Bayesian Nonparametric Dirichlet Process Mixture Modeling in Transportation Safety Studies

by

Shahram Heydari

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Civil Engineering

Waterloo, Ontario, Canada, 2017

EXAMINING COMMITTEE MEMBERSHIP

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Dr. Karim El-Basyouny Associate Professor of Civil Engineering
Supervisor	Dr. Liping Fu Professor of Civil Engineering
Internal Member	Dr. Wei-Chau Xie Professor of Civil Engineering
Internal Member	Dr. Frank Saccomanno Professor Emeritus of Civil Engineering
Internal-external Member	Dr. Chengguo Weng Associate Professor of Actuarial Science

AUTHOR'S DECLARATION

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

STATEMENT OF CONTRIBUTIONS

Chapters 3 and 4 of this thesis mainly consists of three papers that were co-authored by myself, my supervisor, Dr. Dominique Lord, Dr. Luis Miranda-Moreno, Dr. Lawrence Joseph, and Dr. Bani Mallick. I developed and documented the methodology, implemented the method within software, conducted the analyses, and wrote the papers. My supervisor, Dr. Lord and Dr. Miranda-Moreno provided valuable insights relating to transportation engineering aspects. Dr. Joseph and Dr. Mallick assisted with providing guidelines on the statistical analysis and ensuring the validity and accuracy of the proposed models from a mathematical and statistical standpoint. The following peer-reviewed articles have been published by Elsevier:

I. Heydari, S., Fu, L., Joseph, L., Miranda-Moreno, L.F., 2016. Bayesian nonparametric modeling in transportation safety studies: applications in univariate and multivariate settings. Analytic Methods in Accident Research 12, 18-34.

II. Heydari, S., Fu, L., Lord, D., Mallick, B.K., 2016. Multilevel Dirichlet process mixture analysis of railway grade crossing crash data. Analytic Methods in Accident Research 9, 27-43.

III. Heydari, S., Fu, L, Miranda-Moreno, L.F., Joseph, L., 2017. Using a flexible multivariate latent class approach to model correlated outcomes: a joint analysis of pedestrian and cyclist injuries. Analytic Methods in Accident Research 13, 16-27.

ABSTRACT

In transportation safety studies, it is often necessary to account for unobserved heterogeneity and multimodality in data. The commonly used standard generalized linear models (e.g., Poisson-gamma models) do not fully address unobserved heterogeneity, assuming unimodal exponential families of distributions. This thesis illustrates how restrictive assumptions (e.g., unimodality) common to most road safety studies can be relaxed employing Bayesian nonparametric Dirichlet process mixture models. We use a truncated Dirichlet process, so that our models reduce to the form of finite mixture (latent class) models, which can be estimated employing standard Markov chain Monte Carlo methods, emphasizing computational simplicity. Interestingly, our approach estimates the number of latent subpopulations as part of its analysis algorithm using an elegant mathematical framework. We use pseudo Bayes factors for model selection, showing how the predictive capability of models can be affected by different assumptions.

In univariate settings, we extend standard generalized linear models to a Dirichlet process mixture generalized linear model in which the random intercepts density is modeled nonparametrically, thereby adding flexibility to the model. We examine the performance of the proposed approach using both simulated and real data. We also examine the performance of the proposed model in terms of replicating datasets with high proportions of zero crashes. In terms of engineering insights, we provide a policy example related to the identification of high-crash locations, a critical component of the transportation safety management process.

With respect to multilevel settings, this thesis introduces a flexible latent class multilevel model for analyzing crash data that are of hierarchical nature. We extend the standard multilevel model by accounting for unobserved cross-group heterogeneity through multimodal intercepts (group effects). The proposed method allows identifying latent subpopulations (and consequently outliers) at the highest level of the hierarchy (e.g., geographic areas). We evaluate our method on two recent railway grade crossing crash datasets from Canada. This research confirms the need for a multilevel approach for both datasets due to the presence of spatial dependencies among crossings nested within the same region. We provide a novel approach to benchmark different regions

based on their safety performance measures. To this end, we identify latent clusters among different regions that share similar unidentified features, stimulating further investigations to explore reasons behind such similarities and dissimilarities. This could have important policy implications for various safety management programs.

This thesis also investigates inference for multivariate crash data by introducing two flexible Bayesian multivariate models: a multivariate mixture of points and a mixture of multivariate normal densities. We use a Dirichlet process mixture to keep the dependence structure unconstrained, relaxing the usual homogeneity assumptions. We allow for interdependence between outcomes through a Dirichlet process prior on the random intercepts density. The resulting models collapse into a form of latent class multivariate model, an appealing way to address unobserved heterogeneity in multivariate settings. Therefore, the multivariate models that we derive in this thesis account for correlation among crash types through a heterogeneous correlation structure, which better captures the complex structure of correlated data. To our knowledge, this is the first study to propose and apply such a model in the transportation literature.

Using a highway injury-severity dataset, we illustrate how the robustness to homogeneous correlation structures can be examined using a multivariate mixture of points model that relaxes the homogeneity assumption with respect to the location of the dependence structure. We then use the mixture of multivariate normal densities model–relaxing the homogeneity assumption with respect to both the location and the covariance matrix–to investigate the effects of various factors on pedestrian and cyclist safety in an urban setting, modeling both outcomes simultaneously. To our knowledge, this is the first study to conduct a joint safety analysis of active modes at an intersection level, a micro-level, which is expected to provide more detailed insights. We show how spurious assumptions affect predictive performance of the multivariate model and the interpretation of the explanatory variables using marginal effects. The results show that our flexible model specification better captures the underlying structure of pedestrian/cyclist crash data, resulting in a more accurate model that contributes to a better understanding of safety correlates of non-motorist road users. This in turn helps decision-makers in selecting more appropriate countermeasures targeting vulnerable road users, promoting the mobility and safety of active modes of transportation.

ACKNOWLEDGMENTS

I began my PhD studies full of excitement for the research I had planned, hoping to be able to help improve transportation safety practices. Starting out in the iTSS LAB with a research project assigned by Transport Canada was what inspired me to pursue my PhD studies in the first place. I am extremely lucky to have had the chance to collaborate with a remarkable group of researchers, without whom this thesis would not have been completed.

My sincere gratitude goes to my supervisor, Professor Liping Fu, who provided me with the opportunity to join his team. Thank you very much for giving me independence in my research and having confidence in my abilities. This made my PhD experience much more enjoyable and productive. Undoubtedly, working under your supervision has been a great privilege for me.

To Professor Luis Miranda-Moreno for his insightful comments and providing some of the finest datasets used in this research. Thank you for accepting me in your research group at McGill University during my stay in Montreal.

To Professor Lawrence Joseph of McGill University for his outstanding mentorship and patiently answering my numerous questions relating to Bayesian inference. Certainly, working with you has been an invaluable experience from which I will benefit forever.

To Professor Dominique Lord, whose commitment to fundamental research in transportation safety motivated me to focus on theoretical research that contributes significantly to improving road users' safety. I greatly appreciate your valuable guidance and giving me the possibility to closely work under your supervision as a visiting PhD student at Zachry Department of Civil Engineering at Texas A&M University.

