variance was characterized as fixed, varying as a function of site characteristics, and randomly varying. Three datasets including different types of accidents (property damage only, injury-fatality, and total accidents) were used, and influence of various contributing factors, for each type, on mean and variance functions were verified.

In this research study, hierarchical Poisson-Gamma models presented the best fit to all three datasets. Models with random variance structure provided the best fit, followed by those varying as a function of site characteristics and then fixed structures. AADT was the most influential factor in variance functions. Density of horizontal curves was significant in variance functions for modeling property damage only and total accidents while it wasn't bounded away from zero in mean functions. The interaction between precipitation and density of horizontal curves was found to be statistically significant for injury-fatality accidents. However, these contributing factors when considered individually didn't have any important effect on injury-fatality accidents.

Moreover, this study indicated that, similar to the inverse dispersion parameter in Poisson-Gamma models, the inverse of variance in Poisson-Lognormal models can be defined as a function of site characteristics in order to improve estimation precision and goodness-of-fit. In addition, a randomly varying structure for the inverse dispersion parameter and the inverse of variance can be used that cause a noteworthy improvement in model-fitting, especially when the mean function, solely, cannot provide an adequate fit to the dataset. Finally, this study showed that modeling accidents by severity is crucial in identifying contributing factors that affect different accident frequencies.

3.1 INTRODUCTION

Road safety as part of an overall road network management requires a reliable estimation of the safety of each entity (e.g., road segment), not only to guide the decision making process in allocation of funds related to safety improvements but also to provide a safer network to road users. Furthermore, transportation engineering decisions and projects usually cause variations in road network characteristics (e.g. change in geometric design), which in turn, could affect the safety of the network (Hauer, 1997). Therefore, an estimation of the safety effects of such a decision is necessary. Safety of an entity can be expressed as the number of accidents, or accident consequences, by kind and severity, expected to occur during a specified period of time (Hauer, 1997). For the purpose of estimating expected accident frequencies, safety performance functions (SPF), or accident prediction models, can be developed. SPF is a mathematical equation that explains observed accidents based on specific site characteristics. Safety performance modeling relies on historical observations in order to calibrate a functional form that captures interactions between contributing factors and the safety response to local conditions in terms of accident frequency. In this study, we applied Bayesian estimation (Gelman et al, 1995; Gamerman and Lopes, 2006) in order to develop SPF's based on local observations. Bayesian estimation presents some advantages over classical methods (e.g., Maximum Likelihood Estimate) such as capacity to deal with uncertainty associated to contributing factors and to produce more reliable estimates even in cases of small sample size (Mitra and Washington, 2006; Gelman and Hill, 2007, Amador and Mrawira, 2011).

Many transportation engineers and researchers have focused on road safety issues. The main objective has been to examine various factors that may affect expected accident frequencies of transportation facilities, and to develop statistical models that can, accurately, describe accident datasets (Milton et al, 2008; Caliendo et al, 2007). Traditionally, Poisson distribution has been used as a regression approach for modeling accident data (Hauer, 1997). However, equality of the mean and the variance – which is a characteristic of Poisson models – is considered an important disadvantage of this approach (Anastasopoulos and Mannering, 2008; Mitra and Washington, 2006). For instance, Poisson regression does not provide an adequate fit in cases in which data is over dispersed, a common case among accident datasets. Approaches like Poisson-Gamma (Negative Binomial) regression have been adopted to overcome this deficiency (Poch and Mannering, 1996). In fact, the presence of a random effect inside the Poisson-Gamma mean structure allows it to deal with heterogeneity across sites; intersections and road segments. Other researchers suggested the use of generalized negative binomial regression in which the over dispersion parameter is expressed as a function of length and traffic volume (Hauer, 2001; Miaou and Lord, 2003; Geedipally et al, 2009). In addition to the Negative Binomial models, other probability density functions, typically, Poisson mixtures have been adopted like Poisson-Lognormal model (Lord and Miranda-Moreno, 2008; El-Basyouny and Sayed, 2009). The latter also has shown to be a good candidate to substitute simple Poisson regression, because of the presence of a multiplicative random effect in its mean specification that can capture heterogeneity across sites.

Increased computational capabilities have made possible the adoption of hierarchical Bayesian models (Gelman et al, 1995). For this purpose, Markov Chain Monte Carlo

simulation, MCMC, is applied (Gamerman and Lopes, 2006). The use of MCMC methods for estimating hierarchical models, often involve complex data structures, sometimes described as revolutionary development, and has arguably facilitated the fitting of such models (Congdon, 2010). As a matter of fact, some researchers have employed hierarchical Bayesian models in road safety (Miranda-Moreno et al, 2007). Therefore, Poisson mixture models can be adopted in a hierarchical fashion to profit from the additional advantages of their variance function and extra variation in order to account for heterogeneity in data. Indeed, the role of the extra variation in the hierarchical model is not simply to account for the lack of fit of a simpler model, but to use it as a tool to detect irregular patterns and changes in observations (Kim et al, 2002).

Researchers like Miaou and Lord (2003), Mitra and Washington (2007), and Geedipally et al (2009) have investigated various characterization of the dispersion parameter, under Negative Binomial regression, as a function of site characteristics, and provided some comparisons and inferences mainly using non Bayesian methods. In this study, first, we extended the same methodology to hierarchical Poisson-Lognormal models. Second, we examined a randomly varying approach, and finally, all the estimations were obtained in a fully Bayesian framework. Three approaches to specify dispersion parameter in the model were applied in this study: (1) fixed, (2) varying as a function of site characteristics, and (3) randomly varying. Furthermore, for Poisson-Lognormal models a similar methodology, this time for the inverse of variance, was adopted to investigate the effect of the characterization of the variance function on model estimation and fit. In order to increase the number of observations and datasets, three types of accidents (property damage only, injury-fatality, and total accidents) were analyzed separately

under three regression approaches. And, various contributing factors in mean and variance functions were examined to determine statistically significant factors under different model structures and accident types. Accident data used in this study was based on 958 kilometers of rural highway segments in New Brunswick, Canada. The statistical software OpenBUGS was used for running MCMC stochastic simulations applying Gibbs sampler in order to estimate the posterior distributions of models' parameters.

3.2 METHODOLOGY

3.2.1 Safety Performance Function (SPF)

The SPF that represents a non linear mathematical relationship between accident frequencies (expected number of accidents per unit of time) and a vector of contributing factors (e.g. traffic flow and environmental exposure) is used for the purpose of evaluating the safety of road segments or intersections. Having a reliable SPF is important for examining the effect of various contributing factors on expected accident frequencies and also to identify hazardous sites. Three main steps are required in order to develop an SPF: (a) choice of an appropriate model function, (b) choice of a regression approach, and (c) estimation of parameters presented in the model based on local observations.

Equation 3.1 (El-Basyouny and Sayed, 2010) is a widely accepted SPF, applicable on road segments, which was used in this paper. The main contributing factors were annual average daily traffic (AADT), segment length and a series of site characteristics as reported in Table 3.I.

$$\ln(\mu) = \ln(a_0) + a_1 \ln(L) + a_2 \ln(AADT) + \sum ax \quad [3.1]$$

where, μ = expected accident frequency;

$ln(a_0)$ = constant;

a = vector of stochastic parameters to be estimated using Bayesian inference;

L = segment length (km);

$AADT$ = annual average daily traffic (vehicles per day);

x = vector of site characteristics.

3.2.2 Regression Models in a Fully Bayesian Framework

Three different regression approaches were tested in a fully Bayesian framework in order to estimate the parameters (coefficients) of the SPF and to estimate the expected accident frequencies for each site. We assumed that accident data might be explained by: (a) Poisson model, (b) hierarchical Poisson-Gamma model, or (c) hierarchical Poisson-Lognormal model.

3.2.2.1 Poisson model

The assumption of this model is that accidents occur following the Poisson distribution with the mean and the variance being equal (Hauer, 1997). In such a case, the mean value for the expected number of accidents, θ, is only described by known site characteristics; that is, the SPF. A Poisson model is expressed as $k \sim Poisson\ (\theta)$ where k

is the number of observed accidents over an specific period of time. And θ is the function of the contributing factors' vector x and vector of unknown parameters a; $\theta = f(x, a)$. In other words, θ is the mean value obtained from the SPF. Equality of the mean and the variance is a drawback for this model since it cannot deal with the heterogeneity across sites, a typical trait in accident data (Mitra and Washington, 2006). The Poisson model was used as a base case scenario for comparison purposes with Poisson mixture models believed capable of overcoming the heterogeneity issue that usually shows itself in the form of over-dispersion in accident data.

3.2.2.2 Hierarchical Poisson-Gamma model

In this case, the assumption is that accidents within sites are Poisson and unobserved accident heterogeneity across sites is gamma distributed (Washington et al, 2003). Therefore, the expected accident frequency (θ) is described by the SPF and a multiplicative random effect, r, which varies across sites. The model is expressed as $k \sim Poisson(\theta)$ where k is the observed accident frequency, and $\theta = \mu r$ with μ as a function of the contributing factors' vector x and the vector of unknown parameters a; $\mu = f(x, a)$. Random effect r is assumed to follow a Gamma distribution ($r \sim gamma(\varphi, \varphi)$) with mean of 1 and variance of $1/\varphi$; where φ is the inverse dispersion parameter (Anastasopoulos and Mannering, 2008; Miranda-Moreno et al, 2007). In Bayesian hierarchical models (Congdon, 2010), φ is also assumed to have a Gamma distribution with shape and scale parameters a and b, respectively ($\varphi \sim gamma(a, b)$). These parameters are assumed to be identical and can be equal to 0.01, and φ is, therefore, defined as $\varphi \sim gamma(0.01, 0.01)$ with a *mean* = 1 and a large *variance* = 100. This

relatively large variance indicates a non informative hyper-prior (Miaou et al., 2003; Lord and Miranda-Moreno, 2008).

Traditionally, Inverse dispersion parameter has been assumed to be fixed across sites, which means that the variance function of Poisson-Gamma model is only described by the mean of accident counts. However, recent research proved that this parameter may vary across sites as a function of some site characteristics, $\varphi = f(x, a)$, such as segment length and traffic volume (Hauer, 2001; Miaou and Lord; 2003; Winkelmann, 2003; Mitra and Washington, 2006). In this study (for variance function), we found that AADT and density of horizontal curves were statistically significant in modeling (i) property damage only and (ii) total accidents. On the other hand, AADT and segment length were statistically significant in modeling injury-fatality accidents. Hence, Equations 3.2 and 3.3 were used to account for these effects.

$$ln(\varphi_i) = b_0 + b_1(ln(AADT)) + b_2(density\ of\ horizontal\ curves) \quad [3.2]$$

$$ln(\varphi_i) = b_0 + b_1(ln(AADT)) + b_2(length) \quad [3.3]$$

Alternatively, in addition to models developed based on previous approaches, we examined the case in which the inverse dispersion parameter varies randomly across road segments following a Gamma distribution; that is, $\varphi_i \sim gamma\ (0.01, 0.01)$. Thus, in this case, both r and φ vary across sites. This approach was expected to add more flexibility to Bayesian hierarchical models, improving the ability of these models to account for heterogeneity issue, and to provide better model-fitting compare with other structures as explained before. In this study, and under the Poisson-Gamma model, we tested the

above mentioned approaches to specify the inverse dispersion parameter on three different datasets.

3.2.2.3 Hierarchical Poisson-Lognormal model

The Poisson-Lognormal model is also a good alternative in road safety, and has been used and discussed by various researchers (Kim et al, 2002; El-Basyouny and Sayed, 2010). The assumption of this model is that accidents occur following a Poisson distribution with a mean - expected accident frequency - that is log-normally distributed; i.e., $\theta \sim lognormal\ (ln\ (\mu),\ \upsilon)$. In other words, similar to previous model, k, observed accident frequency, is expressed as $k \sim Poisson\ (\theta)$, and $\theta = \mu r$ with μ as a function of the contributing factors' vector x and vector of unknown parameters a; $\mu = f(x, a)$. Random effect r, in this case, is assumed to follow the Lognormal distribution; $r \sim lognormal\ (0,\ \upsilon)$ where in hierarchical Bayesian models υ^{-1}, the inverse of variance, is assumed to follow a Gamma distribution with parameters a and b. These parameters are assumed to be identical and can be equal to 0.01; therefore, υ^{-1} is defined as $\upsilon^{-1} \sim gamma\ (0.01,\ 0.01)$ with a *mean* = 1 and a large *variance* = 100 indicating a vague hyper-prior (Miaou et al., 2003; Lord and Miranda-Moreno, 2008; El-Basyouny and Sayed, 2010).

Under Poisson-Lognormal model, again, we adopted an approach similar to hierarchical Poisson-Gamma model in characterization of the variance function. As mentioned before, in Poisson-Lognormal model, random effect is log-normally distributed with *mean* = 0 and *variance* = υ. This variance, in this study, was assumed to be fixed or varying across sites. Moreover, it might be varying based on a relationship with some site characteristics or randomly varying. Equations 3.4 and 3.5 represented the relationship between the

inverse of variance and some site characteristics, in this study. Contributing factors presented in these equations were found to be statistically significant in the structure of the inverse of variance.

$$ln(\tau_i) = b_0 + b_1(ln(AADT)) + b_2(density\ of\ horizontal\ curves) \quad ; \tau_i = \upsilon_i^{-1} \quad [3.4]$$

$$ln(\tau_i) = b_0 + b_1(ln(AADT)) + b_2(length) \quad ; \tau_i = \upsilon_i^{-1} \quad [3.5]$$

In addition, the inverse of variance could be assumed to vary randomly across sites following a Gamma distribution; that is, $\upsilon_i^{-1} \sim gamma\ (0.01, 0.01)$. We applied the above mentioned approaches for modeling the variance function structure in Poisson-Lognormal models in order to verify the associated outcomes.

3.2.3 Modeling Accidents by Severity

Some researchers have worked on accident modeling by severity suggesting different approaches and methodologies (Saccomano et al, 1996; Ma and Kockelman, 2006; Milton et al, 2008). In this study, accidents were divided into three types; the first type represented the aggregation of all severities and the other two were based on diverse severities. Since the number of fatality accidents were extremely small, fatality and injury accidents were considered under the same severity. Therefore, three different types of accidents: (1) property damage only, (2) injury-fatality, and (3) total accidents were modeled independently, in a fully Bayesian framework. There are some critics to analyzing severity-frequency models separately (Milton et al, 2008); yet, we adopted this method since the main focus of this research was the examination of various model structures in terms of regression approaches and variance function specifications. Thus, by modeling each accident type separately, number of accidents increased by three folds,

and it was possible to test various model structures on three different datasets. Additionally, effects of various contributing factors on each accident type were investigated, individually. Indeed, contributing factors varied for three cases of accident types not only in the mean equation but also in the variance function equation.

