The Dissertation Committee for Prasad Naga Venkata Siva Rama Buddhavarapu certifies that this is the approved version of the following dissertation:

On Bayesian Estimation of Spatial and Dynamic Count Models Using Data Augmentation Techniques: Application to Road Safety Management

Committee:

Jorge A. Prozzi, Supervisor

James G. Scott

Carlos M. Carvalho

Chandra R. Bhat

Stephen Boyles

Andre F. Smit

On Bayesian Estimation of Spatial and Dynamic Count Models Using Data Augmentation Techniques: Application to Road Safety Management

by

Prasad Naga Venkata Siva Rama Buddhavarapu,
B.Tech.C.E; M.S.E.; M.S.Stat.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2015

To my wife, Lakshmi sravani.

Acknowledgments

I wish to express my sincere gratitude to my supervisor Dr.Jorge A. Prozzi for his constant support and encouragement throughout my time at UT Austin. I would like to particularly thank Dr.Prozzi for allowing me to pursue a research topic of my own interest. He has always been very encouraging and supportive to explore new research ideas, and simultaneously provided valuable inputs and guidance. This dissertation would never have been possible without his consistent support and advising. I would like to take this opportunity to acknowledge Dr.Prozzi's efforts that allowed me to attend several conferences across the world. I am indebted to Dr.Prozzi for the constant financial support. I thank Dr.Prozzi for teaching the way to perform good quality research with potential to change the world.

I would like to thank Dr.Carlos M. Carvalho for introducing me to the exciting field of Bayesian Statistics. I thank Dr.Carlos for initiating the interest to pursue this research during his time series course at Statistics department. I would like to thank Dr.James G. Scott for offering great courses and motivating to perform research in Bayesian inference on count data. I sincerely appreciate his availability to discuss several issues during this research. This thesis would have never been possible without the constant guidance and valuable advice from Dr.Carlos and Dr.Scott.

I would also like to thank Dr.Chandra R.Bhat for his valuable inputs during the proposal and meticulous reviews of this dissertation. In particular, I thank Dr.Bhat

for providing a contrasting econometric perspective to the Bayesian estimation methods developed in this study. I would like to thank Dr.Stephen Boyles for identifying the important issues in scaling the proposed framework to larger road networks. I would also like to thank Dr.Boyles for his valuable advice on potential future extensions that may reshape this research as a valuable tool for road safety management.

Last but not least, I would like to thank Dr.Andre F. Smit for the valuable inputs regarding the empirical analysis and providing insights into the practical application of the proposed statistical methods from a practitioner's perspective. I thank Dr.Smit for providing the necessary data sources and several hours of discussions during data cleaning phase. I would like to take this opportunity to thank Dr.Smit for being my mentor at UT apart from Dr.Prozzi.

I would like to thank the large group of my friends and colleagues who made my stay memorable. I do not intend to list the names to save the space as well as not to forget any individual. I would like to thank my parents for their constant support. Finally, I would like to acknowledge Texas Department of Transportation and Department of Statistics & Data Science for providing financial assistance.

On Bayesian Estimation of Spatial and Dynamic Count Models Using Data Augmentation Techniques: Application to Road Safety Management

Prasad Naga Venkata Siva Rama Buddhavarapu, Ph.D.
The University of Texas at Austin, 2015

Supervisor: Jorge A. Prozzi

Over the past several years, roadway safety management has evolved into data-driven or evidence-based science. The corner stone of a data-driven roadway safety management is the knowledge about useful patterns in the complex crash data. Crash data is often difficult to model with several confounding factors and discrete target variables such as crash counts or crash severity. The major goal of this dissertation was to contribute to the methodological realm of roadway safety management.

The research objectives are in two folds: 1) to develop state-of-the-art model specifications for modeling crash data, and 2) to develop a probabilistic model-based site ranking framework. This research addresses methodological issues in crash frequency modeling such as unobserved heterogeneity, spatial correlation, and temporal patterns. Two novel specifications were developed to address these methodological issues: 1) negative binomial spatial with random parameters (NBSRP) modeled as multi-variate normal finite mixture distribution; 2) negative binomial spatial model with dynamic parameters

(NBSDP). The NBSRP with finite-mixture specification allows for identifying the underlying sub-groups of road segments, and for skewness and multi-modality in the underlying random parameter distribution. The NBSDP specification employs dynamic linear model (DLM) formulation of the discrete negative binomial count model by exploiting recently developed polya-gamma data-augmentation techniques. NBSDP model facilitates to investigate the evolution of the model parameters over the time and to make safety predictions for a future year. Both NBSRP and NBSDP models simultaneously accounts for potential spatial correlation of crash counts from neighboring road segments.

Bayesian methods have been widely used for model building and recently gaining further popularity due to the availability of efficient algorithmic techniques for the parameter estimation. Computationally efficient Bayesian estimation frameworks that leverage recent advances in data augmentation techniques were developed in this research to estimate the proposed count specifications. Bayesian estimation methods also facilitate statistical inference on site ranks, thereby allowing for probabilistic ranking. A computationally efficient site ranking framework was developed incorporating the recent probabilistic ranking techniques towards the end of this dissertation. Overall, this dissertation demonstrates the feasibility of designing Bayesian modeling frameworks for probabilistic roadway safety management, which facilitate online learning. The research ideas presented in this dissertation may be extended to bigger networks to test the feasibility of developing a safety management framework that automatically learns from the latest crash data sources over the time.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

The National Highway Traffic Safety Administration (NHTSA) reported an overall decline of 22.7 percent in the occupant fatality rate (per 100,000 population) from 1975 to 1992, which further decreased by 31.1 percent from 1992 to 2012. Although a substantial improvement has been witnessed in terms of road safety during the last 35 years, about 35,000 fatalities and about 1.7 million injuries were still being reported annually in highway vehicle crashes during 2003-2012 [NHTSA, 2014]. In 2010, NHTSA estimated traffic crashes in the United States accounted for over $ 277 billion in economic losses [Blincoe et al., 2014]; this is about $897 per individual distributed among 308.7 million people living in the United States. To place it in perspective, the cost of traffic crashes is reportedly more than two and one-half times the cost of congestion in urban areas [Herbel et al., 2010]. The alarming crash statistics and the associated economic and social costs call for impending safety countermeasures across the United States.

Over the past several years, roadway safety management has evolved into data-driven or evidence-based rather than experience-based science [Herbel et al., 2010]. In 2005, Highway Safety Improvement Program (HSIP), a federally funded, state-administered program was established. HSIP mandated individual states to implement Strategic High-

way Safety Plans (SHSP) that are data-driven and targeted at reducing crash counts and severity on all highways. Furthermore, the enactment of Moving Ahead for Progress in the 21st century act (MAP-21), a federal law approved in 2012, enforced each state-level transportation agency to actively develop and modify the existing SHSP. According to the HSIP manual [Herbel et al., 2010], the goals of a scientific safety management framework are to understand and quantify the changes in the expected crash consequences; the future safety decisions should be reliant on such quantifiable evidence. HSIP's data-driven strategic approach to improve highway safety emphasizes the need for comprehensive database management systems and state-of-the-art data analysis methodologies to identify and prioritize crash prone zones and to establish performance-based goals for optimal utilization of limited safety budgets. Federal Highway Administration's (FHWA) Office of Safety Research and Development with support from Exploratory Advanced Research (EAR) program is actively gathering information on the recent advances in analyzing massive data and data mining. Such recent data analysis technologies may better explain the causes of crash occurrence and offer innovative means of accident prevention. Analytical techniques that allows to learn from integrated datasets comprising information from disparate and often incompatible sources such as big data are needed.

Crash occurrences are highly stochastic and infrequent in nature, therefore, investing safety funds based on the raw crash number is merely "chasing the lightening". The main goals of a rational safety project selection are to avoid personal judgment, to account for the random nature of road crashes, and to reduce the administrative costs associated with the in-field investigations of the promising sites [Deacon et al., 1974]. Based on a survey on the use of hazardous site selection methodologies, the majority

of the state-level transportation agencies are utilizing deterministic methods [Hallmark et al., 2002]; this was further confirmed with local transportation officials at the beginning of this dissertation research. Although crash numbers are highly stochastic in nature, they do represent the safety level of the road network. Probabilistic comparison is indeed necessary to compare randomized crash numbers of different road segments across a road network. Accurate identification of high risk locations will assist transportation agencies to focus the limited safety funding towards most beneficial safety improvement projects. Model based road safety management also allows for a "systemic" safety perspective, which is recently gaining ample attention across highway agencies.

In order to perform probabilistic site ranking, a safety measure of candidate sites that best represents their existing safety level is necessary. Roadway safety is often measured in terms of either crash frequency or crash severity. Crash frequency is defined as the total number of crashes associated with a candidate site during a time period. Crash severity is defined as the injury severity level of the most severely injured individual involved into a road accident. Crash frequency is generally modeled using discrete probabilistic distributions such as Poisson, Negative binomial distributions. Crash severity is expressed in terms of discrete severity levels, and typically modeled using discrete choice models such as ordered probit/logit and multinomial probit/logit. The probability of crash occurrence at a candidate site is of interest while assessing its crash vulnerability. On the other hand, crash severity modeling estimates the probability of a crash to fall under a given severity category conditional on the crash occurrence. Crash frequency modeling is more suitable for site screening applications; crash severity measure becomes relevant while performing individual crash level analysis. Crash rate is another potential

candidate as a safety measure; however, it is sensitive to low traffic volumes and may result in unrealistic safety levels. For instance, an additional crash on a site with very low volume traffic may drastically increase its crash rate, thereby the chances of selecting the site for safety improvement will be high. However, the absolute number of lives saved by investing on the site may be low due to the low exposure, and the resources may be utilized elsewhere.

The safety level of roadway infrastructure is governed by several intertwined factors: geometric, functional, structural and socio-economic characteristics of the candidate site. Therefore, the development of statistical models that best fits the observed crash data, yet producing reliable crash predictions, is a challenging task. The reliability of any predictive model relies largely on the underlying model specification of the dependent variable. In this dissertation crash frequency (or crash count) was selected as the response variable. Mannering and Bhat [2014] presented a historical overview of the evolution of statistical specifications in the roadway safety research and identified several methodological barriers. The authors highlighted that the adoption of new methodologies is indeed essential in the field of roadway safety research to address several statistical issues including unobserved heterogeneity, spatial and temporal correlation that may potentially impact the precision of resulting crash predictions, thereby affect site ranking and budget allocation.

Bayesian predictive methods have been widely used for the analysis of road safety data and recently gaining further popularity due to the availability of efficient algorithmic techniques for the parameter estimation. Bayesian methods allow to conveniently combine prior knowledge on model parameters (if any), and update the prior distributions

of the model parameters by incorporating the observed crash data; this is often desired in road safety management. Bayesian modeling framework facilitates to periodically update the predictive model parameter distributions upon the arrival of the newer data. Such framework is useful in infrastructure management applications as data are collected annually. Bayesian inference also provides full posterior distribution of any function of the model parameters. For instance, Bayesian count specifications facilitate inferences on individual site ranks, which facilitate probabilistic site ranking of road segments.

In summary, the need to incorporate state-of-the-art statistical techniques into the road safety management is evident. The growing roadway safety data resources are of no additional value, unless statistical tools that facilitate learning useful roadway safety information are available. This dissertation aims to contribute to the methodological realm of road safety management. This research emphasizes the need to incorporate advanced model structures into roadway safety management applications to unveil interesting patterns in the crash datasets and to pro-actively prevent accident occurrence. The main objectives and anticipated contributions of this study are provided in the next section.

1.2 Main objectives

The main goal of this research is to incorporate advanced statistical models into roadway safety management applications. To achieve this goal, the study is divided into the following major objectives:

1. To build statistical models for unveiling interesting patterns in the safety data

2. To perform model-based probabilistic ranking of the road segments

This dissertation proposes new statistical models for analyzing historic crash frequency data of road networks. Methodological issues in crash frequency modeling including unobserved heterogeneity, spatial correlation, and temporal pattern extraction are addressed in this research. The proposed crash frequency modeling is divided into the following two scenarios.

Scenario - I: In this scenario, a negative binomial spatial with random parameters (NBSRP) model is progressively developed for analyzing crash data over a given year. First, a negative binomial specification with random parameters is constructed to model crash frequency of contiguous road segments. The unobserved heterogeneity is incorporated using a finite multi-variate normal mixture prior on the random parameters; this allows for non-normality, skewness in the distribution of the random model parameters, facilitates correlation across the model parameters, and relaxes any distributional assumptions. The model extracts the inherent groups of road segments with crash counts that are equally sensitive to the road attributes on an average; the heterogeneity within these groups is also allowed in the proposed framework. The NBSRP model simultaneously accounts for potential spatial correlation of the crash counts from neighboring road segments. The proposed NBSRP is potentially useful in the context of road safety management to identify the road segments that may respond to similar safety treatments and also the strength of the spatial dependence of crash frequencies on contiguous road segments.

Scenario - II: In this scenario, a negative binomial spatial model with dynamic parameters (NBSDP) is developed for analyzing the historical crash data collected over several years. NBSDP model structure allows the regression parameters to vary over the

time using a dynamic linear modeling framework. The proposed model simultaneously allows for spatial correlation of the crash counts on the neighboring road segments, which may vary each year. NBSDP model structure is useful to study the evolution of the negative binomial regression parameters over the time. In addition, NBSDP model also allows to estimate the change in the level of spatial dependence over the time across the road segments.

This research employs a computationally efficient Bayesian estimation framework to perform statistical inference on the proposed models. A Marcov Chain Monte Carlo (MCMC) sampling strategy is proposed that leverages recent theoretical developments on data-augmentation algorithms, and elegantly sidesteps many of the computational difficulties usually associated with Bayesian inference of count models. A probabilistic ranking methodology is also developed towards the end of the dissertation.

In summary, this dissertation research assembles a framework to identify road segments with potential for safety improvement in order to efficiently allocate road safety management funding. This research also documents several empirical findings regarding the influence of road condition and geometric features on crash frequency.

1.3 Dissertation outline

The dissertation is organized into six chapters as follows. Chapter 2 summarizes a comprehensive literature review of most recent methodological advancements in crash count modeling. A discussion on modeling specifications and a variety of estimation techniques, and a general review of road safety management and probabilistic ranking applications are also provided. Chapter 3 describes the NBDRP specification and Gibbs

sampling algorithm for model estimation. A discussion on empirical analysis is also provided towards the end of the chapter. Chapter 4 introduces dynamic linear modeling framework, describes the NBSDP specification and Forward Filtering Backward Sampling (FFBS) algorithm for model estimation. The application of proposed specification in road safety management is further demonstrated using an empirical example. Chapter 5 describes a probabilistic ranking framework for the identification of hazardous road segments. An empirical demonstration of the proposed site ranking framework is also provided using a road network from the Houston area. Chapter 6 summarizes the major contributions of this dissertation research, and points a few potential directions for future research.

Chapter 2

Literature review

This chapter summarizes a comprehensive review of the existing literature on road safety management and crash modeling. A brief historical overview of road safety management is provided at the beginning of the section. Subsequently, a review of the recent probabilistic ranking methods to identify the road segments with potential for safety improvement is provided. Earlier studies that addressed methodological issues in crash frequency modeling such as unobserved heterogeneity and temporal patterns are also discussed.

2.1 Overview of roadway safety management

In 1966, United States Congress enacted Highway Safety Act, a major Federal initiative towards improving roadway safety, that required individual states to establish and monitor a highway safety program in conformity with uniform standards constituted by the Secretary of Transportation. FHWA and NHTSA shared the responsibility of implementing 18 essential standards that were established under the act. The 1966 Highway Safety Act was modified further in 1973 to replace the established standards with five priority safety improvement program areas. In 1978, the Surface Transportation Assistance Act coalesced the five different areas into the Railway-Highway Grade Crossing

and Hazard Elimination Programs. The hazard elimination program was primarily targeted at reducing the frequency of fatalities and serious injuries caused by road crashes on all public roads. The program provided funding for implementing projects to allay or eradicate the hazardous public road segments. Subsequently, individual states required to develop a Safety Management System (SMS) under Intermodal Surface Transportation Efficiency Act (ISTEA) of 1991. SMS promoted the culture of maintaining crash database upon which safety decisions and performance measures may be established. Recently, the Highway Safety Improvement Program (HSIP), fueled by the enactment of roadway safety management has been further driving the safety management into a data-driven or evidence-based science[Herbel et al., 2010].

Any roadway safety management system aims to reduce the frequency and severity of road crashes, however, under unavoidable budgetary constraints. The first and vital stage in State-level safety management plans is to identify road segments or intersections with highest crash potential, or in other words, with most promising safety improvement across the road network. Generally, the proposed safety improvement projects are largely funded by the federal government through HSIP program (up to 90% funding) and the local or State authorities are required to fund the remaining portion. States bear the responsibility of effectively utilizing the massive portion of federal funds to achieve desired safety improvements. Clearly, local transportation agencies require superior analytical tools to identify the suitable safety projects across the road network in order to cost-effectively utilize the federal funds. For instance, in Texas, each district is responsible for identifying sites with safety concerns and for sending the information to Traffic Operations Division office. The office ranks the proposed safety projects using the Safety

Improvement Index (SII), described in TxDOT's HSIP manual [HSIP, 2013], and funds in the order until the allocated funding is depleted. Prior to these cost calculations, a reasonable district level crash ranking tool may assist districts to perform an initial screening exercise to identify the road segments with promising safety improvements.

2.2 Site ranking

The main objectives of the identification of hazardous sections are to avoid personal judgment, to consider the random nature of the crashes into account and to reduce the administrative costs associated with the in-field investigations of the promising sites [Deacon et al., 1974]. Broadly, the site ranking methods may be categorized into deterministic and probabilistic procedures. Hallmark et al. [2002] reported a considerable diversity in the selection of site ranking methods across individual states in US. Based on a survey on the use of site ranking methodologies, it was determined that the majority of the state-level transportation agencies are utilizing deterministic site ranking methods that do not account for the inherent variability of the site ranking exercise [Hallmark et al., 2002]. A probabilistic comparison is indeed necessary to compare the randomized crash numbers of different road segments across a road network. Probabilistic safety management tools assists the transportation agencies to implement reactive as well as proactive safety management. The road segments exhibiting consistent high crash risk may be considered for immediate safety treatments, while the road segments with moderate crash risk may be treated for proactive safety management. Lack of accurate probabilistic network screening tools may lead to ineffective safety investments.

Among the probabilistic methods, the Empirical Bayes (EB) method has been

extensively utilized for site ranking and network screening applications. Recently, hierarchical Bayesian modeling of the crash count data has been utilized extensively in the road safety literature. However, the site ranking methodologies that exploit the power of Fully Bayesian (FB) methods are limited. Given the recent advances, a clear guidance on the selection of appropriate Hierarchical Bayesian probabilistic ranking methods is very helpful to the road safety management professionals and transportation agencies. This dissertation compares several previously proposed probabilistic ranking methods, and provides the merits and demerits. A review of the most recent probabilistic ranking methods that have been implemented or proposed in road safety management applications is provided in this section.

2.2.1 Probabilistic ranking methods

Deacon et al. [1974] refers to hazardous locations as the sites with abnormally severe accident patterns when compared with similar locations elsewhere. Another recent definition by Elvik [2008] is as follows: "*any location that has a higher expected number of accidents than other similar location as a result of local risk factors*". Hauer [1996] introduced a relatively neutral phrase: Sites with Promise (SWiPs), rather than calling the identified sites as black spots or hazardous locations. The majority of the literature on the network screening for SWiPs recommended ranking methods that are reliant on long run crash frequencies. Generally, earlier site ranking methodologies were not probabilistic in nature and did not account for precision of the estimate of underlying decision index (such as predicted crash frequency). On the other hand, probabilistic ranking method provides a framework to construct a list of SWiPs while accounting for the uncertainty

around the decision index.

A wide variety of techniques have been used to screen the sites with potential for improvement. For example, Norden et al. [1956] and Morin [1967] suggested industrial statistical quality control for road safety, which includes an upper bound for accident rate using an assumed level of false detection. Tamburri and Smith [1970] introduced the notion of safety index for site ranking in order to make sensible comparisons. The index was calculated using economic weighting of the expected crash counts in each severity category. McGuigan [1981, 1982] suggested to calculate the difference between the actual number of accidents on road sections and the number of accidents expected for such class of roads under similar traffic volumes. Hauer [1996] provided an overview of the historical and conceptual development of site ranking or identification of black spots. Hauer [1996] emphasized on selecting the underlying decision index (such as observed crash rate, observed frequency, etc.) with respect to the circumstance or target improvements in order to maximize cost effectiveness. The author also identified three different motives behind the site ranking exercise: 1) economic efficiency 2) professional and institutional responsibility and 3) fairness to the road user.

Although the problem of identifying hazardous locations has been widely discussed in the literature, the interest in Bayesian methods to improve the process only originated in the eighties [Hauer, 1996]. Ever since, Empirical Bayes (EB) method has been extensively utilized for site ranking and network screening. To smooth out the random fluctuations in crash count data, the EB method specifies the safety of a site as an estimate of its long-term mean instead of short-term crash count. Persaud et al. [1999], Hauer et al. [2002], Higle and Witkowski. [1988], Hauer [1992], etc. utilized the EB method for

network screening applications. Several safety management software applications (such as SafetyAnalyst and Interactive Highway Safety design Model (IHSDM) software) and the Highway Safety Manual promote the use of EB method to identify SWiPs. Instead of directly utilizing the EB estimate for site ranking,Persaud et al. [1999] proposed the concept of potential for safety improvement. The potential for safety improvement is calculated as the difference between the expected crash counts based on EB estimate using regression models incorporating variables that may contribute to unsafety and expected crash counts based on a model that includes only traffic volume but no treatable variables. Another refinement of potential for safety improvement is the difference between the expected crash counts based on the EB estimate using regression models incorporating variables that may contribute to unsafety and the expected crash counts based on a model that includes typical treatment variables or expected crash count for a base condition. Elvik [2008] compared five different ranking criteria for identifying hazardous locations: upper tail accident count, upper tail accident rate, upper tail accident count and high accident rate, upper tail expected number of accidents (EB estimate), and upper tail EB dispersion criterion. Four years of crash data were utilized to perform the site ranking exercise in order to identify the hazardous locations. Subsequently, the percentage of false positives and false negatives was evaluated using crash data from the subsequent four years to repeat the site ranking exercise. Elvik [2008] concluded that EB estimates provide more reliable identification of the hazardous road locations than traditionally used criteria including accident rates and numbers.

Bayesian hierarchical modeling has been extensively utilized in crash modeling but relatively smaller pool of studies utilized the potential of Fully Bayesian (FB) methods

in the site ranking applications. A study by Miaou and Song [2005] is one of the notable Fully Bayesian based site ranking applications. The study utilized a Poisson-Gamma hierarchical specification that accounts for spatial correlation to identify the vulnerable road segments or intersections. Inferences and site rankings were obtained through computer programs coded in WinBUGS language[1]. Hauer [1996], Bell [1986] also identified the need to identify, if any, underlying clustering of the spatial arrangement of accidents on a road. Mitra [2009] developed a methodology to identify the hot-spots using Bayesian hierarchical modeling and GIS based techniques. The authors concluded that local Moran's spatial autocorrelation method appears to to be a quite satisfactory method for identifying statistically significant crash clusters. Geurts et al. [2006] utilized multi-variate Poisson models to smooth out the randomness of the vectorized crash counts (multiple severity categories); subsequently they combined the crash count vectors into single number (called expected score) using a weighting function. The expected scores of individual sites is further processed to obtain posterior density of the ranks of respective sites using MCMC methods. Brijs et al. [2007] also developed multivariate crash counts and utilized a cost function to combine the vector of crash counts into expected crash cost. The crash costs were calculated in each MCMC iterations followed by the assignment of individual ranks. Subsequently, posterior distributions of individual site ranks were constructed and utilized for probabilistic ranking. According to Mahalel et al. [1982], a road section is identified as black spot if, the probability of estimated crash count (based on a multi-variate model) of a site exceeding the observed number

[1]WinBUGS obtains posterior distribution of the parameters using Metropolis-Hastings algorithm in a typical Markov Chain Monte Carlo simulation framework.

of crashes (at the same site) is smaller than a threshold level (such as 0.05). Davis and Yang [2001] utilized Gibbs sampling and estimated a Hierarchical model to identify the hazardous intersections where the crash risk for a specific driver group is larger than the other groups. Miranda-Moreno et al. [2007] introduced two different Bayesian multiple testing procedures for selecting a list of sites for further engineering inspections with a target error rate. The study emphasizes the importance of incorporating the uncertainty in the model parameters and safety measures, while minimizing the false discovery rate (FDR), in the site selection process.

To implement any site ranking methodology, a robust and accurate crash prediction model is essential. The following section presents a comprehensive literature review discussing several important issues in building reliable crash prediction models.

2.3 Crash frequency modeling

Poisson regression is the most widely used, elementary specification for modeling non-negative discrete crash count data [Cameron and Trivedi, 1998]. However, the equidispersion restriction (equality of mean and variance) disqualifies Poisson regression model to model over-dispersed crash count data. Negative Binomial (NB) models are among the most common probabilistic models utilized by road safety analysts as they account for over-dispersion of the crash count data. The simplicity of the mathematical structure and interpret-ability of the model parameters promoted the use of NB models in road safety [Hauer, 1997]. Mannering and Bhat [2014] presented a historical overview of the evolution of crash count modeling in the roadway safety research and identified several methodological barriers. They highlighted that the adoption of new methodologies

is indeed essential to address several statistical issues such as unobserved heterogeneity, spatial and temporal correlation[Mannering and Bhat, 2014]. A vast body of research studies proposed sophisticated NB likelihood based count models that account for statistical issues such as unobserved heterogeneity, selectivity bias, and spatial and temporal correlations. A comprehensive literature review on incorporating such statistical issues into crash frequency modeling is provided below.

2.3.1 Unobserved heterogeneity

Road segments with identical site-specific attributes often exhibit significantly different crash outcomes. The influence of the site-specific attributes may vary across the road segments due to unobserved reasons; this is termed as unobserved heterogeneity. A wide variety of model specifications have been proposed in the literature to account for unobserved heterogeneity. Unobserved heterogeneity is typically modeled by using fixed and random effects (random parameters) in the econometric literature. Model specifications may be categorized based on the distributional assumption on the random parameters. The following three distributions are generally used in the literature to model random parameters:

- Continuous distribution
- Finite mixture distribution
- Finite mixture of continuous distributions

Continuous distribution:

The parameters of the explanatory variables may be assumed to be randomly generated

according to an underlying continuous probability distribution. In other words, the sensitivity of the outcome to the individual attributes is not identical across the observations; thereby, it introduces taste variation. For instance, Anastasopoulos and Mannering [2009] utilized negative binomial specifications with random regression parameters for modeling the crash frequencies. The regression parameters were assumed to be independent and Gaussian distributed. Wu et al. [2013], Chin and Quddus [2003], Ukkusuri et al. [2011] also employed similar random parameter negative binomial models to capture unobserved heterogeneity while modeling crash count data spanning across multiple years.