To Professor Bani Mallick for his availability and valuable insights. Thank you for your role in improving my understanding of Bayesian nonparametrics during my stay in Texas.

To Professor Luis Amador for his cordiality and encouragement at the very beginning of my PhD studies.

I would like to thank my committee members, Professor Frank Saccomanno, Professor Wei-Chau Xie, Professor Chengguo Weng, and Professor Karim El-Basyouny for their time and effort in evaluating this dissertation. I am very thankful to the University of Waterloo and the Natural Sciences and Engineering Research Council of Canada for financially supporting this research. I am also thankful to the Government of Ontario for providing me the Ontario Graduate Scholarship during my first year at the University of Waterloo.

I would like to extend my appreciation to my colleges at the University of Waterloo for their friendship and support. Thank you to Lalita Thakali, Tae Kwon, Sajad Shiravi, Matthew Muresan, and many others. I would also like to thank the administrative staff of the Department of Civil & Environmental Engineering, especially Mrs. Victoria Tolton.

I am extremely grateful to my parents to whom I dedicate this work for their understanding and encouragement. I should also thank my sisters, especially Mehrnoosh for her unconditional support during my first years in Canada.

Un ringraziamento particolare va ai miei carissimi amici, Mohsen e Isabel. Siete sempre stati pronti ad aiutarmi durante tutti quegli anni a Roma. Se non fosse stato per il vostro sostegno, non avrei potuto arrivare fin qui.

Vorrei infine ringraziare il mio amore di vita Niloofar che ha sempre creduto in me. Ti ringrazio di cuore per il tuo instancabile sostegno morale. Questa tesi è il risultato dei tuoi continui incoraggiamenti. Sono felice di averti incontrato nel mio cammino.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

LIST OF ABREVIATIONS

AADT	Average annual daily traffic
CPO	Conditional predictive ordinates
DIC	Deviance information criterion
LPBF	Log pseudo Bayes factor
LPML	Log pseudo marginal likelihood
MCMC	Markov chain Monte Carlo
MLE	Maximum likelihood estimation
MVN	Multivariate normal density
PBF	Pseudo Bayes factor
USDOT	US Department of Transportation

CHAPTER 1

INTRODUCTION

1.1 Background and Motivation

Traffic safety is a major global health issue since very large proportions of unintentional injuries are caused by traffic-related crashes. According to the Global Health Observatory, 1.25 million fatalities occur on world's roads each year, and traffic-related injuries are major cause of death among people 15 to 29 years old (World Health Organization, 2015; Mannering et al., 2016). In this regard, Transport Canada reports 149,900 injuries, 9,647 serious injuries, and 1,834 fatalities across Canada in 2014 (Transport Canada, 2014). Table 1-1 shows fatalities and injuries sustained by different age groups in Canada, indicating high rate of critical injuries among young people. These numbers obviously indicate the need for further improvements.

Although traffic-related injuries and fatalities have seen a decreasing trend during the past two decades, this reduction has not been drastic. As an example, Fig. 1-1 illustrates reduction in traffic-related fatalities in Canada from 1995 to 2014 (Transport Canada, 2014). Such trend is observed in spite of several improvements made in terms of motor vehicle safety standards/features, traffic safety policies, and road design. Research is thus needed to better understand underlying crash mechanisms. This in turn helps guide safety policy, reducing traffic-related injuries and fatalities. To this end, statistical models play a noteworthy role by defining a relationship between traffic safety

performance measures (crash frequencies or differing injury-severity levels) and a series of factors (explanatory variables) that affect these measures. Section 1.2 provides a quick introduction to traffic safety studies describing its main components.

Table 1-1 Fatalities and injuries by age group

Age group (yrs)	Fatalities	Serious injuries	Injuries (Total)
0–4	17	87	1,984
5–14	30	293	5,957
15–19	146	897	14,015
20–24	194	1,205	17,732
25–34	301	1,725	27,605
35–44	210	1,329	23,051
45–54	265	1,431	23,210
55–64	250	1,185	17,220
65 +	400	1,208	15,047
Unknown	21	287	4,079
Total	1,834	9,647	149,900

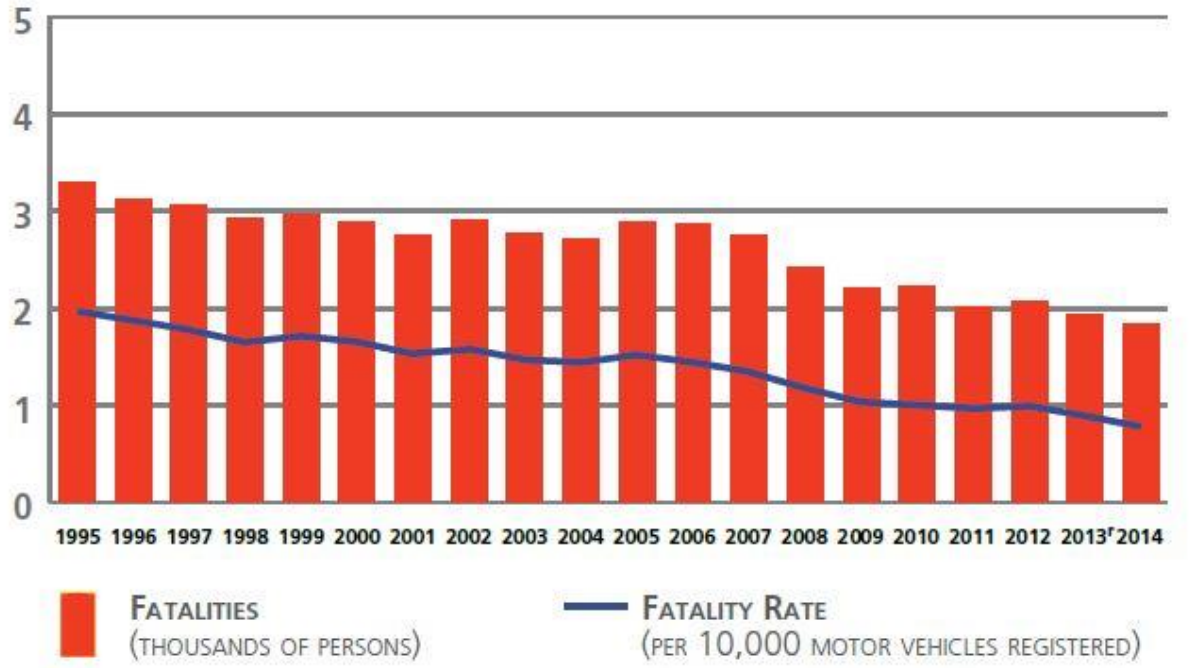


Figure 1-1 Fatalities across Canada from 1995 to 2014
(adopted from https://www.tc.gc.ca/media/documents/roadsafety/cmvtcs2014_eng.pdf)

In addition to the general consensus that further research is needed to mitigate crash risks, one important issue is related to the literature on road safety research. That is while most traffic safety studies have been centered on highway and intersection safety mainly considering motor vehicle accidents, less attention is given to the safety analysis of railway grade crossings, pedestrians, and cyclists. The next two subsections therefore provide an introduction to these under-represented studies, highlighting their importance. Based on the latter observation, besides focusing on methodological aspects, this dissertation also adopts under-represented crash data (railway grade crossings and vulnerable road users) among other crash datasets. The aim is to increase the empirical impact of the dissertation by employing innovative statistical models and providing valuable insights with respect to the under-represented studies as well.