3.2.4 Bayesian Estimation of the Model Parameters

Different methods are available to estimate regression model parameters such as maximum likelihood estimation (Bedford and Cooke, 2001) and Bayesian estimation (Gelman and Hill, 2007). The latter has been used in this study because of its interesting properties, substantial interpretive advantages (Mitra and Washington, 2006), and capacities to deal with uncertainty and randomness related to the contributing factors presented in each SPF. Moreover, Bayesian regression can combine expert criteria with local observations in order to calibrate models based on specified contributing factors for various engineering performance models (Amador and Mrawira, 2011). Bayesian estimation is structured based on prior, likelihood and posterior. The prior distribution, which represents the initial knowledge about a parameter, can be selected based on previous researches, literature, expert criteria, or experience. The likelihood function is represented by data containing local observations, and finally, the posterior distribution can be obtained by mixing these two; prior and likelihood. In particular, posterior distribution can be estimated applying Markov Chain Monte Carlo methods using Gibbs sampler that samples the space of the contributing factors and takes into account the randomness associated to these factors.

3.2.4.1 Goodness-of-fit

The deviance information criterion, DIC (Spiegelhalter et al, 2002), can be used as a goodness-of-fit measure for comparing models in Bayesian statistics. DIC is a generalization of the Akaike information criterion (ACI) based on the posterior distribution of the deviance statistics, and is defined as:

$$DIC = \overline{D} + p_D$$

Here, $\overline{D}$ is the posterior expectation of the deviance, and p_D is the effective number of parameters that captures the complexity of the model (Carlin and Louis, 2009). Models with smaller values of DIC indicate a better fit to the dataset. Differences of more than 10 might definitely rule out the model with the higher DIC, differences between 5 and 10 are substantial, but if the difference in DIC is, say, less than 5, and the models make very different inferences, then it could be misleading just to report the model with the lowest DIC (Spiegelhalter et al, 2002). In addition, DIC can only provide a measure of comparison between models, nested or not, that are applied to the same dataset (Spiegelhalter et al, 2002; Mitra and Washington, 2006).

3.3 CASE STUDY AND DATA DESCRIPTION

A case study of 958 kilometers of rural highway segments in New Brunswick, Canada was used in this study for the application of various models and associated comparisons. This case study consists of 720 observations including 1652 accidents during 3 years. Contributing factors used in this study – based on availability of the data – are reported in Table 3.I, and summary statistics of the dataset is shown in Table 3.II.

Table 3.I Contributing factors list

Contributing factors	*definition*
traffic flow (AADT)	Average annual daily traffic (vehicles per day)
length	Segment length (km)
district	District indicator (1 for districts 5 and 6, 0 otherwise).
undiv	Presence of median; un-divided road indicator (1 if un-divided, 0 otherwise).
precipitation	Average annual precipitation for 2004, 2005 and 2006, in millimeters.
dhc	Density of horizontal curves per km
indp	Precipitation-horizontal curve interaction indicator (1 if precipitation is greater than 1153.4 mm and density of horizontal curves is greater than 0.40 per kilometer, 0 otherwise).

Accident data were aggregated over a period of three years, 2004 to 2006. This aggregation can be justified since it helps to avoid the regression to the mean phenomenon and confounding effects associated to exceptional events observed in a particular year (Lord and Persaud, 2000; Cheng and Washington 2005; El-Basyouny and Sayed 2009). Three types of accidents including property damage only, injury-fatality, and total accidents were taken into account separately to follow objectives of the study. This leads to application and investigation of various models (in terms of regression approach and variance function structure) on three different datasets. Injury and fatality accidents were considered together since the number of fatal accidents was extremely small. In addition, observations of average annual precipitation for the study period, 2004 to 2006, from four weather stations (Canadian National Climate Data and Information Archive) located across highway segments were used in order to examine the effect of environmental exposures on the expected accident frequency. To do so, weather stations were designated to road segments based on their proximity and altitude (see Milton et al, 2008).

Table 3.II Summary statistics of observed data (case study)

Variables	*Mean*	*S.D.*	*Min.*	*Max.*
total Accidents (accident/3 years)	20.6500	13.9203	0	65
property damage only (accident/3 years)	14.4375	9.4719	0	46
injury and fatality (accident/3 years)	6.2125	4.8878	0	20
traffic flow (AADT)	7887.700	3337.916	4435	17550
length	11.9735	4.7270	3.170	19.800
district	0.7500	0.4357	0	1
undiv	0.0750	0.2650	0	1
precipitation	1159.5520	56.3548	1114.300	1256.200
indp	0.3375	0.4758	0	1
dhc	0.4069	0.1392	0.1649	0.7692

3.4 MODEL SPECIFICATION IN OPENBUGS

Statistical software OpenBUGS (for performing Bayesian inference Using Gibbs Sampling) was used for stochastic MCMC simulation in order to estimate the posterior distributions of the models' parameters. A normal distribution (*normal (0, 0.001)*) in OpenBUGS, with a mean value equal to zero and a large variance (non informative prior) was selected as the prior distribution for parameters associated to contributing factors in order to let the data dominate the derivation of the posteriors. Moreover, as stated before, we adopted vague priors for the inverse dispersion parameter and the inverse of variance in Poisson-Gamma and Poisson-Lognormal models, respectively.

Two different chains were considered with different initial values; so that, it was possible to verify the convergence of these chains after running thousands of iterations. An initial portion of the iterations was used to verify the convergence and then excluded from the estimation of the parameters (Burn-in iterations); the rest of the iterations were considered to derive the posterior distributions. In particular, we ran 20000 iterations

from which the first 5000 were discarded as burn-in and the remaining was used to estimate posteriors of parameters.

The convergence was checked using trace plots, iteration history plots, and Gellman Rubin diagrams (Brooks and Gelman, 1998). Furthermore, the stability of the posterior's mean values and the value of the Monte Carlo error that should be less than 5% of the related standard deviations indicated the dependability of the estimates.

3.5 RESULTS AND DISCUSSION

We applied three regression models (Poisson, Poisson-Gamma model and Poisson-Lognormal) to three datasets containing: property damage only, injury-fatality, and total accidents. In Poisson mixture cases, three possible specifications of the variance function were analyzed. The results of the analysis for these estimations are summarized in Tables 3.III to 3.VII. As reported in these tables, the main differences are in DIC values, credible intervals, and variances that indicate how each approach differs from the other in capturing variability, uncertainty, and in fitting data through goodness-of-fit measure. Moreover, for every model, all parameters were positive, except the constant term, which indicated that these contributing factors were positively correlated with accident frequencies.

3.5.1 Comparisons and Inferences Based on Different Regression Approaches

As expected, Poisson regression - because of its limitation to deal with over dispersion as described before - fell short on describing all three datasets used in this study. This can be verified by comparing DIC values for each of three datasets that indicated the greatest DIC value for Poisson model (Tables 3.III, 3.IV, and 3.V). Additionally, in this study,

Poisson-Gamma model fitted all three datasets better than Poisson-Lognormal. However, in some cases DIC values were very close; for instance, under injury-fatality accidents (where variance functions were only described by the mean) DIC differences were only 1.9 (Tables 3.IV and 3.V). Thus, in such cases, choice of either model could be justified.

Table 3.III Estimation results for Poisson models

Contributing factors	*Mean (S.D.)*	*95% C.I.*
Total accidents – dataset 1	*DIC = 648.194*	
constant	-7.7500 (0.790)	-9.3280, -6.2710
a_1: ln(AADT)	0.6091 (0.060)	0.5047, 0.7321
a_2: ln(length)	0.9414 (0.0717)	0.789, 1.0750
a_3: district	0.5103 (0.076)	0.3648, 0.6626
a_4: undiv	0.6018 (0.0752)	0.4527, 0.7474
a_5: precip	0.0216 (0.0047)	0.0128, 0.3051
Property damage only – dataset 2	*DIC = 561.774*	
constant	-6.9620 (0.655)	-8.2490, -5.5830
a_1: ln(AADT)	0.7652 (0.066)	0.6297, 0.8976
a_2: ln(length)	0.9583 (0.076)	0.8103, 1.1150
a_3: district	0.4479 (0.079)	0.2918, 0.6040
a_4: undiv	0.6665 (0.087)	0.4951, 0.8342
Injury & fatality accidents – dataset 3	*DIC = 399.248*	
constant	-4.9590 (1.111)	-6.9860, -2.5330
a_1: ln(AADT)	0.3983 (0.123)	0.1207, 0.6175
a_2: ln(length)	1.204 (0.144)	0.9436, 1.5040
a_3: undiv	0.8566 (0.124)	0.6063, 1.0930
a_4: indp	0.2631 (0.092)	0.0842, 0.4443

3.5.2 Comparisons and Inferences Based on Different Variance Function Characterizations

One should take into account that the effect of the variance function specification on goodness-of-fit is more significant when contributing factors in the mean function are not sufficient or are not able to describe the data appropriately. When comparing models structured with different variance functions, variations in the 95% credible interval band and DIC values were more significant compare with variations in the mean values of parameters (Tables 3.IV and 3.V). Considering both Poisson mixture models, 95% credible interval band was the narrowest in models with φ, or υ varying as a function of site characteristics (φ, or $\upsilon = f(x, a)$), followed by fixed (φ, or υ are fixed) and then randomly varying (φ, or υ vary randomly) structures. Even though the range of the credible interval affects the precision in estimation of model parameters, this does not necessarily imply a better fit. For example, in this study, Poisson model provided the smallest credible interval band with respect to other models; however, it did not produce the best fit (Tables 3.III, 3.IV, and 3.V). For Poisson-Gamma model, Geedipally and Lord (2008) investigated the effect of the varying dispersion parameter as a function of site characteristics on the confidence intervals of estimations and found that models with fixed dispersion parameter produced bigger confidence intervals. The same behavior was observed in this study.

Furthermore, under both Poisson mixture models, when modeling dataset 1 (total accidents), DIC variations were more than 10, for three variance function structures, which indicated noteworthy alterations in goodness-of-fit. Similarly, these variations were greater than 10, in all cases, when comparing fixed structures (φ, or υ are fixed)

with randomly varying structures (φ, or υ vary randomly). Moreover, considering Poisson-Gamma regression, DIC values varied more notably between fixed and varying φ cases. Instead, these differences were smaller, still greater than 5, between two varying φ approaches in which, firstly, the inverse dispersion parameter varied as a function of site characteristics, and secondly, it varied randomly. Again, as explained before, Spiegelhalter et al (2002) state that a DIC difference greater than 5 is substantial. Likewise, under Poisson-Lognormal regression, DIC differences were always greater than 5. Besides, these differences for Poisson-Lognormal models were greater than 10 taking into account, first, dataset 2, property damage only models for the two cases of varying variances (Table 3.V). And second, dataset 3, injury-fatality accident models for the fixed case and varying as a function of site characteristics (Table 3.V).

Finally, this study showed that contributing factors presented in mean functions might vary from those presented in variance functions. Under both Poisson mixture models, for total accidents and property damage only, density of horizontal curves wasn't statistically significant in mean functions, while it was found to be significant in variance functions (Tables 3.IV and 3.VI). Additionally, not necessarily, all statistically significant contributing factors that represented the mean were significant in the variance function. For example, precipitation for total accidents was bounded away from zero in the mean function but it was not significant in the variance function. Similar cases were observed for injury-fatality accidents. Furthermore, the most influential factor in all variance functions was the traffic volume (AADT). Therefore, one should take into account that there is no single functional form or parameterization that is suitable for all datasets

(Geedipally et al, 2009). For more clarifications see Tables 3.VI and 3.VII where additional models are reported.

3.5.3 Comparisons and Inferences Based on Different Types of Accidents

In this study, modeling accidents by type and severity showed similar behaviors in all regression approaches. As summarized in Tables 3.III, 3.IV, and 3.V, the vector of contributing factors in the mean and the variance functions differs for almost all type of accidents. For instance, precipitation that was statistically significant for total accidents was not found to be significant for property damage only and injury-fatality accidents. An interesting finding in this case study was the fact that density of horizontal curves and precipitation was not individually significant in the mean function for injury-fatality accidents. However, their interaction (see Table 3.I) was found to be bounded away from zero for this severity (Table 3.IV). In addition, contributing factors AADT, Length, and the presence of median (Table 3.I) were significant for all three types of accidents.

Table 3.IV Estimation results for hierarchical Poisson-Gamma models

Contributing factors	*φ fixed*		*φ_i varying*			
			φ_i = f (site characteristics)		*φ_i randomly varying*	
	Mean (S.D.)	*95% C.I.*	*Mean (S.D.)*	*95% C.I.*	*Mean (S.D.)*	*95% C.I.*
Total Accidents (including property damage only, injury-fatality accidents) – dataset 1						
	DIC = 514.3		***DIC = 501.4*** [a]		*DIC = 490.2*	
constant	-7.972 (1.739)	-11.43, -0.561	-7.369 (1.495)	-10.28, -4.407	-7.761 (1.796)	-11.29, -0.255
a_1: ln(AADT)	0.5998 (0.194)	0.218, 0.982	0.5804 (0.133)	0.317, 0.839	0.6020 (0.193)	0.222, 0.978
a_2: ln(length)	0.9175 (0.137)	0.651, 1.187	0.9464 (0.116)	0.717, 1.174	1.0260 (0.137)	0.758, 1.294
a_3: district	0.4953 (0.150)	0.197, 0.791	0.5369 (0.122)	0.298, 0.777	0.5055 (0.158)	0.194, 0.819
a_4: undiv	0.6522 (0.198)	0.271, 1.056	0.5352 (0.149)	0.238, 0.828	0.5683 (0.196)	0.188, 0.958
a_5: precip	0.0249 (0.013)	0.001, 0.050	0.0207 (0.010)	0.001, 0.041	0.0207 (0.012)	***-0.002, 0.044*** [b]
			$ln(\varphi_i) = b_0+b_1(ln(AADT))+b_2(dhc)$			
constant			-31.73 (10.490)	-54.70, -13.31		
b_1: ln(AADT)			3.5360 (1.163)	1.4950, 6.123		
b_2: dhc			6.6700 (2.456)	2.0670, 11.780		
Property damage only – dataset 2						
	DIC = 481.0		*DIC = 469.6*		*DIC = 462.8*	
constant	-6.848 (1.669)	-10.17, -0.568	-6.345 (1.334)	-8.920, -3.682	-6.755 (1.844)	-10.40, -3.127
a_1: ln(AADT)	0.7599 (0.177)	0.411, 1.115	0.6977 (0.134)	0.429, 0.952	0.7272 (0.195)	0.345, 1.109
a_2: ln(length)	0.9423 (0.130)	0.685, 1.199	0.9653 (0.111)	0.746, 1.182	1.0220 (0.145)	0.737, 1.307
a_3: district	0.4158 (0.148)	0.118, 0.703	0.4593 (0.116)	0.228, 0.687	0.4651 (0.165)	0.142, 0.791
a_4: undiv	0.6818 (0.193)	0.314, 1.072	0.6270 (0.135)	0.362, 0.897	0.6036 (0.186)	0.236, 0.971
			$ln(\varphi_i) = b_0+b_1(ln(AADT))+b_2(dhc)$			
constant			-27.14 (11.020)	-50.74, -7.337		
b_1: ln(AADT)			2.9570 (1.229)	0.711, 5.556		
b_2: dhc			8.4880 (3.132)	2.864, 15.19		
Injury and fatality accidents – dataset 3						
	DIC = 385.9		*DIC = 376.1*		*DIC = 368.8*	
constant	-5.443 (1.553)	-8.544, -2.420	-5.094 (1.304)	-7.627, -2.514	-4.854 (1.950)	-8.600, -0.964
a_1: ln(AADT)	0.4541 (0.169)	0.121, 0.787	0.4226 (0.148)	0.130, 0.711	0.4008 (0.216)	***-0.027, 0.814*** [c]
a_2: ln(length)	1.1960 (0.158)	0.888, 1.513	1.1710 (0.152)	0.878, 1.474	1.180 (0.212)	0.772, 1.604
a_3: undiv	0.8801 (0.189)	0.507, 1.260	0.8453 (0.167)	0.514, 1.173	0.8200 (0.225)	0.378, 1.267
a_4: indp	0.2786 (0.119)	0.044, 0.518	0.2498 (0.108)	0.036, 0.463	0.2074 (0.153)	***-0.091, 0.510*** [d]
			$ln(\varphi_i) = b_0+b_1(ln(AADT))+b_2(length)$			
constant			-38.39 (17.150)	-72.51, -6.269		
b_1: ln(AADT)			5.6050 (2.039)	1.826, 9.664		
b_2: length			-0.5005 (0.315)	-1.235, -0.006		

[a] See Table 3.VI for two alternative models.
[b] This parameter is statistically significant at 90% C.I. (0.00135, 0.0399); see Table 3.VII for an alternative model.
[c] This parameter is statistically significant at 90% C.I. (0.04209, 0.7499).
[d] This parameter is statistically significant at 80% C.I. (0.01203, 0.4046).