Finite mixture distribution:

A finite mixture distribution assumes the presence of latent groups of subjects that respond similarly to the explanatory variables within a given population [Xiong and Mannering, 2013]. The finite mixture models are semi-parametric, thereby does not require any distributional assumptions for the mixing variable [Deb and Trivedi, 1997]. Deb and Trivedi [1997] describe the finite mixtures as natural representation of the underlying heterogeneity in terms of latent classes, which may be interpreted as inherent "types". Several studies in road safety, marketing, health care etc. employed finite mixture count models. For example, Wedel et al. [1993] proposed a Poisson finite mixture model formulation to model the number of purchases in a marketing research context. The mean parameter of the Poisson distribution varies across a set of finite number of classes, which is modeled using discrete finite mixture distribution. Park and Lord [2009] investigated the possibility of employing finite mixture negative binomial models to capture the unobserved heterogeneity of the crash count data. The proposed models assume that the crash data are generated from a population comprising of several distinct negative binomially

distributed latent groups with distinct parameters. Individual observations belongs to either of these mixture components (or latent groups) with a certain probability, which is essentially the respective mixture weight. The probability of a given subject belonging to a particular latent group may vary across the subjects. Zou et al. [2013] extended the previously proposed (by Park and Lord [2009]) finite mixture negative binomial model by incorporating attribute dependent mixture weights. The authors showed that the finite mixture negative binomial models with varying mixing weights are superior than the models with fixed weights. Zou et al. [2014] further investigated several functional forms to model the mixing weights using site attributes in a two-component finite mixture negative binomial model; they recommended modeling the mixing weights as a function of length raised to a power. Deb and Trivedi [1997]modeled counts of medical care utilization using a finite mixture negative binomial models for a health care research application.

In the context of modeling consumer heterogeneity, Otter et al. [2004] compared the random parameter models and the latent class or finite mixture models using simulated datasets. Random parameter models tend to outperform over the latent class models, if the underlying parameter distribution is strictly continuous. On the other hand, if the underlying distribution is discrete, latent class models tend to outperform over the random parameter models with adequate informative sample size.

Finite mixture of continuous distributions:

The finite mixture models may be extended to incorporate across group heterogeneity by allowing the model parameters to vary within the sub-groups. For instance, a recent study by Xiong and Mannering [2013] incorporated unobserved heterogeneity into crash

severity modeling using random parameters that are distributed as a mixture of multivariate Gaussian kernels. The proposed mixture modeling approach allows for skewness, multimodality, and heavy-tails in the random regression parameter distributions. The model assumes the presence of sub-populations that differ from the global population in terms of the influence of explanatory variables on the model outcome. The mixture of multi-variate Gaussian kernels facilitates modeling component-specific unobserved heterogeneity, while simultaneously allowing for individual-level unobserved heterogeneity within each component. The empirical analysis demonstrated the presence of two mixture components, which represents two distinct driving environments that affect the crash injury severity of the adolescent drivers. Another study by Xiong et al. [2014] employed Markov Switching Random Parameters Ordered Probit (MSRPOP) models for modeling panel crash-injury severity data. The proposed model structure simultaneously handles the unobserved time-varying and time-constant heterogeneity effects on latent crash-injury severity propensity within the underlying ordered probit model structure. The heterogeneity arising within road segments for any given time period is modeled using the random ordered probit regression parameters. The random regression parameters were assumed to be normally distributed at any given time period. The study hypothesizes that road segments switch between two latent safety states according to a first order, but not necessarily stationary, Markov switching process. The earlier studies have utilized the random parameter specifications with finite-mixture distributions while accounting for within-group variation in the case of crash injury severity modeling. Such random parameter specifications are rarely proposed in the road safety literature for modeling crash frequency. In this research, a random parameters negative binomial

count model with finite mixture multivariate normal structure on the random parameters is proposed.

2.3.2 Spatial correlation

Relatively close road segments arguably possess common unobserved features, thereby inducing a correlation between the crash counts of road segments within a neighborhood —spatial correlation. Incorporating spatial correlation among the adjacent road entities significantly improves the model prediction accuracy [Quddus, 2008]. Spatial correlation may be incorporated using a spatially correlated mixing variable. Spatial error correlation specifications allow for the dependence of the outcome variable at a given location on the unobserved attributes of the neighboring spatial units. A vast body of literature employed Gaussian Conditional Autoregressive (CAR) specification for modeling spatially correlated random effects [Miaou and Song, 2005, Aguero-Valverde and Jovanis, 2008, Ahmed et al., 2011, Yu et al., 2013, Noland et al., 2013, Wang and Kockelman, 2013, Zeng and Huang, 2014]. CAR specification involves a weight matrix that controls the extent of spatial dependence; distance-based and neighborhood-based weight matrices are generally used. For example, Noland et al. [2013], Miaou and Song [2005] incorporated the spatial correlation using Conditionally Autoregressive (CAR) model with a spatial weight matrix constructed using a negative exponential decay function based on the distance between block centroids or spatial units. Aguero-Valverde and Jovanis [2008] utilized a neighborhood-based weight matrix to construct CAR specification; the neighbors are defined based on the adjacency of the spatial units. Count models involving CAR specifications are typically estimated through Bayesian inference using MCMC

methods. Although CAR specification induce spatial correlation through spatially correlated random effects, it does not account for the dependence of the outcome variable on the observed attributes of the neighboring spatial units (also referred as spillover effects). CAR models facilitate construction of the convenient Gibbs sampling for posterior inference as described later in Chapter 3.

Another way to induce spatial dependency among the neighboring road segments is through spatial lag based specifications. In addition to random spatial effects, spatial lag specifications also allow for the dependence of the outcome variable on the observed attributes of the neighboring spatial units (spill-over effects). For example, Narayanamoorthy et al. [2013] employed such spatial lag specification for developing a multivariate count model. The model was estimated using composite maximum likelihood method by recasting the count model as a special case of generalized ordered response (GOR) model (see Castro et al. [2012] for details on recasting).

2.3.3 Temporal patterns

Crash count data generated over a time period is arguably simultaneously correlated across time. Moreover, the attributes of the road segments such as road condition, traffic volumes, etc. change with time. The influence of such time-varying attributes may arguably change over the time. The relationship between the time-invariant attributes may also evolve over the time. Dynamic models allow for such temporal variation through time-varying regression parameters. Dynamic linear models are popularly used to model temporal variation of parameter states in time series applications. Dynamic linear models assume a linear Gaussian evolution of the underlying parameter states (or values)

over the time. The estimation of such linear evolution equations becomes difficult in the case of count models with negative binomial likelihoods. A few studies have previously investigated the use of dynamic linear models for modeling parameters of negative binomial likelihoods. For example, Hu et al. [2013] developed dynamic time-series negative binomial regression models to understand the temporal patterns in highway crash counts for senior and non-senior drivers. The computational difficulties in constructing posterior distributions of the model parameters were avoided by employing Integrated Nested Laplace Approximation (INLA). The study is one among few studies that attempted to address the computational difficulties associated with Bayesian inference of count data models.

Malyshkina et al. [2009] employed two-state Markov switching models to study the time-varying accident frequencies, which assume that the actual road safety switches between two unobserved safety states over the time. In each latent state, a different NB data-generating process is assumed. State transitions are modeled using a stationary two-state Markov chain process in time, which is specified through time-independent transition probabilities. The two-state Markov switching NB model was estimated using Bayesian inference through a hybrid MCMC algorithm. Miaou and Song [2005] accounted for temporal correlation across annual crash counts using temporal random effects with AR(1) prior structures.

2.3.4 Estimation issues

Several classical and Bayesian estimation techniques have been utilized in the literature to estimate random parameter and finite mixture models. Simulation-based

Maximum Likelihood Estimation (MLE) with Halton draws [2] have been widely used for estimating the parameters of the aforementioned random parameter models [Anastasopoulos and Mannering, 2009, Ukkusuri et al., 2011, Anastasopoulos et al., 2012, Wu et al., 2013]. As an alternative to Maximum Simulated Likelihood (MSL) based estimation, Bhat [2011] proposed Maximum Approximate Composite Marginal Likelihood (MACML) to circumvent the computational difficulties associated with MSL. Subsequently, Bhat and Sidharthan [2012] employed the MACML approach to estimate a panel multinomial probit model with skew-normally distributed random parameters. In the context of mixture models, Wedel et al. [1993] utilized an Expectation-Maximization (EM) algorithm to estimate the Poisson finite mixture model structure. Park and Lord [2009] mentioned that the traditional EM algorithm may not be very suitable to estimate mixture count models as it requires many different starting values for finding global maximum. In case number of mixture components are unknown and to be estimated, Bayesian method is the only sensible way for mixture model estimation [Richardson and Green, 1997]. Park and Lord [2009] mentioned that full conditional posterior distributions of the model parameters (both dispersion and regression parameters) of the NB mixture model do not belong to any standard distributional family. Therefore, the study utilized random walk Metropolis algorithm with a normal proposal density, and a data augmentation step with a latent component membership indicator variable for performing posterior inference; the Metropolis acceptance rates were reported to be 25% to 45%. To estimate a finite mixture random parameter binary probit model, Xiong and Mannering [2013] employed a data augmentation step, proposed by Albert and Chib [1993], prior to

[2] see Bhat [2003] for details on the Halton draws

the MCMC implementation to improve the computational efficiency while keeping the parameter inferences unchanged. The augmented posterior conditionals offer tractable full conditional distributions for Gibbs sampling.

In this research, two recently proposed data augmentation techniques are employed to derive the full conditional distributions of the model parameters. First, a Polya-Gamma distribution based technique is utilized to transform the discrete negative binomial likelihood into a conditionally Gaussian likelihood; this allows to construct a conditionally Gaussian posterior distribution for the vector of regression parameters. Second, a compound Poisson representation of the negative binomial likelihood is utilized to derive a closed-form update equations for the dispersion parameter. The proposed data-augmentation techniques thus allow to construct a Gibbs sampler with closed-form update equations using conjugate non-informative prior distributions. Windle et al. [2013] explored the feasibility of employing such Polya-Gamma data augmentation techniques for posterior inference on logistic likelihoods involving dynamic parameters.

2.4 Influence of road features

A few earlier studies investigated the influence of road features such as pavement condition, geometric, and traffic attributes on the respective crash counts. International Roughness Index (IRI), surface rutting, median barrier presence, interior shoulder width, horizontal curve's degree of curvature, and AADT were generally reported as the significant predictors in crash frequency and crash rate models [Anastasopoulos et al., 2012, Shively et al., 2010, Anastasopoulos and Mannering, 2009, Ihs et al., 2002]. Anastasopoulos and Mannering [2009] emphasized the need for incorporating unobserved heterogene-

ity as the effect of these variables varied across the road segments. They reported that lower crash frequencies were associated with the road segments carrying lower traffic volume, while a minor portion of such road segment witnessed higher crashes. Shively et al. [2010] reported a non-linear increasing relationship between AADT and the expected crash counts. Presence of median barrier was reportedly associated with reduced number of crashes, while wider shoulders were generally associated with larger number of accidents [Anastasopoulos and Mannering, 2009, Shively et al., 2010]. Road surfaces with inferior ride quality (high IRI) values were reported to be generally associated with higher crash frequency and higher crash rates [Anastasopoulos and Mannering, 2009, Anastasopoulos et al., 2012, Ihs et al., 2002]. Surface rutting was reported to be positively associated with crash rates [Anastasopoulos et al., 2012]; on the other hand, the road segments with significant rutting were associated with lower crash frequencies [Anastasopoulos and Mannering, 2009]. Anastasopoulos et al. [2012] reported a negative correlation between crash counts and overall road condition.

Chapter 3

Modeling unobserved heterogeneity in count data

In this chapter, a negative binomial spatial random parameters (NBSRP) count model is proposed to simultaneously address three important characteristics of road crash count data:over-dispersion, spatial correlation, and unobserved heterogeneity. A hierarchical specification that accommodates these issues is progressively developed as follows. First, the Poisson specification is modified by introducing Gamma-distributed random effects thereby specifying a Poisson-Gamma mixture likelihood, which is re-parametrized as a negative binomial likelihood. Subsequently, a finite mixture multi-variate normal distributed random parameters are specified to accommodate the unobserved heterogeneity. Spatial dependence is simultaneously incorporated using spatially correlated random effects that are generated through Intrinsic Conditional Autoregressive (ICAR) prior.

3.1 Model development

Let y_{it} denote the observed uni-variate crash count on i^{th} road segment ($i \in \{1, 2, ...n\}$) during t^{th} year ($t \in \{1, 2, ...T\}$); n represents the total number of road segments and T represents the number of years. We define $X_{it} = [x_{it1}, x_{it2}, ..., x_{itk}]$ as the $k \times 1$ vector of time-varying attributes corresponding to the crash count observation y_{it}. Time-invariant attributes are also denoted using the same notation, although the value is

constant across the time for the sake of notation. First, we assume that the crash counts are Poisson distributed and fully characterized by the mean parameter (λ_{it}). Incorporating a random effect into the crash rate parameter removes the underlying equidispersion restriction (equality of mean and variance) imposed by Poisson regression model and induces over-dispersion. Assuming strictly positive Gamma distributed random effects (ϵ_{it}) produces a Poisson-Gamma mixture model, which turns out to be negative binomial model upon marginalizing the random effects. Let p_{it} and r be the probability and dispersion parameters of the negative binomial likelihood corresponding to y_{it}. We model the probability parameter p_{it} as a function of road segment specific attributes (X_{it}). The site-specific attributes may be divided into two groups based on their effect on the respective crash count. Let $X_{it}^F = [x_{it1}^F, x_{it2}^F, ..., x_{itp}^F]$ denote the $p \times 1$ attribute vector with fixed coefficients $\gamma = [\gamma_1, \gamma_2, ...\gamma_p]$ respectively. And, $X_{it}^R = [x_{it1}^R, x_{it2}^R, ..., x_{itq}^R]$ denote the $q \times 1$ attribute vector with random coefficients $\beta_i = [\beta_{i1}, \beta_{i2}, ...\beta_{iq}]$ respectively. The total number of attributes, $p+q$ is equal to k. Note that the random parameters are estimated for each i^{th} road segment using T data records from each year. In this specification, we assume independence of crash counts across years. To justify the absence of significant temporal trends, we used only four years of crash data. Additionally, a spatially correlated random effect (ϕ_i) is also included while modeling the probability parameter. A detailed description of the distributional assumptions on both random parameters and spatial random effects is described later. The proposed distribution of crash counts on the contiguous road segments is shown in Equation 5.1.

$$y_{it} \sim NB(r, p_{it}) \tag{3.1}$$

$$p_{it} = \frac{e^{\psi_{it}}}{1 + e^{\psi_{it}}}; \qquad \psi_{it} = X_{it}^{F'} \gamma + X_{it}^{R'} \beta_i + \phi_i$$

3.1.1 Finite mixture multivariate normal distribution

The vector of random parameters β_i corresponding to an observation i is hierarchically modeled as a finite mixture of multivariate normal distributions as shown in Equation 3.2. Let μ_c and Σ_c be the component specific mean vector and covariance matrix corresponding to c^{th} component; $c \in \{1, 2, ...C\}$, where C is the total number of mixture components. In this dissertation, we explored specifications with C values ranging from 2 to 4. η_c denote the mixture weight corresponding to c^{th} component.

$$\beta_i \sim \sum_{c=1}^{C} \eta_c N(\mu_c, \Sigma_c); s.t. \sum_{c=1}^{C} \eta_c = 1 \tag{3.2}$$

where, $\mu_c = [\mu_{c1}, \mu_{c2}, ...\mu_{cq}]$ is component mean of the random parameters corresponding to q attributes (X^R) with random parameters; Σ_c is the component covariance matrix $(q \times q)$ of the random parameters.

Modeling the random parameters using the proposed finite mixture of multivariate normal distributions circumvents the need to make distributional assumptions on the random parameters. The model assumes the presence of C sub-groups of road segments each with identical distributions on the random parameters; it is potentially useful to identify latent "types" of road segments. Each road segment belongs to either of these sub-groups with a probability η_c. Identifying the presence of latent sub-groups is potentially useful in the context of road safety management to discover the road segments that may respond to similar safety treatments. In addition to the heterogeneity arising due to the such grouping, the road segments may respond differently to the identical attributes within each sub-group. The vector of random parameters corresponding to road segments within each sub-groups are assumed to be jointly generated by respective multivariate

normal distributions to allow the within group heterogeneity. By specifying a joint distribution, the proposed model allows for correlation across different random parameters (off-diagonal elements are not restricted to zero). Thus, the proposed model structure exploits the advantages of both finite mixture models and the random parameter models for modeling unobserved heterogeneity.

3.1.2 Intrinsic conditional autoregressive priors

Relatively closer road segments may be correlated due to common unobserved reasons —spatial correlation. Spatial correlation may be incorporated using spatially correlated random effects as shown in 5.1. Spatial random effects (ϕ_i) are typically generated form normal distribution with real support ($\phi_i \in (-\infty, \infty)$). Conditionally Auto Regressive (CAR) priors, originally proposed by Besag [1974] are increasingly being used in the context of hierarchical spatial models to generate spatially correlated random effects[Banerjee et al., 2004]. Employing CAR priors facilitate relatively easier and computationally efficient implementation of the Gibbs sampling, particularly with Gaussian likelihoods. The Polya-Gamma data augmentation scheme transforms negative binomial likelihood into conditionally Gaussian likelihoods thereby eases the incorporation of CAR priors (further described in model estimation section). A Gaussian or autonormal CAR prior (shown in Equation 3.3) specifies prior probability distribution of the spatial random effect corresponding to i^{th} road segment, given that of the remaining road segments.

$$p(\phi_i|\phi_{-i}) \sim N\left(\sum_j b_{ij}\phi_j, \tau_i^2\right), \qquad i \in \{1, 2, ...n\} \tag{3.3}$$

A proximity matrix (W) that governs the extent of spatial dependence across the road network is defined. Either a neighborhood-based (binary) or distance-based proximity matrix (need not be a row-stochastic matrix) with zero diagonal elements is typically utilized. In order to ensure the symmetry of the covariance matrix (i.e. $(I - B)^{-1}D$), $b_{ij} = \frac{w_{ij}}{w_{i+}}$ and $\tau_i^2 = \frac{\tau_c^2}{w_{i+}}$ are commonly used. A neighborhood-based proximity matrix is adopted in this study. $w_{ij} = 1$, if i and j are first-order[1] neighbors; $w_{ij} = 1/2$, if i and j are second-order neighbors and so on; otherwise $w_{ij} = 0$. The intuition behind such weight matrix is trivial; the closer road segments have greater influence than the farther road segments.

The n-dimensional joint prior distribution of the spatial random effects is obtained by combining their individual full conditional prior distributions (shown in 3.3) using Brooks Lemma (see Equation 3.4).

$$p(\phi) \propto exp\left(-\frac{1}{2}\phi^T D^{-1}(I - B)\phi\right) \tag{3.4}$$

where, $\phi = [\phi_1, \phi_2, ...\phi_n]^T$, D is a diagonal matrix with $D_{ii} = \tau_i^2$ and $B = \{b_{ij}\}$, and $\tau_i^2 = \frac{\tau_c^2}{w_{i+}}$. Hyper prior can be used to learn τ_c rather than providing a fixed prior value, which is often unavailable.

The propriety of the joint density of the spatial random effects is another concern as the matrix $D^{-1}(I-B)$ is clearly singular, so that the covariance matrix does not exist. Many earlier studies in road safety have introduced ρ parameter in the full conditional

[1]First-order neighbors are immediate neighbors of a road segments, while second-order neighbors are neighbors of the immediate neighbors)

mean[2] of the spatial random effects to circumvent the issue of improper[3] joint distribution. Also, the ρ parameter has been extensively utilized as a proxy for the strength of underlying spatial correlation among the crash counts across road network. On the other hand, it is reported that the ρ parameter does not compare well with other descriptive measures of spatial association such as Moran's I or Geary's C (see Banerjee et al. [2004] page 78 for discussion). Banerjee et al. [2004] mentioned that ρ can mislead the analyst regarding the inferences on the strength of spatial correlation, particularly in the context of CAR priors. Interestingly, Banerjee et al. [2004] did not provide any guidelines on the inclusion of the ρ parameter, but remained neutral. Also, introduction of the ρ results in a prior mean that is only a proportion of the neighborhood weighted average; there is no intuitive reason behind such prior mean. We opted not to include the ρ parameter; specifically, spatial random effects are modeled using an improper CAR (often termed as Intrinsic CAR (ICAR)) prior.

The absence of ρ parameter appears to complicate the measurement of the strength of spatial correlation among the crash frequencies of the road segments across the road network. It is to be noted that τ_c (or $1/P_c$) is associated with the conditional distribution of the spatial random effects and does not represent strength of the underlying spatial correlation. Upon multiplying the crash counts (y_i) by a constant c, τ_c^2 becomes $c\tau_c^2$; however, the strength of the spatial correlation remains unaffected [see Banerjee et al., 2004, pg. 78 for discussion]. We use the empirical proportion of variability (α) in the expected crash counts due to clustering as the strength of the spatial clustering

[2] $b_{ij} = \frac{W_{ij}}{W_{i+}}$ is replaced with $b_{ij} = \rho\frac{W_{ij}}{W_{i+}}$

[3] An improper probability distribution has a precision matrix that is not of full rank

[see Banerjee et al., 2004, pg. 160]. The remaining portion $(1-\alpha)$ of the variability is due to unstructured heterogeneity (due to Gamma distributed random effects). The spatial random effects enter the crash rate through an exponential function, while the random effects due to the unstructured heterogeneity are directly multiplied with the deterministic component. The random effects due to clustering are exponentiated before calculating α to ensure a sensible comparison with unstructured random effects.

$$\alpha = \frac{\sigma_\phi}{\sigma_\phi + \sigma_\epsilon} \tag{3.5}$$

σ_ϕ is the empirical standard deviation of the exponentiated spatial random effect posterior draws in each MCMC iteration. σ_ϵ is the standard deviation of the unstructured random effects generated by the Gamma mixing in the negative binomial model. Thus, the empirical posterior distribution of α is constructed and utilized for the Bayesian inference of the strength of spatial correlation.

Bayesian estimation is utilized to perform statistical inference on the parameters of the proposed NBSRP model; the model parameters are treated as random variables in a Bayesian framework. Any Bayesian model specification is complete with the specification of the prior beliefs on the model parameters. We chose to utilize conjugate priors on the model parameters in order to be able to derive analytically tractable full conditional posterior distributions as described in the next subsection. In summary, the following hierarchical specification is utilized for crash frequency modeling within this study.

$$\begin{aligned} y_{it} &\sim NB(r, p_{it}), & i \in \{1,2,...n\}; t \in \{1,2,...T\} \\ p_{it} &= \frac{e^{\psi_{it}}}{1+e^{\psi_{it}}}; & \psi_{it} = X_{it}^{F'}\gamma + X_{it}^{R'}\beta_i + \phi_i \end{aligned}$$

$$\beta_i \sim \sum_{c=1}^{C} \eta_c MVN_q(\mu_c, \Sigma_c); \qquad c \in \{1, 2, ...C\};$$

$$s.t. \sum_{c=1}^{C} \eta_c = 1;$$

$$\eta_c \sim Dirichilet(\alpha_0, ...\alpha_0);$$

$$\mu_c \sim MVN_q(b_0, B_0); \Sigma_c^{-1} = \Lambda_c \sim Wish^4(\nu_0, V_0);$$

$$\phi_i | \phi_{-i} \sim N\left(\sum_j \frac{w_{ij}}{w_{i+}} \phi_j, \frac{\tau_c^2}{w_{i+}}\right)$$

$$\gamma \sim MVN_p(g_0, G_0); \qquad r \sim Gamma^5(r_0, h); \qquad h \sim Gamma(ha_0, hb_0)$$

$$1/\tau_c^2 = P_c \sim Ga(c_0, d_0)$$

A multivariate normal prior is used for the fixed parameter vector γ and the component specific random parameter mean vectors mu_c. The component specific covariance matrices Σ_c are re-parametrized into respective precision matrices Λ_c and a Wishart prior is assumed for obtaining posterior inference. A Dirichilet prior is imposed on the mixture weight vector η. A Gamma prior is utilized for posterior inference on the dispersion parameter r along with a hyperprior to tune the scale parameter h.

3.2 Bayesian posterior inference

A Gibbs sampling algorithm is utilized to perform MCMC simulations for constructing the joint distribution of the model parameters. MCMC samples are generated

[4] $Wish(\nu, V)$ is Wishart distribution with mean νV

[5] $Gamma(\alpha, \beta)$ is Gamma distribution with mean α/β

iteratively drawing from full conditional posterior distributions of the individual parameters (see Gamerman and Lopes [2006] for details on MCMC simulation). Full conditional posterior distributions of the model parameters (both dispersion and regression parameters) of the negative binomial model do not belong to any standard distributional family [Park and Lord, 2009]. Although, Metropolis-Hastings (M-H) algorithm is a great alternative sampling technique (used by WinBUGS), it is often associated with slow mixing thereby delaying attainment of stationarity in MCMC chains (see Van Dyk and Meng [2001] for discussion). To avoid M-H algorithm, we used data augmentation techniques in this research and constructed analytical posterior probability distributions. Data Augmentation technique involving intermediate latent random variables is a technique to construct analytically tractable posteriors (only up to proportionality constant) in the statistics literature (see Van Dyk and Meng [2001] for discussion on the art of data augmentation). In this research, we simultaneously utilized three data augmentation techniques for deriving full conditional distributions. The analytical full conditional distributions are derived for the model parameters to facilitate Gibbs sampling as described below.

3.2.1 Data Augmentation

Bayesian inference of r:

Zhou et al. [2012] proposed a data-augmentation based Bayesian inference procedure for the dispersion parameter (r) of negative binomial likelihoods using compound Poisson representation. Negative binomial random variables can also be generated using sums of Logarithmic random variables under compound Poisson distribution [Quenouille, 1949].

Data augmentation is achieved by introducing a Poisson distributed random variable L_i (see Appendix A for details on L_i). The full conditional posterior distribution of $r|L$ is constructed using the conjugacy of the Poisson likelihood and non-informative Gamma prior. Details on the derivation of the Gamma distributed conditional posterior distribution of r are shown in Appendix B.

Introducing latent component labels:

The proposed finite mixture random parameter structure is parametrized in terms of the component-specific means $(\mu_1, ..., \mu_C)$ and covariance matrices $(\Sigma_1, ..., \Sigma_C)$. A latent $n \times 1$ component label vector (G) representing the component membership of the n^{th} road segment is introduced into the estimation process. In other words, $G_i = \{c : c \in \{1, 2...C\}\}$ if the i^{th} road segment belongs to the c^{th} component. Latent component labels facilitate the Bayesian estimation of the component-specific parameters. Upon conditioning on the G vector, the random effect vectors (β_i) essentially becomes independent multivariate Gaussian draws within each mixture component; thus, a component specific likelihood is constructed.