1.1.1 Railway grade crossings

The Canadian rail network is the fifth most extensive globally, moving a significant number of passengers and more than 70% of surface goods in Canada each year (Railway Association of Canada, 2012). Rail transportation therefore plays a significant role in maintaining the quality of life of all Canadians and the vitality of Canada's economy. When compared to road transportation (vehicles, trucks, etc.), rail transportation also has a carbon footprint that is considerably lower and is thus a sustainable alternative with lower greenhouse gas emissions. In fact, rail transportation produces only 3.4 percent of the total greenhouse gas emissions produced by the transportation sector in Canada (Railway Association of Canada, 2012).

The presence of a vast railway network in Canada imparts some risk to road/rail users and to residents living around railway lines. As reported by the Transportation Safety Board of Canada, 11,998 rail accidents of various forms have been observed over a ten-year period from 2006 to 2015 (Transportation Safety Board of Canada, 2015). According to these data, around 17% of all rail accidents have occurred at railway grade crossings. Despite improvements in the recent years, the number of grade crossing crashes remains high; and therefore, grade crossing safety is still a significant concern for transportation authorities and Canadian society as a whole. For example, the Transportation Safety Board of Canada reports 1,138 crossing accidents for the years 2008-2013, causing 155 fatalities, and 166 serious injuries (Transportation Safety Board

of Canada, 2014). The significant concern is also due to the great monetary (property damage, derailment, service delay, etc.) and non-monetary (psychological consequences, grief, pain, etc.) costs that usually result from grade crossing accidents. Fig. 1-2 illustrates the number of crossing accidents for the period 2006 to 2015, indicating a non-drastic reduction. Further research is therefore needed to better understand the complex crash mechanisms at railway grade crossings. For instance, spatial spread of Canadian railway crossings may lead to significant variation in unknown spatial attributes (e.g. climate) of crossings. This in turn may have a bearing on safety. Consequently, an enhanced understanding of grade crossing safety issues will lend itself to safety policy, resulting in more cost-effective safety improvement programs.

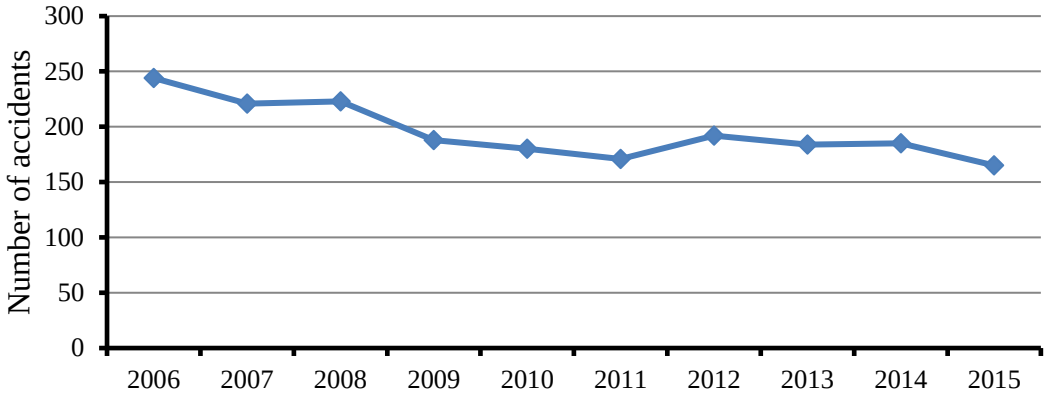


Figure 1-2 Crossing accidents across Canada from 2006 to 2015 (Transport Canada)

1.1.2 Vulnerable road users: pedestrians and cyclists

Health benefits of cycling and walking, referred to as active modes of transport, have been documented in several studies (Khattak and Rodriguez, 2005; Li et al., 2005; Krizek and Johnson, 2006; Saelens and Handy, 2008; de Hartog et al., 2010; Forsyth and Oakes, 2015). Further benefits, from a larger perspective, can be achieved due to a shift from motorized modes of transportation to cycling and walking, two environmentally friendly modes. This modal shift results in a decrease in greenhouse gas emissions and air pollution (de Hartog et al., 2010), benefitting entire communities. For these reasons, many municipalities have been aiming at promoting active modes of transport

particularly during the last decade. Examples include the installation and development of bicycle sharing systems and bicycle paths in cities such as Montreal, Toronto, Boston, Seattle, Chicago, New York, etc.

Walking and cycling in environments shared with motorized traffic, however, impart some risk to road users, in particular, pedestrians and cyclists. Fig. 1-2 displays fatalities for pedestrians and cyclists across Canada from 1975 to 2011. This figure implies no drastic decreasing trend in the last decade. Each year 7,500 cyclists sustain serious injuries across Canada, with 64% of deaths among cyclists, due to traffic crashes, on urban roads with a speed limit of less than 70 km/h (Canadian Automobile Association, 2016). To this end, improving safety is a vital factor to promoting active modes of transport (Fuzhong et al., 2005; Khattak and Rodriguez, 2005; Moudon et al., 2005; McMillan, 2007; Winters et al., 2011; Narayanamoorthy et al., 2013; Chataway et al., 2014; Braun et al., 2016). It is thus important to examine factors that correlate most strongly with pedestrian and cyclist injury frequencies, both vulnerable road users. Such studies can help decision-makers to identify high-crash locations and to select countermeasures that can mitigate crash and injury risk among pedestrians and cyclists.

Due to similarities between walking and cycling, both being non-motorized modes of transport, similar observed and unobserved or unmeasured factors may affect the safety of cyclists and pedestrians simultaneously. For example, drivers and cyclists in walk/cycle friendly neighborhoods or municipalities may have less hostile attitudes towards sharing the road, in part because these drivers (or their family members) may be pedestrians or cyclists themselves on other occasions (Aldred, 2016). Being more accepting of vulnerable road users can reduce the likelihood of crash with pedestrians and cyclists concurrently. Consequently, when crash data are available for both pedestrians and cyclists, a joint analysis is expected to provide richer insights into the key influences on safety dynamics of active modes of transport. In turn, the joint analysis helps transportation authorities in the implementation of appropriate countermeasures that can affect safety of both walking and cycling modes. This can lead to a more cost-effective allocation of funds while promoting the safety and mobility of all vulnerable road users.