Table 3.V Estimation results for hierarchical Poisson-Lognormal models

Contributing factors	*τ fixed*		*τ_i varying*			
			τ_i = f (site characteristics)		*τ_i randomly varying*	
	Mean (S.D.)	*95% C.I.*	*Mean (S.D.)*	*95% C.I.*	*Mean (S.D.)*	*95% C.I.*
Total Accidents (including property damage only, injury-fatality accidents) – dataset 1						
	DIC = 524.1		*DIC = 512.4*		*DIC = 492.4*	
constant	-8.6250(1.675)	-11.91, -5.304	-8.4290 (1.515)	-11.46, -5.536	-8.230 (1.922)	-12.14, -4.580
a_1: ln(AADT)	0.6289 (0.185)	0.261, 0.991	0.6510 (0.132)	0.393, 0.904	0.6310 (0.210)	0.212, 1.039
a_2: ln(length)	0.9019 (0.132)	0.641, 1.158	0.9349 (0.113)	0.715, 1.155	1.0520 (0.146)	0.768, 1.342
a_3: district	0.4894 (0.148)	0.198, 0.777	0.5567 (0.124)	0.309, 0.800	0.5297 (0.172)	0.198, 0.877
a_4: undiv	0.6652 (0.198)	0.287, 1.070	0.5493 (0.155)	0.240, 0.854	0.5627 (0.217)	0.161, 0.995
a_5: precip	0.0280 (0.012)	0.004, 0.053	0.0239 (0.009)	0.005, 0.044	0.0215 (0.012)	***-0.004,0.046*** [a]
			$ln(\tau_i) = b_0 + b_1(ln(AADT)) + b_2(dhc)$			
constant			-32.360 (11.340)	-56.48, -12.060		
b_1: ln(AADT)			3.6160 (1.241)	1.400, 6.237		
b_2: dhc			6.4780 (3.027)	1.098, 13.080		
Property damage only – dataset 2						
	DIC = 487.6		*DIC = 480.0*		*DIC = 464.9*	
constant	-7.241 (1.641)	-10.53, -4.031	-7.0100 (1.284)	-9.531, -4.490	-7.082 (2.044)	-11.17, -3.174
a_1: ln(AADT)	0.7971 (0.173)	0.458, 1.145	0.7649 (0.128)	0.512, 1.015	0.7501 (0.213)	0.341, 1.175
a_2: ln(length)	0.9447 (0.129)	0.691, 1.197	0.9720 (0.108)	0.756, 1.185	1.0530 (0.156)	0.748, 1.363
a_3: district	0.3973 (0.147)	0.106, 0.688	0.4671 (0.114)	0.241, 0.686	0.4729 (0.184)	0.125, 0.842
a_4: undiv	0.7255 (0.187)	0.358, 1.101	0.6570 (0.130)	0.401, 0.915	0.6156 (0.195)	0.212, 0.997
			$ln(\tau_i) = b_0 + b_1(ln(AADT)) + b_2(dhc)$			
constant			-27.730 (12.660)	-53.86, -4.142		
b_1: ln(AADT)			2.9540 (1.453)	0.138, 5.875		
b_2: dhc			10.7400 (5.741)	2.966, 28.530		
Injury and fatality accidents – dataset 3						
	DIC = 387.8		*DIC = 377.3*		*DIC = 371.6*	
constant	-5.675 (1.542)	-8.799, -2.743	-5.4040 (1.261)	-7.866, -2.934	-5.2880(2.080)	-9.337, -1.193
a_1: ln(AADT)	0.4759 (0.168)	0.155, 0.818	0.4643 (0.143)	0.184, 0.742	0.4279 (0.230)	***-0.029, 0.872*** [b]
a_2: ln(length)	1.1950 (0.157)	0.889, 1.508	1.1370 (0.149)	0.845, 1.430	1.2330 (0.234)	0.781, 1.702
a_3: undiv	0.9002 (0.187)	0.541, 1.279	0.8818 (0.152)	0.579, 1.177	0.8382 (0.248)	0.345, 1.319
a_4: indp	0.2702 (0.119)	0.035, 0.507	0.2538 (0.106)	0.046, 0.463	0.2073 (0.165)	***-0.122, 0.528*** [c]
			$ln(\tau_i) = b_0 + b_1(ln(AADT)) + b_2(length)$			
constant			-37.070 (21.010)	***-80.62, 1.423*** [d]		
b_1: ln(AADT)			6.5550 (2.569)	1.8720, 11.930		
b_2: length			-1.0340 (0.544)	-2.323, -0.1479		

[a] This parameter is statistically significant at 90% C.I. (0.0007, 0.0418).
[b] This parameter is statistically significant at 90% C.I. (0.0469, 0.8034).
[c] This parameter is statistically significant at 50% C.I. (0.0985, 0.3181).
[d] This parameter is statistically significant at 90% C.I. (-72.110, -4.155).

Table 3.VI Estimation results for alternative Poisson-Gamma models (Total Accidents); density of horizontal curves in the mean function

Contributing factors	*φ_i fixed*		*φ_i = f (site characteristics)*	
	Mean (S.D.)	*95% C.I.*	*Mean (S.D.)*	*95% C.I.*
	DIC = 514.6		*DIC = 501.8*	
constant	-7.7780 (1.741)	-11.230, -4.3640	-7.4990 (1.490)	-10.4300, -4.5770
a_1: ln(AADT)	0.5701 (0.195)	0.1886, 0.9540	0.5622 (0.134)	0.2930, 0.8230
a_2: ln(length)	0.9214 (0.135)	0.6588, 1.1890	0.9602 (0.115)	0.7340, 1.1870
a_3: district	0.4637 (0.154)	0.1563, 0.7610	0.5046 (0.128)	0.2570, 0.7610
a_4: undiv	0.5935 (0.206)	0.1979, 1.0130	0.4682 (0.163)	0.1410, 0.7880
a_5: precip	0.02404 (0.012)	-0.0003, 0.0491	0.0216 (0.010)	0.0020, 0.0420
a_6: dhc	0.4553 (0.405)	***-0.3288, 1.249*** [a]	0.3826 (0.345)	***-0.2940, 1.0620*** [b]
			$ln(\varphi_i) = b_0 + b_1(ln(AADT)) + b_2(dhc)$ [c]	
constant			-32.810 (10.41)	-56.090,-14.3100
b_1: ln(AADT)			3.6800 (1.154)	1.6300,6.2450
b_2: dhc			6.1980 (2.679)	1.4470,11.8100

[a] density of horizontal curves is not significant in mean function. This parameter is significant at 50% C.I. (0.1802, 0.7624).
[b] density of horizontal curves is not significant in mean function. This parameter is significant at 50% C.I. (0.1529, 0.6118).
[c] density of horizontal curves is significant in the variance function.

Table 3.VII Estimation results for alternative Poisson-Gamma models (Total Accidents)

Contributing factors	*φ_i randomly varying*		*φ_i = f (site characteristics)*			
	Mean (S.D.)	*95% C.I.*	*Mean (S.D.)*	*95% C.I.*	*Mean (S.D.)*	*95% C.I.*
	DIC = 490.9 [a]		*DIC = 509.2* [b]		*DIC = 508.3* [b]	
constant	-6.723 (1.661)	-10.00, -3.50	-7.616 (1.556)	-10.650, -4.515	-7.589 (1.583)	-10.7, -4.40
a_1: ln(AADT)	0.739 (0.175)	0.400,1.090	0.568 (0.142)	0.291, 0.845	0.564 (0.143)	0.280, 0.840
a_2: ln(length)	1.109 (0.129)	0.860,1.360	0.920 (0.125)	0.680, 1.169	0.915 (0.125)	0.670, 1.160
a_3: district	0.424 (0.146)	0.140,0.720	0.511 (0.132)	0.255, 0.769	0.514 (0.132)	0.250, 0.770
a_4: undiv	0.680 (0.191)	0.330,1.080	0.584 (0.219)	0.167, 1.032	0.569 (0.210)	0.170, 0.990
a_5: precip			0.024 (0.011)	0.003, 0.046	0.024 (0.011)	0.003,0.050
			$ln(\varphi_i)$ $=b_0+b_1(ln(AADT))+b_2(length)$		*$ln(\varphi_i) = b_0+b_1 ln(AADT)$*	
constant			-22.28 (8.639)	-40.500, -6.341	-23.29 (8.262)	-40.300, -8.100
b_1: ln(AADT)			2.816 (0.979)	1.032, 4.887	2.855 (0.941)	1.120, 4.780
b_2: length			-0.049 (0.061)	***-0.17, 0.065*** [c]		

[a] Similar DIC respect to model presented in Table 3.IV where precipitation was among contributing factors (not being statistically significant under φ_i randomly varying). And smaller DIC respect to model presented in Table 3.IV under φ fixed specification.
[b] Higher DIC respect to model presented in Table IV where $ln(\varphi_i) = b_0 + b_1(ln(AADT)) + b_2(dhc)$.
[c] This parameter is not statistically significant at *95%* credible Interval.

3.6 SUMMARY AND CONCLUSIONS

This paper had five main objectives: (1) application of Poisson and hierarchical Poisson mixture models, in a Bayesian framework, to examine probable differences in parameter estimation and goodness-of-fit, (2) different characterization of variance functions under hierarchical Poisson mixture models to investigate on associated outcomes and to compare results related to parameter estimation and goodness-of-fit, (3) to investigate a randomly varying approach in which the inverse dispersion parameter and the inverse of variance (in hierarchical Poisson-Gamma and hierarchical Poisson-Lognormal models, respectively) were allowed to vary randomly across sites, (4) identification of statistical significance of various site characteristics in the variance function specifications for hierarchical Poisson mixture models, and (5) specification of models by severity, and accident type, to examine the statistical significance of various contributing factors in each case. To do so, for Poisson-Gamma models, the inverse dispersion parameter, φ, was incorporated in the models as (a) fixed, (b) varying as a function of site characteristics, and (c) randomly varying. Likewise, for Poisson-Lognormal approach the same procedure was adopted. However, in the latter, instead of the inverse dispersion parameter, the inverse of variance, υ^{-1}, was characterized as fixed, varying as a function of site characteristics, and randomly varying.

To address such objectives we used three datasets. Over 35 models were developed from which 26 are presented in Tables 3.III to 3.VII. In this study, hierarchical Poisson-Gamma models provided the best fit, having the smallest DIC values, for all cases, followed by hierarchical Poisson-Lognormal and then Poisson models. Considering variance function characterization, in terms of goodness-of-fit, randomly varying

structure under both Poisson mixture models provided the smallest DIC values; hence, the best fit. DIC differences were greater than 5 in all cases and greater than 10 in some others (Tables 3.IV and 3.V). Nevertheless, some contributing factors which were statistically significant under the fixed dispersion parameter models were not found to be significant under randomly varying models; for instance, precipitation in modeling total accidents (Table 3.IV). Furthermore, we observed that goodness-of-fit increased significantly when the inverse dispersion parameter in Poisson-Gamma and the inverse of variance in Poisson-Lognormal were defined as a function of site characteristics as compare to those models in which these parameters were fixed. A similar situation was still more obvious when comparing fixed with randomly varying specifications. In fact, in this study for some cases, we observed that by specifying a randomly varying dispersion parameter still the best fit was obtained even if fewer contributing factors were presented in the SPF. For instance, when modeling total accidents under Poisson-Gamma models (see Tables 3.IV and 3.VII). Therefore, when facing lack of data regarding some contributing factors, it may be still possible to have an adequate model-fitting by adopting this structure that can account for heterogeneity across sites. Consequently, by applying a randomly varying structure practitioners would still be able to obtain an SPF that reflects data accurately even though a few contributing factors are available. Obviously, goodness-of-fit and level of significance of contributing factors in an SPF should be taken into consideration together to choose a model over others. Another justification in using randomly varying framework may be the fact that identifying contributing factors in an SPF or variance function and proving their true presence in the model is somehow time consuming, and in some situations not realistic because of the

discrepancies in datasets. So, by adopting this approach a simpler mean function could be developed without penalizing goodness-of-fit of the model.

An additional aspect of this research was to identify contributing factors that might affect the variance function. In fact, in this study, traffic volume, density of horizontal curves, and segment length were presented in variance functions, being bounded away from zero with AADT as the most important factor. Moreover, the density of horizontal curves that was not statistically significant in mean functions was significant in variance functions when modeling property damage only and total accidents. Additionally, we observed that each type of accidents was described by a different vector of contributing factors. For instance, interaction between precipitation and density of horizontal curves was bounded away from zero for injury-fatality accidents while it was not statistically significant for other accident types. Finally, adopting other case studies - road segments and intersections - is recommended in order to verify methodologies discussed in this research study.

CHAPTER 4

Sensitivity of Safety Performance Functions to Different Prior Specifications in Poisson-Gamma Models Applying Bayesian Hierarchical Methods

This chapter aims to explore Bayesian accident data analysis from a practical perspective; that is, the choice of priors for model parameters.

Abstract

This paper aims to explore Bayesian accident data analysis from a practical perspective; that is, the choice of priors for model parameters. The use of Bayesian statistics in road safety has recently become popular by researchers and practitioners who mainly apply Poisson-Gamma models using non-informative priors to calibrate safety performance functions. Bayesian modeling requires the specification of priors for model parameters. In this paper, we, firstly, determined a series of informative, semi-informative, and non-informative priors for model parameters. Then, we examined the effect of prior choices on the accuracy of outcomes in terms of parameter estimates, hotspot identification, and goodness-of-fit. A case study consisting of 958 km of rural highway segments in New Brunswick has been used to obtain the true estimates for model parameters. From this case study, three different sample sizes having two different mean values, high and relatively low mean, have been examined in the analyses. For each case, in a simulation

framework, 100 datasets have been replicated, and consequently, calibrated under three different prior specifications.