Bayesian inference of $\beta_{1:n}$***,*** γ***,*** $\mu_{1:C}$***,*** $\Sigma_{1:C}$***:***

A recently developed data augmentation strategy for fully Bayesian inference in models with negative binomial likelihoods using Polya-Gamma random variables is adopted in this study (see Polson et al. [2013]). Polya-Gamma random variables (ω_{it}) are introduced into the hierarchical specification, which in turn assist in building analytical conditional posterior of negative binomial regression coefficients. Polson et al. [2013] proved that binomial likelihoods parametrized by log-odds can be written as mixtures of Gaussians with respect to Polya-Gamma distribution. The finding is very useful as it translates

the discrete negative binomial model to convenient Gaussian form upon conditioning on the latent Polya-Gamma random variables. A brief description of Polya-Gamma random variables and the details on transformation of negative binomial likelihood into conditionally Gaussian likelihood are provided in Appendix A. The conditionally Gaussian likelihood combined with a conjugate prior structures allows to construct respective analytically tractable posterior distributions.

The posterior samples of the fixed and random parameters, and component specific means are blocked together in each MCMC iteration. In other words, the posterior draws of $\{\beta_{1:n}, \gamma, \mu_{1:C}\}$ are sampled from a joint full conditional distribution. The joint full conditional distribution of $\beta_{1:n}, \gamma, \mu_{1:C}$ is constructed by employing the aforementioned Polya-Gamma data augmentation; the conditionally Gaussian likelihood is combined with respective multivariate normal conjugate priors (see Appendix B). The blocking of random effects, component specific mean vectors, and fixed regression parameters improves the convergence speed by accelerating the mixing of the MCMC chains [Frhwirth-Schnatter et al., 2004, Xiong and Mannering, 2013]. The implementation of blocked posterior samples requires marginalization of the random effects as described in Appendix B. The analytically tractable full conditional distributions of the component specific covariance matrices are subsequently derived by assuming respective conjugate and non-informative prior structures. Details on the derivation of the component specific full conditional distributions are provided in Appendix B.

Model identification:

The component specific labels are not identified in a finite mixture model, which leads to so-called label switching problem. It is important to resolve the label-switching problem,

where component-specific inferences contain interesting insights; however, the issue may not important for prediction purposes. Frhwirth-Schnatter et al. [2004] discussed the unidentifiability and label switching issues in the case of linear finite mixture models. The authors emphasize the importance of employing a permutation sampler using a constraint during MCMC simulation. In this research, a constraint is selected upon exploring the unconstrained MCMC parameter draws. A random parameter with significantly distinct components in the posterior draws was identified rather than assuming arbitrary constraints (such as labeling by mixture weights). The selected constraint is imposed numerically by re-labeling the current component labels in each MCMC iteration (see Appendix B for details).

Bayesian inference of ϕ:

Transforming the negative binomial likelihood into a conditionally Gaussian likelihood also facilitates the construction of analytically tractable posterior for spatial random effects (as shown in Appendix B). The impropriety of the prior joint distribution of spatial random effects, arising due to the absence of the ρ parameter, does not effect the propriety of the posterior distribution. The posterior precision is the sum of the likelihood and the prior precision, thereby ensuring the propriety of the posterior probability distribution of ϕ. However, ICAR being a pairwise difference prior identifies random effects only upto an additive constant[6]; a sum-to-zero constraint[7]is one way to evade the issue. The constraint is numerically imposed by re-centering the draws of ϕ_i around its own mean in each Gibbs sampling iteration.

[6] Addition of a constant to the all the spatial random effects does not change the joint density function

[7] $\sum_{i=1}^{n} \phi_i = 0$

The detailed derivations of analytical full conditional posteriors for each model parameter are provided in Appendix B . In summary, the model parameters are estimated by iteratively drawing from the derived full conditional posteriors of the individual parameters using a Gibbs sampling framework. The necessary full conditional distributions of the model parameters turned out to be typical probability distributions for which efficient sampling techniques do exist. Sampling from Polya-Gamma distribution remains an exception; we utilized an R package —*BayesLogit* containing efficient Polya-Gamma samplers developed by Polson et al. [2012].

3.2.2 Measuring spatial correlation

The absence of ρ parameter appears to complicate the measurement of the strength of spatial correlation among the crash frequencies of the road segments across the road network. It is to be noted that τ_c (or $1/P_c$) is associated with the conditional distribution of the spatial random effects and does not represent strength of the underlying spatial correlation. Upon multiplying the crash counts (y_i) by a constant c, τ_c^2 becomes $c\tau_c^2$; however, the strength of the spatial correlation remains unaffected [see Banerjee et al., 2004, pg. 78 for discussion]. We use the empirical proportion of variability (α) in the expected crash counts due to clustering as the strength of the spatial clustering [see Banerjee et al., 2004, pg. 160]. The remaining portion $(1 - \alpha)$ of the variability is due to unstructured heterogeneity (due to Gamma distributed random effects). The spatial random effects enter the crash rate through an exponential function, while the random effects due to the unstructured heterogeneity are directly multiplied with the deterministic component. The random effects due to clustering are exponentiated before

calculating α to ensure a sensible comparison with unstructured random effects.

$$\alpha = \frac{\sigma_\phi}{\sigma_\phi + \sigma_\epsilon} \tag{3.6}$$

σ_ϕ is the empirical standard deviation of the exponentiated spatial random effect posterior draws in each MCMC iteration. σ_ϵ is the standard deviation of the unstructured random effects generated by the Gamma mixing in the negative binomial model. Thus, the empirical posterior distribution of α is constructed and utilized for the Bayesian inference of the strength of spatial correlation.

3.2.3 Model selection

A model selection criterion should discount for the number of parameters or model complexity. We use Deviance Information Criterion (DIC) (see Spiegelhalter et al. [2002]), a commonly used model selection tool particularly in Bayesian Hierarchical modeling. DIC accounts for over-fitting by penalizing for additional model parameters; models with better goodness of fit are associated with smaller DIC values. DIC contains two components: 1) a measure of the fit of a model, 2) complexity of the model. $\bar{D}$ represents the posterior mean deviance, while p_D represents the effective number of parameters or model complexity.

$$DIC = \bar{D} + p_D \tag{3.7}$$

$$\overline{D} = E_{\theta|y}[D] = -\int 2\log p(y|\theta)d\theta,$$

$$p_D = \overline{D} - D(E_{\theta|y}[\theta]) = \overline{D} - (-2\log p(y|\overline{\theta}))$$

In a hierarchical specification, the model generating the observable data is used for the calculation of DIC. It is to be noted that $\overline{\theta}$ is the posterior mean of the parameter vector of the data generating model in a hierarchical specification. Spiegelhalter et al. [2002]) suggested to use a difference in DIC value of 7 to statistically distinguish two different models.

3.2.4 Convergence diagnostics

MCMC simulation involves building a multi-variate Markov chain for the vector of model parameters until the chain attains stationarity with respect to the target joint posterior distribution of the model parameters. Gelman and Rubin diagnostic [Gelman and Rubin, 1992] is one of the most commonly used tool for testing the closeness of the simulated Markov chain convergence to the stationary distribution. This diagnostic requires simulation of multiple Markov chains with varying starting values that are over dispersed throughout the parameter space. The procedure calculates a weighted average of the within (W) and between (B) chain variances, which is an estimate of the variance of the stationary distribution. Subsequently, Potential Scale Reduction Factor(PSRF or $\hat{R}$) is calculated, which is defined as the square root of the ratio of the estimated variance of the stationary distribution and the within chain variation. Scale reduction factors values that are larger than 1.1 or 1.2 indicate that the Markov chain need to be run for longer period to achieve desired stationarity. On the other hand, we may conclude that the simulated observations are close to the target/stationary distribution if PSRF values that are closer to 1.0. Intuitively, the Gelman-Rubin diagnostic tool reflects the magnitude of the level of disparity across multiple Markov chains (with different initial

values) with reference to the within chain variation.

It is to be noted that the PSRF is calculated for each of the model parameters using uni-variate MCMC chains. We have also considered multi-variate version of the PSRF (MPSRF), originally proposed by Brooks and Gelman [1998], that accounts for any potential correlations across individual parameter MCMC chains. The MPSRF is particularly useful in this study given the potential for correlated parameters. For instance, the intercept term in the regression coefficients vector (β_0) is expected to be correlated with the dispersion parameter (r); this is due to the specific parametrization of the negative binomial model. MPSRF, a direct analogue of the univariate PSRF in higher dimensions, measures the closeness of the estimated variance-covariance matrix of the target/stationary distribution and within chain variance-covariance matrix (see Equation (3.8)). The MPSRF is calculated using the following equations (see Lemma 2 in Brooks and Gelman [1998] for derivation of these equations). We simulate m multivariate Markov chains each of length 2n, and discard the first n draws.

$$\hat{R} = \frac{n-1}{n} + \left(\frac{m+1}{m}\right)\lambda \tag{3.8}$$

Where, λ is the largest eigenvalue of the symmetric, positive definite matrix $W^{-1}B/n$ and,

$$W = \frac{1}{m(n-1)} \sum_{i=1}^{m} \sum_{t=1}^{n} (\theta_{it} - \bar{\theta_{it}})(\theta_{it} - \bar{\theta_{it}})^T$$

$$B = \frac{n}{(m-1)} \sum_{i=1}^{m} (\theta_{i.} - \bar{\theta_{..}})(\theta_{i.} - \bar{\theta_{..}})^T$$

where, θ_{it} represent the vector of t^{th} parameter draw of the i^{th} chain.

3.3 Empirical Analysis

This section describes an empirical analysis conducted to demonstrate the potential application of the proposed model structure in modeling crash counts. A comprehensive dataset was prepared by integrating accident and pavement databases.

3.3.1 Data collection

In this study, we modeled the crash counts of contiguous road segments of a selected road network. The crash counts are sourced from the motor vehicle Crash Record Information System (CRIS) database maintained by Texas Department of Transportation (TxDOT). The road segment attributes such as extent of surface distresses, ride quality, traffic volumes, geometric features, etc. were obtained from TxDOT's Pavement Management Information System (PMIS) database. The PMIS database allows TxDOT to store, retrieve, analyze, and report network-level pavement surface condition information that is essential for decision making in pavement maintenance and rehabilitation [TxDOT, 1994]. Road segments and crashes are physically identified using Texas Reference Marker (TRM) system. The CRIS database was integrated with the PMIS database using Texas Reference Marking (TRM) system. TRMs are used to map crashes onto respective road segments and subsequently the crash counts are calculated. A comprehensive description of the dataset utilized for the empirical analysis is provided below.

3.3.2 Data description

The dataset constituted crash counts and relevant attribute information corresponding to 1158 contiguous road segments from eleven different road facilities in the Houston area during the years 2007-10. Figure 3.1 shows a map of the road segments included in the empirical analysis. The contiguous nature of the road segments potentially induce spatial correlation across the respective crash counts. Total number of crash counts are utilized for this empirical study; the crash counts varied from 0 to 318 with a mean value of 17.3 within the dataset across the four years. The analysis may be extended to high or low severity crash counts using a similar model estimation procedures. Descriptive statistics (see Table 3.1) indicate that crash counts are clearly over-dispersed, which may be modeled using a negative binomial likelihood. Crash counts are collected from the following types of facilities: 1) Interstate Highways (IH), 2) US highways, and 3) State Highways (SH and SL) roads. Traffic[8] or exposure level is one of the important and most intuitive predictors in crash count modeling. The dataset covers a wide range of traffic levels on the road segments as per the descriptive statistics. The influence of the Truck traffic is also included in the dataset as it may potentially influence the crash counts. The truck traffic percentage varied between 2.6% to 34.3% with a mean value of 10.56%. The speed limit of the road segments varied between 55 and 70 miles/hour with a mean value of 61 miles/hour. Table 3.1 also provides descriptive statistics corresponding to left and right shoulder widths and total surface way width.

Distress score represents the extent of surface distresses such as cracking, rutting,

[8]Annual Average Daily Traffic (AADT).

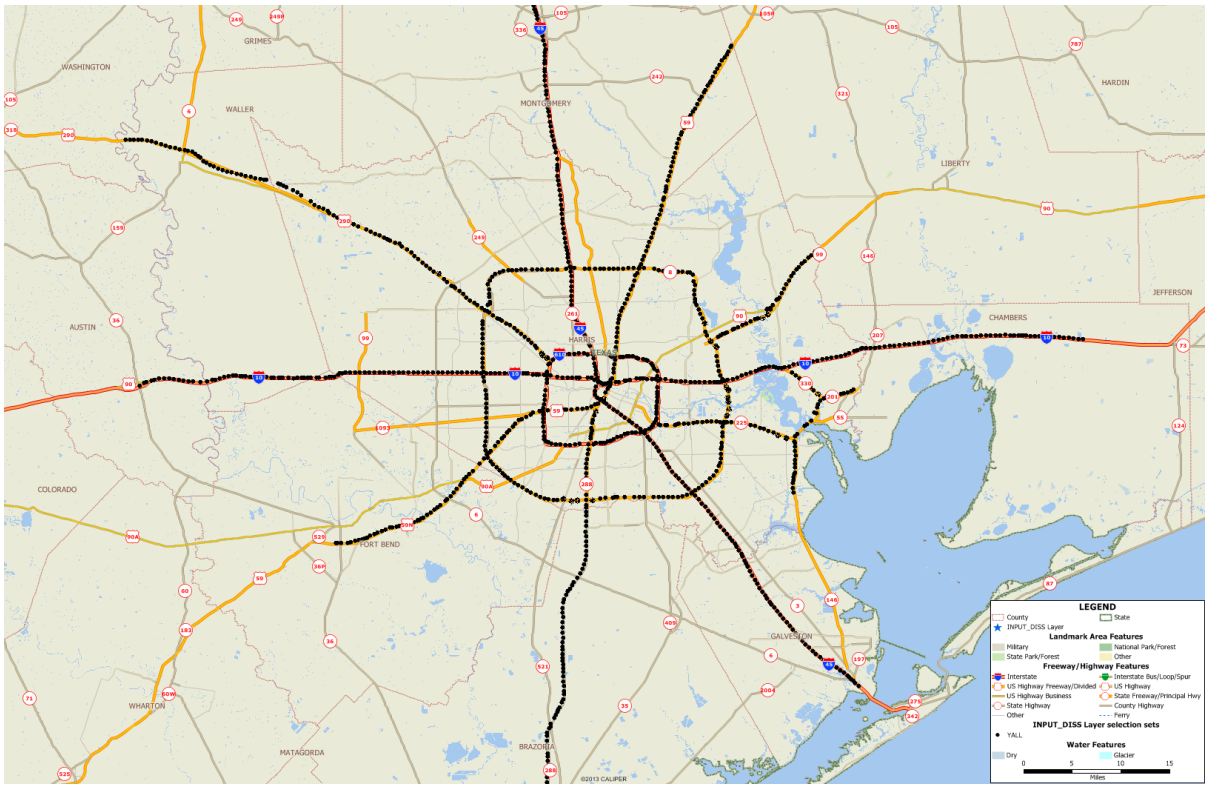

Figure 3.1: Road segment locations

edge drop offs, raveling, etc. within a road segment. The distress score ranges between 1 (worst pavement condition with major distresses) and 100 (best pavement condition with minor distresses). Table 3.1 reports a mean distress score of 93 with a standard deviation of 14. International Roughness Index (IRI), a measure of pavement ride quality, is annually collected across the entire Texas road network using Laser-based inertial profilers. Lower IRI values are associated with road segments with superior ride quality. Table 3.1 shows that the average IRI values of the road segments ranges between 35 and 319 inch/mile with a mean value of 116 inch/mile. Condition score represents the overall condition of a road segment in terms of both distresses and ride quality. The condition

score also ranges between 1 (worst pavement condition) and 100 (best pavement condition). About 74% of the road segments are continuously reinforced concrete pavements (CRCP), which is expected for Houston area. As shown in Table 3.1, about 45% percent of the road segments belong to Interstate facilities. Table 3.1 shows other important descriptive statistics of the dataset.

3.3.3 Model estimation

The main goal of any specification building process is to develop a hierarchical model that best represents the dataset. Several specifications, using the road features provided in Table 3.1 as predictors in the mean function of the negative binomial likelihood, were considered. First, aspatial model specification with fixed parameters, and followed by an equivalent spatial specification were estimated. The number of neighbors required to construct the spatial weight matrix was computed using model selection. Subsequently, the spatial specification was extended to incorporate the unobserved heterogeneity by allowing for random parameters. As mentioned earlier, the random parameters were modeled using a mixture of multi-variate normal distributions. The number of mixture components required to adequately capture the unobserved heterogeneity was identified using model selection. A final specification was cautiously chosen among several alternative specifications. DIC was utilized throughout the specification refinement for model selection and comparison.

The posterior distributions of the model parameters are constructed iteratively drawing from the full conditional posterior distributions of the individual parameters (provided Appendix B). The simulations were carried out on a Windows-based machine

with Intel Core i7 CPU with 1.73GHz and 8GB RAM, and coded in the R language [R Core Team, 2013]. Multiple MCMC chains differing in the initial parameter values were constructed. A burn-in period of 5000 iterations was used, and the model parameters were estimated based on samples obtained from 15000 iterations after the burn-in period during the MCMC simulation. An average run time of about 7 hours (20,000 iterations) was required to attain stationarity of the multi-variate MCMC chain, and to collect the required parameter samples. Gelman & Rubin convergence diagnostics were performed to ensure convergence of the MCMC chains (of the model parameters) to stationary target posterior distributions. Posterior probability of a coefficient to be positive ($P(\beta > 0)$)was used to determine the statistical significance of regression parameters; the values closer to 1 or 0 indicate statistical significance of the respective coefficient.

Based on the DIC model selection criterion, the spatial finite-mixture random parameter negative binomial model (with two components) surpassed the alternative specifications with fixed parameters as well as specifications with random parameters (without finite-mixtures). Figure 3.2 shows the scatter plots of posterior means of random parameters. As the multivariate random parameter posterior means are difficult to visualize, bivariate plots of selected pairs of random parameters are shown in the figure. The figure indicates the presence of two components in the selected dimensions within the joint distribution of posterior mean random parameters. Table 3.2 provides the posterior summaries of the model parameters corresponding to final refined specification, and necessary model estimation details. An intuitive interpretation of posterior summaries is provided below to understand the empirical associations between the road features and crash counts.

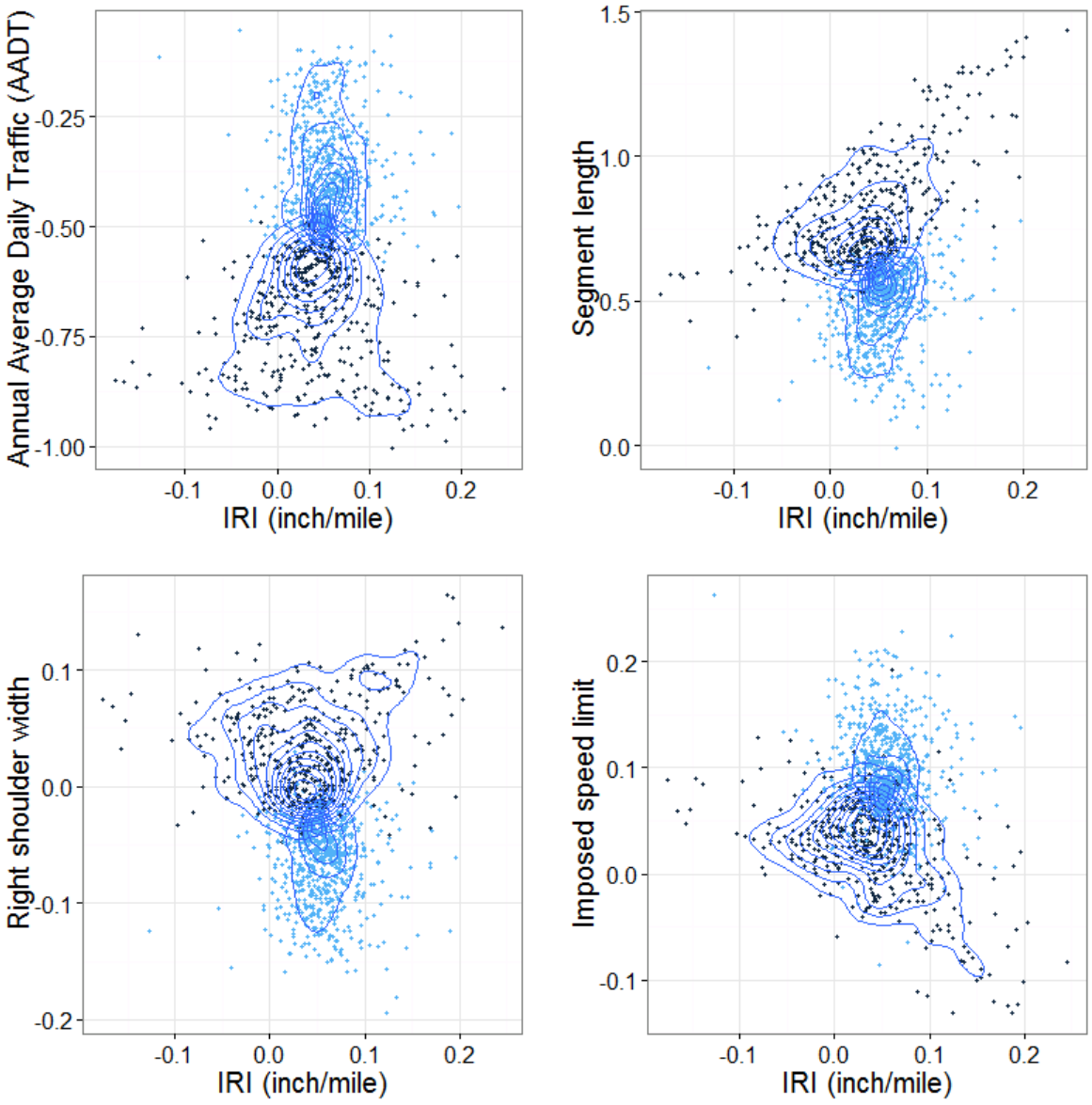

Figure 3.2: Posterior means of random parameters (The contours are kernel density estimates of the distribution of posterior means)

3.3.4 Discussion

Data suggests that the road features including facility type, pavement type, road segment location, shoulder type were significantly associated with the crash counts; the parameters corresponding to these features were fixed in the specification. The negative posterior mean of Interstate Highway (IH) facility indicator suggests that the road segments on IH facilities witness lower mean crash count relative to the other facilities. The mean crash count on the road segments in the outskirts of the Houston area is found to be generally lower than that of the road segments in the urban area. The asphalt road segments and/or segments with asphalt shoulders witnessed higher mean crash counts relative to the concrete road segments and/or segments with other types of shoulders respectively. The strength of spatial correlation is very significant, which indicates the importance of accounting for the spatial correlation among the crash counts. Perhaps, the spatial correlation was caused by unobserved socio-demographic factors such as Household income, local communities, magnitude of roadside advertising, etc., which are arguably shared among neighboring road segments.

The unobserved heterogeneity is incorporated by allowing random parameters on the selected[9] road features including road condition, segment length, traffic load estimate, truck traffic percentage, AADT, IRI, shoulder width, and imposed speed limit. As mentioned earlier, the random parameters are modeled using a two-component multivariate normal mixture prior. The posterior summaries of component specific means and standard deviations are provided in Table 3.2. The model specification allows for

[9]The feature selection is based on the DIC model selection criterion.

covariance among the regression parameters, and the covariance elements were significant. The finite-mixture specification allows for identifying the underlying sub-groups of road segments, and for skewness and multi-modality in the underlying random parameter distribution. The posterior means of the component mixture weights (η_1, η_2) indicate that about 45% of the road segments belong to the component1, and the remaining 55% belong to the component2. The latent segmentation of road segments represents the heterogeneity of road segments in terms of the relationship between the road features and the crash counts. The latent segmentation may not translate to spatial clustering of the road segments, but represents the clustering of road segments in terms of their safety response to a given set of road attributes. The component specific means represents an average influence of road features on the crash counts within the respective sub-group of road segments. Road segments arguably respond very differently to similar safety treatments depending on the site-specific unobserved factors. The latent segmentation of road segment may be helpful to identify the groups of road segments that respond fairly similar to safety treatments.

The mean crash count on the road segments in good condition[10] was found to be lower than that of road segments in inferior condition. In other words, data suggests that the road condition positively influences the road safety in both the sub-groups of road segments as shown by the negative component specific posterior means. The finding highlights the safety benefits of improving overall road condition across the study area on an average. However, the magnitude of safety benefits for a given road condition improvement may vary across the road segments due to heterogeneity, which is represented by

[10]Road segments with condition score greater than 90.

the respective significant standard deviation. The influence of improving the road condition on the mean crash count may also effect the influence of other explanatory variables such as traffic characteristics, and shoulder width on the crash counts; this is represented by respective significant covariance parameters[11]. Positive component specific posterior means corresponding to IRI represents an increase in mean crash count with a decrease in the ride quality across both sub-groups of road segments. A significant component specific standard deviations associated with the IRI indicate the heterogeneity in terms of the influence of IRI on the mean crash count.

The posterior results show that the traffic characteristics significantly influences the crash counts. However, the influence of traffic characteristics may be different across the two underlying (latent) sub-groups of road segments. For instance, the traffic load is adversely affecting the safety level of road segments in component-1, while an opposite effect is observed in component-2[12]. A significant heterogeneity in the influence of heavier vehicles on crash counts within each group was also evident as indicated by significant standard deviation of the respective regression parameters. The truck traffic percentage was found to be negatively associated with the mean crash count on the road segments that belong to component-1 on an average. On the other hand, the positive posterior mean[13] component-2 on the truck traffic percentage indicates an opposite relationship. A negative posterior mean on both component specific means corresponding to logarithm of AADT indicate that road segments with higher traffic volumes are associated with lower mean crash count across both sub-groups. The road segments with larger traffic volumes

[11]The estimated component specific covariance parameters are not shown due to space restrictions.
[12]However, 36% of the posterior mass of the parameter is in the positive region.
[13]However, 29% of the posterior mass is in the positive region.

may often be congested with reduced actual travel speeds, which may decrease the crash counts. A significant heterogeneity was evident on the magnitude of the influence of AADT on the mean crash count as represented by the respective component specific significant standard deviations.

A positive posterior component-specific mean on the regression parameter corresponding to segment length indicates that larger mean crash count is associated with longer segments in both the sub-groups of road segments. It should be noted that the segment length may be interpreted as an exposure variable rather than a causal variable. Posterior results indicate that the left and right shoulder widths are positively associated with the mean crash count across the road segments within the component-1 on an average. On the other hand, the higher shoulder widths are associated with reduction in the mean crash count across the road segments that belong to latent component-2. The road segments in different latent groups may potentially react differently to a treatment such as an increase in shoulder width on an average. Increase in the imposed speed limit was found to be associated with higher mean crash counts across the road segments belonging to latent component-2; data suggests an opposite behavior[14] regarding the road segments in component-1. Moreover, significant component specific standard deviations indicate the presence of within group heterogeneity regarding the influence of left and right shoulder widths, and imposed speed limit on the mean crash count.