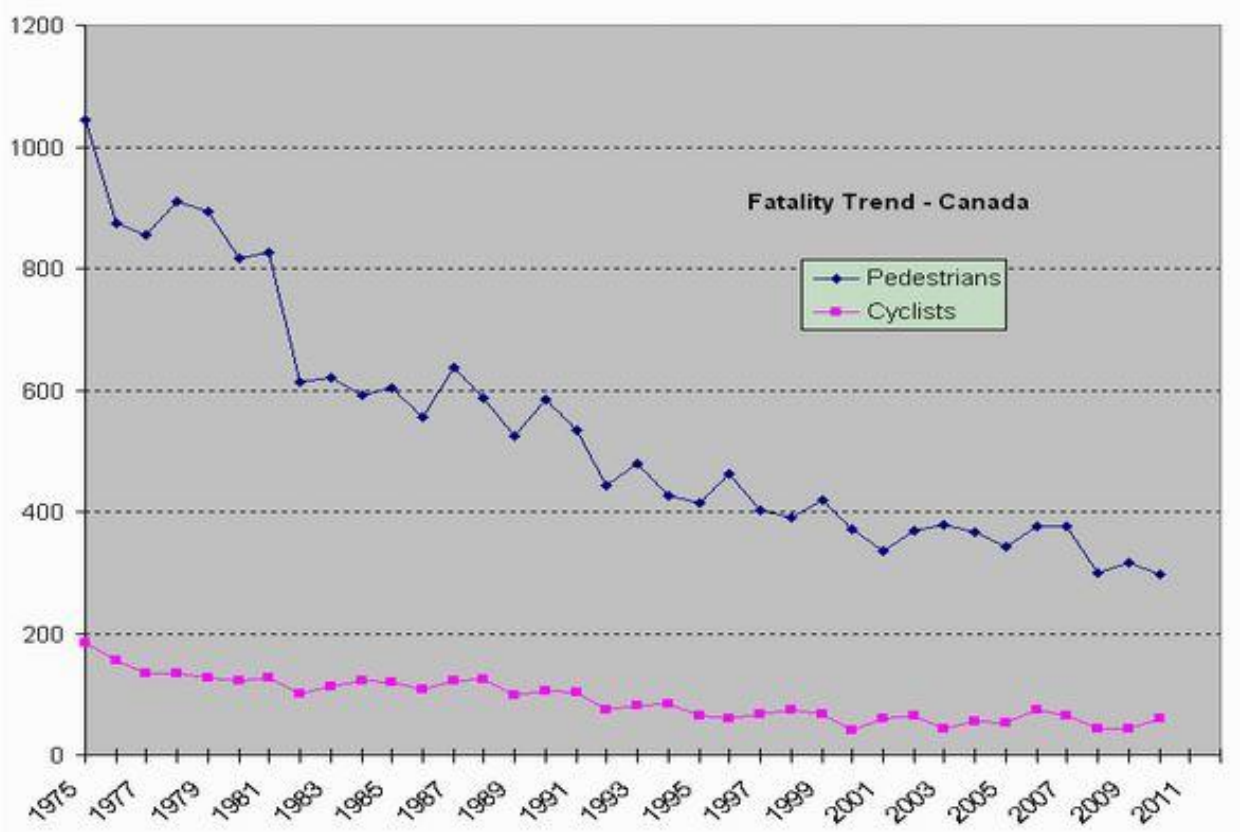


Figure 1-3 Fatalities for pedestrians and cyclists across Canada
(adopted from http://www.vehicularcyclist.com/fatals.html)

1.2 Road Safety Analysis at a Glance

The safety of a site (e.g., grade crossing, road intersection or highway segment) is usually measured by accident (crash) frequency and/or severity (Hauer, 1997). By definition, accident frequency is the number of observed accidents in a specific period of time (e.g., 5 years) at a site. Accidents can be divided into different types (e.g., head-on or rear-end accidents) and severity levels (e.g., minor injury, serious injury, and fatality). In transportation safety engineering, the primary aim is to improve the safety of transportation facilities. Based on the definition of safety, the main goal is thus to reduce accident frequencies and severities.

The process of improving the safety of a site often involves three basic steps. Firstly, it is necessary to quantify the safety condition at that site. Secondly, existing safety

problems should be determined. Finally, potential safety improvement treatments (countermeasures) should be examined for implementation. To undertake these tasks in a scientific way, the concept of transportation safety management arises. This is also because budget constraints require a rigorous effort to optimize the allocation of available funds to improvement projects. When there are numerous candidate sites (e.g., intersections in a city or railway crossings across Canada) to be considered for safety improvements with a limited budget, it is necessary to select a subset of sites for improvement that would yield maximum benefits. This can usually be done by ranking sites according to their safety performances and identifying those that have unusually high accident frequencies and/or severities. The identified sites are commonly referred to as high-crash locations (hotspots).

Once hotspots and their safety issues are identified, the next step is to select the best possible safety countermeasure(s) with consideration to the expected benefits and costs, regulations, logistics, and monetary constraints. At this stage, an accurate estimation of the effectiveness of potential countermeasures is required. In fact, countermeasure assessment is perhaps a key component of the road safety management process. The estimation of countermeasures effectiveness affects both safety outcomes (i.e., accident frequencies and severities) and economic appraisal. All the above tasks mainly rely on accident models, the accuracy of which has a significant impact on the safety management process.

1.2.1 Accident modeling

In accident modeling, a mathematical relationship between the number of accidents (of any type or severity level) and some contributing factors, mainly, site characteristics, is built. The resulting models are generally referred to as safety performance functions. The aim of the modeling process is to explain the occurrence of accidents based on a series of known or observed explanatory variables while accounting for the randomness associated with this occurrence in a probabilistic framework. Hence, the analysis includes a probabilistic component that assumes the occurrence of accidents follows a specific probability density function. In this regard, the most commonly used statistical models for accident-frequency analysis are presented in Chapter 2.

1.3 Limitations of Existing Methods and Practices

In traffic safety studies, there are almost inevitable concerns about the unobserved heterogeneity problem. Crash data are often limited since many unobserved or unmeasured factors that affect crash likelihood may not be available, causing the unobserved heterogeneity problem. In effect, some factors related to driver behavior, vehicle characteristics, site attributes, and environmental conditions may be missing in crash databases. In road safety studies, therefore, it is necessary to account for unobserved heterogeneity in order to obtain reliable statistical inferences. Another important concern relates to restrictive distributional assumptions that are common in analyzing crash data. When a model assumption does not hold, the true structure of the data is not reflected by the model. Consequently, that model is highly likely to draw misleading statistical inferences. In general, more flexible statistical models are thus needed to better capture the underlying structure of crash data. Based on the crash modeling literature, this dissertation focuses on the following interrelated methodological and empirical limitations.

1.3.1 Methodological limitations

- Often unobserved heterogeneity manifests itself in the form of multimodality in crash data, meaning that data are not generated from a unique density. In fact, crash data may be a collection of widely differing subpopulations. The commonly used standard or over-dispersed generalized linear models (e.g., Poisson-gamma models) do not fully address unobserved heterogeneity, assuming that crash frequencies follow unimodal exponential families of distributions.
- Random effects and random parameter models are limited in accounting for unobserved heterogeneity as the analyst should usually specify groupings in crash databases. However, unknown groupings that might exist due to unobserved features of crash data are ignored. In addition, sensitivity to parametric distributional assumptions may be a concern. Such restrictive assumptions may be in contrast with the structure of crash databases.
- Finite mixture models overcome the above issues by identifying latent subpopulations (components) of data based on data attributes. However, an

important limitation to finite mixture models is that the number of latent components must be prespecified before analyzing the data, but the analyst often does not know the underlying structure of the data a priori. To select the optimal number of components, different models with varying numbers of components must be fit to the data and the one providing the best fit chosen. In practice, a limited number of latent components are usually considered in finite mixture modeling, and the exact number of components may remain uncertain, both of which can compromise the results.