We observed that introducing an informative prior for the inverse dispersion parameter dramatically improved estimates, especially, when modeling datasets characterized by low sample mean and small sample size. We also observed that regression parameters were less sensitive to prior choice compare with the inverse dispersion parameter. However, as the sample size or the sample mean decreases, an informative prior specification provides more precise estimates also for regression parameters. Finally, prior specification didn't have any significant impact on hotspot identification and goodness-of-fit.

4.1 INTRODUCTION

Researchers in transportation engineering have recently employed Bayesian statistics, particularly hierarchical models, in road safety to develop safety performance functions (SPF) and to identify hotspots (hazardous sites) that, eventually, would be subject to safety treatments (Miaou and Song, 2005; Miranda-Moreno at al., 2007; Mitra and Washington, 2006; El-Basyouny and Sayed, 2009). Bayesian inference (Gelman et al, 1995) consists of three main elements (i) prior, (ii) likelihood, and (iii) posterior, in which the posterior distribution is drawn from the prior and the likelihood. The prior distribution may provide some sort of information about an unknown parameter based on previous studies, expert criteria or experiences, and the likelihood is represented by the data itself. Bayesian inference requires significant amount of computation that is not, any more, of a major concern since the computational capacities of personal computers have

increased dramatically in recent years. This computation, which cannot be done analytically, takes advantage of Markov Chain Monte Carlo (MCMC) methods (Gelman et al, 1995; Gamerman and Lopes, 2006; Carlin and Louis, 2009) to obtain the posterior distribution from the prior and likelihood distributions. The use of MCMC methods for estimating hierarchical models, often involve complex data structures, sometimes described as revolutionary development, and has arguably facilitated the fitting of such models (Congdon, 2010).

Bayesian estimation has some advantages over traditional methods (e.g., maximum likelihood estimation), such as interesting interpretive capacities in providing the probability of the null hypothesis being true using the credible interval concept (Mitra and Washington, 2006; Carlin and Louis, 2009). And when the sample size is relatively small, Bayesian inference is still able to provide reliable estimates for the model parameters (Mitra and Washington, 2006; Gelman and Hill, 2007; Amador and Mrawira, 2011). In addition, as explained by Congdon (2010), using the Bayesian estimation is relevant when facing complex data, involving hierarchical nesting of subjects, spatially configured data, and repeated measures on subjects. Finally, the Bayesian approach can easily accommodate hierarchical models.

The Bayesian estimation requires specification of priors in order to be able to approximate posteriors of the model parameters. So that where this prior knowledge exists, in e.g. based on previous studies or expert opinion, Bayesian inference of the posterior distributions can take advantage of this known knowledge to estimate unknown parameters. Adopting this methodology, usually, leads to more reliable inferences on posteriors especially in limited datasets. Incorporating the prior knowledge in the model

varies in terms of the precision and the level of knowledge available about that prior. In general, priors that introduce very small amounts of information about a parameter are known as non-informative (vague) priors, and those introducing considerable amount of information are known as informative priors. Several studies in different fields like Reliability Engineering and Epidemiology have been conducted in order to verify the effect of informative priors on the Bayesian analysis outcomes (Lambert et al. 2005; Van Dongen, 2006). However, in the road safety community this type of research has been rare, and researchers have mainly focused on the development of statistical models, and identification of hotspots and contributing factors that affect accident frequencies.

The most common regression approach used in road safety is the Poisson-Gamma (Negative Binomial) model (Miranda-Moreno et al., 2005) that can be defined in a hierarchical fashion under the Bayesian context. The structure of the mean in Poisson-Gamma models contains a multiplicative random effect that follows a Gamma distribution by identical shape and scale parameters, called the inverse dispersion parameter. Since such a parameter may have a great impact on the model estimates, its characterization has attracted the attention of some researchers (Hauer, 2001; Miaou and Lord, 2003; Geedipally et al, 2009). Adopting the MLE approach, Lord (2006) has investigated the effect of low sample mean and small sample size on the estimation of the fixed dispersion parameter in Poisson-Gamma models. The author concluded that for the dispersion parameter, the probability of an unreliable estimate increases significantly as sample mean and sample size decrease. Considering the Bayesian approach and relating to the prior choice, Lord and Miranda-Moreno (2008) stated that a dataset characterized by a low sample mean combined with a small sample size can seriously affect the

estimation of the posterior mean of the inverse dispersion parameter when a non-informative prior specification is used for the gamma hyper-parameter. And that by choosing an appropriate prior for the inverse dispersion parameter the accuracy of estimates will increase significantly. Additionally, Miranda-Moreno et al. (2008) have examined the incorporation of an informative prior for the inverse dispersion parameter, in the analysis, considering different sample sizes (and years of data) and found that this type of priors provided more reliable estimates for the posterior mean of the inverse dispersion parameter.

The objective of this paper is to study the influence of prior specifications, in hierarchical Poisson-Gamma models, on: (1) the estimation of the model parameters (the inverse dispersion parameter and regression parameters), (2) credible intervals of estimates, (3) hotspot identification, and (4) goodness-of-fit. Basically, a series of informative and semi-informative priors have been determined from previous studies, respectively, for the inverse dispersion parameter and regression parameters. Consequently, results have been compared in terms of parameter estimates, hotspot identification, and goodness-of-fit. For parameter estimation comparisons, the mean value and 95% credible intervals have been calculated. In addition, Spearman's correlation coefficient has been used to compare ranking of sites for hotspot identification. And finally, DIC values were computed as a Bayesian measure of goodness-of-fit for model-fitting comparisons.

For these objectives, a simulation framework has been employed to replicate 100 datasets for various samples produced based on a case study which consists of accident data for 958 km of rural highway segments in New Brunswick, Canada – from 2004 to 2006. These replicated data were then used to investigate on the outcomes of various models

under three prior specification structures taking into account different sample sizes and sample means. All estimations were obtained by applying the Bayesian approach using MCMC methods. Finally, statistical software (i) Stata, (ii) OpenBUGS, (iii) R2OpenBUGS, and (iv) R were used for the computational purposes in this study.

4.2 METHODOLOGY

4.2.1 Hierarchical Poisson-Gamma Model

Poisson-Gamma models are very popular in road safety. This is because of their interesting property in dealing with heterogeneity across sites that make them adequate for accident data. The Poisson-Gamma model is mathematically described as follows:

$Y_i|\theta_i \sim Poisson\ (\theta_i),$

$\theta_i = \mu_i r_i,$

$\mu_i = f(x_i, a),$

$r_i \sim Gamma(\varphi,\varphi).$

Where, θ_i is the expected accident frequency on site *i*. μ_i is function of the contributing factors' vector x and the vector of unknown parameters *a* for site *i*; in other words, μ_i is the mean obtained from the SPF. r_i is a multiplicative random effect that is usually assumed to be gamma distributed with a mean of 1 and a variance of $1/\varphi$; where φ is the inverse dispersion parameter (Anastasopoulos and Mannering, 2008). When adopting a hierarchical model the inverse dispersion parameter, φ, is in turn assumed to follow a hyper-prior, usually, the Gamma distribution. That is $\varphi \sim gamma\ (a, b)$ where *a* and *b* are shape and scale parameters, respectively (Lord and Miranda-Moreno, 2008).

Instead of simple Poisson regression, Poisson-Gamma models can overcome the over dispersion issue, which is common in many accident datasets. In fact, by the presence of the multiplicative random effect in the mean structure, Poisson-Gamma models can account for heterogeneity - unobserved factors that may vary across sites (Anastasopoulos and Mannering, 2008). Thus, the expected accident frequency is then described by the SPF and a multiplicative random effect *r*.

4.2.2 Simulation Framework (Replication of Datasets)

Based on the case study, which basically represents the entire population of observations, two different types of accidents, Injury-fatality and total accidents, with different mean values have been used in the analysis. So that it was possible to verify differences between datasets characterized with a high and a relatively low sample means. For each of the high and the relatively low mean data, three different sample sizes consisting of 20, 50, and 80 observations were replicated 100 times under the Poisson-Gamma structure. Additionally, three different approaches in the specification of the prior distribution for model parameters have been examined. These specifications were (1) non-informative priors for all parameters, (2) an informative prior only for the inverse dispersion parameter, and (3) an informative prior for the inverse dispersion parameter and semi-informative priors for regression parameters. Consequently, for each replicated dataset all model parameters (regression parameters and the inverse dispersion parameter) have been estimated, in a fully Bayesian framework, under the hierarchical Poisson-Gamma approach considering different prior specifications. The results have then been compared with the true estimates; the parameter estimation results obtained from the analysis of the case study. In fact, the case study has been calibrated using the MLE to obtain the true

(real) estimates using the standard Negative Binomial (Poisson-Gamma) model. Here, the aim was to determine, for each case of the sample size and the sample mean, the prior specification approach that provides better estimates. And basically, a better estimate is the one that is closer to the true estimates. In addition to monitoring parameter estimation accuracies, we obtained 95% credible intervals for parameter estimates so that it was possible to compare credible intervals for three prior specification approaches.

Furthermore, models have been compared in terms of hotspot identification. For this comparison purpose a ranking criterion based on the Posterior distribution of θ, or expected accident frequency, was adopted to rank sites (Rao, 2003). To do so, first, we obtained true ranks for each replicated data after generation of that data. Second, for each replicated data, true ranks have been compared - using the Spearman's correlation coefficient (Ruppert, 2010) - with those ranks obtained from Bayesian inference considering various prior specifications. Finally, goodness-of-fit has been the third comparison measure in this study to investigate the impact of the prior specification on the associated outcomes. To summarize, the following steps have been followed in this study:

1. Estimation of parameters for the case study, using the MLE, to obtain the true estimates;
2. Application of the true estimates in the SPF to obtain μ_i;
3. Generation of the multiplicative random effect r_i based on the true estimates of φ, obtained from step 1, using the Gamma distribution; $r_i \sim Gamma(\varphi,\varphi)$;
4. Calculation of θ_i^{true} (the true expected accident frequency) based on μ_i and r_i obtained from previous steps; that is $\theta_i^{true} = \mu_i r_i$;

5. Preparation of the true rank of sites based on θ_i^{true} calculated in the previous step;
6. Generation of synthetic accidents based on the mean of real observed accidents for each case of high and relatively low sample means; i.e., $Y_i|\theta_i \sim Poisson\ (\theta_i)$;
7. Replication of datasets using accidents from the previous step together with site attributes (characteristics) obtained from the case study;
8. Application of Bayesian inference considering three prior specifications to obtain posteriors of parameters (regression parameters, φ, and θ_i) and goodness-of-fit;
9. Evaluation of the outcomes of the previous step with respect to the true estimates and the true rank of sites, and comparison of credible intervals and the model-fitting.

4.2.3 Safety Performance Function (SPF)

An SPF is a mathematical equation that represents the relationship between accident frequencies and a series of site characteristics such as traffic flow, segment length, and environmental exposure. The SPF adopted in this study is presented in the Equation 4.1. This SPF is a basic function used in the literature to model the safety of roadway segments based on the traffic flow (AADT) and the segment length (Highway Safety Manual, 2010).

$$\mu_i = a_0 L_i (AADT_i)^{a_1} \qquad [4.1]$$

where,

μ_i = expected accident frequency for road segment i;

a_0 = constant;

a_1 = parameter associated to traffic flow;

L_i = segment length (km) for road segment i;

$AADT_i$ = annual average daily traffic (vehicles per day) for road segment i.

4.2.4 Prior Specifications

Three prior specifications used in this study are described as follows.

a) Non-informative priors: The most used approach in the road safety literature when applying Bayesian statistics specifies priors as non-informative. This way, the importance of the likelihood becomes more significant. In other words, the data itself will lead to the parameter estimations and the contribution of the prior distribution is then minimized. In the road safety literature, for regression parameters, such as those representing the constant and the traffic flow, a normal distribution with a mean of 0 and a large variance of 1000, $a \sim normal\ (0, 1000)$, has been commonly used (Mitra and Washington, 2006). Moreover, to specify the prior for the inverse dispersion parameter in the hierarchical Poisson-Gamma model a Gamma distribution with the shape and the scale parameters identical and equal to 0.001, $\varphi \sim Gamma(0.001, 0.001)$, can be assumed (El-Basyouny and Sayed, 2009). Therefore, this specified prior for φ has a mean equal to 1 and a large variance equal to 1000. In fact, in the Gamma distribution the mean and the variance are calculated as:

$\varphi \sim Gamma(a, b)$ where a is the shape parameter and b is the scale parameter.

$Mean\ (\varphi) = a/b$

$variance\ (\varphi) = a/b^2$

b) Informative prior only for the inverse dispersion parameter: As explained above, a non-informative prior is usually used for the inverse dispersion parameter; however,

adopting an informative prior is also possible. In this study, we determined such an informative prior based on previous studies. To do so, we explored the road safety literature to obtain some values that have been reported for the inverse dispersion parameter, for roadway segments, by various researches. These values are tabulated in Table 4.I.

Table 4.I Reported values for the inverse dispersion parameter (φ)

	Inverse dispersion parameter (φ)	
Previous studies	*Total accidents*	*Injury-Fatality accidents*
Persaud et al, 2004	2.703, 1.695	3.030
Caliendo et al, 2007	4.227, 3.623	6.339, 2.625
Lord et al, 2008	2.638	3.244

From these studies the inverse dispersion parameter has a mean and a variance equal to 2.905 and 0.863, respectively, for total accidents. And it has a mean and variance equal to 3.8095 and 2.9096, respectively, for injury-fatality accidents. From these values the scale and the shape parameters for the Gamma distribution have been calculated; thus, informative priors for φ adopted in this study were:

Total accidents: $\varphi \sim Gamma\ (9.7787,\ 3.3660)$

Injury-fatality accidents: $\varphi \sim Gamma\ (4.9877,\ 1.3092)$

c) Informative prior for the inverse dispersion parameter & semi-informative priors for regression parameters: In this case, an informative prior for the inverse dispersion parameter has been defined as indicated in the previous approach. Additionally, a

methodology similar to that adopted for the inverse dispersion parameter was used to characterize priors for regression parameters. Based on previous studies parameter estimates for the constant and the traffic flow are reported in Table 4.II. For total accidents, the mean and the variance for the constant term are -6.186 and 11.264, respectively. Moreover, the mean and the variance for traffic flow are 0.788 and 0.098, respectively. For injury-fatality accidents, the constant has a mean and a variance equal to -6.495 and 12.613, respectively. And, the traffic flow has a mean and a variance equal to 0.778 and 0.089, correspondingly. In order to specify an informative prior for these parameters we assumed a normal distribution as stated in the case of non-informative priors. Under this approach, we adopted an informative prior with a large variance (say, semi-informative prior) because of the fact that informative priors with small to medium variances were penalizing the estimation results. Hence, for regression parameters mean values have been specified in the analysis as calculated above with a large variance of 1000. To summarize, following priors have been used for regression parameters.