[14]However, 36% of the posterior mass is supporting a consistent behavior across the sub-groups.

Table 3.1: Descritive statistics

Category	Description	Minimum	Maximum	Mean	Std.Dev.
Time Span	Year	2007	2010	-	-
Crashes	Crash count	0	318	17.3	25.3
Traffic	Annual Average Daily Traffic (AADT)	5520	167625	51692	37184
	Traffic load estimate	120	197	169	18
	Truck traffic percentage	2.6	34.3	10.56	6.52
	Imposed speed limit	55	70	61	5
Geometrics	Number of lanes (per traffic direction)	2	6	3.16	1.01
	Total surface roadway width	31	110	55.44	15.36
	Left shoulder width	0	27	8.53	3.04
	Right shoulder width	0	13	9.35	2.26
	Indicator Variable: Asphalt shoulder	0	1	0.585	-
	Segment length	0	2	0.46	0.16
Pavement	Condition score	3	100	87	19
	Distress score	8	100	93	14
	Ride score	1.1	4.9	3.48	0.57
	Avg International Roughness Index (IRI inch/mile)	35	319	116	35
	Left International Roughness Index (IRI inch/mile)	32	343	113	34
	Right International Roughness Index (IRI inch/mile)	37	338	118	38
	Maintenance Cost	0	215175	1066	4826
	Indicator Variable: Asphalt pavement	0	1	0.15	-
	Indicator Variable: CRCP pavement	0	1	0.74	-
	Indicator Variable: JCP pavement	0	1	0.11	-
	Indicator Variable: Shoulder type - Asphalt	0	1	0.58	-
Location	Indicator Variable: Facility-IH	0	1	0.45	-
	Indicator Variable: Facility-SH	0	1	0.15	-
	Indicator Variable: Facility-US	0	1	0.26	-
	Indicator Variable: Facility-Other	0	1	0.14	-
	Indicator Variable: Rural Area	0	1	0.25	-

Routes: IH0010, IH0045, IH0610, SH0146, SH0225, SH0228, SL0008, SS0330, US0059, US0090, US0290

Table 3.2: Posterior estimation results for spatial NB finite-mixture random parameters model

Fixed parameter posterior summaries						
	Posterior mean	Posterior Std.Dev.	P($\beta > 0$)			
Intercept	-0.046	0.033	0.08			
Indicator Variable: Facility-IH	-0.070	0.044	0.05			
Indicator Variable: Asphalt pavement	0.193	0.059	1.00			
Indicator Variable: Rural Area	-0.055	0.045	0.11			
Indicator Variable: Asphalt shoulder	0.096	0.045	0.98			
r	7.125	0.220	1.00			
τ_c	1.268	0.129	1.00			
Spatial strength	0.659	0.030	1.00			
	Random parameter posterior summaries					
	Component 1			Component 2		
	Posterior mean	Posterior Std.Dev.	P($\beta > 0$)	Posterior mean	Posterior Std.Dev.	P($\beta > 0$)
Indicator Variable: Good road condition	-0.129	0.072	0.04	-0.164	0.066	0.01
	(0.221)	(0.188)	(1)	(0.118)	(0.101)	(1)
Segment length	1.052	0.087	1.00	0.226	0.089	1.00
	(0.306)	(0.225)	(1)	(0.184)	(0.174)	(1)
Traffic load estimate	0.201	0.040	1.00	-0.011	0.031	0.36
	(0.202)	(0.121)	(1)	(0.104)	(0.077)	(1)
Truck traffic percentage	-0.118	0.086	0.08	0.043	0.076	0.71
	(0.179)	(0.125)	(1)	(0.116)	(0.107)	(1)
Annual Average Daily Traffic (AADT)	-0.922	0.164	0.00	-0.150	0.098	0.06
	(0.14)	(0.116)	(1)	(0.161)	(0.149)	(1)
International Roughness Index (Inch/mile)	0.045	0.055	0.79	0.053	0.043	0.88
	(0.157)	(0.119)	(1)	(0.12)	(0.102)	(1)
Left shoulder width	0.117	0.070	0.96	-0.061	0.046	0.08
	(0.236)	(0.176)	(1)	(0.099)	(0.081)	(1)
Right shoulder width	0.101	0.063	0.96	-0.134	0.048	0.00
	(0.111)	(0.089)	(1)	(0.098)	(0.069)	(1)
Imposed speed limit	-0.030	0.097	0.36	0.154	0.075	0.99
	(0.135)	(0.117)	(1)	(0.162)	(0.147)	(1)
η_1	0.456	0.077				
η_2	0.544	0.077				
DIC	26105.9					
PSRF (max)	1.14					
MPSRF	1.16					
Number of iterations	20000					
Number of road segments	1158					
Number of years of data	4					

Chapter 4

Modeling temporal patterns in count data

In this chapter, a dynamic negative binomial spatial model is proposed to identify the underlying temporal patterns of the regression parameters while simultaneously allowing for spatial correlation. A Bayesian estimation methodology that leverages the recent advances in data augmentation methods and Forward Filtering Backward Sampling (FFBS) algorithm for estimating the proposed model is also described.

4.1 Model development

We assume a crash count dataset comprising of n contiguous road segments and crash count data collected over T time units. Let $Y = \{y_{it}\}$ denote a vector of crash counts, where y_{it} represents the crash counts on i^{th} road segment and t^{th} year; $\{i \in \{1, 2....., n\}\}$ and $\{t \in \{1, 2....., T\}\}$. The crash counts y_{it} are modeled as a nagative binomial population with parameters p_{it} (probability parameter) and r (dispersion parameter). Let $Z_i = [Z_{i1}, Z_{i2}, ..., Z_{ip}]$ denote a $p \times 1$ vector of time-invariant attributes and $X_{it} = [X_{it1}, X_{it2}, ..., X_{itq}]$ denote a $q \times 1$ vector of time-varying attributes. Note that the matrices Z and X are $nT \times p$ and $nT \times q$ dimensional matrices respectively. The probability parameter is modeled as a function of both time-varying and time-invariant attributes. Let $\beta = \{\beta_1, \beta_2, ..., \beta_p\}$ denote the vector of the fixed regression coefficients

corresponding to the time-invariant attributes (Z_i), and $\theta_t = \{\theta_{t1}, \theta_{t2}, ..., \theta_{tq}\}$ denote the vector of time-varying regression coefficients corresponding to the time-varying attributes (X_{it}). Note that the attribute value may remain fixed over the time, however the effect of the attribute is allowed vary with time. The proposed specification facilitates to model the variation of the regression coefficients corresponding to time-invariant as well as time-varying attributes. The temporal variation of the regression coefficients θ_t is modeled as a dynamic linear model. The crash counts of contiguous road segments are potentially spatially correlated and the magnitude of spatial correlation may change over the time. The proposed specification allows for time-varying spatial correlation through time-specific spatial random effects. A vector of spatial random effects $\phi^{(t)}$ generated using Intrinsic Conditional Autoregressive (ICAR) prior structure is utilized to induce spatial correlation across the crash counts at each time t.

$$y_{it} \sim NB(r, p_{it}); i \in \{1, 2, ...n\}; t \in \{1, 2, ...T\} \tag{4.1}$$

$$p_{it} = \frac{\psi_{it}}{1 + e^{\psi_{it}}}; \qquad \psi_{it} = Z_i^T \beta + X_{it}^T \theta_t + \phi_i^{(t)}$$

where, $\phi_i^{(t)}$ is the spatial random effect corresponding to i^{th} road segment in t^{th} time unit.

4.1.1 Dynamic linear models

Time series data analysis typically involve a series of noisy observations over a given period of time. The underlying process that is causing the temporal variation of the outcome variable is of interest upon filtering out the observation noise. Dynamic regression facilitates to investigate the underlying signal by allowing the variation of the regression parameters according to any specified state-space structure. Dynamic Linear

Models (DLM) are formulated by assuming linear operators while specifying the system equations. DLMs are extensively used in time series applications for extracting the underlying states that are driving the temporal changes in the outcome of interest. In this context, we intend to study an underlying crash count generating system, which may be represented using a DLM comprising an observation equation and a state equation as shown below.

$$\zeta_t = F_t\theta_t + \nu_t; \nu_t \sim MVN(0, V_t)$$

$$\theta_t = G_t\theta_{t-1} + w_t; w_t \sim MVN(0, W_t)$$

ζ_t represents a $n \times 1$ vector of noisy observations of the system at time t. We pretend ζ_t instead of y_t ($n \times 1$ vector of crash counts at time t) as the noisy observation. As mentioned before, y_t is spatially correlated negative binomially distributed crash count vector. A data augmentation technique is employed to transform the negative binomial observation vector y_t into a multivariate Guassian distributed random variable z_t (this is further discussed in the model estimation section). The attribute matrix, F_t is constructed using the observable attributes of the system i.e. X_t ($n \times q$ matrix). As mentioned earlier, the regression parameter of fixed attributes may be modeled as time-varying parameters[1]. The vector of observations ζ_t are generated by latent regression parameter state vector θ_t after transformation using the attribute matrix F_t. A zero-centered multivariate Gaussian noise term is defined using the covariance matrix V_t.

The latent parameter vector θ_t is assumed to be generated according to the linear state equation. θ_t is generated by the transformation of θ_{t-1} using the system operator

[1]Fixed attributes (Z_{ik}) may be included in the attribute matrix F_t by repeating at different times.

matrix G_t $(q \times q)$. G_t may be designed such that θ_t is generated through an auto-regressive (AR (1)) process. DLM model structure allows for specifying any other general auto-regressive structure. In this research, we assumed individual AR(1) processes for evolution of θ_t. G_t matrix may be specified such that a current state θ_{tk} is dependent on another previous state $\theta_{tk'}$ (where $k \neq k'$). In addition, G_t may be designed as a time-varying system operator matrix; however, we assume a time-invariant G_t in this study. In summary, the DLM framework is adequately flexible to investigate several underlying temporal patterns.

4.1.2 Time-varying intrinsic conditional autoregressive priors

As mentioned earlier, the negative-binomial likelihood of the observed crash counts is transformed into conditionally Gaussian likelihood. Assuming a Gaussian prior on spatial random effects is useful to exploit the conjugacy. Intrinsic conditional autoregressive(ICAR) priors are Gaussian spatial prior structures in Bayesian hierarchical modeling. ICAR prior generates the spatially correlated random effects based on a neighborhood or distance based correlation matrix. We utilize a neighborhood weight matrix W defined as follows: $w_{ij} = 1/k$, if i and j are k-order neighbors. The spatial dependence is assumed to be proportional to the closeness of the neighboring road segments. The spatial correlation may vary with time, which can be modeled by allowing for temporal variation in the ICAR parameters. The time-specific τ_t parameter controls the extent of spatial correlation at time t. The τ_t parameter does not provide any information about the strength of spatial correlation. We compute the strength of spatial correlation α_t (see 3.5) parameter for each year t. ICAR prior defines the distribution of spatial random

effect of i^{th} road segment at time t (ϕ_i^t) conditional on spatial random effects of other road segments (ϕ_{-i}^t).

$$\phi_i^t|\phi_{-i}^t \sim N\left(\sum_j \frac{w_{ij}}{w_{i+}}\phi_j^t, \frac{\tau_t^2}{w_{i+}}\right) \tag{4.2}$$

To facilitate Bayesian estimation, non-informative conjugate priors are imposed on the model parameters. A multivariate normal prior is imposed on β and θ_0. Non-informative Gamma prior is imposed on parameters τ_t, r and h. The proposed model structure is described below.

$$y_{it} \sim NB(r, p_{it}); i \in \{1, 2, ...n\}; t \in \{1, 2, ...T\}$$

$$p_{it} = \frac{1}{1 + e^{-\psi_{it}}}; \qquad \psi_{it} = Z_i^T\beta + X_{it}^T\theta_t + \phi_i^{(t)}$$

$$\theta_t = G_t\theta_{t-1} + W_t; W_t \sim N(0, \Sigma)$$

Where,

$$\theta_t = \begin{bmatrix}\theta_{1t}\\ \theta_{2t}\\ ...\\ \theta_{qt}\end{bmatrix} G_t = \begin{bmatrix}\rho & 0 & .. & .. & 0\\ 0 & \rho & .. & .. & 0\\ .. & .. & .. & .. & ..\\ 0 & 0 & .. & .. & \rho\end{bmatrix} \Sigma = \begin{bmatrix}\sigma_{\theta_0}^2 & .. & .. & .. & 0\\ 0 & \sigma_{\theta_1}^2 & .. & .. & 0\\ .. & .. & .. & .. & ..\\ 0 & 0 & .. & .. & \sigma_{\theta_T}^2\end{bmatrix}$$

$$\beta \sim N(b_0, B_0); \theta_0 \sim N(m_0, V_0)$$

$$\phi_i^t|\phi_{-i}^t \sim N\left(\sum_j \frac{w_{ij}}{w_{i+}}\phi_j^t, \frac{\tau_t^2}{w_{i+}}\right)$$

$$1/\tau_t^2 = P_t \sim Ga(c_0, d_0); \qquad r \sim Ga(r_0, h); h \sim Ga(ha_0, hb_0)$$

4.2 Bayesian inference

The model parameters are estimated using Bayesian inference through MCMC simulation. A Gibbs sampling algorithm involving full conditional posterior distributions for the model parameters is described in this section. Negative binomial model parameters do not enjoy the conjugacy with any distributional forms in the context of posterior inference; therefore, the tractable analytical full conditional distributions are not available. We employed recently developed data-augmentation techniques for constructing analytical full conditional distributions for the negative binomial model parameters. Forward Filtering Backward Sampling (FFBS) algorithm is commonly employed for posterior inference of DLM parameters with Gaussain distributed observations [Prado and West, 2010]. The data augmentation facilitated embedding the FFBS algorithm into the Gibbs sampling framework for performing posterior inference of the dynamic parameters θ_t with negative binomial distributed observations.

4.2.1 Data augmentation

Data augmentation involves introducing latent random variables into the specification that are useful to derive analytically tractable full conditional posteriors. Polson et al. [2013] developed a data-augmentation technique to transform logistic likelihoods into Gaussian likelihood conditional on Polya-Gamma latent random variables[2]. The details on the transformation of negative binomial likelihood into Gaussian likelihood conditioning on Polya-Gamma random variables are provided in Appendix A. Analytically tractable full conditional posteriors are derived for the fixed parameter vector β

[2]Details on Polya-Gamma random variables is provided in Appendix A

using the conjugacy of the conditionally Gaussian likelihood and a multivariate normal prior(see Appendix C for details). Another data augmentation technique proposed by Zhou et al. [2012] was also employed for constructing the full conditional distribution for the dispersion parameter. A Poisson distributed latent indicator variable is introduced into the model structure. The full conditional posterior distribution of the dispersion parameter turns out to be Gamma; see Appendix C for details.

The Polya-Gamma data augmentation also allows to construct a tractable analytical full conditional posterior for spatial random effects. Time-dependent ICAR priors are improper prior probability distributions on the vector of spatial random effects [Banerjee et al., 2004]; the absence of the ρ parameter causes the impropriety. In addition, ICAR prior is a pairwise difference prior for a given time t, which identifies the spatial random effects $\phi^{(t)}$(specific to time t) only upto an additive constant. The full conditional posterior turns out to be a proper probability distribution with invertible covariance distribution. We impose a sum-to-zero constraint for each time t by recentering the draws of $\phi^{(t)}$ in each MCMC iteration; i.e. $\phi_i^{(t)} = \phi_i^{(t)} - \sum_{i=1}^{n} \phi_i^{(t)} \; \forall i \in \{1, 2, ..., n\}; t \in \{1, 2, ..., T\}$.

4.2.2 Forward Filtering Backward Sampling (FFBS)

Bayesian inference of time-varying regression parameters $\theta_{1:T}$ requires the conditional density $\pi(\theta_{1:T}|D_{1:T}, \beta, \phi^{(1):(T)}, r)$. FFBS is a commonly used algorithm (originally proposed by Frhwirth-Schnatter [1994], Carter and Kohn [1994]) for posterior sampling of states in dynamic linear models. FFBS algorithm simultaneously produces posterior draws of the state vector $\theta_{1:T}$ through forward sampling followed by backward smoothing in each MCMC iteration. FFBS algorithm is not capable of handling a non-linear

model (such as negative binomial) with linear Gaussian evolution equations [Windle et al., 2013]. However, Polya-Gamma data augmentation transformed the negative binomially distributed data vector y_t into a conditionally Gaussian distributed data vector z_t; where $z_t = \frac{y_t - r}{2\omega_t}$ (see Appendix A for details). The negative binomial likelihood shown in Equation 4.1 may be equivalently written as

$$\pi(z_{it}|\omega_{it}) \sim N(\psi_{it}, \frac{1}{\omega_{it}})$$

Where, $\psi_{it} = Z_i^T\beta + X_{it}^T\theta_t + \phi_i^{(t)}$. Now, the evolution equations of the system may be written as follows using the transformed data z_t.

$$z_t = Z\beta + \phi_t + X_t\theta_t + \nu_t$$

$$z_t - Z\beta - \phi_t = X_t\theta_t + \nu_t$$

$$\zeta_t = F_t\theta_t + \nu_t; \nu_t \sim N(0, V_t)$$

$$V_t = \begin{bmatrix} \frac{1}{\omega_{1t}} & .... & 0 & 0 \\ 0 & \frac{1}{\omega_{2t}} & .... & 0 \\ 0 & 0 & .... & \frac{1}{\omega_{nt}} \end{bmatrix}$$

$$\theta_t = G_t\theta_{t-1} + W_t$$

We employed the FFBS algorithm on the transformed data conditioning on the Polya-Gamma random variables. A detailed description of the FFBS recursions for drawing posterior state vectors ($\theta_{1:T}$) is provided in Appendix C.

The full conditional analytical posteriors were derived for each model parameter and a Gibbs sampling framework is provided for Bayesian inference. The detailed derivations along with the full conditional distributions are included in the Appendix C.

The Gibbs sampling framework was mainly implemented in R language, while coding a few portions of the code in Rcpp [Eddelbuettel and Franois, 2011] to gain additional computational speed. An efficient Polya-Gamma sampler provided as part of an R package —*BayesLogit* [Polson et al., 2012] was employed for generating Polya-Gamma random variables. Model selection was based on Deviation Information Criterion (DIC)(see Spiegelhalter et al. [2002]). Gelman and Rubin diagnostic [Gelman and Rubin, 1992] was utilized to verify the convergence of multi-variate Markov chains.

4.3 Empirical Analysis

This chapter describes an empirical analysis conducted to demonstrate the potential application of the proposed dynamic spatial negative binomial model structure in modeling crash counts. A comprehensive dataset was prepared by integrating accident and pavement databases.

4.3.1 Data description

The historical crash counts of contiguous road segments of Houston's road network were modeled in this empirical analysis. The data sources were very similar to that of previously described empirical analysis in Chapter 3; however, the study period is extended to 9 years. The crash counts were sourced from the motor vehicle Crash Record Information System (CRIS) database maintained by Texas Department of Transportation (TxDOT). The road segment attributes such as extent of surface distresses, ride quality, traffic volumes, geometric features, etc. were obtained from TxDOT's Pavement Management Information System (PMIS) database. The CRIS database was integrated

with the PMIS database using Texas Reference Marking (TRM) system. TRMs were used to map crashes onto respective road segments and subsequently the annual crash counts were calculated. A few interesting features of the dataset are described below.

The dataset constituted crash counts and relevant attribute information corresponding to 1158 contiguous road segments from eleven different road facilities in the Houston area during the years 2003-11. As shown in Tables 4.1,4.2, & 4.3, , data suggests a decreasing trend in the mean crash count of the study road network, which indicates an improvement in safety over time. The tables also show descriptive statistics of several time-varying and time-constant road features. For instance, the Annual Average Daily Traffic (AADT) increased every year over the time across the road segments on an average. The proportion of truck traffic remained constant for initial few years followed by a slight reduction during the recent years within the study period. The average imposed speed limit across the road segments was slightly changed during 2003 to 2004 and remained consistent until the end of the study period. The average number of lanes, and average left and right shoulder widths were fairly consistent across the study period. The average condition and distress scores indicate that the road condition of the study road network (on an average) was slightly improved over the time during the analysis period; this reflects TxDOT's pavement maintenance efforts in managing the existing road network. The ride quality (measured by IRI) of the study road network seems to be slightly improved over the time during the analysis period. The proportion of continuously reinforced concrete pavements (CRCP) pavements was slightly increased over the time across the study road network. About 45% of the road segments belong to interstate highways (IH) as shown in Tables 4.1 through 4.3. The proportion of the road

segments with rural area flag has slightly reduced over the time, which may indicate an urban sprawl in the study area.

Table 4.1: Descriptive statistics for years 2003 - 2005

Category	**Description**	**Mean (St.Dev)**		
Time Span	Year	**2003**	**2004**	**2005**
Crashes	Crash count	21.2 (33.8)	17.3 (25.6)	19.8 (27.4)
Traffic	Annual Average Daily Traffic (AADT)	44706 (32102)	49566 (35998)	50820 (38000)
	Traffic load estimate	130 (12)	173 (20)	119 (16)
	Truck traffic percentage	11.4 (7.3)	11.1 (6.8)	11.1 (6.3)
	Imposed speed limit	62 (6)	59 (6)	61 (6)
Geometrics	Number of lanes (per traffic direction)	3 (1)	3 (1)	3 (1)
	Total surface roadway width	52.5 (14.3)	53.3 (14.9)	54.2 (15.4)
	Left shoulder width	8.4 (2.4)	8.2 (2.6)	8.4 (2.7)
	Right shoulder width	8.8 (2.3)	8.9 (2.3)	9 (2.3)
	Segment length	0.5 (0.1)	0.5 (0.1)	0.5 (0.1)
Pavement	Condition score	82 (24)	83 (22)	84 (20)
	Distress score	88 (21)	88 (19)	90 (16)
	Ride score	3.4 (0.6)	3.5 (0.5)	3.4 (0.6)
	Avg International Roughness Index (IRI inch/mile)	118 (35)	113 (33)	118 (36)
	Left International Roughness Index (IRI inch/mile)	117 (35)	111 (32)	116 (36)
	Right International Roughness Index (IRI inch/mile)	119 (36)	116 (35)	120 (36)
	Maintenance Cost	921 (2026)	1133 (2191)	834 (1286)
	Indicator Variable: Asphalt pavement	0.21	0.19	0.20
	Indicator Variable: CRCP pavement	0.65	0.68	0.68
	Indicator Variable: JCP pavement	0.14	0.13	0.13
	Indicator Variable: Shoulder type - Asphalt	0.61	0.61	0.60
Location	Indicator Variable: Facility-IH	0.45	0.45	0.45
	Indicator Variable: Facility-SH	0.15	0.15	0.15
	Indicator Variable: Facility-US	0.26	0.26	0.26
	Indicator Variable: Rural Area	0.28	0.27	0.27

4.3.2 Model estimation

The main goal of model specification is to identify the set of road attributes that potentially explain the variation in the crash counts across the study area during the analysis period. The specification refinement was carried out by estimating several model structures using different combinations of road attributes. The model selection was

Table 4.2: Descriptive statistics for years 2006 - 2008

Category	**Description**	**Mean (St.Dev)**		
Time Span	Year	**2006**	**2007**	**2008**
Crashes	Crash count	19.4 (26.9)	19.2 (25.8)	15.2 (20.6)
Traffic	Annual Average Daily Traffic (AADT)	50098 (36601)	51049 (37334)	51911 (37074)
	Traffic load estimate	142 (10)	166 (14)	167 (14)
	Truck traffic percentage	11.2 (6.3)	10.5 (6.7)	10.4 (6.5)
	Imposed speed limit	61 (5)	61 (5)	61 (5)
Geometrics	Number of lanes (per traffic direction)	3 (1)	3 (1)	3 (1)
	Total surface roadway width	54.3 (15.4)	54.5 (15.2)	55 (15.3)
	Left shoulder width	8.5 (2.8)	8.6 (2.8)	8.7 (2.8)
	Right shoulder width	9 (2.3)	9 (2.3)	9.1 (2.3)
	Segment length	0.5 (0.1)	0.5 (0.1)	0.5 (0.1)
Pavement	Condition score	84 (21)	85 (20)	87 (18)
	Distress score	90 (17)	92 (15)	93 (14)
	Ride score	3.5 (0.6)	3.4 (0.6)	3.5 (0.6)
	Avg International Roughness Index (IRI inch/mile)	114 (37)	117 (36)	114 (34)
	Left International Roughness Index (IRI inch/mile)	112 (36)	117 (35)	108 (34)
	Right International Roughness Index (IRI inch/mile)	116 (37)	118 (38)	121 (38)
	Maintenance Cost	887 (1834)	1446 (9154)	991 (1969)
	Indicator Variable: Asphalt pavement	0.19	0.17	0.17
	Indicator Variable: CRCP pavement	0.69	0.71	0.72
	Indicator Variable: JCP pavement	0.12	0.12	0.11
	Indicator Variable: Shoulder type - Asphalt	0.60	0.60	0.58
Location	Indicator Variable: Facility-IH	0.45	0.45	0.45
	Indicator Variable: Facility-SH	0.15	0.15	0.15
	Indicator Variable: Facility-US	0.26	0.26	0.26
	Indicator Variable: Rural Area	0.27	0.27	0.27

Table 4.3: Descriptive statistics for years 2009 - 2011

Category	**Description**	**Mean (St.Dev)**		
Time Span	Year	**2009**	**2010**	**2011**
Crashes	Crash count	14.2 (19.2)	17.4 (23.7)	14.5 (21.6)
Traffic	Annual Average Daily Traffic (AADT)	52504 (37918)	50889 (36505)	51047 (36330)
	Traffic load estimate	157 (12)	187 (15)	185 (15)
	Truck traffic percentage	10.7 (6.6)	10.8 (6.4)	10.6 (5.2)
	Imposed speed limit	61 (5)	61 (5)	61 (5)
Geometrics	Number of lanes (per traffic direction)	3 (1)	3 (1)	3 (1)
	Total surface roadway width	56 (15.6)	56.3 (15.4)	56.2 (15.4)
	Left shoulder width	8.4 (3.3)	8.5 (3.2)	8.3 (3.5)
	Right shoulder width	9.8 (2.1)	9.6 (2.3)	9.5 (2.6)
	Segment length	0.5 (0.1)	0.5 (0.1)	0.5 (0.1)
Pavement	Condition score	86 (19)	88 (18)	88 (17)
	Distress score	93 (14)	94 (13)	94 (12)
	Ride score	3.5 (0.6)	3.5 (0.6)	3.5 (0.6)
	Avg. International Roughness Index (IRI inch/mile)	117 (36)	113 (34)	113 (34)
	Left International Roughness Index (IRI inch/mile)	115 (35)	112 (33)	107 (31)
	Right International Roughness Index (IRI inch/mile)	119 (39)	114 (36)	119 (40)
	Maintenance Cost	993 (1963)	879 (1632)	943 (2020)
	Indicator Variable: Asphalt pavement	0.14	0.12	0.12
	Indicator Variable: CRCP pavement	0.75	0.79	0.81
	Indicator Variable: JCP pavement	0.11	0.08	0.07
	Indicator Variable: Shoulder type - Asphalt	0.58	0.57	0.57
Location	Indicator Variable: Facility-IH	0.45	0.45	0.45
	Indicator Variable: Facility-SH	0.15	0.15	0.15
	Indicator Variable: Facility-US	0.26	0.26	0.26
	Indicator Variable: Rural Area	0.27	0.20	0.20

performed using Deviance Information Criterion (DIC) (see Spiegelhalter et al. [2002]). The simulations were carried out on a Windows-based machine with Intel Core i7 CPU with 1.73GHz and 8GB RAM, and coded in the R language [R Core Team, 2013]. A burn-in period of 2000 iterations was deemed sufficient, and the model parameters were estimated based on samples obtained from 1000 iterations after the burn-in period during the MCMC simulation. An average run time of about 10 hours (3000 iterations) was required to attain stationarity of the multi-variate MCMC chain, and to collect the required parameter samples. The parameter estimation algorithm was tested on several

simulated datasets (similar in size to the real-world datasets) to assess the efficiency of the parameter retrieval. The average percentage bias was less than 5% for majority of the time-varying and time-invariant model parameters. It was found that the estimation of the AR(1) standard deviations corresponding to the model parameters is difficult using limited temporal data spanning across only a few years. Posterior probability of a coefficient to be positive ($P(\beta > 0)$)was used to determine the statistical significance of regression parameters; the values closer to 1 or 0 indicate statistical significance of the respective coefficient. Table 4.4 shows the posterior summaries of the time-invariant model parameters, and standard deviation of the AR(1) time-varying model parameters. Figures 4.1 to 4.3 show the posterior summaries of the time-varying model parameters. A comprehensive discussion on the posterior estimation results including the implications to the safety management is provided below.