- Random effects and random parameter models are often employed to account for cross-group unobserved heterogeneity when analyzing data characterized by a hierarchical structure (observations nested within multiple groups or levels). Finite mixture (latent class) models are known as a viable alternative to account for unobserved heterogeneity; however, the application of finite mixture models in multilevel road safety studies is rare if non-existent.
- While providing valuable insights that help our understanding of crash mechanisms in the presence of correlated outcomes, most previous multivariate traffic safety studies have not considered whether their assumptions relating to the dependence structure reflects the true structure of the data. In effect, despite the general consensus that restrictive assumptions (e.g., homogeneity) in dependence structure may have an adverse effect on the accuracy of estimates, studies addressing the sensitivity of the results to these assumptions in multivariate settings are surprisingly rare in transportation safety studies.

1.3.2 Empirical limitations

- While a number of studies have examined spatial dependencies among observations in the crash literature, to our knowledge, no attempt has been made so far, especially, in Canada to accommodate spatial dependencies in railway grade crossing crash data. Overlooking spatial dependencies may result in spurious statistical inferences.
- Due to similarities between walking and cycling, both being non-motorized modes of transport, it is reasonable to hypothesize that both observed and unobserved (or unmeasured) site attributes may affect the safety of cyclists and pedestrians simultaneously. A few instances of analyzing pedestrian and cyclist

safety simultaneously at a macro-level (e.g., neighborhood level) exist; nevertheless, studies on the joint analysis of pedestrian and bicyclist injuries at a micro-level (e.g., intersections) are rare if non-existent. A macro-level safety analysis is valuable in terms of zone level policy decision making (Hadayeghi et al., 2010; Lee et al., 2015). Nevertheless, a micro-level modeling approach usually provides superior predictive performance, more specific high-crash location identification, and more detailed insight on factors that affect traffic safety; therefore, allowing decision makers to select safety improvement programs more properly (Huang et al., 2016). Furthermore, the quality of statistical inferences in macro-level modeling may be compromised due to the aggregation of data (Davis, 2004; Osama and Sayed, 2016).

- In modeling non-motorist safety, many studies lacked detailed motorist and non-motorist exposure information and used proxy measures instead. Depending on how these proxy measures thoroughly reflect traffic exposure, statistical inferences may be biased to some extent.

1.4 Research Objectives

Based on the above limitations, the principal objective of this thesis is to provide a novel methodological framework to overcome the unobserved heterogeneity problem for different types of crash data, with a prime focus on the use of Bayesian nonparametric Dirichlet process mixture models. Specifically, one general objective of this research is to explore the use of Dirichlet process mixture models in univariate, multilevel, and multivariate settings, the three most common settings often encountered in traffic safety studies. In accordance with the limitations highlighted in Section 1.3, this thesis centers on the following specific methodological and empirical objectives.

1.4.1 Methodological objectives

- Show how Dirichlet process mixture models can be used to examine and relax restrictive distributional assumptions, and eventually capture unobserved heterogeneity in univariate settings; and compare the proposed models with some of the most commonly used models for count data such as the Poisson-

gamma (negative binomial) model, the finite-mixture Poisson-gamma model, and the random intercepts model.

- Introduce a latent construct into the multilevel modeling framework to allow the identification of latent subpopulations at the highest levels of the hierarchy such as regions. The aim is thus to account for cross-group unobserved heterogeneity through a flexible latent class multilevel model.
- Investigate departures from restrictive dependence structures in multivariate settings and demonstrate how the robustness to standard assumptions can be verified; and propose a flexible multivariate model that allows for heterogeneous dependence structures in the joint analysis of correlated outcomes.

1.4.2 Empirical objectives

- Investigate the presence of spatial dependencies among railway grade crossings nested within the same geographic area using a flexible multilevel model developed based on Dirichlet process mixing.
- Verify the application of the derived flexible generalized linear model for data with excess zero counts such as railway grade crossing data.
- Investigate the joint analysis of pedestrian and cyclist safety at a micro-level using detailed motorist and non-motorist exposure measures; and examine the form of the dependence structure using a flexible multivariate model developed based on Dirichlet process mixing.
- Provide policy examples to highlight the practical advantages of the proposed methods. To this end, adopt the identification of high-crash locations, the detection of latent regional subpopulations, and the estimation of marginal effects, and provide comparison examples between the flexible and conventional models.

1.5 Dissertation Outline

This thesis is organized in five chapters:

- Chapter 1 introduces a brief background relating to transportation safety studies and highlights a number of empirical and methodological limitations in the crash literature.
- Chapter 2 provides a literature review presenting various statistical models and approaches used in traffic safety studies.
- Chapter 3 discusses the proposed methods to overcome the limitations reported in previous chapters, extending the conventional statistical models.
- Chapter 4 focuses on the analysis and the results. This chapter applies the proposed models to several datasets characterized by different characteristics in univariate, multilevel, and multivariate settings. For each setting, a policy exercise is conducted to show the advantages of our models.
- Chapter 5 concludes the thesis reporting our major contributions and future research directions.

CHAPTER 2

LITERATURE REVIEW

This chapter reviews the most important statistical models and approaches used to analyze crash frequencies. We also provide a review of previous studies relating to the safety analysis of railway grade crossings and vulnerable road users, being under-represented in the crash literature compared to roadway segment or intersection vehicle crash studies.

2.1 Statistical Analysis of Crash Data

A brief introduction of the most commonly used statistical models in road safety literature is given in the subsequent sections.

2.1.1 Poisson model

Poisson regression has traditionally been used in modeling accident data because of the nature of accidents being random count events (Hauer, 1997). Some instances of accident-frequency analysis using Poisson regression can be found in Gustavsson and Svensson (1976), Joshua and Garber (1990), and Miaou (1994). For a group of sites ($i=1,2,\ldots,n$) and for a specific period of time, given the observed and expected number of accidents, y_i and μ_i respectively, the occurrence of accidents can be assumed to be Poisson distributed independently over all sites:

$$y_i|\mu_i \sim Poisson(\mu_i) \tag{2-1}$$

The Poisson probability density function has the following form:

$$p(y_i|\mu_i) = \frac{e^{-\mu_i}}{y_i!}\,\mu_i^{y_i} \tag{2-2}$$

where, $p(y_i|\mu_i)$ is the probability of having y accidents in a specific time period for site i given the expected accident frequency μ_i. The Poisson parameter μ_i is the mean or average accident frequency, i.e., $E[y_i]$, and is assumed to be a function of the vector of site characteristics X such as traffic flow. This function usually has an exponential form:

$$\mu_i = f(\boldsymbol{X}, \boldsymbol{\beta}) = e^{\boldsymbol{\beta X}} \tag{2-3}$$

where $\boldsymbol{\beta}$ is the vector of coefficients including a constant to be estimated. An important characteristic (assumption) of the Poisson distribution is that the mean and variance are equal (Winkelman, 2008).

$$E[y_i|\mu_i] = VAR[y_i|\mu_i] = \mu_i \tag{2-4}$$

Clearly, Eq. 2.4 imposes a restriction on the flexibility of the Poisson density function. In fact, the Poisson assumption is often violated because the variance is usually larger than the mean in many accident data, resulting in over-dispersed data (the variance can also be smaller than the mean, i.e., under-dispersion). Therefore, alternative statistical models derived mostly from standard Poisson models are often used to relax the Poisson assumption. These alternative models are introduced in the following sections.