For Total accidents:

- Constant; $a_0 \sim Normal\ (-6.186,\ 1000)$
- Traffic flow; $a_1 \sim Normal\ (0.788,\ 1000)$

For Injury-Fatality accidents:

- Constant; $a_0 \sim Normal\ (-6.495,\ 1000)$
- Traffic flow; $a_1 \sim Normal\ (0.778,\ 1000)$

4.2.5 Goodness-of-fit

In this study, models obtained from different types of prior specifications have been also monitored in terms of goodness-of-fit. So that it was possible to verify the effect of the

prior specification on the model-fitting. For this purpose, the deviance information criterion, DIC (Spiegelhalter et al., 2002), were computed to compare model-fitting. DIC is a Bayesian goodness-of-fit measure, and generally, a smaller DIC value indicates a better fit. However, one should take into consideration that as stated by Spiegelhalter et al. (2002), differences of more than 10 might definitely rule out the model with the higher DIC, differences between 5 and 10 are substantial, but if the difference in DIC is, say, less than 5, and the models make very different inferences, then it could be misleading just to report the model with the lowest DIC.

Table 4.II Reported values for constant and traffic flow (previous studies)

	Constant	
Previous studies	*Total accidents*	*Injury-Fatality*
Persaud et al, 2004	-7.432, -6.541, -6.973, -5.817	-8.770
Caliendo et al, 2007	0.539, 0.035	-1.353, -1.330
Lord et al, 2008	-8.108, -8.459	-7.960, -8.139
Geedipally et al, 2008	-6.750, -9.240, -2.990	-
HSM, 2010	-9.025, -9.653	-8.505, -9.410
	Traffic flow	
Previous studies	*Total accidents*	*Injury-Fatality*
Persaud et al, 2004	0.933, 0.844, 0.803, 0.811	0.945
Caliendo et al, 2007	0.221, 0.323	0.419, 0.391
Lord et al, 2008	1.028	0.858
Geedipally et al, 2008	0.72, 1.12, 0.43	-
HSM, 2010	1.049, 1.176	0.958, 1.094

4.2.6 Computations

In this study the replication of 100 datasets for each case of the sample mean and the sample size, obtaining site ranks, and the calculation of Spearman's correlation coefficients have been processed in the statistical software R (R Development Core Team, 2004). Then OpenBUGS (for performing Bayesian inference Using Gibbs Sampling) were used for running MCMC simulations to estimate model parameters (Spiegelhalter et al., 2003). In fact, since the number of datasets to be analyzed was large, R2OpenBUGS (Sturtz et al., 2005) was used to connect R and OpenBUGS together for the convenience of the computational purpose. In OpenBUGS two different chains were defined for producing samples. The total number of iterations was 7000 from which initial 3000 iterations were discarded as burn-in; therefore, 4000 iterations were used to compute posteriors. Finally, for analyzing the case study in order to obtain the true estimates, the statistical software Stata (StataCorp LP) has been utilized for the MLE.

4.3 CASE STUDY AND DATA DESCRIPTION

The case study employed in this paper consists of 958 kilometers of rural highway segments in New Brunswick, Canada. Two types of accidents considered here were injury-fatality and total accidents, which basically represent two different sample means. For total and injury fatal accidents, the sample means are 20.65 (acc./3 years) and 4.36 (acc./2 years), respectively. Total accident data covers a period of three years, 2004 to 2006. And Injury-fatality accident data represents a two year period, 2004 and 2005. Table 4.III shows the summary statistics of the data. The case study has been used to create different samples and to represent the true parameter estimates. Accordingly, three

samples in terms of size have been chosen from the case study; each sample included randomly selected observations.

Table 4.III Summary statistics of observed data

Variables	*Mean*	*Standard deviation*	*Minimum*	*Maximum*
Total accidents	20.650	13.920	0	65
Injury-Fatality accidents	4.360	3.566	0	14
Traffic flow (AADT)	7887.700	3337.916	4435	17550
Length	11.974	4.727	3.170	19.800

4.4 RESULTS AND DISCUSSIONS

4.4.1 Parameter Estimates and Associated Credible Intervals

Estimation results in terms of the mean and the 95% credible interval are reported in Tables 4.IV and 4.V for total accidents and injury-fatality accidents, respectively. Basically, these values represent the mean of 100 replications. In addition, true estimates, obtained by the MLE for observed data, are tabulated in above mentioned tables. The comparison between true estimates and different prior characterization estimates shows that a non-informative prior on the inverse dispersion parameter (φ) provides more inaccurate estimates for this parameter, especially, when analyzing data characterized by a low sample mean, a small sample size, or a combination of these two (in e.g., injury-fatality accidents with a sample size of 20). As sample size decreases, the difference in φ estimates increases (Table 4.IV and 4.V). These differences are still more significant under the relatively low sample mean data (Table 4.V). For instance, while φ estimate for

the non-informative prior specification in a data with a high sample mean (total accidents) and 20 observations varied from 4.032 to 6.312, this estimate varied from 4.608 to 84.307 in a data with a relatively low sample mean (injury-fatality accidents). Therefore, in such cases adopting an informative prior approach for φ is fundamental in order to obtain accurate estimates. This study also shows that there is a slight variation in φ estimates for two approaches in which informative priors are adopted (Tables 4.IV and 4.V). Furthermore, one can observe that credible intervals of φ in the non-informative approach are greater than those in two informative approaches. The latter, indicates that φ estimations using an informative prior can give more precise estimates by providing a smaller boundary around the mean. With regard to credible intervals for φ, as it is usually expected, the confidence interval of true estimate is smaller than the credible interval bound. One should take into account that confidence and credible intervals have different meanings (Carlin and Louis, 2009). That is, certain mean for a population with an specific percentage (like 95%) credible interval directly indicates the probability for that population being in that credible interval boundary while confidence interval does not imply the same implication.

As indicated in Tables 4.IV and 4.V, differences in regression parameters estimates (constant and traffic flow) are not as sensitive as φ estimates to three prior specification structures. However, differences between true estimates and different prior characterization estimates are more obvious as sample size and sample mean decrease. Indeed, for both sample mean cases, total accidents and injury-fatality accidents, with a small sample size (20 observations) the non-informative prior approach provides the least accurate estimates (Tables 4.IV and 4.V). In these cases, the approach in which φ has an

informative prior and regression parameters have a semi-informative prior provides the best estimates in terms of accuracy - similarity to true estimates. Therefore, the use of such an approach should be taken into account, particularly, in modeling datasets with a low sample mean and a small sample size. For credible intervals of regression parameter estimates, the non-informative prior approach had the smallest interval followed by that with only φ informative, for all cases. This is in contrast with φ credible intervals for which the non-informative prior approach provided the largest intervals.

Table 4.IV Estimation results for high sample mean data, Total accidents (T = 3 years), sample mean = 20.65 (acc./T)

		ln(a_0)			*Traffic flow*			*Inverse dispersion parameter (φ)*		
True Value (MLE)		-3.014[a] (-6.101, 0.074)[b]			0.397 (0.0512, 0.743)			4.032 (2.703, 6.024)		
Sample Size		Inf.1[c]	Inf.2[d]	Ninf.[e]	Inf.1	Inf.2	Ninf.	Inf.1	Inf.2	Ninf.
20	mean	-2.746	-2.685	-2.633	0.367	0.360	0.353	3.555	3.556	6.213
	2.5%[f]	-9.485	-9.374	-8.606	-0.353	-0.360	-0.292	1.775	1.776	2.154
	97.5%[g]	3.710	3.767	3.152	1.123	1.110	1.022	6.127	6.128	16.230
50	mean	-2.916	-2.875	-2.863	0.385	0.381	0.379	3.827	3.827	4.493
	2.5%	-7.083	-7.056	-6.849	-0.069	-0.073	-0.054	2.395	2.394	2.620
	97.5%	1.163	1.200	1.030	0.852	0.849	0.825	5.716	5.719	7.201
80	mean	-3.108	-3.095	-3.069	0.408	0.406	0.403	3.930	3.932	4.339
	2.5%	-6.290	-6.238	-6.117	0.061	0.061	0.069	2.688	2.689	2.861
	97.5%	-0.011	-0.011	-0.087	0.765	0.759	0.746	5.502	5.508	6.304

[a] Mean value
[b] 95% Confidence Interval
[c] Inf.1: φ has an informative prior, and regression parameters have semi-informative priors
[d] Inf.2: only φ has an informative prior
[e] Ninf.: all parameters have a non-informative prior
[f] Lower limit, 2.5 percentile, for the 95% Credible Interval
[g] Upper limit, 97.5 percentile, for the 95% Credible Interval

Table 4.V Estimation results for relatively low sample mean data, Injury-Fatality accidents (T = 2 years), sample mean = 4.36 (acc./T)

		ln(a_0)			*Traffic flow*			*Inverse dispersion parameter (φ)*		
True Value (MLE)		-4.915[a] (-8.641, -1.188)[b]			0.435 (0.018, 0.852)			4.608 (2.208, 9.615)		
Sample Size		Inf.1[c]	Inf.2[d]	Ninf.[e]	Inf.1	Inf.2	Ninf.	Inf.1	Inf.2	Ninf.
20	mean	-4.738	-4.603	-4.585	0.412	0.397	0.394	3.961	3.949	84.307
	2.5%[f]	-13.175	-13.024	-11.991	-0.520	-0.534	-0.436	1.458	1.448	1.995
	97.5%[g]	3.601	3.735	2.845	1.356	1.340	1.223	8.574	8.530	638.117
50	mean	-4.884	-4.854	-4.848	0.432	0.428	0.427	4.349	4.354	22.616
	2.5%	-9.991	-9.975	-9.666	-0.131	-0.134	-0.103	2.127	2.129	2.471
	97.5%	0.169	0.201	-0.083	1.000	0.999	0.964	8.102	8.112	159.642
80	mean	-5.174	-5.156	-5.143	0.464	0.462	0.461	4.670	4.664	19.038
	2.5%	-9.014	-8.972	-8.774	0.041	0.038	0.057	2.529	2.524	2.934
	97.5%	-1.388	-1.359	-1.529	0.893	0.889	0.866	8.126	8.132	125.608

[a] Mean value

[b] 95% Confidence Interval

[c] Inf.1: φ has an informative prior, and regression parameters have semi-informative priors

[d] Inf.2: only φ has an informative prior

[e] Ninf.: all parameters have a non-informative prior

[f] Lower limit, 2.5 percentile, for the 95% Credible Interval

[g] Upper limit, 97.5 percentile, for the 95% Credible Interval

4.4.2 Hotspot Identification Comparisons

Table 4.VI shows calculated Spearman's correlation coefficients between the true ranks and those ranks based on three different approaches adopted in the specification of priors for model parameters. From this table, it can be clearly inferred that the effect of various prior characterizations in the identification of hazardous sites was not significant.

However, the informative prior structure performed slightly better compare with the non-informative prior specification. These differences were somewhat greater when working on data with a small sample size, a low sample mean, and particularly, the combination of these two.

Table 4.VI Spearman's correlation coefficients (Hotspot identification)

High sample mean data, Total accidents (T = 3 years)			
Sample size	Inf.1[a]	Inf.2[b]	Ninf.[c]
20	0.9048	0.9051	0.9039
50	0.9403	0.9403	0.9402
80	0.9428	0.9428	0.9427
Relatively low sample mean data, Injury-Fatality accidents (T = 2 years)			
Sample size	Inf.1	Inf.2	Ninf.
20	0.7672	0.7671	0.7474
50	0.8360	0.8357	0.8297
80	0.8375	0.8375	0.8318

[a] Inf.1: φ has an informative prior, and regression parameters have semi-informative priors
[b] Inf.2: only φ has an informative prior
[c] Ninf.: all parameters have a non-informative prior

4.4.3 Goodness-of-fit Comparisons

As reported in Table 4.VII, differences in DIC values between three prior specification approaches were smaller than 5 in all cases of the sample mean and the sample size. Non-informative characterizations provided the biggest DIC value. However, DIC differences of smaller than 5 are usually thought of as hardly worth mentioning (Carlin and Louis,

2009). Thus, based on the assumptions and methodology used in this study it can be concluded that goodness-of-fit was not sensitive to the prior specification.

Table 4.VII Goodness-of-fit, DIC values

High sample mean data, Total accidents (T = 3 years)			
Sample size	Inf.1[a]	Inf.2[b]	Ninf.[c]
20	128.877	128.882	129.338
50	321.759	321.748	321.949
80	510.707	510.670	510.957
Relatively low sample mean data, Injury-Fatality accidents (T = 2 years)			
Sample size	Inf.1	Inf.2	Ninf.
20	90.149	90.145	91.665
50	225.153	225.180	226.744
80	353.978	354.022	355.372

[a] Inf.1: φ has an informative prior, and regression parameters have an informative prior with a large variance
[b] Inf.2: only φ has an informative prior
[c] Ninf.: all parameters have a non-informative prior

4.5 SUMMARY AND CONCLUSIONS

In this paper, in the Bayesian paradigm by applying the hierarchical Poisson-Gamma structure, the impact of various prior specifications (Informative, semi-informative, and non-informative) on the model validity has been examined considering the following aspects: (1) regression parameters estimates, (2) the inverse dispersion parameter estimates, (3) hotspot identification through the ranking of sites based on the posterior distribution of θ and Spearman's correlation coefficient, and (4) goodness-of-fit criterion using DIC values.

A case study of 958 km of rural highway segments in New Brunswick, Canada has been used to present the observed data and to obtain true (real) estimates applying the MLE. Three different sample sizes (including 20, 50, and 80 observations) with two different sample means (i.e., data characterized by high mean values and relatively low mean values) have been analyzed. In fact, these sample means have been introduced in the analysis using injury-fatality and total accidents for a period of 2 years and 3 years, respectively. For each case, a series of 100 simulated data has been produced. Then, three different prior specifications have been employed to estimate model parameters - regression parameters and the inverse dispersion parameter. In order to generate and analyze datasets we used the Poisson-Gamma model. The analysis has been done in a hierarchical fashion by adopting a full Bayes framework through MCMC methods.

The results indicated that the specification of an informative prior for the inverse dispersion parameter, introduced in the analysis based on previous studies, has a noteworthy impact on this parameter estimates. In particular, when working on data characterized with small sample size and low sample mean. We found that regression parameters are less sensitive to prior specifications compare with the inverse dispersion parameter. For regression parameters, similarly, a non-informative specification was the most inaccurate when modeling limited data. This founding was also valid for data characterized with a high sample mean. In other words, data characterized by high mean values and small number of observations was still affected by prior choice. Considering regression parameters, the approach in which regression parameters were semi-informative and the inverse dispersion parameter was informative provided the best estimates in terms of the estimation accuracy. In general, we observed that as sample size

and/or sample mean increase, parameter estimates approach the true estimates. Moreover, hotspot identification wasn't found to be affected considerably by different prior specifications; yet, models with informative priors, slightly, performed better. Similarly, DIC values didn't show any significant improvement in goodness-of-fit when using informative priors for model parameters. Based on this study, we recommend the use of informative priors especially in modeling data characterized by a small sample size and a low sample mean. Lastly, the future research should investigate the effect of prior choice on other road transportation facilities such as intersections. Moreover, the methodology used in this study should also focus on other regression approaches like hierarchical Poisson-Lognormal models to explore the impact of different prior specifications.