Table 4.4: Posterior estimation results for dynamic spatial NB model

Description	**Posterior mean**	**Posterior Std.Dev.**	$P(\beta > 0)$
Time-invariant parameters			
Intercept	0.591	0.062	1.00
Indicator Variable: Asphalt pavement	0.270	0.046	1.00
r	1.415	0.033	1.00
AR (1) standard deviation for Time-varying parameters			
Intercept	0.104	0.051	1.00
Indicator Variable: Facility-IH	0.218	0.138	1.00
Indicator Variable: Rural Area	0.110	0.055	1.00
Indicator Variable: Shoulder type - Asphalt	0.070	0.034	1.00
Segment length	0.052	0.020	1.00
Truck traffic percentage	0.065	0.031	1.00
Annual Average Daily Traffic	0.058	0.026	1.00
Avg International Roughness Index (IRI inch/mile)	0.050	0.020	1.00
Total shoulder width	0.052	0.025	1.00
Imposed speed limit	0.047	0.020	1.00

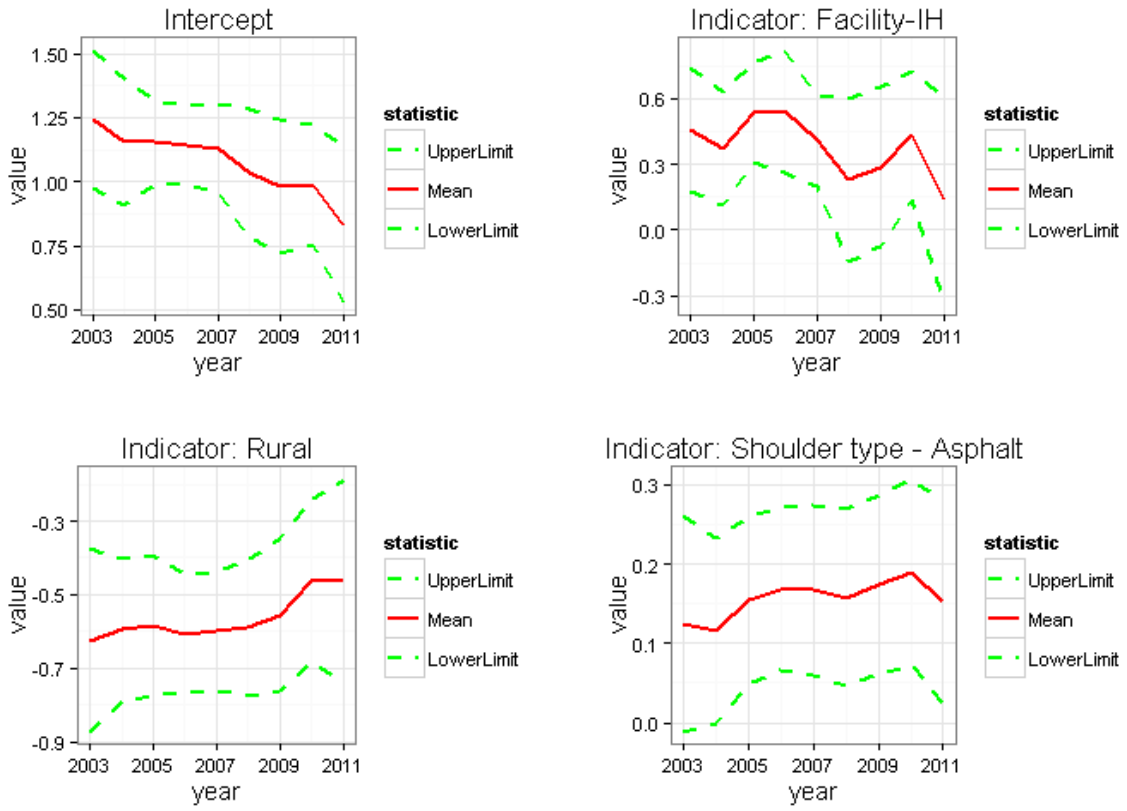

Figure 4.1: Posterior estimates of temporal parameters

4.3.3 Discussion

The model estimation results indicate that majority of the regression parameters significantly changed over time. The posterior mean of the model parameter corresponding to the Interstate Highway (IH) facility indicator indicates that the interstate highways witnessed higher mean crash count than the other facilities in each year across the study period (see Figure 4.1). The average difference in the mean crash count between the IH and the other routes had been generally decreasing over the time as shown in Figure 4.1.

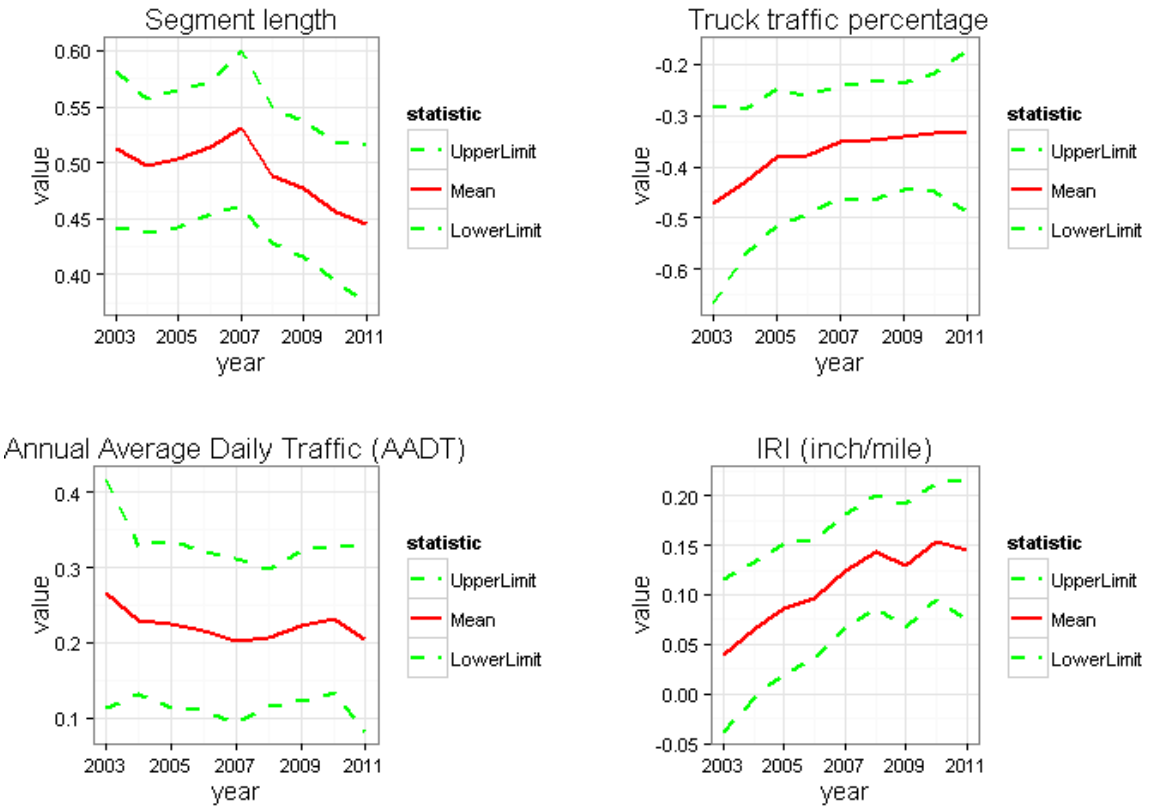

Figure 4.2: Posterior estimates of temporal parameters

Data suggests that road segments with asphalt pavements were associated with higher mean crash counts compared to that of concrete pavements, and the difference in mean crash counts between the two pavement types remained constant over the time (modeled as a fixed parameter. See Table 4.4). Posterior results suggests that road segments in rural areas witnessed lower mean crash counts compared to that of non-rural road segments, and the difference in mean crash counts between rural and non-rural road segments has decreased over the time indicating the effect of urban sprawl. The pavements

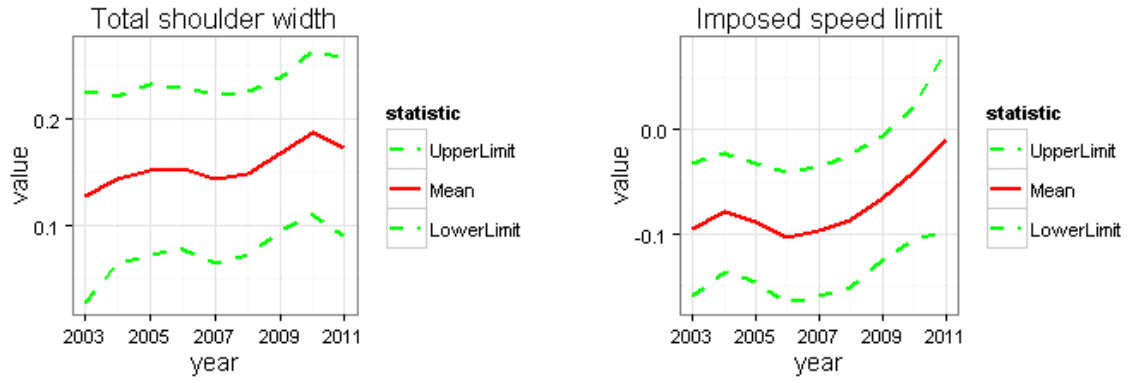

Figure 4.3: Posterior estimates of temporal parameters

with asphalt shoulders experienced larger mean crash count relative to the other types of shoulders throughout the study period; the magnitude of difference in mean crash count increased over the time as shown in the Figure 4.1. The longer segments were associated with higher mean crash counts in each year; however, the mean crash count per unit length had declined over the time as shown in the Figure 4.2. The effect of annual average daily traffic (AADT) on mean crash count remained fairly constant throughout the study period. A positive posterior mean indicates that larger traffic volumes were associated with higher crash counts. It is to be noted that the traffic volume and segment length should be viewed as exposure variables rather than causal factors. The influence of truck traffic volume on mean crash count had consistently decreased over the time; higher truck volumes were associated with lower crash counts throughout the study period for a given traffic volume.

Data suggests that the roughness of road segments was positively associated with the mean crash counts during each year throughout the study period. The influence of

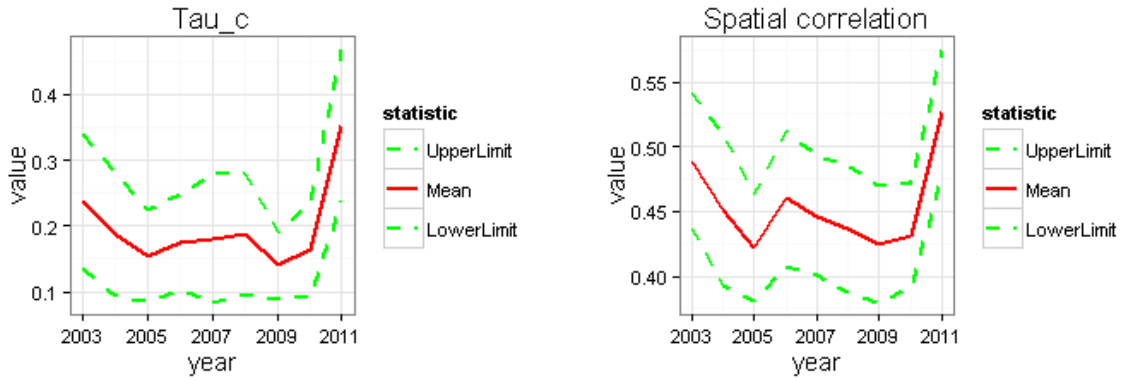

Figure 4.4: Posterior estimates of τ_c and α

the roughness on the crash count had been increasing over the time as shown in Figure 4.2. The results highlight that the road roughness is becoming more important over the time regarding the road safety. In other words, the safety benefits of improving the ride quality of road segments are increasing over the time. As shown in Figure 4.3 the influence of total shoulder width on mean crash count was also increasing over the time. The shoulder width was positively associated with the mean crash count throughout the study period. The road users may feel comfortable driving over the road segments with larger shoulder widths, thereby increasing the level of false safety. Posterior results suggest that the road segments with higher posted speed limits were associated with lower mean crash counts during the study period. The safety impact of increasing speed limit by one mile per hour is reducing over the time; in other words, the speed limit change is becoming less important within the study road network.

Figure 4.4 shows the posterior summaries of the spatial covariance parameter τ_c. As mentioned earlier, τ_c parameter does not reveal information regarding the spatial cor-

relation of crash counts of neighboring road segments. An estimate of spatial correlation was computed during the MCMC simulation, and the posterior summaries are shown in Figure 4.4. The magnitude of spatial correlation slightly varied across the time, which is expected for the Houston's road network with increasing urban population density. Overall, the empirical findings of the proposed dynamic model highlight the importance of time-evolving model parameters in crash count specifications. The posterior predictive distributions of the model parameters may be constructed, which can be utilized to predict the crash counts corresponding to a future year. However, a long crash history would be necessary to reliably estimate the AR(1) process parameters of the individual model parameters, thereby to accurately predict the model parameters corresponding to a future year.

Chapter 5

Probabilistic site ranking

This chapter introduces the concepts behind the probabilistic ranking of individual road segments, and describes a few probabilistic ranking methods employed in this research. An empirical demonstration of the proposed probabilistic ranking methods for the identification of sites with promise is also provided towards the end of this chapter.

5.1 Probabilistic ranking concepts

The main goal of a safety management framework is to identify the hazardous sites with safety concerns or sites with promise (in Ezra Hauer's words, Hauer [1996]). A naïve site ranking method may utilize the observed crash counts to identify the hazardous road segments with relatively higher crash occurrences. Such naïve ranking may be ineffective due to issues such as stochastic nature of the crash occurrence and regression to mean (RTM) bias. As mentioned in the literature review section, model-based site ranking methods offer a solution to circumvent the aforementioned issues. Model-based site ranking utilizes expected crash counts or long-run crash rates to perform site-ranking. A sound statistical model that closely represent the underlying population of crash counts from different road segments is necessary to accurately estimate the long-run crash rates. Such statistical model may be developed based on the recent historical crash data; the

model can be used for predicting the safety level of the subsequent year, and to perform probabilistic ranking of road segments.

Miaou and Song [2005] discussed the concept of decision parameter for site ranking applications. The decision parameter is defined as an estimator of site safety level such as long-run crash rates[1], and depends on the parameters of the underlying statistical model. Several decision parameters may be designed to reflect different decision mechanisms. For instance, site ranking may be performed using the deviation of long-run crash frequency of a given site from that of the peers with similar features. Decision parameter plays a vital role in the allocation of road safety funds. Selection of decision parameter is dependent on the level of risk that transportation agency is willing to undertake to reap the safety returns. A couple of different decision parameters will be introduced and compared in the next section to provide more insights into the selection of decision parameter.

As mentioned earlier, a decision parameter is an estimator, therefore a random variable. Any decision parameter is a function of model parameters; Bayesian estimation framework allows to construct the posterior distribution for any function of model parameters. The uncertainty or precision associated with an estimate of decision parameter should be incorporated into the site ranking exercise. The hazardous sections may not be accurately identified by merely comparing the corresponding posterior mean value of the decision parameter. Alternatively, road segments may be ranked during Bayesian simulation, thereby constructing a posterior distribution for the road segment ranks. A detailed description of proposed site ranking procedures is provided in the later part of

[1]An example for decision parameter.

this chapter. The proposed site ranking framework is feasible due to the implementation of Bayesian estimation framework; classical estimation procedures involving Maximum likelihood estimation do not facilitate such site ranking procedure.

5.2 Model building and decision parameters

Assume a road network with n contiguous road segments and a crash data collected over T consecutive years. Let y_{it} denote a uni-variate crash count on i^{th} road segment ($i \in \{1, 2, ...n\}$) during t^{th} year ($t \in \{1, 2, ...T\}$). We define $X_{it} = [x_{it1}, x_{it2}, ..., x_{itk}]$ as the $k \times 1$ vector of time-varying or time-invariant attributes[2] corresponding to the crash count observation y_{it}. The crash counts corresponding to the contiguous road segments are potentially spatially correlated and should be incorporated in the model specification for accurate predictions. The following model specification was selected to develop a predictive model for crash counts.

$$y_{it} \sim NB(r, p_{it}) \tag{5.1}$$

$$p_{it} = \frac{e^{\psi_{it}}}{1 + e^{\psi_{it}}}; \qquad \psi_{it} = X_{it}{}'\gamma + \phi_i$$

$$\phi_i | \phi_{-i} \sim N\left(\sum_j \frac{w_{ij}}{w_{i+}}\phi_j, \frac{\tau_c^2}{w_{i+}}\right)$$

Where, $\gamma = [\gamma_1, \gamma_2, ...\gamma_q]$ is $k \times 1$ vector of regression coefficients. ϕ_i is a spatial random effect term corresponding to i^{th} road segment; an ICAR prior structure[3] with a neighborhood-based weight matrix was assumed as a prior for spatial random effects.

[2]Time-invariant attributes are also denoted using the same notation, although the value is constant across the time for the sake of notation.

[3]The ICAR spatial prior structure was described in detail in the earlier chapters.

The model parameters are estimated using Bayesian inference methods involving Markov Chain Monte Carlo (MCMC) simulation.

We utilized historical crash data from T recent years corresponding to n road segments to build the predictive model. The model may be used to predict the safety level of road segments in the subsequent year. As mentioned earlier, the safety level of road segments is measured using any decision parameter, which is a function of the model parameters. Site ranking is largely affected by the selection of the underlying decision parameter. In order to select the appropriate decision criteria, we refer to the theoretical definition of a hazardous road location mentioned by Elvik [2008]: *"a hazardous road location is any location that (a) has a higher expected number of accidents (b) than other similar locations (c) as a result of local risk factors."*

The following decision parameters (ξ_i) may be potentially considered in site ranking: 1)the expected crash count or ψ_{it}; 2)ϕ_i; and 3)$X_{it}{}'\beta$. Site ranking based on the expected crash count or ψ_{it} suggests directing safety investments to hazardous road segments with potential to witness larger number of crashes in the subsequent year. Road segments with larger crashes are often associated with higher traffic volumes; therefore, employing ψ_i as a decision parameter may not be able to prioritize the road segments that are relatively riskier per unit traffic volume. Site ranking based on the spatial random effect ϕ_i prioritizes the road segments that are relatively riskier among the pool of other similar road segments (with similar features). In other words, the road segments with larger number of crashes for unobserved reasons over the last few years are expected to be prioritized as hazardous locations. Although employing ϕ_i as the decision criteria identifies the sites with unusual crashes, the unusual crash patterns are less likely to con-

tinue over next few years. Such uncertainty translates to increased financial risk to the highway agency to make safety investment on the identified hazardous road segments. On the other hand, the site ranking based on $X_{it}{}'\beta$ prioritizes the hazardous sites higher crash rates for a given set of local attributes on average. The chance of reaping the safety benefits may be arguably higher by selecting the hazardous sites based on $X_{it}{}'\beta$. In summary, the issue of selecting a decision criterion is analogous to classical risk-return trade-off.

5.3 Ranking methods

A decision parameter is a function of underlying model parameters. Bayesian estimation allows to construct the posterior distribution of any function of model parameters; therefore, it is possible to obtain the posterior draws for the selected decision parameter during the MCMC simulation and to construct the posterior distribution. A posterior distribution is constructed corresponding to the decision parameters of each individual road segment in the study road network. A naïve approach is to rank the posterior means of decision parameter estimates (or $E(\xi_i|y, X)$) corresponding to individual road segments to identify the hazardous road segments. However, posterior mean does not reflect the uncertainty associated with the estimated decision parameters. Miaou and Song [2005] utilized posterior mean rank to incorporate the uncertainty in estimating the decision parameter (ξ_i). The posterior draws of the decision parameter ξ_i corresponding to individual road segments are ranked during each MCMC iteration. Subsequently, a posterior distribution of site rank $R(\xi_i)$[4] is constructed for each road segment. Road

[4]Site rank may be regarded as a function of model parameters.

segments with highest posterior mean rank ($E(R(\xi_i)|y, X)$) are regarded as hazardous road segments that require safety investments.

Alternatively, the uncertainty associated with the estimates of decision parameters (ξ_i) may also be be incorporated by estimating the probability of a road segment to be most unsafe. Miaou and Song [2005] computed the probability of a site to be the worst site ($P(\xi_i > \xi_j) \forall i \neq j$) among all the sites to perform site ranking. The probability of becoming the most unsafe site corresponding to each road segment is compared to identify the hazardous road segments. However, a majority of road segment may possess a very low probability to be the most unsafe site; this may be due to the presence of road segments with considerably higher crash counts relative to the other road segments in the network. As a result, the site selection based on the probability of a site to be the worst site may yield limited number of hazardous sites. Schmidt [2012] extended the procedure to compute the probability of a site to be in top m unsafe sites ($R(\xi_i) \leq m$). The probability of a site to be in top m unsafe sites (for $m > 1$) is expected to be larger than the probability of a site to be the most unsafe site in the network. As we increase the value of m, the probability of a site to be in top m sites increases; the probability reaches the maximum value of 1 in the limiting case of setting m value equal to total number of road segments (n). Depending on the availability of safety funding, highway agency may select the value of m for the site ranking exercise.

Reasonable thresholds are necessary in conjunction with the aforementioned site ranking methods to identify an appropriate number of road segments with safety concerns. The aforementioned site ranking procedures were implemented using a computationally efficient Bayesian estimation algorithm using data augmentation methods in this

research. The proposed algorithm is expected to perform at a reasonable speed upon implementing at the network level; relatively smaller network is utilized to demonstrate the site ranking concepts in this research.

5.4 Empirical example

A road network comprising of 10 different routes in the Houston area is selected to empirically demonstrate the probabilistic ranking methods. The road network was divided into 1158 contiguous road segments of lengths ranging from 0.2 to 2.0 miles. Annual crash count data and other road feature information was obtained over four consecutive years (2007 to 2010) from crash and pavement management databases. The descriptive statistics of the road features are provided in Table 3.1; a detailed discussion on the interesting elements of the dataset is also provided in section 3.3.2.

As mentioned earlier, a reasonably accurate crash prediction model is necessary to perform model-based probabilistic site ranking. Although a multitude of road feature information was collected from the relevant databases, only a carefully selected set of road features were included in the model building process based on the following criteria. First, a road feature should be time-invariant or predictable for a future year to be included in the model as an explanatory variable; this is to ensure the availability of road feature information to incorporate into the predictive model. Second, the posterior confidence intervals or Bayesian credible intervals corresponding to a road feature should not contain zero; in other words, the regression parameters should be statistically significant. A computationally efficient data-augmentation based Bayesian estimation algorithm using MCMC simulation was employed to construct the posterior distribution

of the model parameters. A Gibbs sampling algorithm that involves sampling from full-conditional posterior distributions of the model parameters is described in Appendix D. The Appendix also includes the methodology to construct the individual site ranks for the road segments, which is embedded into the posterior simulation framework.

Table 5.1: Posterior estimation results of crash predictive model for site ranking

Description	**Posterior mean**	**Posterior Std.Dev.**	$P(\beta > 0)$
Intercept	-0.603	0.115	0.00
Indicator Variable: Facility-IH	0.973	0.157	1.00
Indicator Variable: Asphalt pavement	0.253	0.054	1.00
Segment length	0.454	0.027	1.00
Annual Average Daily Traffic	-0.299	0.044	0.00
Total shoulder width	0.061	0.031	0.96
r	6.735	0.356	1.00
τ_c	2.464	0.134	1.00
Spatial correlation estimate (α)	0.666	0.023	1.00

The probabilistic ranking is expected to be affected by the underlying crash prediction model; site ranking may not significantly change between different models producing similar predictions. A crash predictive model was developed using the appropriate explanatory variables as per the aforementioned criterion using the crash data obtained over 2007-10 period. Several specifications were estimated, and model selection was performed using deviance information criterion (DIC). Table 5.1 shows the posterior summaries of the statistically significant model parameters. Posterior probability of a coefficient to be positive ($P(\beta > 0)$)was used to determine the statistical significance of regression parameters; the values closer to 1 or 0 indicate statistical significance of the respective coefficient. The results shows an evidence for significant spatial correlation across the crash counts of neighboring road segments.

The model was utilized to predict the crashes in the year 2011 on the road segments from the training dataset. The accuracy of the prediction was assessed prior to employing the predictions in probabilistic ranking. Figure 5.1 shows the crash prediction residuals of individual road segments, which are defined as the difference between the posterior mean prediction and the observed crash count. The prediction residual is reasonably low for majority of the road segments as shown in the figure. The neighboring residuals tend to cluster in the case of spatial specification as shown in Figure 5.1. A conditional quantile plot may better reflect the prediction accuracy for an over-dispersed data. Figure 5.2 shows the conditional quantile plot, which highlights that the distributions of the predictions and the observed values are reasonably close.