2.1.2 Poisson-gamma (negative binomial) model

As stated above, most of the accident data are over-dispersed, which is mainly caused by unobserved heterogeneity in data (Cameron and Trivedi, 1998; Maher and Summersgill, 1996; Mitra and Washington, 2007). The Poisson model can be extended by assuming its mean to follow a gamma distribution with mean 1 and variance α (Lord and Mannering, 2010). The resulting model is called Poisson-gamma model, which is most commonly known as Negative Binomial model. The Poisson-gamma model allows the variance to be greater than the mean and has become the most common statistical

model used in road safety literature (Persaud, 1994; Milton and Mannering, 1998; Karlaftis and Tarko, 1998; Heydeker and Wu, 2001; El-Basyouny and Sayed, 2006; Lord and Bonneson, 2007). The mathematical form of this model is defined as follows:

$$y_i|\theta_i \sim Poisson(\theta_i)$$

$$\theta_i = \mu_i e^{\varepsilon_i} \tag{2-5}$$

$$e^{\varepsilon_i} \sim gamma\,(\varphi, \varphi);\ E[e^{\varepsilon_i}] = 1$$

Consequently, the mean function can be written as follows:

$$\theta_i = f(\boldsymbol{X}, \boldsymbol{\beta}) = e^{(\boldsymbol{\beta X}+\varepsilon)} \tag{2-6}$$

The variance of the Poisson-gamma model is

$$VAR(y_i) = E(y_i) + \alpha[E(y_i)]^2 \tag{2-7}$$

where, α is called the over-dispersion parameter and is also defined as a function of the inverse dispersion parameter φ, that is, $\alpha=1/\varphi$ (the variance of the gamma distributed error term). In a hierarchical fashion, it is assumed that φ ~ gamma(a, b); where a and b are hyper-parameters (see Lord and Miranda-Moreno (2008) for a detailed discussion).

2.1.3 Poisson-lognormal model

The Poisson-lognormal model is similar to the Poisson-gamma model, except it assumes a lognormally distributed error term. This model is more flexible than Poisson-gamma in accommodating the multivariate nature of outcomes and spatial correlation (Aguero-Valverde and Jovanis, 2008; Aguero-Valverde and Jovanis, 2009; Ma et al., 2008). In several applications, it has been shown that the Poisson-lognormal model performs better than the Poisson-gamma model in terms of model fitting (Winkelman, 2008). The Poisson-lognormal model can be easily employed under the Bayesian paradigm; thus, the majority of studies employing this model use Bayesian statistics for computational convenience (Lord and Miranda-Moreno, 2008; El-Basyouny and Sayed, 2009a). The mathematical expression for the Poisson-lognormal model can be defined as

$$y_i|\theta_i \sim Poisson(\theta_i)$$

$$\theta_i = \mu_i e^{\varepsilon_i}$$

$$e^{\varepsilon_i} \sim lognormal\ (0, v)\ or\ \varepsilon_i \sim normal(o, v);\ E[e^{\varepsilon_i}] = 1$$

$$v^{-1} \sim gamma\ (a, b) \qquad (2\text{-}8)$$

$$E(y_i) = \mu_i e^{(0.5v)}$$

$$VAR(y_i) = E(y_i) + [E(y_i)]^2(e^v - 1)$$

where, a and b (in the Bayesian approach) are hyper-parameters to be defined by the analyst (Lord and Miranda-Moreno, 2008).

2.1.4 Crash models in the form of generalized linear models

Generalized linear models (McCullagh and Nelder, 1989; Zeger and Karim, 1991) have been extensively used in analyzing road safety data, conveniently handling crash data through a linear relationship between covariates and log-transformed outcomes such as crash frequencies. Indeed, over-dispersed generalized linear models such as Poisson mixtures (e.g., negative binomial or Poisson-gamma, Poisson-lognormal, etc.) constitute the mainstream approach to accounting for heterogeneity in crash data (Persaud, 1994; Hauer, 1997; Milton and Mannering, 1998; Karlaftis and Tarko, 1998; Shankar et al., 2003; Ukkusuri et al., 2012). The over-dispersed generalized linear model assumes that crash data follow a unique exponential density. Nevertheless, crash data may arise from a collection of widely differing subpopulations, so over-dispersed generalized linear models do not fully account for unobserved heterogeneity.

A generalized linear model in its simplest form for count data can be described as follows. Let y_i be the observed outcome of interest (e.g., observed crash frequency) for site i. Let $\boldsymbol{X}$ and $\boldsymbol{\beta}$ be the vectors of covariates (i.e., site characteristics) and the respective regression coefficients excluding the intercept β_0. Then, the model outcome mean λ_i can be related to the covariates using a logarithmic link function $g(.)$,

$$y_i|\boldsymbol{X}_i \sim Poisson\ (\lambda_i)$$

$$log(\lambda_i) = \beta_0 + \boldsymbol{\beta X}_i \qquad (2\text{-}9)$$

2.1.4.1 Over-dispersed (random effects) generalized linear models

The above model does not account for over-dispersion and unobserved heterogeneity. Therefore, an extension can be applied to handle over-dispersion. The most common way to overcome heterogeneity is to include an additive error term ε_i,

$$y_i|\boldsymbol{X}_i, \varepsilon_i \sim Poisson\,(\lambda_i)$$
$$log(\lambda_i) = \beta_0 + \boldsymbol{\beta X}_i + \varepsilon_i \quad (2\text{-}10)$$

The above model is an over-dispersed generalized linear model, which is also referred to as the random effects model. Depending on the distributional assumption for the error term, the above model results in different Poisson mixture settings. Two common Poisson mixtures often used in road safety literature are the Poisson-gamma (negative binomial) model and the Poisson-lognormal model that are respectively obtained by assuming

$e^{\varepsilon_i}|\varphi \sim gamma(\varphi, \varphi);\ where\ \varphi \sim gamma(.)$

$\varepsilon_i|v_\varepsilon \sim normal(0, v_\varepsilon);\ where\ v_\varepsilon^{-1} \sim gamma(.)$

2.1.5 Random parameter models

Another approach to overcoming unobserved heterogeneity in crash data is based on random parameter models such as a random parameter negative binomial model (Anastasopoulos and Mannering, 2009; Venkataraman et al., 2014; Wu et al., 2013; Chen and Tarko, 2014; Mannering and Bhat, 2014, Barua et al., 2015, Coruh et al., 2015). In random parameter models, different sets of parameters are estimated for different observations or groups of observations. Therefore, the effects of covariates (contributing factors) are not fixed across all data; instead, they are assumed to have a distribution across heterogeneous subsets. While standard random parameter models are limited in their restrictive distributional assumptions, further extensions such as the heterogeneity-in-means approach (Venkataraman et al., 2014) are possible to better address heterogeneity. As discussed in Mannering and Bhat (2014), however, an important limitation to random parameter models is that the analyst must prespecify groupings of observations across which parameters vary. As a consequence, unknown

groupings that might exist due to unobserved features are ignored. Based on the previous notation, a generic random parameter model can be specified as

$$
\begin{aligned}
&y_i|\boldsymbol{X}_i, \varepsilon_i \sim Poisson\,(\lambda_i)\\
&log(\lambda_i) = \beta_{0i} + \boldsymbol{\beta_i X}_i + \varepsilon_i\\
&\boldsymbol{\beta}_i|v_{\boldsymbol{\beta}} \sim normal\big(\boldsymbol{\beta}, v_{\boldsymbol{\beta}}\big)\\
&\beta_0|v_\beta \sim normal\big(\beta_0, v_\beta\big)
\end{aligned}
\qquad (2\text{-}11)
$$

In the above model, besides the error term, regression coefficients vary between observations (in univariate settings) or between groups of observations when groupings exist in the data.