CHAPTER 5

SUMMARY AND CONCLUSIONS

This chapter consists of three sections. First section summarizes the thesis and provides the contributions of this research study. Second section, suggests various steps that researchers and practitioners can follow to benefit from this work. Recommendations for future research are discussed in the last section.

5.1 SUMMARY AND MAJOR CONTRIBUTIONS

In this thesis, a case study comprising 958 km of highway segments in New Brunswick have been adopted to examine some of the most important aspects of road safety using the Bayesian approach. The focus was on accident modeling, model-fitting, and the development of reliable SPFs that are used for hotspot identification and countermeasure assessment. Furthermore, an important part of the thesis explored the effect of prior specifications in Bayesian analysis of accident data.

In Chapter 3, three datasets representing different accident severities were analyzed. A series of comparisons were presented for Poisson and hierarchical Poisson mixture models. Different characterizations of the inverse dispersion parameter in hierarchical Poisson-Gamma models were studied; that is, this parameter were introduced in the analysis as fixed, varying as a function of site characteristics, and randomly varying across sites. Similarly, in a novel approach the inverse of variance in the hierarchical Poisson-Lognormal model was characterized as a function of site characteristics. In addition, considering different severities, the presence of various contributing factors in

the mean and the variance functions was examined. For the above mentioned objectives, more than 35 Bayesian models have been analyzed, running the MCMC simulation to make posterior inferences, form which 26 models were presented in Chapter 3.

The results showed that, for the case study, hierarchical Poisson-Gamma models provided the best model-fitting for all three datasets: property damage only, injury-fatality, and total accidents. Regarding the variance function characterizations, fixed inverse dispersion parameter and fixed inverse of variance presented the worst fit to all datasets. The introduction of the inverse of variance as a function of site characteristics in hierarchical Poisson-Lognormal models performed adequately improving goodness-of-fit; similar to hierarchical Poisson-Gamma models. Additionally, from this research study, it can be inferred that modeling accidents by severity is crucial for the identification of the contributing factors. In fact, it was found that the interaction between precipitation and the density of horizontal curves was statistically significant in modeling injury-fatality accidents; however, these variables when considered separately weren't affecting the injury-fatality accidents. Finally, the results demonstrated that contributing factors presented in the mean and the variance functions were not necessarily the same.

In Chapter 4, a data simulation framework based on the previous case study was used in order to verify how the specification of prior in hierarchical Poisson-Gamma models affects the results - parameter estimates, hotspot identification, and goodness-of-fit. The Poisson-Gamma model is the most common model in road safety and is widely used and accepted in a variety of studies. When applying Bayesian inference in accident analysis, almost all studies have used non-informative or vague priors for model parameters. Here, the sensitivity of the analysis to prior choice was tested. Since, in the road safety

literature, it has been demonstrated that the MLE provides inaccurate results for data characterized by a low sample mean and a small sample size, different sample sizes and sample means were taken into account in this thesis applying the Bayesian estimation. Therefore, a high sample mean data and a relatively low sample mean data were extracted from the case study. High sample mean dataset was presented by total accidents for a period of three years. And relatively low sample mean dataset was presented by injury-fatality accidents for a period of two years. Consequently, three sample sizes consisting of 20, 50, and 80 observations have been, randomly, chosen from the case study, for each sample mean dataset. Then, all synthetic datasets were analyzed using three diverse prior specifications. So in total 18 models were developed for this chapter's objectives. The prior specification included: (a) non-informative priors as commonly used in the road safety literature, (b) an informative prior for the inverse dispersion parameter, and (c) an informative prior for the inverse dispersion parameter and semi-informative priors for regression parameters. Furthermore, these priors have been defined based on previous studies as explained in detail in Chapter 4.

The model outcomes indicated that introducing informative priors improved the parameter estimation accuracy, particularly, for the inverse dispersion parameter. Moreover, this improvement was more obvious as sample size and sample mean decreased. Therefore, in cases of low mean problem (low sample mean) and limited data, the use of an informative prior specification approach is strongly recommendable. To understand the importance of this research one should take into consideration that typical accident data are usually limited in size and characterized by low mean problem. This

study also showed that the choice of priors doesn't have any outstanding influence on hotspot identification and goodness-of-fit.

5.2 PRACTICAL SUGGESTIONS FOR PRACTITIONERS

Practitioners and researchers wishing to implement the methodologies discussed in this thesis can follow the bellow mentioned steps to analyze accident datasets:

1. Verifying data availability related to observed accidents and site characteristics;
2. Explanatory data analysis to identify the most important contributing factors that may affect accident frequencies (property damage only, injury-fatality, and total accidents);
3. Choosing an appropriate SPF;
4. Selecting a regression approach (hierarchical Poisson mixture models are recommended);
5. Verifying if data is characterized by low mean problem and/or small sample size;
6. In the case of low mean problem and/or small sample size, it is recommended to estimate informative or semi-informative priors, depending on the type of parameters, from previous studies. Otherwise, use non-informative priors as indicated in the road safety literature;
7. Selecting a variance function characterization approach. For instance, the inverse dispersion parameter in hierarchical Poisson-Gamma models may be fixed, varying as function of site characteristics, and randomly varying across sites.
8. Applying Bayesian statistics to obtain the model parameters.

5.3 RECOMMENDATION FOR FUTURE WORK

This research study provides some practical methods in order to improve SPFs reliability in terms of goodness-of-fit and parameter estimation accuracy, which in turn can lead to more dependable hotspot identification and safety countermeasure assessment. In this thesis, the research focus was on road segments. However, future work can focus on applying these methodologies to other road facilities. Discussions related to model comparisons, variance function characterizations, severity modeling considering interaction between site characteristics (e.g., road alignment and weather condition interaction), and the choice of prior should be applicable to rural or urban intersections.

In this research study, the effect of prior specification on parameter estimates, hotspot identification, and goodness-of-fit was examined in Poisson-Gamma models. Future research can investigate the impact of prior choice on other Poisson mixture regression approaches such as hierarchical Poisson-Lognormal models.

Lastly, future research can provide specific guidelines for practitioners (1) to identify the amount of data (e.g., in terms of years) required to obtain accurate estimates, particularly when using a non-informative prior specification approach and (2) to provide a series of informative priors for the most important contributing factors such as traffic flow and segment length based on the type of the road facility and accident severity.

REFERENCES

1. Amador, L., Mrawira, D. (2011) Bayesian regression in pavement deterioration modeling: revisiting the AASHO road test rut depth model. Proceedings of the 90th Annual Meeting of the Transportation Research Board, Washington DC, USA.
2. Anastasopoulos, P., Mannering, F. (2008) A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention, 41, 153-159.
3. Bedford, T., Cooke, R. (2001) Probabilistic risk analysis. Cambridge University Press, Cambridge, UK.
4. Brooks, S.P., Gelman, A. (1998) Alternative methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics. 7, 434-455.
5. Carlin, B.P., Louis, T.A. (2009) Bayesian methods for data analysis. Chapman & Hall/CRC. Taylor & Francis Group. Boca Raton, Florida.
6. Cheng, W., Washington, S. (2005) Experimental evaluation of hotspot identification methods. Accident Analysis and Prevention. 37, 870-881.
7. Congdon, P.D. (2010) Applied Bayesian hierarchical methods. Chapman & Hall/CRC. Taylor & Francis Group. Boca Raton, Florida.
8. El-Basyouny, K., Sayed, T. (2010) Safety performance functions with measurement errors in traffic volume. Safety Science. 48, 1339-1344.
9. El-Basyouny, K., Sayed, T. (2009) Accident prediction models with random corridor parameters. Accident Analysis and Prevention. 41, 1118-1123.
10. Gamerman, D., Lopes, H.F. (2006) Markov Chain Monte Carlo stochastic simulation for Bayesian inference, 2nd ed. Chapman & Hall/CRC. London, UK.

11. Geedipally, S.R., Lord, D., Park, B.J. (2009) Analyzing different parameterization of the varying dispersion parameter as a function of segment length. Transportation Research Record. 2103, 108-118.
12. Geedipally, S.R., Lord, D. (2008) Effects of the varying dispersion parameter of Poisson-gamma models on the estimation of confidence intervals of crash prediction models. Transportation Research Record. 2061, 46-54.
13. Gelman, A., Hill, J. (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge University Press. Cambridge, UK.
14. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (1995) Bayesian data analysis. Chapman & Hall. London, UK.
15. Hauer, E. (1997) Observational before-after study in road safety. Pergamon Press, Elsevier Science Ltd. Oxford, UK.
16. Hauer, E. (2001) Over-dispersion in modeling accidents on road sections and in empirical Bayes estimation. Accident Analysis and Prevention. 33, 799-808.
17. Hauer, E, and Persaud, B.N. (1984) Problem of identifying hazardous locations using accident data. Transportation Research Record. 975, 1984.
18. Heydecker, B.G., and Wu, J. (2001) Identification of sites for accident remedial work by Bayesian statistical methods: an example of uncertain inference. Advances in Engineering Software. 32, 859-869.
19. Highway safety manual (2010). American Association of State Highway and Transportation Officials. Washington, DC, USA.
20. Hinde, J., Demetrio, C.G.B. (1998). Over-dispersion: model and estimation. Computational Statistics and Data Analysis. 27 (2), 151–170.

21. International Road Assessment Program (2011) Road deaths in developing countries. Available from http://www.irap.net/library/cat_view/4-research-and-technical-papers.html [accessed in January 2012].
22. Kim, H., Sun, D., Tsutakawa, R.K. (2002) Lognormal vs. gamma: extra variations. Biomedical Journal. 44, 305–323.
23. Lambert, P., Sutton, A., Burton, P., Abrams, K., Jones, D. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Statistics in Medicine. 24, 2401–2428.
24. Lan, B., Persaud, B., Lyon, C., Bhim, R. (2009) Validation of a full Bayes methodology for observational before-after road safety studies and application to evaluation of rural signal conversions. Accident Analysis and Prevention. 41, 574-580.
25. Lord, D., Geedipally, S., Persaud, B., Washington, S., Van Schalkwyk, I., Ivan, J., Lyon, C., Jonsson, T. (2008). Methodology to predict the safety performance of rural multilane highways. National Cooperative Highway Research Program, NCHRP, Web Document 126.
26. Lord, D., Miranda-Moreno, L.F. (2008) Effects of the low sample mean values and the small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models: a Bayesian perspective. Safety Science. 46, 751-770.
27. Lord, D., Persaud, B. (2000) Accident prediction models with and without trend: application of the generalized estimating equation. Transportation Research Record. 1717, 102–108.

28. Ma, j., Kockelman, K. (2006) Bayesian multivariate Poisson regression for models of injury count, by severity. Transportation Research Record. 1950, 24-34.
29. Miaou, S.P., Hu, P., Wright, T., Rathi, A.K., Davice, S.C. (1992) Relationships between truck accidents and highway geometric design: a Poisson regression approach. Proceedings of the 71st Annual Meeting of the Transportation Research Board, Washington DC, USA.
30. Miaou, S.P., Lord, D. (2003) Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. Transportation Research Record. 1840, 31–40.
31. Miaou, S.P., and Song, J.J. (2005) Bayesian ranking of sites for engineering safety improvement: decision parameter, treatability concept, statistical criterion and spatial dependence. Accident Analysis and Prevention. 37, 699-720.
32. Milton, J.C., Shankar, V.N., Mannering, F.L. (2008) Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accident Analysis and Prevention. 40, 260-266.
33. Miranda-Moreno, L.F., Fu, L., Saccomano, F., and Labbe, A. (2005) Alternative risk models for ranking locations for safety improvement. Journal of the Transportation Research Board. 1908, 1-8.
34. Miranda-Moreno, L.F., Labbe, A., Fu, L. (2007) Multiple Bayesian testing procedures for selecting hazardous sites. Accident Analysis and Prevention. 39, 1192-1201.

35. Miranda-Moreno, L.F., Lord, D., Fu, L. (2008) Bayesian road safety analysis: incorporation of past experience and effect of hyper-prior choice. Proceedings of the 87th Annual Meeting of the Transportation Research Board. Washington DC, USA.
36. Mitra, S., Washington, S.P. (2006) On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis and Prevention. 39, 459–468.
37. National Climate Data and Information Archive (Canada). Available from http://climate.weatheroffice.gc.ca/Welcome_e.html [accessed in October 2011].
38. Oh, J., Washington, S.P., Nam, D. (2006) Accident prediction model for Railway-highway interfaces. Accident Analysis and Prevention. 38, 346–356.
39. Persaud, B., Lyon, C., and Nguyen, T. (1999) Empirical Bayes procedure for ranking sites for safety investigation by potential for safety improvement. Transportation Research Record. 1665, 7-12.
40. Persaud, B., Lan, B., Lyon, C., Bhim, R. (2010) Comparison of empirical Bayes and full Bayes approaches for before-after road safety evaluations. Accident Analysis and Prevention. 42, 38-43.
41. Poch, M., Mannering, F.L. (1996) Negative binomial analysis of intersection accident frequencies. Journal of Transportation Engineering. 122, 105-113.
42. Rao, J.N. (2003) Small area estimation. John Wiley and Sons.
43. R Development Core Team (2004). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http: //www.R-project.org.
44. Ruppert, D. (2010) Statistics and data analysis for financial engineering. Springer; 1st Edition.

45. Saccomano, F., Nassar, S., Shortreed, J. (1996) Reliability of statistical road accident injury severity models. Transportation Research Record. 1542, 14-23.
46. Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). Journal of the Royal Statistical Society, Series B. 64, 583-616.
47. Spiegelhalter, D.J., Thomas, A., Best, N.G., Lunn, D. (2003). WinBUGS version 1.4 users manual." MRC Biostatistics Unit, Cambridge. http://www.mrc-bsu.cam.ac.uk/bugs/.
48. StataCorp LP. http://www.stata.com/company/.
49. Sturtz, S., Ligges, U., Gelman, A. (2005) R2WinBUGS: a package for running WinBUGS from R. Journal of Statistical Software. 12(3). R2OpenBUGS was adapted from R2WinBUGS by Neal Thomas.
50. Transport Canada (2007) Analysis and estimation of the social cost of motor vehicle collisions in Ontario, TP 14800 E (08-2007), Available from http://www.tc.gc.ca/media/documents/roadsafety/TP14800E.pdf [accessed in February 2012].
51. Transport Canada (2011) Canadian motor vehicle traffic collision statistics, TP 3322 (05-2011), collected in cooperation with the Canadian council of motor vehicle administrators. Available from http://www.tc.gc.ca/media/documents/roadsafety/tp3322-2009_eng.pdf [accessed in February 2012].
52. Van Dongen, S. (2006) Prior specification in Bayesian statistics: three cautionary tales. Journal of Theoretical Biology. 242, 90–100.
53. Washington, S.P., Congdon, P., Karlaftis, M., Mannering, FL. (2003) Statistical and econometric methods for transportation data analysis. Chapman & Hall/CRC. USA.
54. Winkelmann, R. (2003) Econometric analysis of count data. Springer-Verlag, Berlin, Germany.