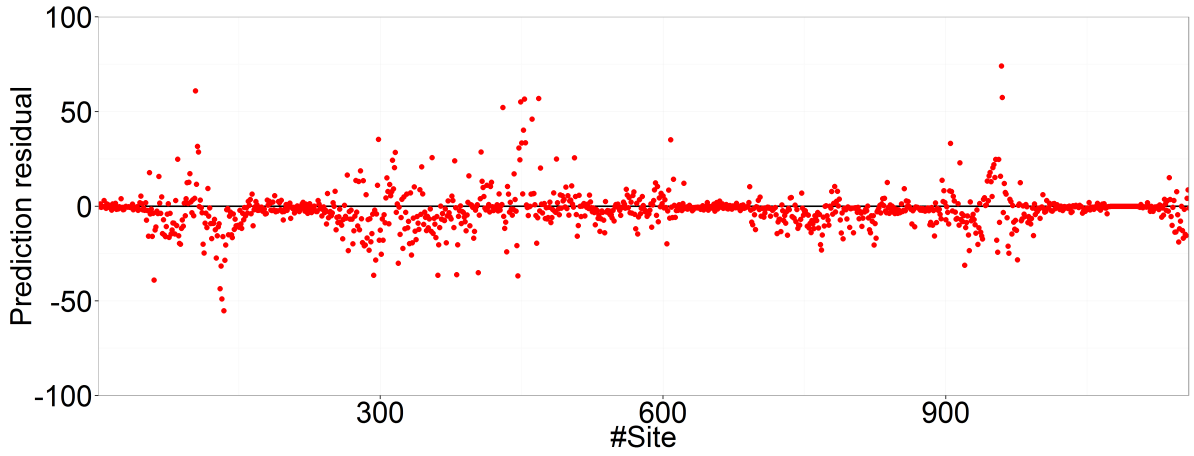

Figure 5.1: Prediction residuals

The decision parameters of individual road segments are predicted for a future year (2011) using the estimated model parameters. A probabilistic ranking is performed using the predicted decision parameters to identify the potentially hazardous road segments. We computed individual site ranks based on the expected crash rate or ψ_{it} and site-

specific unobserved spatial random effect ϕ_i. The former decision parameter represents the overall mean crash count of the respective road segment, while the later reflects the mean crash count due to unobserved factors. The probability of a given road segment to be in top m worst sites was computed for various values of m corresponding to individual road segments. Figures 5.3 & 5.4 depict the variation of the probability of a site to be in top m sites with increasing values of m for the road segments in this study; each line represents the data corresponding to a road segment. Figures 5.3 & 5.4 are probabilities calculated using expected crash counts (ψ_i) and spatial random effects (ϕ_i) respectively as the decision parameters. The probability of a road segment to be in top m sites increases with increasing values of m, and asymptotically reaches unity. The probability increases drastically and reaches the asymptotic value for a few road segments, whereas it increases at a slower rate for the other. The road segments with a steeper increase in the probability values are typically associated with larger number of observed crashes, and may be ranked top in a probabilistic site ranking.

The value of m may be selected by the highway agency depending on the availability of safety funding. For example, Figure 5.5 shows the probability of a road segment to be in top m sites for $m = 1, 5, 20, 50$ based on the expected crash counts (ψ_i) as the decision parameter, along with the observed crash counts (the bottom most figure) from the year 2011. Figure 5.6 shows a similar plot based on the site-specific spatial random effects (ϕ_i) as the decision parameter. The decision parameter plays an important role, and changing the decision parameter leads to different road segment ranking. The road segments may be selected using a threshold probability or by simply ranking the respective probabilities. For a given threshold probability of a site to be in top m sites, the

number of selected road segments increases with the increasing value of m as evident in the figures. Figure 5.5 & 5.6 also highlight that the probabilities are able to identify the crash sites with higher observed crashes reasonably well, which qualifies the probabilistic ranking as a rational approach. In other words, the magnitude of the probability to be in top m sites is higher for road segments with higher number of observed crashes.

The advantage of adopting a probability based site ranking scheme is to circumvent the need for noisy observed crash count data while performing site ranking. Moreover, the proposed approach may be adopted as a reasonable site screening tool to prioritize road segments for a future year prior to observing the crash count data. The probabilistic ranking framework may be adopted by any highway agency upon the availability of relevant data sources. It is recommended to adopt a crash predictive model that is developed using the data sources from the same highway network rather than employing any existing crash prediction models. In addition the crash prediction models needs to be updated frequently using recent historical crash databases.

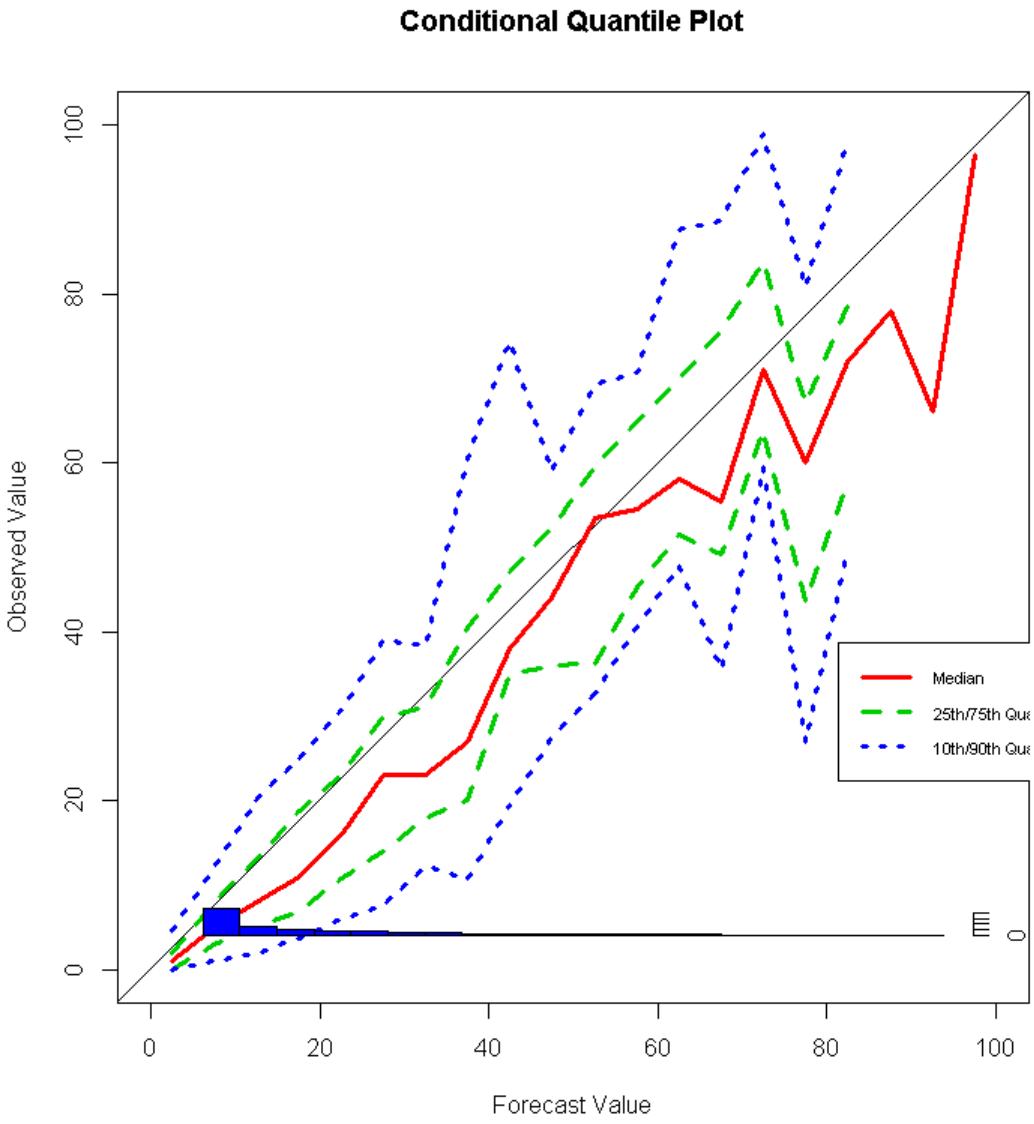

Figure 5.2: Conditional quantile plot

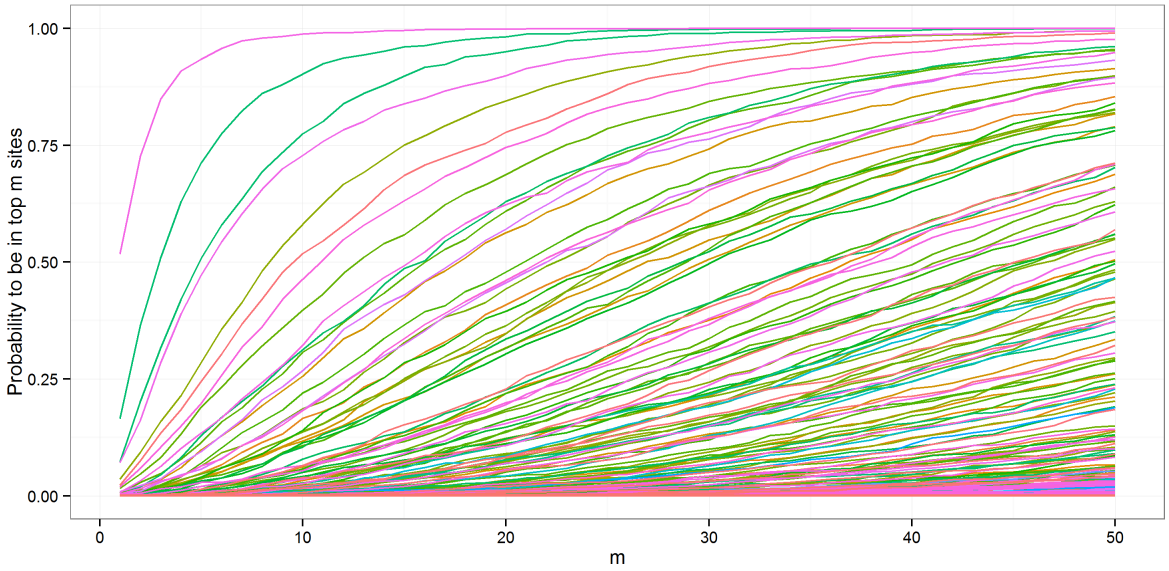

Figure 5.3: Probability of a site to be in top m sites based on expected crash count ψ_i

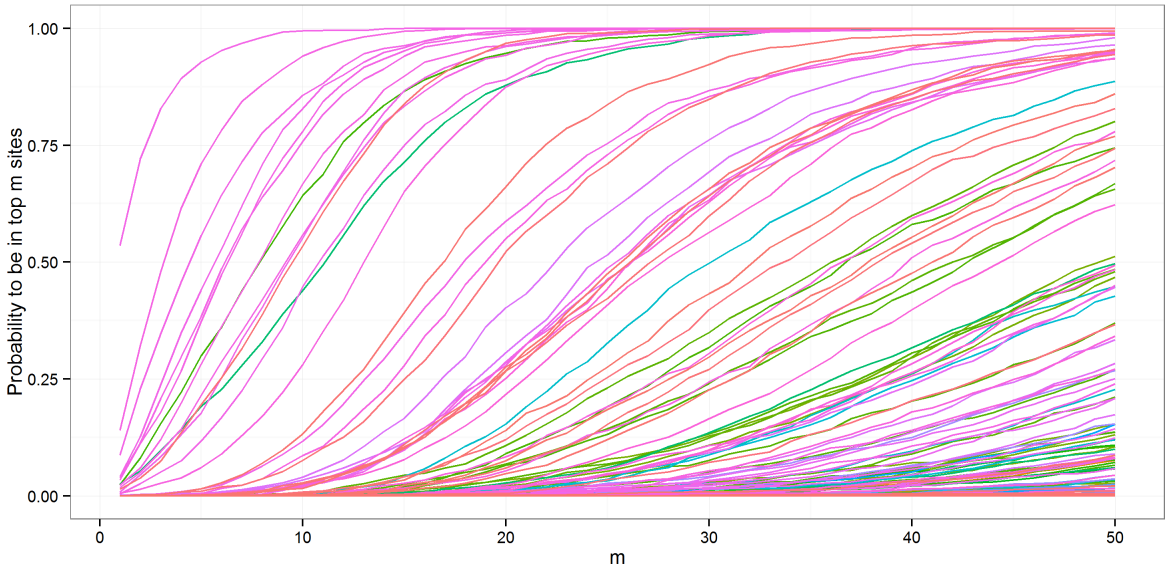

Figure 5.4: Probability of a site to be in top m sites based on spatial random effects ϕ_i

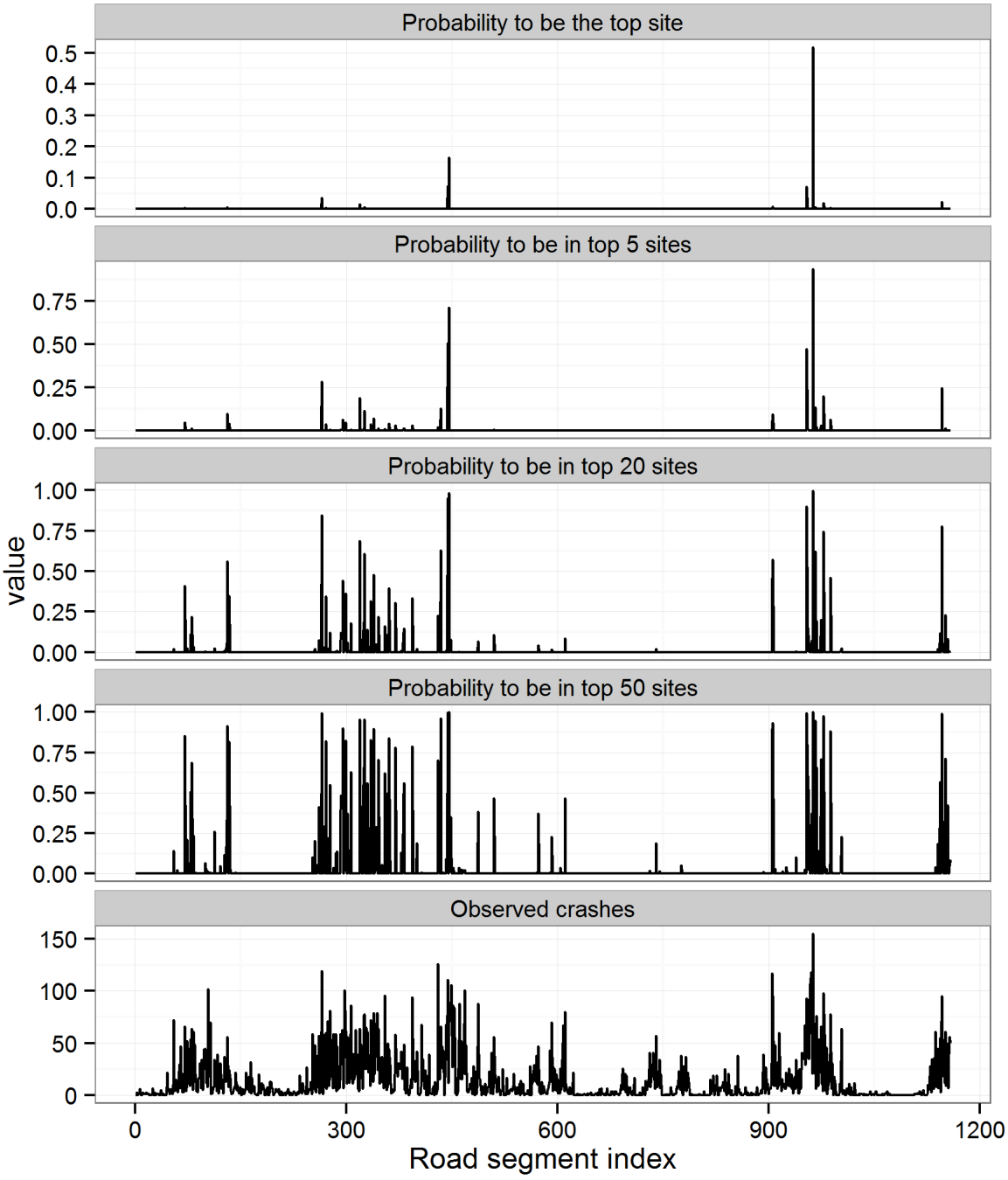

Figure 5.5: Site ranking based on expected crash counts

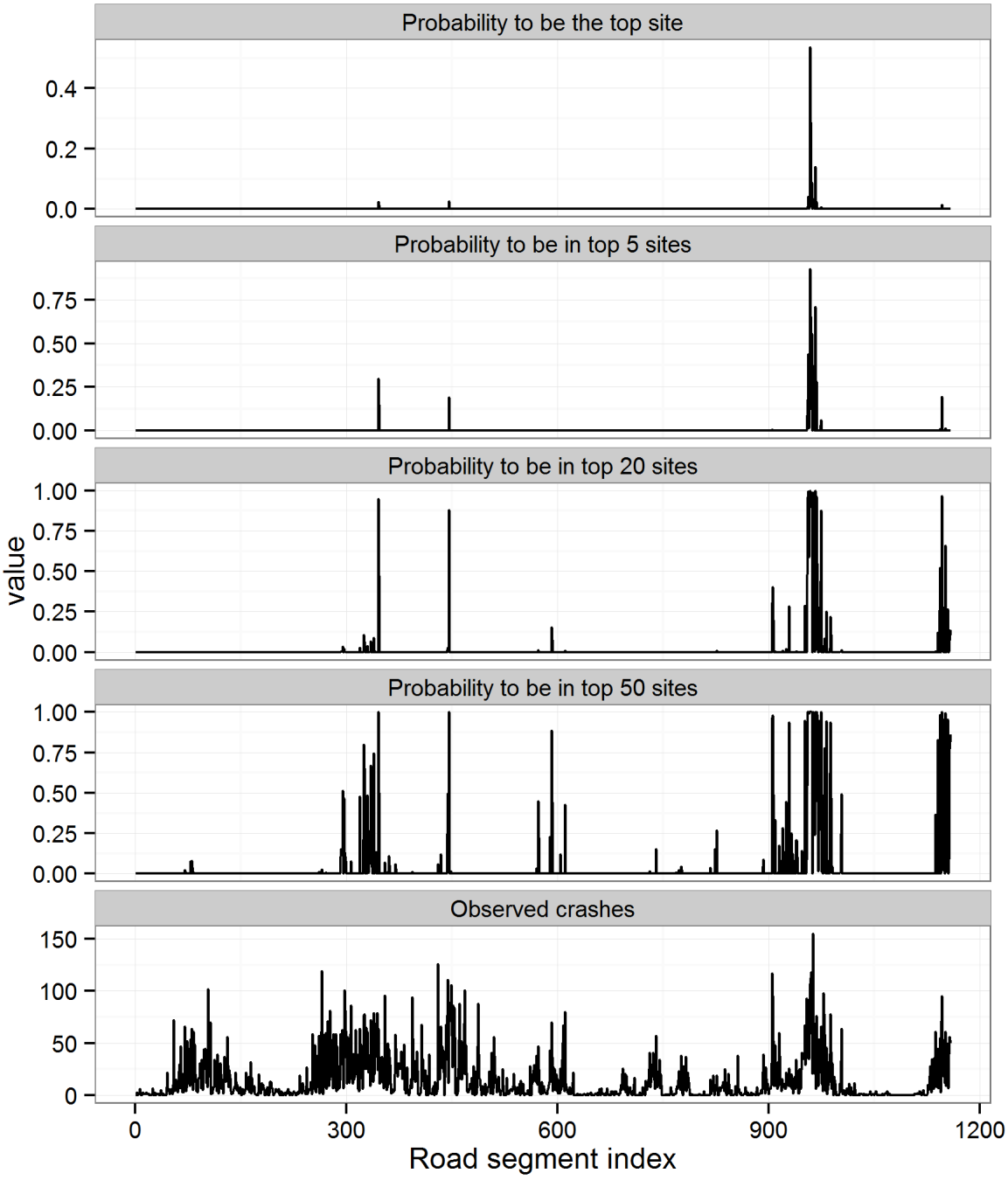

Figure 5.6: Site ranking based on spatial random effects

Chapter 6

Conclusions and future work

In this chapter the major contributions of this dissertation are documented. A summary of methodological and empirical findings at various stages of this dissertation research are provided. The limitations of the current research methodologies have been identified, and future research directions to address the limitations are provided.

The major goal of this dissertation was to contribute to the methodological realm of road safety management. The dissertation objectives are divided into two folds: 1) to develop state-of-the-art model specifications, and 2) to develop a probabilistic model-based site ranking framework. This research addresses methodological issues in crash frequency modeling (as identified by Mannering and Bhat [2014]) such as unobserved heterogeneity, spatial correlation, and temporal patterns. Two novel specifications were developed to address these methodological issues: 1) random parameter spatial negative binomial count model with finite mixtures; 2)spatial negative binomial count model with dynamic parameters. Computationally efficient Bayesian estimation frameworks that leverage recent advances in data augmentation techniques were developed to estimate the proposed count specifications. Bayesian estimation methods facilitate statistical inference of site ranking. A computationally efficient site ranking framework was developed incorporating the recent probabilistic ranking techniques towards the end of this

dissertation.

Empirical examples were constructed to demonstrate the applicability of the proposed specifications in the road safety management. State governments maintain historical crash information that typically includes characteristics of the involved vehicles, people and the respective crash sites. Pavement information such as extent of surface distresses, ride quality, traffic volumes, geometric features, etc. is also collected and documented across the network. Database applications that can effectively merge the crash and pavement information facilitate the incorporation of safety into annual road management programs and assist the agency in project prioritization. In this research, crash data is sourced from the motor vehicle Crash Record Information System (CRIS) database maintained by TxDOT, and road information was obtained from TxDOTs Pavement Management Information System (PMIS). An integrated database was developed by merging the crash data with road condition and feature information. The empirical analysis was performed using the integrated database containing records from 11 different routes in Houston area spanning across 9 years. This dissertation emphasizes on the harmony across several data sources, and encourages exploiting the benefits of integrated databases.

6.1 Conclusions

6.1.1 Modeling unobserved heterogeneity

In discrete count data modeling literature, unobserved heterogeneity has been handled by modeling the parameters of the underlying count specification using random parameters or finite-mixtures models(for example, see Park and Lord [2009], Anasta-

sopoulos and Mannering [2009]). Recently, Xiong and Mannering [2013] proposed a finite-mixture random parameters approach for modeling unobserved heterogeneity in crash severity modeling. The model parameters were modeled as finite multivariate normal mixture distribution; this allows for non-normality, skewness in the distribution of the random model parameters, facilitates correlation across the model parameters, and relaxes any distributional assumptions. The specification allows to investigate the underlying groups of observations (as in the case of a finite-mixture model), while allowing for within group heterogeneity. Such model specifications are not available to model the unobserved heterogeneity for discrete count data; this may be due to the computational difficulties associated with the Bayesian model estimation algorithms for Poisson or negative binomial likelihoods. This dissertation addresses such computational difficulties associated with the Bayesian estimation of count specifications (with negative binomial likelihoods) by employing recently developed polya-gamma data-augmentation technique. A negative binomial specification with finite-mixture random parameters that simultaneously accounts for potential spatial correlation was developed. A Gibbs sampling framework was developed and full conditional posterior distributions of the model parameters were derived. The proposed Bayesian estimation framework circumvents the need for Metropolis-Hastings (M-H) algorithm while performing Markov Chain Monte Carlo (MCMC) simulations. The Gibbs sampling algorithm was coded in R language [R Core Team, 2013] using Rcpp package (a C++ extension; [Eddelbuettel and Franois, 2011]) to enhance the computational speed of the posterior estimation. Simulation experiments were conducted to ensure the accuracy of posterior estimation algorithm in terms of retrieving true model parameters.

An empirical case study was developed to demonstrate the proposed specification utilizing crash data spanning across 2007-10. The dataset comprises of crash counts from contiguous road segments from different routes in the Houston area. A two-component finite-mixture model was deemed sufficient for the dataset based on the model selection criterion. The posterior estimation results suggested the presence of two underlying latent sub-groups of road segments. Traffic volumes, road type and condition, shoulder type and width, and imposed speed limit were identified as significant predictors. The influence of a few explanatory variables on crash counts varied across the sub-groups or components. For instance, larger shoulder widths were associated with larger crashes in a sub-group of road segments on an average, while the smaller shoulder widths were dangerous in the other sub-group. The empirical results demonstrate the possibility of varying safety outcomes as a result of similar treatments (such as improving shoulder width). On the other hand, a few variables influence the crash counts in a similar fashion across both latent sub-groups of road segments. Better road condition was always associated with the lower crash counts, which highlights the safety benefits of a well-maintained road network. A significant spatial correlation was evident emphasizing the need to simultaneously account for the spatial correlation across the neighboring road segments. In summary, the empirical results highlight the need to understand the extent of unobserved heterogeneity associated with a road feature prior to selecting the relevant safety treatment. Highway agencies should seek the treatable road features with relatively lower variability in terms of its influence (or lower unobserved heterogeneity on the relevant model parameter) on the crash outcomes.

6.1.2 Modeling temporal patterns

Crash data is typically collected annually and potentially temporally correlated. Many earlier research studies have accounted for such temporal correlation using auto-regressive prior structures (for example, see Miaou and Song [2005]). The explanatory variables such as road features are temporally changing along with crash count data. Moreover, the influence of the time-varying or time-invariant road features may arguably change over the time. Such scenarios may be modeled using dynamic models that allow for time-varying parameters. A few earlier studies have exploited the benefits of dynamic count specifications for modeling crash count data with temporally evolving road features. For example, Hu et al. [2013] developed dynamic time-series negative binomial regression models and estimated using Integrated Nested Laplace Approximation (INLA) technique. The study is one among few studies that attempted to address the computational difficulties associated with Bayesian inference of count data models. This study utilized traffic volume as the only explanatory variable with dynamic parameter and estimated several such independent models for different scenarios. In this dissertation, a flexible dynamic negative binomial count specification that simultaneously account for temporally varying spatial correlation was developed. The proposed specification allows for both time-varying and time-invariant parameters. A dynamic linear model (DLM) formulation of the discrete negative binomial count model was employed by exploiting recently developed polya-gamma data-augmentation techniques. A Gibbs sampling framework was developed that involves the Forward Filtering and Backward Sampling (FFBS) algorithm for performing posterior inference on dynamic parameters. Full conditional distributions were derived for all the model parameters, and computationally

efficient sampling and matrix algebra techniques were employed to perform the MCMC simulations within reasonable estimation time. The accuracy of retrieving true model parameters was ensured using simulation experiments of the proposed posterior estimation algorithm. The algorithm was coded in coded in R language [R Core Team, 2013] using Rcpp package (a C++ extension; [Eddelbuettel and Franois, 2011]) to enhance the computational speed of the posterior estimation.