2.1.6 Finite mixture (latent class) models

As discussed in Mukhopadhyay and Gelfand (1997), compared to over-dispersed generalized linear models, a more comprehensive approach to addressing heterogeneity would be the finite mixture or latent class models. As Park and Lord (2009) stated, "the mixture model can help provide the nature of the over-dispersion in the data." Accordingly, a number of road safety studies have recently employed finite mixture models to analyze crash frequency data or differing injury-severity levels (Park and Lord, 2009; Xiong and Mannering, 2013; Zou et al., 2014, Cerwick et al., 2014; Shaheed and Gkritza, 2014).

One important limitation to finite mixture models is that the number of latent components must be prespecified before analyzing the data, but the analyst often does not know the underlying structure of the data a priori. To select the optimal number of components, different models with varying numbers of components must be fit to the data and the one providing the best fit chosen. In practice, a limited number of latent components are usually considered in finite mixture modeling, and the exact number of components may remain uncertain, both of which can compromise the results. In this regard, Behnood at al. (2014) argue that such a limited number of components may result in inadequate approximation of the heterogeneity. For further discussion related to finite mixture modeling, see Mannering et al. (2016).

Studies in fields such as econometrics have employed finite mixture random parameter models to overcome some of the above issues. This approach relaxes the homogeneity assumption in each latent component of the mixture. In other words, model parameters can vary within each latent component. To our knowledge, such an approach has not been employed in modeling crash frequency data. In road safety literature, Xiong and Mannering (2013) adopted a finite mixture random parameter model to examine the effects of guardian supervision on adolescent driver-injury severities. While such an approach captures unobserved heterogeneity, similar to finite mixture models, the need to prespecifying latent components and the limited number of components are shortcomings. For a comprehensive discussion on unobserved heterogeneity in road safety data see Mannering et al. (2016).

Let f_Y be a density of interest for observations y, $\boldsymbol{\beta}$ be the vector of model coefficients, w_r be the weight of component r in the mixture, and C be the total number of components, a finite mixture model can be defined as

$$f_Y(y|w_r, \boldsymbol{\beta}_r) = \sum_{r=1}^{C} w_r f_r(y|\boldsymbol{\beta}_r) \qquad (2\text{-}12)$$

$$\sum_{r=1}^{C} w_r = 1;\ w_r > 0$$

For instance, a mixture of negative binomial densities for count data can be defined by substituting f_r in Eq. 2-12 as follows:

$$f_r = NB(\lambda_{ir}, \varphi_r)$$

where φ is the over-dispersion parameter as described in Section 2.1.2.

2.1.7 Multilevel (hierarchical) models

Crash data are often characterized by a multilevel (hierarchical) structure in which observations at the lower level(s) of the hierarchy are nested in different groups (e.g., vehicles, sites, geographical areas, etc.) at the higher level(s) (Huang and Abdel-Aty, 2010; Dupont et al., 2013). Due to unobserved group-specific factors, such a hierarchical structure challenges the basic assumption of independency between residuals since observations nested in the same groups usually share similar unknown and/or

unmeasured traits and are thus correlated (Heydari et al., 2014a). In fact, if the hierarchical structure of the data is not accounted for through adequate statistical techniques, the estimated standard errors could be underestimated, resulting in erroneously estimated narrow confidence intervals (Lenguerrand et al., 2006; Dupont et al., 2013). Given the importance of the problem, instances of multilevel modeling in road safety have been numerous over the last decade; see, for example, Jones and JØrjensen (2003), Kim et al. (2007), Yannis et al. (2007), Huang et al. (2008), Helai et al. (2008), Cruzado and Donnell (2010); Heydari et al. (2014a), Islam and El-Basyouny (2015). Readers are referred to Huang and Abdel-Aty (2010) and Dupont et al. (2013) for a comprehensive review of multilevel modeling in road safety literature.

In road safety, the multilevel structure of the data is often due to the nesting of observations in various geographical areas (Yannis et al., 2007; Yannis et al., 2008; Huang and Abdel-Aty, 2010; Dupont et al. 2013; Papadimitriou et al., 2014). In such circumstances, it is quite plausible to speculate that sites such as railway grade crossings situated in the same regions share a number of similar unknown characteristics. For instance, these characteristics can be generated as a result of regional traffic regulations, driver demography and behavior, climate-related features, etc. Therefore, spatial dependencies may exist among sites sampled from similar geographical areas. In this regard, for example, Papadimitriou et al. (2014) investigated motorcycle riding under the influence of alcohol in 19 European countries and found significant regional variations.

With respect to the spatial concept, it should be noted that the conditional autoregressive model incorporating structured spatial random effects is one of the major spatial models used in road safety literature (Aguero-Valverde, 2013; Wang and Kockelman, 2013; Barua et al., 2014). It is important to highlight that the conditional autoregressive model does not differentiate between separate geographical areas, whereas it estimates spatial random effects (neighborhood effects) to account for the proximity of sites (e.g., intersections) that might share similar unobserved covariates (Aguero-Valverde, 2013; Dupont et al., 2013). For that reason, when the interest is in explicitly modeling the effect of geographical areas (or separation of geographical areas), as in this paper, the multilevel framework is a viable technique to accommodate spatial dependencies in the analysis (Huang and Abdel-Aty, 2010; Dupont et al. 2013).

In the next section, we present a generic parametric (standard) random intercept multilevel model.

In multilevel data, as discussed earlier, it is essential to account for group-specific effects. Three main approaches have been proposed in literature to address this need: random effects models, random parameters models, and latent class or finite mixture models. Random effects models assume fixed parameters associated with the covariates but random intercept or error term (Kim et al., 2007; Heydari et al., 2014a). In contrast to random effects models, in multilevel settings, random parameters models allow model covariates to vary across groups of observations to account for cross-group heterogeneity in data (El-Basyouny and Sayed, 2009b; Chen and Tarko, 2014; Islam and El-Basyouny; 2015). In general, random parameters models constitute therefore a more comprehensive way of overcoming unobserved heterogeneity in crash data including multilevel crash data, in comparison to random effects models. The higher quality and performance of random parameters models obviously comes with a higher cost in terms of computational complexity compared to random effects models (Chen and Tarko, 2014; Venkataraman et al., 2014). For a discussion related to random effects models and random parameters models, see Anastasopoulos and Mannering (2009), Lord and Mannering (2010), and Chen and Tarko (2014).