APPENDICES

1) Precipitation data for weather stations across highway segments

Precipitation observations used in Chapter 3 are reported in Table A.I. The mean values were introduced in the analysis to verify the effect of environmental exposure.

Table A.I Precipitation in mm

	Weather Stations			
Year	*Moncton*	*Fredericton*	*Woodstock*	*St. Leonard*
2004	1132.2	779.0	908.6	930.3
2005	1412.2	1364.5	1615.4	1381.8
2006	1204.2	1199.5	1256.2	1148.0

2) Spearman's Correlation Coefficient

Spearman's correlation coefficient represents the correlation between two sets of statistical ranks. This coefficient varies from -1 to +1 indicating the perfect negative and the perfect positive correlations, respectively. One should take into consideration that Spearman's coefficient does not necessarily imply that variables associated to these compared ranks are correlated. This coefficient is computed based on Equation A.1.

$$s = 1 - \frac{6\sum d^2}{n(n^2-1)} \qquad [A.1]$$

where

s is the Spearman's correlation coefficient;

d is difference between corresponding ranks;

n is number of elements to be ranked.

The calculated value should then be compared with critical values of Spearman's correlation coefficient in order to be accepted or rejected.

3) Example of History Plots - OpenBUGS

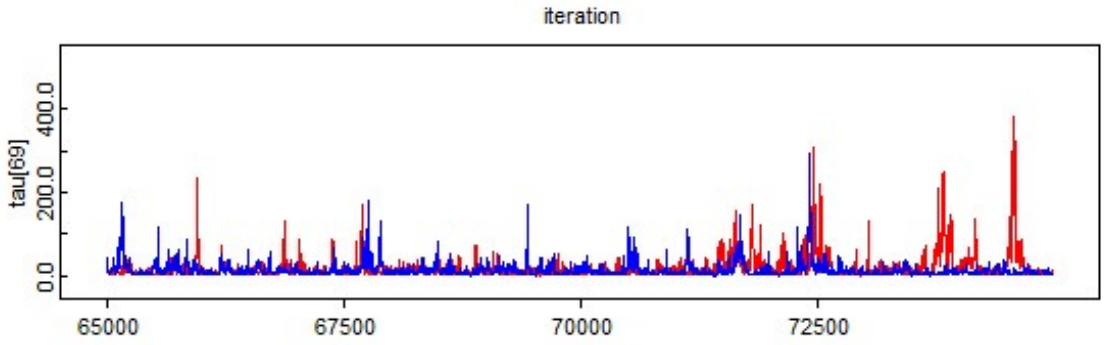

Fig. A.I Unstable chains – convergence not reached

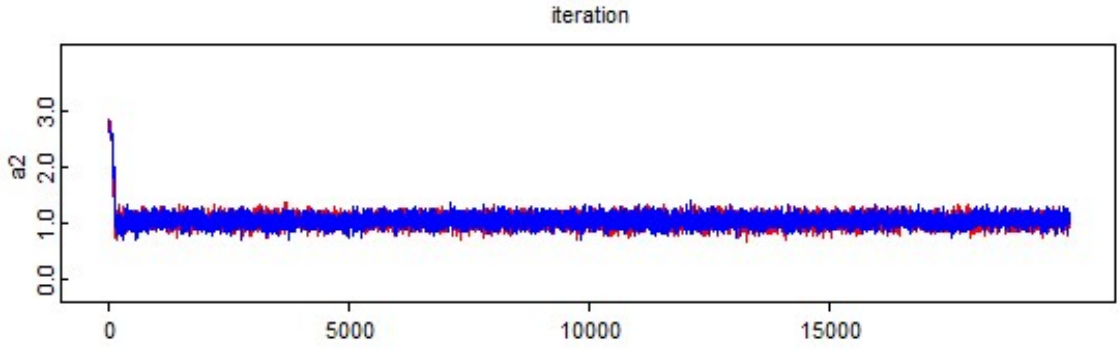

Fig. A.II Stable chains – convergence reached

4) Example of Density Plots – OpenBUGS

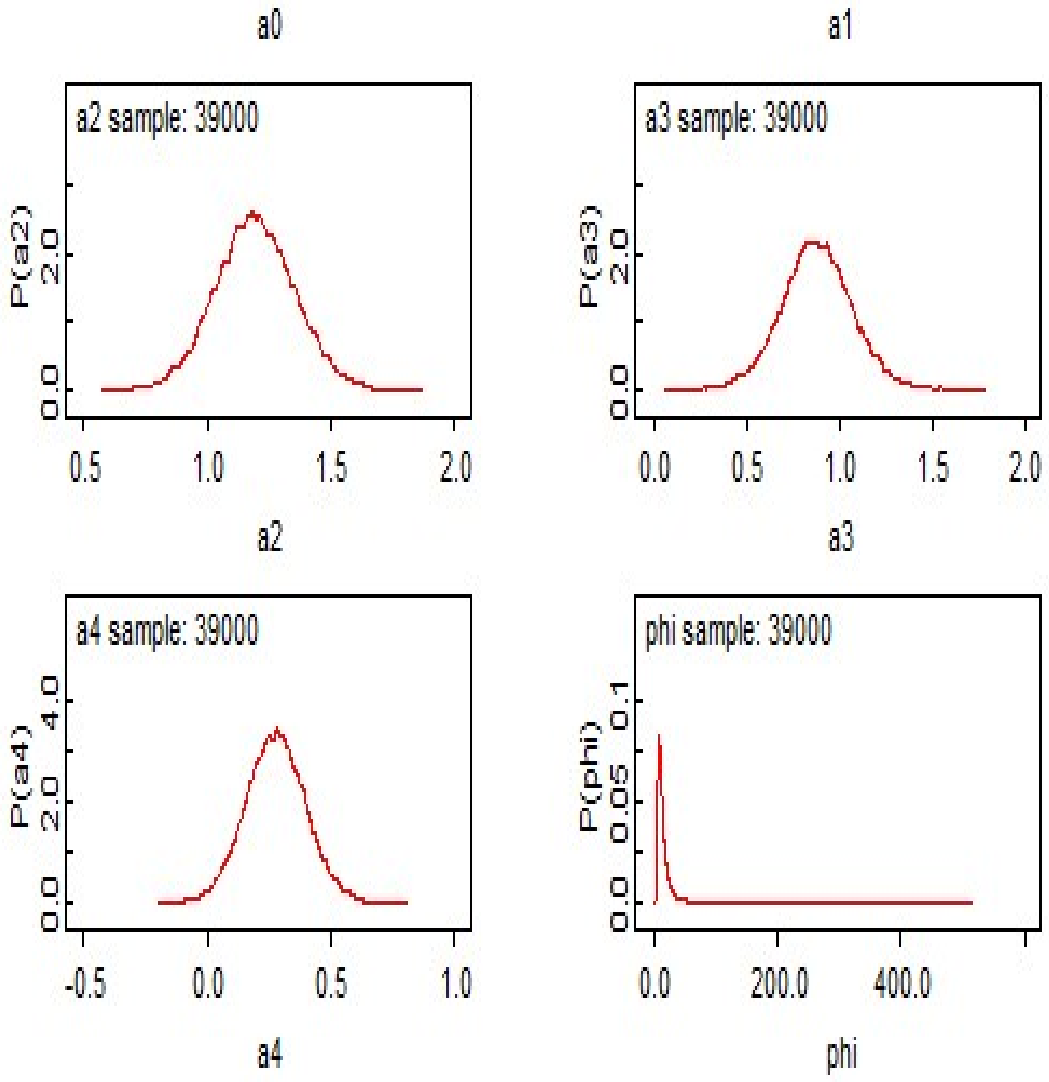

Fig. A.III Density Plots

5) R code for prior specification comparisons

```
#   non informative
nipg_a0 <- rep.int(0,100)
nipg_a1 <- rep.int(0,100)
nipg_DIC <- rep.int(0,100)
nipg_phi <- rep.int(0,100)
nipg_a02.5 <- rep.int(0,100)
nipg_a12.5 <- rep.int(0,100)
nipg_phi2.5 <- rep.int(0,100)
nipg_a097.5 <- rep.int(0,100)
nipg_a197.5 <- rep.int(0,100)
nipg_phi97.5 <- rep.int(0,100)
nipg_a0sd <- rep.int(0,100)
nipg_a1sd <- rep.int(0,100)
nipg_phisd <- rep.int(0,100)
#  all informative
ipg_a0 <- rep.int(0,100)
ipg_a1 <- rep.int(0,100)
ipg_DIC <- rep.int(0,100)
ipg_phi <- rep.int(0,100)
ipg_a02.5 <- rep.int(0,100)
ipg_a12.5 <- rep.int(0,100)
ipg_phi2.5 <- rep.int(0,100)
ipg_a097.5 <- rep.int(0,100)
ipg_a197.5 <- rep.int(0,100)
ipg_phi97.5 <- rep.int(0,100)
```

```
ipg_a0sd <- rep.int(0,100)
ipg_a1sd <- rep.int(0,100)
ipg_phisd <- rep.int(0,100)
#  only phi informative
ipg2_a0 <- rep.int(0,100)
ipg2_a1 <- rep.int(0,100)
ipg2_DIC <- rep.int(0,100)
ipg2_phi <- rep.int(0,100)
ipg2_a02.5 <- rep.int(0,100)
ipg2_a12.5 <- rep.int(0,100)
ipg2_phi2.5 <- rep.int(0,100)
ipg2_a097.5 <- rep.int(0,100)
ipg2_a197.5 <- rep.int(0,100)
ipg2_phi97.5 <- rep.int(0,100)
ipg2_a0sd <- rep.int(0,100)
ipg2_a1sd <- rep.int(0,100)
ipg2_phisd <- rep.int(0,100)
# spearman all informative and non informative against true ranking
spearman.pginf <- rep.int(0,100)
spearman.pginf2 <- rep.int(0,100)
spearman.pgninf <- rep.int(0,100)
# mean of synthetic datasets
acc <- rep.int(0,100)
mean.acc <- rep.int(0,100)
# DIC differences
delta.DIC <- rep.int(0,100)
delta.DIC2 <- rep.int(0,100)
```

```
for (j in 1:100)
{
acctot<- read.csv("data20sitesacctot3years.csv", header=TRUE)
    #variables: x1=AADT; x2=length#
x1 <- acctot$x1
x2 <- acctot$x2
    #----------Poisson-Gamma----------#
n <- nrow(acctot)
a0 <- 0.04912    # ln (a0) = -3.0135
a1 <- 0.39735
mu <- a0*(x2)*((x1)^a1)
phi <- 4.03
    #simulation
r <- rgamma(n,shape=phi,rate=phi)
theta <- mu*r
acc <- rpois(n,theta)
mean.acc[j] <- mean(acc)
acctotpg <- read.csv ("data20sitesacctot3years.csv", header=TRUE)
N <- nrow(acctotpg)
y <- acc
x1 <- acctotpg$x1
x2 <- acctotpg$x2
data <- list("N","y","x1","x2")
parameters <- c("a0","a1","phi","theta")
inits <- function(){list (a0=0,a1=0,phi=1,r=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1))}
# Bayesian Inference; non-informative prior #
acctotpg.ni <- bugs(data, inits, parameters, model.file="ni_acctotpg.txt",
n.chains=2,n.burnin=3000,n.iter=7000)
```

```
sim1 <- acctotpg.ni$summary

# Bayesian Inference; informative prior 1 #

acctotpg.in <- bugs(data, inits, parameters, model.file="inew_acctotpg.txt", n.chains=2,n.burnin=3000, n.iter=7000)

sim2 <- acctotpg.in$summary

# Bayesian Inference; informative prior 2 #

acctotpg.in2 <- bugs(data, inits, parameters, model.file="i_acctotpgphi.txt", n.chains=2,n.burnin=3000, n.iter=7000)

sim3 <- acctotpg.in2$summary

                                    # non-informative estimation #

nipg_DIC[j] <- acctotpg.ni$pD

 # mean parameters #

nipg_a0[j] <- sim1[1,1]

nipg_a1[j] <- sim1[2,1]

nipg_phi[j] <- sim1[3,1]


 # credible interval #

nipg_a02.5[j] <- sim1[1,3]

nipg_a12.5[j] <- sim1[2,3]

nipg_phi2.5[j] <- sim1[3,3]

nipg_a097.5[j] <- sim1[1,7]

nipg_a197.5[j] <- sim1[2,7]

nipg_phi97.5[j] <- sim1[3,7]


 # standard deviation #

nipg_a0sd[j] <- sim1[1,2]

nipg_a1sd[j] <- sim1[2,2]

nipg_phisd[j] <- sim1[3,2]
```

```
# informative estimation all informative#
ipg_DIC[j] <- acctotpg.in$pD
# mean parameters #
ipg_a0[j] <- sim2[1,1]
ipg_a1[j] <- sim2[2,1]
ipg_phi[j] <- sim2[3,1]

# credible interval #
ipg_a02.5[j] <- sim2[1,3]
ipg_a12.5[j] <- sim2[2,3]
ipg_phi2.5[j] <- sim2[3,3]
ipg_a097.5[j] <- sim2[1,7]
ipg_a197.5[j] <- sim2[2,7]
ipg_phi97.5[j] <- sim2[3,7]

 # standard deviation #
ipg_a0sd[j] <- sim2[1,2]
ipg_a1sd[j] <- sim2[2,2]
ipg_phisd[j] <- sim2[3,2]

# informative estimation only phi informative#
ipg2_DIC[j] <- acctotpg.in2$pD
# mean parameters #
ipg2_a0[j] <- sim3[1,1]
ipg2_a1[j] <- sim3[2,1]
ipg2_phi[j] <- sim3[3,1]
```

```
# credible interval #
ipg2_a02.5[j] <- sim3[1,3]
ipg2_a12.5[j] <- sim3[2,3]
ipg2_phi2.5[j] <- sim3[3,3]
ipg2_a097.5[j] <- sim3[1,7]
ipg2_a197.5[j] <- sim3[2,7]
ipg2_phi97.5[j] <- sim3[3,7]

 # standard deviation #
ipg2_a0sd[j] <- sim3[1,2]
ipg2_a1sd[j] <- sim3[2,2]
ipg2_phisd[j] <- sim3[3,2]

 # Ranking based on posterior mean
theta.ni<-sim1[4:23,1]
rank.theta.ni <- (n+1)-rank(theta.ni)
theta.i<-sim2[4:23,1]
rank.theta.i <- (n+1)-rank(theta.i)
theta.i2<-sim3[4:23,1]
rank.theta.i2 <- (n+1)-rank(theta.i2)
# True rank
rank.true <- (n+1)-rank(theta)
 #rank comparison between inf & non inf with true rank
spearman.pgtrue.inf <- 1-((6*(sum((rank.theta.i-rank.true)^2)))/(n*(n^2-1)))
spearman.pginf[j] <- spearman.pgtrue.inf
spearman.pgtrue.inf2 <- 1-((6*(sum((rank.theta.i2-rank.true)^2)))/(n*(n^2-1)))
spearman.pginf2[j] <- spearman.pgtrue.inf2
```

```
spearman.pgtrue.noninf <- 1-((6*(sum((rank.theta.ni-rank.true)^2)))/(n*(n^2-1)))

spearman.pgninf[j] <- spearman.pgtrue.noninf

 # compare DIC between informative and non-informative

delta.DIC[j] <- nipg_DIC[j]-ipg_DIC[j]

delta.DIC2[j] <- nipg_DIC[j]-ipg2_DIC[j]

}

# Saving results as a CSV file

write.csv(cbind(nipg_a0,nipg_a0sd,nipg_a02.5,nipg_a097.5,nipg_a1,nipg_a1sd,nipg_
a12.5,nipg_a197.5,nipg_phi,nipg_phisd,nipg_phi2.5,nipg_phi97.5,nipg_DIC,ipg_a0,i
pg_a0sd,ipg_a02.5,ipg_a097.5,ipg_a1,ipg_a1sd,ipg_a12.5,ipg_a197.5,ipg_phi,ipg_ph
isd,ipg_phi2.5,ipg_phi97.5,ipg_DIC,ipg2_a0,ipg2_a0sd,ipg2_a02.5,ipg2_a097.5,ipg2
_a1,ipg2_a1sd,ipg2_a12.5,ipg2_a197.5,ipg2_phi,ipg2_phisd,ipg2_phi2.5,ipg2_phi97.
5,ipg2_DIC,
spearman.pgninf,spearman.pginf,spearman.pginf2,delta.DIC,delta.DIC2,mean.acc),"
Result_pg20sitesacctot3years2infnew.csv",quote=F)
```

6) The inverse dispersion parameter values estimated for 100 datasets.