The proposed negative binomial spatial model was estimated using 9 years of historical crash dataset (2003 to 2011). Both time-varying and time-invariant road features were identified from the PMIS database and incorporated into the specification. A majority of the parameters corresponding to both time-varying and time-invariant explanatory variables exhibited significant temporal patterns. For example, the magnitude of positive parameter corresponding to road roughness was increasing over the time, which suggests the increased importance of maintaining superior ride quality over the time. In other words, the safety benefits of a unit improvement in the ride quality were increasing over the time. Posterior results suggest that the number of crashes per unit road segment length were decreasing over the time within the study period. The safety impact of increasing speed limit by one mile per hour was reducing over the time; in other words, the speed limit change was becoming less important within the study road network. The safety benefits of altering shoulder width were also reducing over the time. The mean difference in the expected crash counts between the IH facilities and the other routes was decreasing over the time. Possible increase in road congestion over the time could be a potential reason for reduction in the influence of several important road features. The empirical findings also suggest that the variables that influence crash counts of road

segments may not remain the same over the time. In summary, the proposed modeling framework facilitates to study the evolution of the relationship between the road features and the crash counts.

6.1.3 Probabilistic site ranking

Road safety management requires identification of the hazardous road segments or sites with promise for safety benefits to invest the highway safety funds. Observed crash counts of road segments are typically noisy, and safety ranking of road segments using the observed data is arguably not reliable. A model-based site ranking approach sidesteps the issue by utilizing the expected crash count instead of observed crash data. The raw comparison of the expected crash counts of road segments does not account for the uncertainty associated with the model estimation. This dissertation explored several probabilistic ranking methods that account for the uncertainty associated with the model estimates. Probability of a site to be in top m sites was computed for the study road network and hazardous road segments were identified.

The probabilistic ranking was performed on an example road network of 11 different routes in the Houston area. A crash prediction model was developed using crash data and road feature information from years 2007-10. Crash counts of the road segments were predicted for the subsequent year using the estimated model parameters. A probabilistic ranking was performed on the predicted crash counts and compared with the observed data (but unused in during the model estimation) to evaluate the accuracy of the probabilistic ranking framework. The results indicate that the road segments with larger observed crash counts exhibited relatively higher probability to be in top m

sites, which emphasizes the rationality of the proposed probabilistic ranking approach. The spatial random effects were also utilized to perform site ranking instead of expected crash counts as the decision parameter. It was found that the decision parameter plays an important role and significantly affects the road segment ranking.

6.2 Future work

This dissertation sheds light into several lines of research, particularly in the methodological aspects of crash data modeling. For instance, the proposed finite-mixture random parameters model assumed the knowledge of number of mixture components. The number of mixture components (k) was selected based on model selection using Deviance Information Criterion (DIC). However, the proposed Bayesian estimation framework allows to treat k as a model parameter. Posterior inference may be performed on the model parameter k using reversible jump MCMC technique, which involve Metropolis-Hastings algorithm. Perhaps, a larger and informative datasets would be necessary to reliably learn the number of mixture components (k). The temporal evolution of unobserved heterogeneity may be studied by allowing for dynamic parameters in the finite-mixture random parameters model.

The accuracy of the estimated auto-regressive model parameters may be improved by using a sufficiently large and informative temporal data. The temporal auto-regressive process parameters may be learned by utilizing crash data from several years or splitting the existing crash counts and modeling monthly counts. The dynamic negative binomial model parameters may be utilized to generate the predicted model parameters corresponding to a future year. A crash prediction model will thus be formulated us-

ing the predicted model parameters for the future year. Probabilistic ranking may be performed using the crash count prediction of the model structure corresponding to the future years. Crash counts may be categorized according to the severity level, which recognizes the multi-variate nature of the crash count data. The proposed uni-variate models may be extended to multi-variate settings to incorporate correlation across the cash counts of different severity levels. Subsequently, a safety index may be designed as a weighted combination of predicted crash counts (per unit exposure or traffic) of multiple severity levels on a given road segment at any given time; the weights may be selected based on relative economic losses associated with the corresponding severity level. Probabilistic ranking may be performed on such weighted safety index to better reflect the economics associated with different crash severity levels. This dissertation demonstrates the feasibility of designing Bayesian modeling frameworks for probabilistic roadway safety management, which facilitate online learning. The proposed index may be designed as a self-updating index and the predictions are expected to get better with time as it accumulates more data under Bayesian framework.

A cost allocation framework that is compatible with the proposed probabilistic ranking framework may be developed and deployed to effectively allocate the road safety funds. An optimization problem may be formulated to design such cost allocation framework. The research ideas presented in this dissertation may be extended to bigger networks to test the feasibility of developing a safety management framework that automatically learns from the latest crash data sources over the time. A feasibility study to evaluate the possibility of adapting the proposed probabilistic ranking framework to a larger network should be performed as a future research.

Appendices

Appendix A

A.1 Polya-Gamma data augmentation

A random variable ω has a PG(b,c) distribution (Polya-Gamma distribution with parameters b and c) if

$$\omega \overset{D}{=} \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k-1/2)^2 + c^2/(4\pi^2)} \tag{A.1}$$

Where, $g_1, g_2, ...g_k, ...$ are independent random variables and identically distributed with $Gamma(b, 1)$.

The density of the Polya-Gamma random variables (see Polson et al. [2013] for other distributional properties) is shown in Equation (A.2).

$$f(x|b, c) = \{cosh^b(c/2)\} \frac{2^{b-1}}{\Gamma(b)} \sum_{n=0}^{\infty} (-1)^n \frac{\Gamma(n+b)}{\Gamma(n+1)} \frac{(2n+b)}{\sqrt{2\pi x^3}} e^{-\frac{(2n+b)^2}{8x}} \tag{A.2}$$

The Polya-Gamma distributed random variables can be generated from an infinite sum of weighted $i.i.d$ Gamma distributed random variables. We refer the interested readers to Polson et al. [2013] for details on the sampling methods and efficiency of drawing Polya-Gamma random variables.

Mathematically, the negative binomial likelihood parametrized by log-odds can

be written as,

$$\begin{aligned} P(y_i|\psi_i, r) &= \frac{\Gamma(y_i + r)}{\Gamma(r)y_i!} \frac{(e^{\psi_i})^{y_i}}{(1 + e^{\psi_i})^{r+y_i}} \\ &= \frac{\Gamma(y_i + r)}{\Gamma(r)y_i!} 2^{-(r+y_i)} e^{(y_i - r)\psi_i/2} \int_0^\infty e^{-\omega_i \psi_i^2/2} p(\omega_i) d\omega_i \\ &= \frac{\Gamma(y_i + r)}{\Gamma(r)y_i!} 2^{-(r+y_i)} e^{(y_i - r)\psi_i/2} E_{\omega_i}[e^{-\omega_i \psi_i^2/2}] \end{aligned} \tag{A.3}$$

where, $\psi_i = x_i^T \beta + \phi_i$ and $\omega_i \sim PG(y_i + r, 0)$

After algebraical rearrangement of terms, the negative binomial likelihood (a function of β) becomes Gaussian likelihood conditional on the Polya-Gamma random variable ω_i, r and ψ_i as shown below.

$$\begin{aligned} P(y|\psi, r, w) &= \prod_{i=1}^{n} P(y_i|\psi_i, r, w_i) \\ &= \prod_{i=1}^{n} \frac{\Gamma(y_i + r)}{\Gamma(r)y_i!} 2^{-(r+y_i)} e^{\frac{(y_i - r)}{2}\psi_i} e^{-\frac{\omega_i \psi_i^2}{2}} \\ &= K e^{-\frac{1}{2}(z-\psi)^T \Omega (z-\psi)} \end{aligned}$$

Where, $z_i = \frac{y_i - r}{2\omega_i}$ for $i \in \{1, 2, ...n\}$ and the constant K is independent of the ψ_i that contains the regression coefficients. It is to be noted that we pretend z_i is observed instead of y_i. Thus, we obtained a Gaussian likelihood form of crash counts in terms of z_i.

A.2 Compound Poisson representation

Negative binomial random variables can be expressed as sums of Logarithmic random variables under compound Poisson distribution [Quenouille, 1949].

$$y_i = \sum_{k=1}^{L_i} u_k, \qquad u_k \overset{iid}{\sim} Logarithmic(p_i), \qquad L_i \sim Poisson(-r \ln(1 - p_i)) \tag{A.4}$$

Where, $i \in \{1, 2, ...n\}$. The probability density of the Logarithmic distributed random variable u, is given by

$$f_U(U = t) = -\frac{p^t}{t\ln(1-p)}, t \in \{1, 2, 3, ...\}$$

Appendix B

Gibbs sampling for Scenario - I

A Gibbs sampling algorithm for posterior inference of the parameters of the proposed model structure is provided below. Let y denote a $nT \times 1$ vector of crash counts, X^F denote a $nT \times p$ matrix, and X^R denote a $nT \times q$ matrix.

Posterior sampling of dispersion parameter (r):

We need full conditional distributions of $r|L_{it}$ and $L_{it}|r$ for inference of the fixed parameter vector r. First, $r|L_{it}$ is derived.

$$\begin{aligned} P(r|L,p,-) &\propto \prod_{i=1}^{n}\prod_{t=1}^{T} P(L_{it}|r)P(r) \\ &\propto \prod_{i=1}^{n}\prod_{t=1}^{T} Poisson(-r\ln(1-p_{it}))Gamma(r_0,h) \\ &\sim Gamma\left(r_0+\sum_{i=1}^{N}\sum_{t=1}^{T} L_{it}, h-\sum_{i=1}^{N}\sum_{t=1}^{T} ln(1-p_{it})\right) \end{aligned} \tag{B.1}$$

Hyper parameter h is also learned by constructing the full conditional posterior.

$$\begin{aligned} p(h|r,-) &\propto P(r|h,-)p(h) \\ &\sim Gamma(r_0+ha_0, r+hb_0) \end{aligned} \tag{B.2}$$

Now, $L_{it}|r$ is to be derived. Zhou et al. [2012] derived a closed form expression for the conditional posterior of the L_i. The likelihood of observing y_i given a particular

realization of L_i is constructed using Probability Generating Function (PGF) of the Logarithmic distribution, which is combined with the aforesaid Poisson prior with respect to the L_i. We refer interested reader to Zhou et al. [2012] for a detailed derivation of the conditional posterior of L_i. The analytical closed form expression for the posterior of discrete random variable L_i is written in a matrix form as shown below.

$$P(L_{it} = j|r, y) = R(y_{it}, j), \qquad j \in \{0, 1, ...y_{it}\} \tag{B.3}$$

$$R(l,m) = \begin{cases} 1 & l = 0; m = 0 \\ F(l,m)r^m \Big/ \sum_{j=1}^{l} F(l,j)r^j & l \neq 0; m \neq 0 \end{cases}$$

$$F(m,j) = \begin{cases} 1 & m = 1 \& j = 1 \\ 0 & m < j \\ \frac{(m-1)}{m}F(m-1,j) + \frac{1}{m}F(m-1,j-1) & 1 \leq j \leq m \end{cases}$$

Posterior sampling of ($\beta_{1:N}, \mu_{1:C}, \gamma$):

The Polya-Gamma data augmentation allows to transform the negative binomial likelihood to a Gaussian likelihood (see Appendix A). The joint conditional posterior is constructed using the conditionally Gaussian likelihood using the respective conjugate prior structures. The crash counts y_i are transformed to $z_{it} = \frac{y_{it}-r}{2\omega_{it}}$ as described in Appendix A. It is to be noted that we pretend z_{it} is observed instead of y_i. Thus, we obtained a Gaussian likelihood form of crash counts in terms of z_i. The random parameters, component specific means, and fixed parameters are sampled jointly, or in blocks, to improve the mixing of MCMC chain and to accelerate the convergence. The blocked sampling is performed in two steps as follows.

Goal: $p(\beta_{1:N}, \gamma, \mu_{1:C}, \omega|, -)$

We iterate between $p(\beta_{1:N}, \gamma, \mu_{1:C}|\omega, z^N, -)$ and $p(\omega|z^N, -)$. First we derive, $p(\beta_{1:N}, \gamma, \mu_{1:C}|\omega, z^N, -)$

as below.

$p(\beta_{1:N}, \gamma, \mu_{1:C}|\omega, z^N, -) \propto \prod_{i=1}^{n} p(\beta_i|\gamma, \mu_{1:C}, z^N, -)p(\gamma, \mu_{1:C}|z^N, -)$

- I: $p(\beta_{1:N}|\gamma, \mu_{1:C}, z^N, -)$
- II: $p(\gamma, \mu_{1:C}|z^N, -)$

I: Sampling $p(\beta_{1:N}|\gamma, \mu_{1:C}, z^N, -)$*:*

$$
\begin{aligned}
p(\beta_{1:N}|\gamma, \mu_{1:C}, z^N, -) &\propto \prod_{i=1}^{n} \mathrm{p}(\beta_i|\gamma, \mu_{1:C}, z^N, -) \\
P(\beta_i|y, X^F, X^R, \phi, r, \omega, G) &\propto \prod_{t=1}^{T} P(y_{it}|\psi_{it}, r, w_{it}, G_i)P(\beta_i) \\
&\propto exp\left(-\frac{1}{2}(z_i - \psi_i)^T \Omega((z_i - \psi_i)\right) * \\
&\quad exp\left(\prod_{c=1}^{C}\left(-\frac{1}{2}(\beta_i - \mu_c)^T \Sigma_c(\beta_i - \mu_c)\right)^{I(G_i=c)}\right) \\
&\propto Normal(m_{\beta_i}, V_{\beta_i})
\end{aligned}
\tag{B.4}
$$

$m_{\beta_i} = V_{\beta_i}(X_i^{R'}\Omega_i(z_i - X_i^{F'}\gamma - \phi_i) + \prod_{c=1}^{C}(\Sigma_c^{-1}\mu_c)^{I(G_i=c)}); V_{\beta_i} = (X_i^{R'}\Omega_i X_i^R + \prod_{c=1}^{C}(\Sigma_c^{-1})^{I(G_i=c)})$

where, z^N is vector of transformed crash counts. X_i^F is a $T \times p$ matrix;X_i^R is a $T \times q$ matrix; Ω is a $T \times T$ diagonal matrix; $\Omega_{pp} = \omega_{it}$ for $t \in \{1, 2, ...T\}$; z_i, ψ_i, and ϕ_i are $T \times 1$ vectors.

II: Sampling $p(\gamma, \mu_{1:C}|z^N, -)$*:*

To construct the conditional posterior $p(\gamma, \mu_{1:C}|z^N, -)$, the random parameters are marginalized as follows. Transformed likelihood is

$$z_i = X_i^{F'}\gamma + X_i^{R'}\beta_i + \phi_i + \epsilon_i; \epsilon_i \sim N(0, diag(1/\omega_{it}))$$

$$\delta_i = X_i^{F'}\gamma + \sum_{c=1}^{C} X_i^{R'} D_i^{(c)} \mu_c + \epsilon_i^*$$

Where, $\epsilon_i^* \sim N(0, V_i)$ and $V_i = \Omega_i + X_i^R \Sigma_{G_i} X_i^{R'}$

$$p(\gamma, \mu_{1:C} | z^N, -) \propto \prod_{c=1}^{C} p(\mu_c | \gamma, z^N, -) p(\gamma | z^N, -)$$

- II-1: $p(\mu_{1:C} | \gamma, z^N, -)$
- II-2: $p(\gamma | z^N, -)$

II-1: sampling from $p(\mu_{1:C} | \gamma, z^N, -)$

$$p(\mu_{1:C} | \gamma, z^N, -) \propto \prod_{c=1}^{C} \mathrm{p}(\mu_c | \gamma, z^N, -)$$

$$P(\mu_c | \gamma, z^N, -) \sim Normal(m_{\mu_c}, V_{\mu_c}) \tag{B.5}$$

$$m_{\mu_c}(\gamma) = V_{\mu_c} \left(\sum_{i=1}^{n} X_i^{R'} V_i^{-1} D_i^{(c)} (z_i - \phi_i - X_{it}^{F'} \gamma) + B_0^{-1} b_0 \right)$$

$$V_{\mu_c} = \left(\sum_{i=1}^{n} X_i^{R'} V_i^{-1} X_i^R D_i^{(c)} + B_0^{-1} \right)^{-1}$$

II-2: sampling from $p(\gamma | z^N, -)$

$$P(\gamma | z^N, -) \sim Normal(m_\gamma, V_\gamma) \tag{B.6}$$

$$m_\gamma = V_\gamma \left(\sum_{i=1}^{n} X_i^{F'} V_i^{-1} (z_i - \phi_i - X_i^R m_{\mu_{G_i}}(0)) + G_0^{-1} g_0 \right)$$

$$V_\gamma = \left(\sum_{i=1}^{n} X_i^{F'} V_i^{-1} (V_i - X_i^R V_{\mu_{G_i}} X_i^{R'}) V_i^{-1} X_i^F + G_0^{-1} \right)^{-1}$$

Second, the full conditional distribution $p(\omega|z^N, -)$ turns out to be a distribution in the Polya-Gamma class (see Polson et al. [2013] for derivations).

$$P(\omega_{it}|\gamma, r, y, X^F, X^R, \phi) \propto PG(y_{it} + r, \psi_{it}) \tag{B.7}$$

where, $PG(b, c)$ represents the Polya-Gamma distribution with the density indicated in Equation (A.2).

Posterior sampling of component specific covariance:

The following full conditional distribution is used to iteratively draw the component specific covariance matrices $\Sigma_1, \Sigma_2, ...\Sigma_C$.

$$\begin{aligned} P(\Sigma_c|G, \beta, \mu_c) &\propto \prod_{i=1}^{n} P(\beta_i|\mu_c, \Sigma_c, G_i) P(\Sigma_c) \\ &\propto Wish(\nu_{\Sigma_c}, V_{\Sigma_c}) \end{aligned} \tag{B.8}$$

$$\nu_{\Sigma_c} = \nu_0 + \sum_{i=1}^{n} S_c; V_{\Sigma_c} = \left(V_0^{-1} + \sum_{i=1}^{n} I(G_i = c)(\beta_i - \mu_c)(\beta_i - \mu_c)^T \right)^{-1}$$

where, β is $n \times q$ random parameter matrix; S_c is $n \times 1$ vector and $S_c = I(G = c)$;

Posterior sampling of latent indicator vector (G_i):

We assume a non-informative categorical prior on the latent indicator vector, which results in a categorical posterior. The support of the categorical variable $G_i = c : c \in \{1, 2....., C\}$. We iteratively generate n posterior component indicators using the following full conditional distribution.

$$P(G_i = c|z_i, \gamma, \mu_c, \Sigma_c, -) \propto \phi_q(z_i; m_{G_i}, V_{G_i})\eta_c$$

Where, $m_{G_i} = X_i^F\gamma + X_i^R\mu_c + \phi_i$ and $V_{G_i} = X_i^R\Sigma_c X_i^{R'} + \Omega_i$ and $\Omega_i = diag(1/\omega_{it})$

$$P(G_i = c|\beta, \mu_c, \Sigma_c) = \frac{\phi_q(z_i; m_{G_i}, V_{G_i})\eta_c}{\sum_{c=1}^{C} \phi_q(z_i; m_{G_i}, V_{G_i})\eta_c}$$

where, $\phi_q(.)$ is a q-dimensional multivariate Gaussian density.

The mixture weight vector $\eta = [\eta_1, \eta_2, ..., \eta_C]$ are drawn using the following full conditional distribution.

$$\begin{aligned} P(\eta|G) &\propto P(G|\eta)P(\eta) \\ &\propto Dirichilet(\alpha_0 + g_1, \alpha_0 + g_2, ..., \alpha_0 + g_C) \end{aligned} \tag{B.9}$$

where, $g_c = \sum_{i=1}^{n}(G_i = c)$.

Posterior sampling of spatial random effects (ϕ_i):

$$\begin{aligned} \mathrm{p}(\phi_i|\phi_{-i}, z_i, -) &\propto \prod_{t=1}^{T} P(z_{it}|\phi_{-i}, -)P(\phi_i|\phi_{-i}) \\ &\propto N(m_\phi, V_\phi) \end{aligned} \tag{B.10}$$

$$m_\phi = \left(\sum_{t=1}^{T}(z_{it} - X_{it}^{F'}\gamma + X_{it}^{R'}\beta_i)\omega_{it} + \left(\sum_j \frac{w_{ij}}{w_{i+}}\phi_j\right)\frac{w_{i+}}{\tau_c^2}\right)V_\phi$$

$$V_\phi = \left(\sum_{t=1}^{T}\omega_{it} + \frac{w_{i+}}{\tau_c^2}\right)^{-1}$$

Mean centering: $\phi = \phi - \bar{\phi}$

$$p(1/\tau_c^2|\phi, -) \sim Gamma\left(c_0 + \frac{n}{2}, d_0 + \sum_{i=1}^{n}\frac{w_{i+}}{2}\left(\phi_i - \sum_j b_{ij}\phi_j\right)^2\right) \tag{B.11}$$

An MCMC simulation is performed by iteratively drawing posterior samples using equations C.4 through B.11. To avoid potential label switching, the latent component labels are reassigned at the end of each iterations using the constraint: $\mu_{1k} < \mu_{2k} < ..... < \mu_{Ck}$; where, μ_{ck} is the k^{th} element of the c^{th} component specific mean. A random parameter (denoted by k) with considerably higher variation and clearly defined components is identified using unconstrained sampling and utilized for re-labeling.

Appendix C

Gibbs sampling for Scenario - II

Gibbs sampling scheme:

A Gibbs sampling scheme to iteratively draw from full conditional distributions of individual model parameters is provided.

Posterior sampling of fixed regression coefficients (β):

Let D represent the total data (including $\{y_{it}\}$ and $\{X_{it}\}$) across different time periods.

Step 1: Sample β

$$\begin{aligned} P(\beta|\theta_{1:T}, D, \phi, r, \omega) &\propto \prod_{t=1}^{T}\prod_{i=1}^{n} P(y_{it}|\psi_{it}, r, w_{it})P(\beta) \\ &\propto exp\left(-\frac{1}{2}(z-\Lambda-\phi)-Z\beta)^T\Omega((z-\Lambda-\phi)-Z\beta)\right) * \\ &\quad exp\left(-\frac{1}{2}(\beta-b_0)^T B_0(\beta-b_0)\right) \\ &\propto Normal(m_\beta, V_\beta) \end{aligned} \tag{C.1}$$

$$m_\beta = V_\beta(Z^T\Omega(z-\Lambda-\phi)+B_0^{-1}b_0) \qquad V_\beta = (Z^T\Omega Z + B_0^{-1})^{-1}$$

Where, $z_{it} = \frac{y_{it}-r}{2\omega_{it}}$ and $\Lambda_{it} = \theta_t X_{it}$

Step 2: Sample ω

$$P(\omega_{it}|\beta, r, y, x, \phi) \propto PG(y_{it}+r, \psi_{it}), \qquad i \in \{1, 2, ...n\} \tag{C.2}$$

Posterior sampling of dispersion parameter (r):

Step 3: Sample r

$$\begin{aligned} P(r|L,p,-) &\propto \prod_{t=1}^{T}\prod_{i=1}^{n} P(L_{it}|r)P(r) \\ &\sim Gamma\left(r_0 + \sum_{t=1}^{T}\sum_{i=1}^{N} L_{it}, h - \sum_{t=1}^{T}\sum_{i=1}^{N} ln(1-p_{it})\right) \end{aligned} \tag{C.3}$$

Step 4: Sample L

$$P(L_{it} = j|r,y,x,\psi,\omega) = R(y_{it},j), \qquad j \in 0,1,...y_{it} \tag{C.4}$$

Where,

$$R(l,m) = \begin{cases} 1 & l=0; m=0 \\ F(l,m)r^m \Big/ \sum_{j=1}^{l} F(l,j)r^j & l \neq 0; m \neq 0 \end{cases}$$

$$F(m,j) = \begin{cases} 1 & m = 1 \& j = 1 \\ 0 & m < j \\ \frac{(m-1)}{m}F(m-1,j) + \frac{1}{m}F(m-1,j-1) & 1 \leq j \leq m \end{cases}$$

Posterior sampling of state vector ($\theta_{1:T}$):

The evolution equations of the system are written as follows using the transformed data z_t.

$$\zeta_t = F_t\theta_t + \nu_t; \nu_t \sim N(0, V_t)$$

$$V_t = \begin{bmatrix} \frac{1}{\omega_{1t}} & .... & 0 & 0 \\ 0 & \frac{1}{\omega_{2t}} & .... & 0 \\ 0 & 0 & .... & \frac{1}{\omega_{nt}} \end{bmatrix}$$

$$\theta_t = G_t\theta_{t-1} + W_t$$

FFBS algorithm is performed in two steps: Forward filtering and backward smoothing. The following recursions are performed in each MCMC iteration within the Gibbs sampling framework. We start the recursions by drawing a state vector at $t = 0$ (i.e. θ_0)

from a non-informative state prior distribution at $t = 1$ (which is the state posterior at time $t = 0$).

Step 5: Sample $\theta_{1:T}$

Forward Filtering: Draw $\theta \sim N(m_0, T_0)$

from t = 1 to T:

Posterior at t-1:

$$p(\theta_{t-1}|D_{t-1}) \sim N(m_{t-1}, C_{t-1})$$

Prior at t:

$$p(\theta_t|D_{t-1}) \sim N(a_t, R_t)$$

$$a_t = G_t m_{t-1}; R_t = G_t C_{t-1} G_t^T + W_t$$

Predictive at t:

$$p(\zeta_t|D_{t-1}) \sim N(f_t, Q_t)$$

$$f_t = F_t a_t; Q_t = F_t R_t F_t^T + V_t$$

Posterior at t:

$$p(\theta_t|D_t) \sim N(m_t, C_t)$$

$$m_t = a_t + R_t F_t^T Q_t^{-1}(\zeta_t - f_t)$$

$$C_t = R_t - R_t F_t^T Q_t^{-1} F_t R_t$$

The above computations involve inversion of $n \times n$ matrix Q_t, which places an increasing computational burden with number of road segments. However, we use the following established result form matrix algebra to circumvent the issue.

$$Q_t^{-1} = (F_t R_t F_t^T + V_t)^{-1} = V_t^{-1} - V_t^{-1} F R (I_q + F_t^T V_t^{-1} F R)^{-1} F^T V^{-1}$$

We continue the recursions until T, then draw the state vector at time T using $\theta_T|D_T \sim N(m_T, C_T)$. Subsequently, we recursively draw the remaining states $\theta_{1:T-1}$ by backward smoothing using the following equations.

Backward Smoothing:

$$p(\theta_t|\theta_{t+1}, D_t) \sim N(h_t, B_t)$$

$$h_t = m_t + C_t G_{t+1}^T R_{t+1}^{-1}(h_{t+1} - a_{t+1})$$

$$B_t = C_t - C_t G_{t+1}^T R_{t+1}^{-1}(R_{t+1} - B_{t+1}) R_{t+1}^{-1} G_{t+1}^T C_t$$

Now, a single draw of the complete state vector $\theta_1, ..\theta_T$ is available and we proceed to the next MCMC iteration.

Posterior sampling of spatial random effects ($\phi^{(t)}$):

$$\begin{aligned} p(\phi_i^{(t)}|\phi_{-i}^{(t)}, z_{it}, -) &\propto P(z_{it}|\phi_{-i}^{(t)}, -) P(\phi_i^{(t)}|\phi_{-i}^{(t)}) \\ &\propto exp\left(\frac{-\omega_i}{2}\left(z_{it} - (Z_i^T\beta + X_{it}^T\theta_t + \phi_i^{(t)})\right)^2\right) * \\ &\quad exp\left(\frac{-1}{2\tau_t^2}\left(\phi_i^{(t)} - \sum_j b_{ij}\phi_j^{(t)}\right)^2\right) \end{aligned}$$

Step 6: Sample $\phi^{(t)}$

$$p(\phi_i^{(t)}|\phi_{-i}^{(t)}, z_{it}, -) \sim Normal(m_\phi^{(t)}, V_\phi^{(t)}) \tag{C.5}$$

$$m_\phi^{(t)} = \left((\zeta_{it} - Z_i^T\beta - X_{it}^T\theta_t)\omega_i + \left(\sum_j \frac{w_{ij}}{w_{i+}}\phi_j^{(t)}\right)\frac{w_{i+}}{\tau_t^2}\right) V_\phi^{(t)}$$

$$V_\phi^{(t)} = \left(\omega_{it} + \frac{w_{i+}}{\tau_t^2}\right)^{-1}$$

Mean centering is performed to accommodate the identification issue: $\phi^{(t)} = \phi^{(t)} - \bar{\phi^{(t)}}$

Step 7: Sample $1/\tau_t$

$$p(1/\tau_t|\phi^{(t)}, -) \sim Gamma\left(c_0 + \frac{n}{2}, d_0 + \sum_{i=1}^{n} \frac{w_{i+}}{2}\left(\phi_i^{(t)} - \sum_j b_{ij}\phi_j^{(t)}\right)^2\right) \tag{C.6}$$

We used a non-informative prior on $1/\tau_t$. A hyper prior may be used to learn the parameters of the Gamma process generating the τ_t terms.

An MCMC simulation is performed by iteratively drawing posterior samples using steps 1 through 7.

Appendix D

Bayesian probabilistic ranking procedure

A Gibbs sampling algorithm for posterior inference of the parameters of the spatial negative binomial model is provided below. Let y denote a $nT \times 1$ vector of crash counts, X denote a $nT \times k$ matrix.

Posterior sampling of dispersion parameter (r):

The full conditional analytical posterior distributions are shown below. The detailed derivations of full-conditional posterior distribution of dispersion parameter r are provided in Appendix B.

$$P(r|L,p,-) \sim Gamma\left(r_0 + \sum_{i=1}^{N}\sum_{t=1}^{T} L_{it}, h - \sum_{i=1}^{N}\sum_{t=1}^{T} ln(1-p_{it})\right) \tag{D.1}$$

Where, the probability parameter $p_{it} = \frac{e^{X_{it}\gamma+\phi_i}}{1+e^{X_{it}\gamma+\phi_i}}$. Hyper parameter h is also learned by constructing the full conditional posterior.

$$p(h|r,-) \sim Gamma(r_0 + ha_0, r + hb_0) \tag{D.2}$$

The analytical closed form expression for the posterior of discrete random variable L_i is shown in Equation B.3 in Appendix B.

Posterior sampling of regression coefficients (γ):

We need full conditional distributions of $\gamma|\omega_{it}$ and $\omega_{it}|\gamma$ for inference of the regression parameter vector γ.

$$\begin{aligned} P(\gamma|y, X^F, X^R, \phi, r, \omega) &\propto \prod_{i=1}^{n}\prod_{t=1}^{T} P(y_{it}|\gamma, X_{it}, \phi_i, r, \omega_{it})P(\gamma) \\ &\propto Normal(m_\gamma, V_\gamma) \\ m_\gamma = V_\gamma(X^{'}\Omega(z-\phi) + G_0^{-1}g_0); V_\beta &= (X^{'}\Omega X + G_0^{-1})^{-1} \end{aligned} \tag{D.3}$$

where, z and ϕ are $nT \times 1$ vectors; $z_{it} = \frac{y_{it}-r}{2\omega_{it}}$ for $i \in \{1, 2, ...n\}$ and $t \in \{1, 2, ...T\}$; Ω is a $nT \times nT$ diagonal matrix; $\Omega_{pp} = \omega_{it}$ for $i \in \{1, 2, ...n\}$ and $t \in \{1, 2, ...T\}$

Second, the full conditional distribution $p(\omega|z^N, -)$ turns out to be a distribution in the Polya-Gamma class (see Polson et al. [2013] for derivations).

$$P(\omega_{it}|\gamma, r, y, X^F, X^R, \phi) \propto PG(y_{it} + r, \psi_{it}) \tag{D.4}$$

where, $PG(b, c)$ represents the Polya-Gamma distribution with the density indicated in Equation (A.2); $\psi_{it} = X_{it}\gamma + \phi_i$

Posterior sampling of spatial random effects (ϕ_i):

$$p(\phi_i|\phi_{-i}, z_i, -) \propto N(m_\phi, V_\phi) \tag{D.5}$$

$$m_\phi = \left(\sum_{t=1}^{T}(z_{it} - X_{it}{}^{'}\gamma)\omega_{it} + \left(\sum_j \frac{w_{ij}}{w_{i+}}\phi_j\right)\frac{w_{i+}}{\tau_c^2}\right)V_\phi$$

$$V_\phi = \left(\sum_{t=1}^{T}\omega_{it} + \frac{w_{i+}}{\tau_c^2}\right)^{-1}$$

Mean centering: $\phi = \phi - \bar{\phi}$

$$p(1/\tau_c^2|\phi, -) \sim Gamma\left(c_0 + \frac{n}{2}, d_0 + \sum_{i=1}^{n}\frac{w_{i+}}{2}\left(\phi_i - \sum_j b_{ij}\phi_j\right)^2\right) \tag{D.6}$$

An MCMC simulation is performed by iteratively drawing posterior samples using equations D.1 through D.6.

Probabilistic ranking

To perform probabilistic ranking, the predicted decision parameter ξ_i corresponding to a future year is calculated in each MCMC iteration. For example, the following decision parameters are calculated in this research.

- $\xi_i = X_{i\tau}\gamma + \phi_i$
- $\xi_i = X_{i\tau}\gamma$
- $\xi_i = \phi_i$

Where, $X_{i\tau}$ is the vector of attributes corresponding to i^th road segment in a future year τ. The decision parameters of the road segments are ranked in each MCMC iteration. The MCMC draws of $n \times 1$ rank vector are collected to perform posterior inference on the site ranks. Subsequently, posterior mean rank ($E(R(\xi_i)|y, X)$) is estimated. To estimate the probability of a site to be in top m unsafe sites, a vector of $n \times 1$ indicator variables is created as shown below.

$$I_i^m = \begin{cases} 1 & R(\xi_i) \in T_m \\ 0 & R(\xi_i) \notin T_m \end{cases}$$

Where, T_m is set of top m decision parameters. The probability of a site to be in top m unsafe sites is $E(I_i^m|y, X)$.

Bibliography

J. Aguero-Valverde and P. P. Jovanis. Analysis of Road Crash Frequency with Spatial Models. *Transportation Research Record: Journal of the Transportation Research Board*, 2061(-1):55–63, Dec. 2008. ISSN 0361-1981. doi: 10.3141/2061-07. URL `http://trb.metapress.com/openurl.asp?genre=article&id=doi:10.3141/2061-07`.

M. Ahmed, H. Huang, M. Abdel-Aty, and B. Guevara. Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. *Accident Analysis & Prevention*, 43(4):1581–1589, July 2011. ISSN 00014575. doi: 10.1016/j.aap.2011.03.021. URL `http://linkinghub.elsevier.com/retrieve/pii/S0001457511000728`.

J. H. Albert and S. Chib. Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, 88(422):669–679, June 1993. ISSN 0162-1459. doi: 10.2307/2290350. URL `http://www.jstor.org/stable/2290350`.

P. C. Anastasopoulos and F. L. Mannering. A note on modeling vehicle accident frequencies with random-parameters count models. *Accident Analysis & Prevention*, 41(1):153–159, Jan. 2009. ISSN 00014575. doi: 10.1016/j.aap.2008.10.005. URL `http://linkinghub.elsevier.com/retrieve/pii/S0001457508001954`.

P. C. Anastasopoulos, F. L. Mannering, V. N. Shankar, and J. E. Haddock. A study of factors affecting highway accident rates using the random-parameters tobit model.

Accident Analysis & Prevention, 45:628–633, Mar. 2012. ISSN 00014575. doi: 10.1016/j.aap.2011.09.015. URL http://linkinghub.elsevier.com/retrieve/pii/S0001457511002521.

S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical modeling and analysis for spatial data*. Number 101 in Monographs on statistics and applied probability. Chapman & Hall/CRC, Boca Raton, Fla, 2004. ISBN 158488410X.

R. Bell. A Linear Filter Statistic for Roadway Accident Research. Technical Report Project MTR-86-006A, NHTSA, U.S. Department of Transportation, Apr. 1986.

J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of Royal Statistical Society*, 36(Ser.B):192–236, 1974.

C. R. Bhat. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research Part B: Methodological*, 37 (9):837–855, Nov. 2003. ISSN 0191-2615. doi: 10.1016/S0191-2615(02)00090-5. URL http://www.sciencedirect.com/science/article/pii/S0191261502000905.

C. R. Bhat. The maximum approximate composite marginal likelihood (MACML) estimation of multinomial probit-based unordered response choice models. *Transportation Research Part B: Methodological*, 45(7):923–939, Aug. 2011. ISSN 0191-2615. doi: 10.1016/j.trb.2011.04.005. URL http://www.sciencedirect.com/science/article/pii/S019126151100049X.

C. R. Bhat and R. Sidharthan. A new approach to specify and estimate non-normally mixed multinomial probit models. *Transportation Research Part B: Methodological*,

46(7):817–833, Aug. 2012. ISSN 0191-2615. doi: 10.1016/j.trb.2012.02.007. URL http://www.sciencedirect.com/science/article/pii/S019126151200032X.

L. Blincoe, T. Miller, E. Zaloshnja, and B. Lawrence. The Economic Impact of Motor Vehicle Crashes, 2000, May 2014. URL http://www-nrd.nhtsa.dot.gov/Pubs/811659.pdf.

T. Brijs, D. Karlis, F. Van den Bossche, and G. Wets. A Bayesian model for ranking hazardous road sites. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(4):1001–1017, 2007. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-985X.2007.00486.x/full.

S. P. Brooks and A. Gelman. General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, Dec. 1998. ISSN 1061-8600, 1537-2715. doi: 10.1080/10618600.1998.10474787. URL http://www.tandfonline.com/doi/abs/10.1080/10618600.1998.10474787.

A. C. Cameron and P. K. Trivedi. *Regression analysis of count data*. Number 30 in Econometric society monographs. Cambridge University Press, Cambridge, UK ; New York, NY, USA, 1998. ISBN 0521632013.

C. K. Carter and R. Kohn. On Gibbs Sampling for State Space Models. *Biometrika*, 81 (3):541–553, Aug. 1994. ISSN 0006-3444. doi: 10.2307/2337125. URL http://www.jstor.org/stable/2337125.

M. Castro, R. Paleti, and C. R. Bhat. A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting

crash frequency at intersections. *Transportation Research Part B: Methodological*, 46 (1):253–272, Jan. 2012. ISSN 01912615. doi: 10.1016/j.trb.2011.09.007. URL http://linkinghub.elsevier.com/retrieve/pii/S0191261511001366.

H. C. Chin and M. A. Quddus. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. *Accident Analysis & Prevention*, 35(2):253–259, Mar. 2003. ISSN 0001-4575. doi: 10.1016/S0001-4575(02)00003-9. URL http://www.sciencedirect.com/science/article/pii/S0001457502000039.

G. A. Davis and S. Yang. Bayesian identification of high-risk intersections for older drivers via Gibbs sampling. *Transportation Research Record: Journal of the Transportation Research Board*, 1746(1):84–89, 2001. URL http://trb.metapress.com/index/53UL1X0044Q6208N.pdf.

J. A. Deacon, C. V. Zeger, and R. C. Deen. Identification of hazardous rural highway locations. Technical Report Research report 410, Kentucky Bureau of Highways, Nov. 1974.

P. Deb and P. K. Trivedi. Demand for medical care by the elderly: a finite mixture approach. *Journal of applied Econometrics*, 12(3):313–336, 1997. URL http://users.stat.umn.edu/~sandy/courses/8053/Data/trevedi/debtrevedi.pdf.

D. Eddelbuettel and R. Franois. Rcpp: Seamless R and C++ Integration. *Journal of Statistical Software*, 40(8):1–18, 2011. URL http://www.jstatsoft.org/v40/i08/.

R. Elvik. Comparative Analysis of Techniques for Identifying Locations of Hazardous Roads. *Transportation Research Record: Journal of the Transportation Research Board*, 2083(-1):72–75, Dec. 2008. ISSN 0361-1981. doi: 10.3141/2083-08. URL http://trb.metapress.com/openurl.asp?genre=article&id=doi:10.3141/2083-08.

S. Frhwirth-Schnatter. Data Augmentation and Dynamic Linear Models. *Journal of Time Series Analysis*, 15(2):183–202, Mar. 1994. ISSN 1467-9892. doi: 10.1111/j.1467-9892.1994.tb00184.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9892.1994.tb00184.x/abstract.

S. Frhwirth-Schnatter, R. Tchler, and T. Otter. Bayesian Analysis of the Heterogeneity Model. *Journal of Business & Economic Statistics*, 22(1):2–15, Jan. 2004. ISSN 0735-0015. URL http://www.jstor.org/stable/27638777.

D. Gamerman and H. F. Lopes. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. Number 68 in Texts in statistical science series. Taylor & Francis, Boca Raton, 2nd ed edition, 2006. ISBN 1584885874.

A. Gelman and D. B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472, 1992. doi: 10.1214/ss/1177011136. URL http://dx.doi.org/10.1214/ss/1177011136.

K. Geurts, G. Wets, T. Brijs, K. Vanhoof, and D. Karlis. Ranking and selecting dangerous crash locations: Correcting for the number of passengers and Bayesian ranking plots. *Journal of Safety Research*, 37(1):83–91, Jan. 2006. ISSN 00224375. doi: 10.1016/j.jsr.2005.10.020. URL http://linkinghub.elsevier.com/retrieve/pii/S0022437506000089.

S. Hallmark, R. Basavaraju, and M. Pawlovich. Evaluation of the Iowa DOTs Safety Improvement Candidate List Process. Technical report, Iowa State Univ. Center for Trans. Res. and Educ./Iowa Dept. of Trans. Office of Traffic and Safety, Ames, Iowa., 2002.

E. Hauer. Empirical bayes approach to the estimation of unsafety: The multivariate regression method. *Accident Analysis & Prevention*, 24(5):457–477, Oct. 1992. ISSN 00014575. doi: 10.1016/0001-4575(92)90056-O. URL http://linkinghub.elsevier.com/retrieve/pii/000145759290056O.

E. Hauer. Identification of sites with promise. *Transportation Research Record: Journal of the Transportation Research Board*, 1542(1):54–60, 1996. URL http://trb.metapress.com/index/64J3308L150J5122.pdf.

E. Hauer. *Observational before–after studies in road safety: estimating the effect of highway and traffic engineering measures on road safety*. Pergamon, Oxford, OX, U.K. ; Tarrytown, N.Y., U.S.A, 1st ed edition, 1997. ISBN 0080430538.

E. Hauer, J. Kononov, B. Allery, and M. S. Griffith. Screening the road network for sites with promise. *Transportation Research Record: Journal of the Transportation Research Board*, 1784(1):27–32, 2002. URL http://trb.metapress.com/index/30nj038822412857.pdf.

S. Herbel, L. Laing, and C. McGovern. Highway Safety Improvement Program (HSIP) Manual. *Report No. FHWA-SA-09-029*, Jan. 2010. URL http://www.kutc.ku.edu/pdffiles/HSIP.pdf.

J. Higle and J. Witkowski. Bayesian Identification of Hazardous Locations. *In Transportation Research Record*, 1185:24–36, 1988.

HSIP. Highway Safety Improvement Program Manual. Manual, Texas Deaprtment of Transportation, May 2013.

S. Hu, J. N. Ivan, N. Ravishanker, and J. Mooradian. Temporal modeling of highway crash counts for senior and non-senior drivers. *Accident Analysis & Prevention*, 50: 1003–1013, Jan. 2013. ISSN 00014575. doi: 10.1016/j.aap.2012.08.001. URL `http://linkinghub.elsevier.com/retrieve/pii/S0001457512002886`.

A. Ihs, H. Velin, and M. Wiklund. The influence of road surface condition on traffic safety: Data from 1992-1998. *VTI Meddelanden*, (909), 2002. ISSN 0347-6049. URL `http://trid.trb.org/view.aspx?id=685126`.

D. Mahalel, A. S. Hakkert, and J. N. Prashker. A system for the allocation of safety resources on a road network. *Accident Analysis & Prevention*, 14(1):45–56, Feb. 1982. ISSN 0001-4575. doi: 10.1016/0001-4575(82)90006-9. URL `http://www.sciencedirect.com/science/article/pii/0001457582900069`.

N. V. Malyshkina, F. L. Mannering, and A. P. Tarko. Markov switching negative binomial models: An application to vehicle accident frequencies. *Accident Analysis & Prevention*, 41(2):217–226, Mar. 2009. ISSN 00014575. doi: 10.1016/j.aap.2008.11.001. URL `http://linkinghub.elsevier.com/retrieve/pii/S0001457508002170`.

F. L. Mannering and C. R. Bhat. Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*, 1:1–22, Jan.

2014. ISSN 22136657. doi: 10.1016/j.amar.2013.09.001. URL http://linkinghub.elsevier.com/retrieve/pii/S2213665713000031.

D. McGuigan. The Use of Relationships Between Road Accidents and Traffic Flow in Black-Spot Identification. *Traffic Engineering and Control*, pages 448–453, Sept. 1981.

D. McGuigan. Nonjunction Accident Rates and Their Use in Black-Spot Identification. *Traffic Engineering and Control*, pages 60–65, Feb. 1982.

S.-P. Miaou and J. J. Song. Bayesian ranking of sites for engineering safety improvements: Decision parameter, treatability concept, statistical criterion, and spatial dependence. *Accident Analysis & Prevention*, 37(4):699–720, July 2005. ISSN 00014575. doi: 10.1016/j.aap.2005.03.012. URL http://linkinghub.elsevier.com/retrieve/pii/S0001457505000497.

L. F. Miranda-Moreno, A. Labbe, and L. Fu. Bayesian multiple testing procedures for hotspot identification. *Accident Analysis & Prevention*, 39(6):1192–1201, Nov. 2007. ISSN 0001-4575. doi: 10.1016/j.aap.2007.03.008. URL http://www.sciencedirect.com/science/article/pii/S0001457507000516.

S. Mitra. Spatial Autocorrelation and Bayesian Spatial Statistical Method for Analyzing Intersections Prone to Injury Crashes. *Transportation Research Record: Journal of the Transportation Research Board*, 2136(-1):92–100, Dec. 2009. ISSN 0361-1981. doi: 10.3141/2136-11. URL http://trb.metapress.com/openurl.asp?genre=article&id=doi:10.3141/2136-11.

D. Morin. Application of Statistical Concepts to Accident Data. *In Highway Research Record*, 188, 1967.

S. Narayanamoorthy, R. Paleti, and C. R. Bhat. On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. *Transportation Research Part B: Methodological*, 55:245–264, Sept. 2013. ISSN 01912615. doi: 10.1016/j.trb.2013.07.004. URL http://linkinghub.elsevier.com/retrieve/pii/S0191261513001197.

NHTSA. Traffic Safety Facts 2012. Technical Report DOT HS 812 032, National Highway Traffic Safety Administration, Washington, DC, 2014. URL http://www-nrd.nhtsa.dot.gov/Pubs/811659.pdf.

R. B. Noland, N. J. Klein, and N. K. Tulach. Do lower income areas have more pedestrian casualties? *Accident Analysis & Prevention*, 59:337–345, Oct. 2013. ISSN 0001-4575. doi: 10.1016/j.aap.2013.06.009. URL http://www.sciencedirect.com/science/article/pii/S0001457513002376.

N. Norden, J. Orlansky, and H. Jacobs. Application of Statistical Quality-Control Techniques to Analysis of Highway Accident Data. Bulletin 117, National Research Council, Washington, D.C., 1956.

T. Otter, R. Tchler, and S. Frhwirth-Schnatter. Capturing consumer heterogeneity in metric conjoint analysis using Bayesian mixture models. *International Journal of Research in Marketing*, 21(3):285–297, Sept. 2004. ISSN 01678116. doi: 10.1016/j.ijresmar.2003.11.002. URL http://linkinghub.elsevier.com/retrieve/pii/S0167811604000308.

B.-J. Park and D. Lord. Application of finite mixture models for vehicle crash data analysis. *Accident Analysis & Prevention*, 41(4):683–691, July 2009. ISSN 00014575. doi: 10.1016/j.aap.2009.03.007. URL http://linkinghub.elsevier.com/retrieve/pii/S0001457509000542.

B. Persaud, C. Lyon, and T. Nguyen. Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Safety Improvement. *Transportation Research Record*, 1665(1):7–12, Jan. 1999. ISSN 0361-1981. doi: 10.3141/1665-02. URL http://trb.metapress.com/openurl.asp?genre=article&id=doi:10.3141/1665-02.

N. Polson, J. G. Scott, and J. Windle. BayesLogit. Explanation of the Polya-Gamma latent variable method., 2012. URL http://arxiv.org/abs/1205.0310.

N. G. Polson, J. G. Scott, and J. Windle. Bayesian inference for logistic models using Plya-Gamma latent variables. *Journal of the American Statistical Association*, page 130808174755007, Aug. 2013. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.2013.829001. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.2013.829001.

R. Prado and M. West. *Time Series: Modeling, Computation, and Inference*. CRC Press, May 2010. ISBN 9781439882757.

M. A. Quddus. Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data. *Accident Analysis & Prevention*, 40 (4):1486–1497, July 2008. ISSN 0001-4575. doi: 10.1016/j.aap.2008.03.009. URL http://www.sciencedirect.com/science/article/pii/S0001457508000523.

M. Quenouille. A Relation between the Logarithmic, Poisson, and Negative Binomial Series. *Biometrics*, 5(2):162–164, June 1949.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/. ISBN 3-900051-07-0.

S. Richardson and P. J. Green. On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4):731–792, Jan. 1997. ISSN 1467-9868. doi: 10.1111/1467-9868.00095. URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00095/abstract.

K. Schmidt. *Modeling crash frequency data*. PhD thesis, Iowa state unviersity, Ames, Iowa., 2012.

T. S. Shively, K. Kockelman, and P. Damien. A Bayesian semi-parametric model to estimate relationships between crash counts and roadway characteristics. *Transportation Research Part B: Methodological*, 44(5):699–715, June 2010. ISSN 01912615. doi: 10.1016/j.trb.2009.12.019. URL http://linkinghub.elsevier.com/retrieve/pii/S0191261509001635.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639, 2002. URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00353/full.

T. Tamburri and R. Smith. The Safety Index: Method of Evaluating and Rating Safety Benefits. *In Highway Research Record*, 332, 1970.

f. TxDOT. Pavement Management Information Systems Users Manual, 1994.

S. Ukkusuri, S. Hasan, and H. M. A. Aziz. Random Parameter Model Used to Explain Effects of Built-Environment Characteristics on Pedestrian Crash Frequency. *Transportation Research Record: Journal of the Transportation Research Board*, 2237 (-1):98–106, Dec. 2011. ISSN 0361-1981. doi: 10.3141/2237-11. URL `http://trb.metapress.com/openurl.asp?genre=article&id=doi:10.3141/2237-11`.

D. A. Van Dyk and X.-L. Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1), 2001. URL `http://amstat.tandfonline.com/doi/full/10.1198/10618600152418584`.

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

M. Wedel, W. S. DeSarbo, J. R. Bult, and V. Ramaswamy. A latent class Poisson regression model for heterogeneous count data. *Journal of Applied Econometrics*, 8(4):397–411, 1993. URL `http://onlinelibrary.wiley.com/doi/10.1002/jae.3950080407/full`.

J. Windle, C. M. Carvalho, J. G. Scott, and L. Sun. Polya-Gamma Data Augmentation for Dynamic Models. *arXiv preprint arXiv:1308.0774*, 2013. URL `http://arxiv.org/abs/1308.0774`.

Z. Wu, A. Sharma, F. L. Mannering, and S. Wang. Safety impacts of signal-warning flashers and speed control at high-speed signalized intersections. *Accident Analysis & Prevention*, 54:90–98, May 2013. ISSN 0001-4575. doi: 10.1016/j.aap.2013.01.016. URL http://www.sciencedirect.com/science/article/pii/S0001457513000298.

Y. Xiong and F. L. Mannering. The heterogeneous effects of guardian supervision on adolescent driver-injury severities: A finite-mixture random-parameters approach. *Transportation Research Part B: Methodological*, 49:39–54, Mar. 2013. ISSN 01912615. doi: 10.1016/j.trb.2013.01.002. URL http://linkinghub.elsevier.com/retrieve/pii/S0191261513000131.

Y. Xiong, J. L. Tobias, and F. L. Mannering. The analysis of vehicle crash injury-severity data: A Markov switching approach with road-segment heterogeneity. *Transportation Research Part B: Methodological*, 67:109–128, Sept. 2014. ISSN 01912615. doi: 10.1016/j.trb.2014.04.007. URL http://linkinghub.elsevier.com/retrieve/pii/S0191261514000630.

R. Yu, M. Abdel-Aty, and M. Ahmed. Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. *Accident Analysis & Prevention*, 50:371–376, Jan. 2013. ISSN 00014575. doi: 10.1016/j.aap.2012.05.011. URL http://linkinghub.elsevier.com/retrieve/pii/S0001457512001637.

Q. Zeng and H. Huang. Bayesian spatial joint modeling of traffic crashes on an urban road network. *Accident Analysis & Prevention*, 67:105–112, June 2014. ISSN

0001-4575. doi: 10.1016/j.aap.2014.02.018. URL http://www.sciencedirect.com/science/article/pii/S0001457514000633.

M. Zhou, L. Li, D. Dunson, and L. Carin. Lognormal and gamma mixed negative binomial regression. *arXiv preprint arXiv:1206.6456*, 2012. URL http://arxiv.org/abs/1206.6456.

Y. Zou, Y. Zhang, and D. Lord. Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis. *Accident Analysis & Prevention*, 50:1042–1051, Jan. 2013. ISSN 0001-4575. doi: 10.1016/j.aap.2012.08.004. URL http://www.sciencedirect.com/science/article/pii/S0001457512002916.

Y. Zou, Y. Zhang, and D. Lord. Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models. *Analytic Methods in Accident Research*, 1:39–52, Jan. 2014. ISSN 2213-6657. doi: 10.1016/j.amar.2013.11.001. URL http://www.sciencedirect.com/science/article/pii/S2213665713000079.