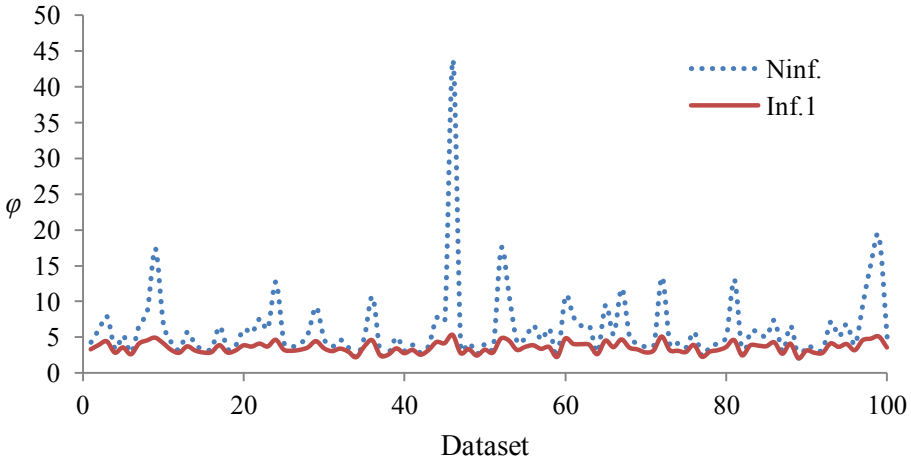

Fig. A.IV Inverse dispersion parameter (φ) - High mean data, 20 observations.

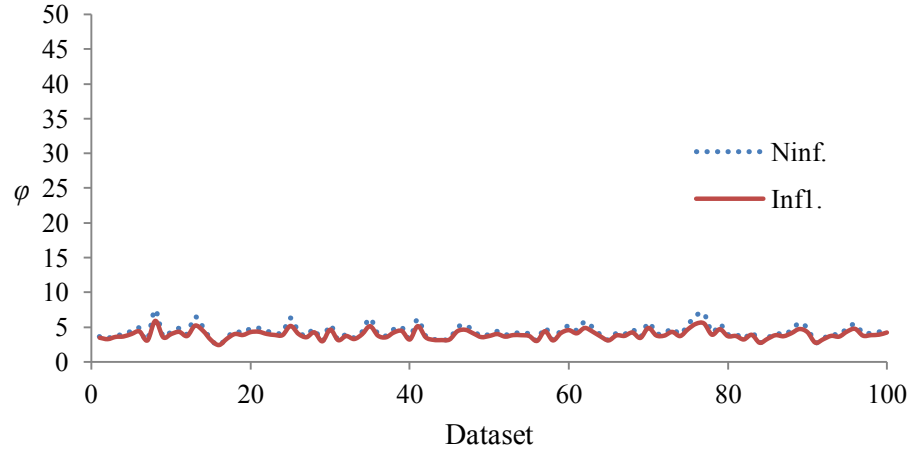

Fig. A.V Inverse dispersion parameter (φ) - High mean data, 80 observations.

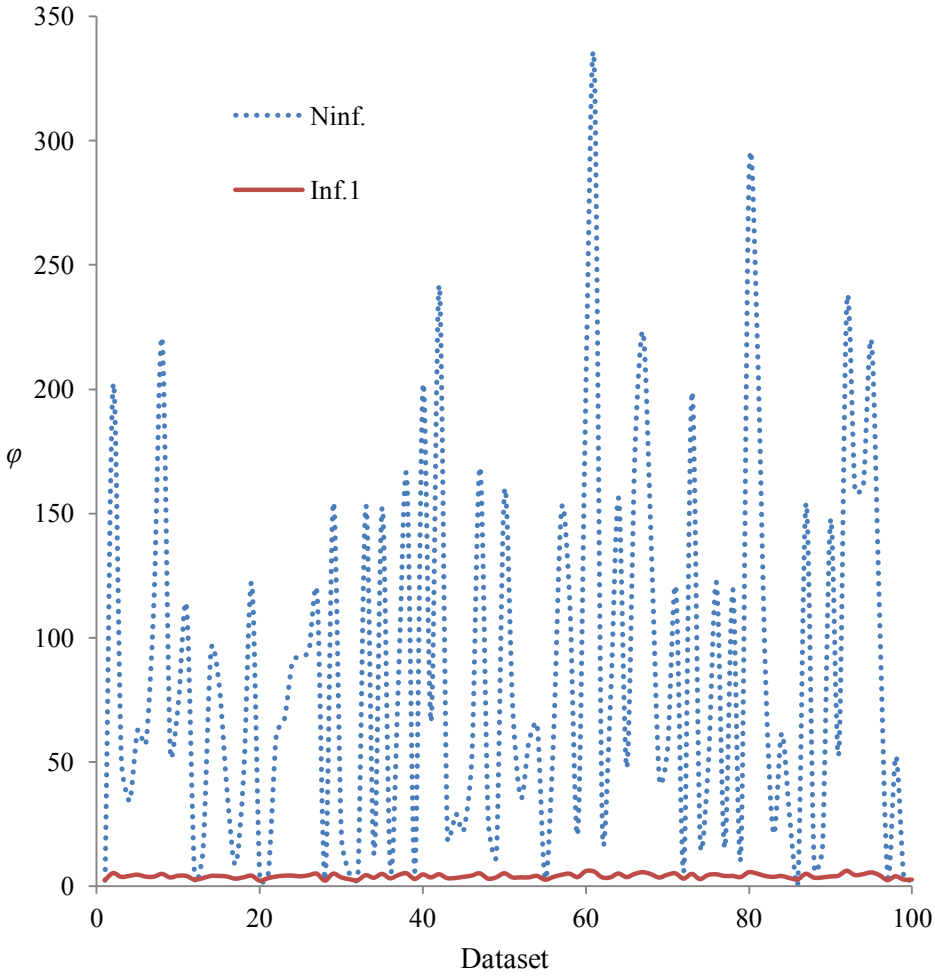

Fig. A.VI Inverse dispersion parameter (φ) - Low mean data, 20 observations.

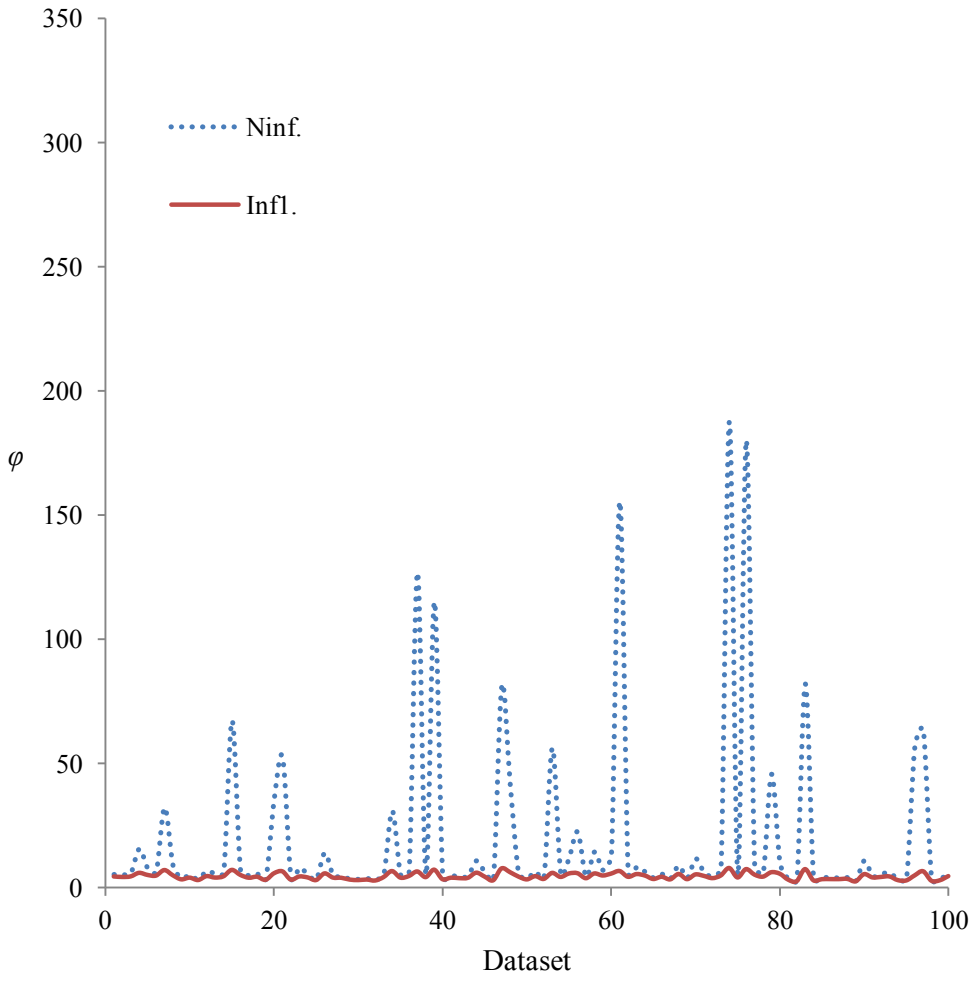

Fig. A.VII Inverse dispersion parameter (φ) - Low mean data, 80 observations.

7) Spearman's correlation coefficients estimated for 100 datasets.

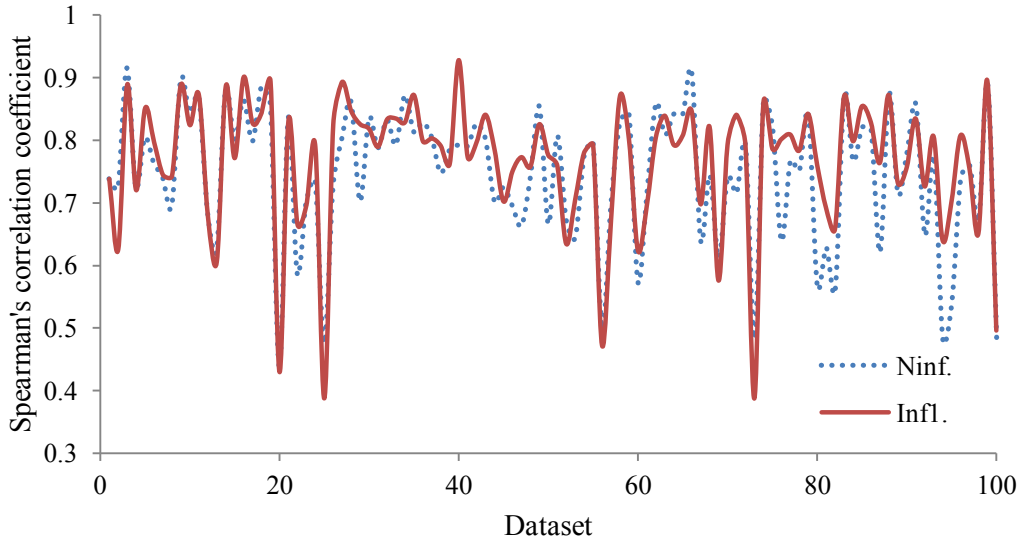

Fig. A.VIII Spearman's correlation coefficient - Low mean data, 20 observations.

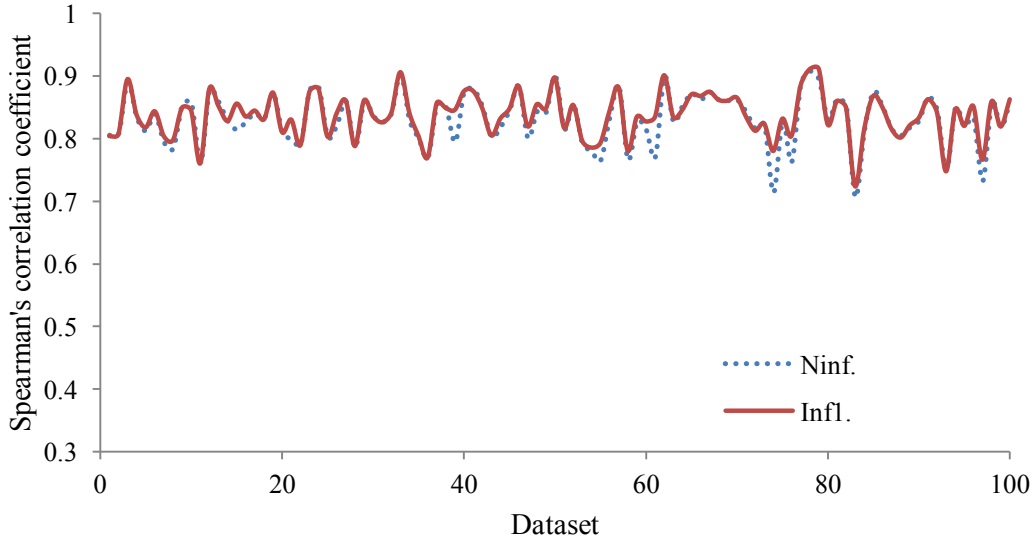

Fig. A.IX Spearman's correlation coefficient - Low mean data, 80 observations.