The finite mixture modeling approach (Greene and Hensher, 2003; Park and Lord, 2009; Zou et al., 2012; Xiong and Mannering, 2013; Zou et al., 2014) is another alternative to overcome unobserved heterogeneity in crash data. However, to our knowledge, the application of finite mixture models in multilevel traffic safety studies has been limited in contrast to single-level safety studies. For a comparison between random parameter models and finite mixtures or latent class models, interested readers are referred to Behnood et al. (2014) and Mannering and Bhat (2014).

2.1.7.1 Random intercepts (random effects) multilevel model

Let r denotes groupings, given the notation presented in Section 2.1.5, a typical multilevel model with random intercepts across groupings can be obtained by extending the previously discussed simple Poisson-lognormal model as follows:

$$
\begin{aligned}
&y_{ri}|\boldsymbol{X}_{ri}, \boldsymbol{\beta}, \varepsilon_{ri}, \eta_r \sim Poisson\,(\lambda_{ri}) \\
&log(\lambda_{ri}) = \eta_r + \boldsymbol{\beta}\boldsymbol{X}_{ri} + \varepsilon_{ri} \\
&\eta_r|m_\eta, v_\eta \sim normal(m_\eta, v_\eta) \\
&\varepsilon_{ri}|v_\varepsilon \sim normal(0, v_\varepsilon)
\end{aligned}
\tag{2-13}
$$

where m_η and v_η are, respectively, the mean and the variance for the random intercepts η_r. The intra-group correlation γ can be obtained from

$$\gamma = v_\eta/(v_\eta + v_\varepsilon) \tag{2-14}$$

It can be seen that the random intercept Poisson-lognormal model assumes a common normally distributed random intercept at the grouping level. Considering different groupings in data, it is plausible to doubt that all of them are generated from a single distribution. In fact, one can question the presence of latent subpopulations among these groupings.

2.1.8 Multivariate models

A larger body of transportation safety literature stresses the importance of accounting for dependence among correlated crash types (Ye et al., 2009; Lord and Mannering, 2010; Mannering and Bhat, 2014; Lee et al., 2015; Mothafer et al., 2016). For example, Ye et al. (2009) analyzed different crash types such as head-on, sideswipe, rear-end, and angle crashes simultaneously. The authors state that "there is a need to be able to model the expected frequency of crashes by collision type at intersections to enable the detection of problems and the implementation of effective design strategies and countermeasures. Statistically, it is important to consider modeling collision type frequencies simultaneously to account for the possibility of common unobserved factors affecting crash frequencies across crash types". Similarly, Lord and Mannering (2010) argue that neglecting the correlation among crash counts (by type or severity level) results in losses in estimation efficiency.

In fact, research on joint or multivariate modeling of correlated outcomes (differing injury-severity levels or crash types) has recently proliferated in transportation safety studies (Park and Lord, 2007; Ma et al., 2008; Aguero-Valverde and Jovanis, 2009; El-

Basyouny and Sayed, 2009a; Anastasopoulos et al., 2012; Dong et al., 2014; Zhan et al., 2015; Anastasopoulos, 2016; Barua et al., 2016; Serhiyenko et al., 2016). These studies have focused on different empirical and theoretical aspects of multivariate modeling employing both Bayesian and frequentist estimation techniques. In road safety research, when modeling crash frequencies of different types, most studies have employed multivariate Poisson-lognormal models that can account for both negative and positive correlation, whereas multivariate negative-binomial models only accommodate positive correlation (Winkelmann, 2008; Zhan et al., 2015). For example, Aguero-Valverde and Jovanis (2009) employed multivariate Poisson-lognormal models to analyze differing crash severities and to identify high-crash locations. Zhan et al. (2015), using multivariate Poisson log-normal models, provided a parallel sampling scheme that improves the customary Markov chain Monte Carlo approach, reducing run times. Similarly, Serhiyenko et al. (2016) introduced integrated nested Laplace approximations for Bayesian estimation of multivariate crash counts. Under the frequentist framework, Narayanamoorthy et al. (2013) adopted a composite marginal likelihood approach to make statistical inferences for their multivariate setting. See Narayanamoorthy et al. (2013), Mannering and Bhat (2014), and Zhan et al. (2015) for a more elaborate discussion on multivariate models and different available formulations.

Modeling correlated outcomes in transportation safety studies is not limited to analyzing different crash types or injury-severity levels. For instance, previous research has investigated the simultaneity (correlation among outcomes) in modeling traffic safety, health care services, and motorization (Anwaar et al., 2012). Therefore, modeling correlated outcomes in the context of transportation safety is often encountered, which makes its application of a paramount importance. Consequently, to improve the crash safety analysis framework, addressing methodological barriers relating to modeling correlated outcomes are warranted.

To more effectively overcome heterogeneity not accounted for by observed covariates, multivariate random parameter models have been recently employed in a few road safety studies, generally outperforming classical multivariate models (Russo et al., 2014; Anastasopoulos, 2016; Anastasopoulos and Mannering, 2016; Barua et al., 2016). For example, Russo et al. (2014) compare factors that affect injury-severity in angle crashes using a random parameter bivariate ordered Probit model. The latter study accounts for within-crash correlation between pairs of crash-involved drivers by fault status, taking

advantage of random parameter models to account for unobserved heterogeneity. Nevertheless, instances of multivariate latent class models in the crash literature are rare (Buddhavarapu et al., 2016; Heydari et al., 2016a). Buddhavarapu et al. (2016) employed a conventional finite mixture model with two components to formulate their multivariate latent class model. Instead, Heydari et al. (2016a) employed a form of such model to jointly analyze differing injury-severity levels for highway segments in Ontario using a flexible Bayesian semiparametric Dirichlet process mixing approach. The latter study has the advantage of inferring the number of components as a stochastic parameter like other model parameters.

Given the importance of latent class models in dealing with unobserved heterogeneity, it would be interesting to develop and examine multivariate models that not only account for dependence across outcomes, but also identify latent subpopulations in data. In this regard, Mannering et al. (2016) state that "A particularly appealing way to combine unobserved heterogeneity effects with a multivariate outcome context (with the outcomes being of different types, including continuous, count, nominal, ordered, and grouped outcomes) is based on identifying stochastic latent constructs (for example, unobserved driver- specific psychological factors)". Few examples of integrating latent psychological constructs into modeling correlated outcomes exist in the transportation literature (Bhat and Dubey, 2014).

2.1.8.1 Multivariate Poisson-lognormal model

Let y_{ik} and λ_{ik} denote, respectively, the observed and the expected crash frequencies of crash type k (here $k = (1, 2)$) for site $i = (1, 2, \ldots, n)$; $X = (X_1, X_2, \ldots, X_m)$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_m)$ denote the vectors of m covariates (here, intersection characteristics) and their respective regression coefficients; β_0 denotes a fixed intercept across sites for crash type k; Σ denotes the covariance matrix; ε denotes correlated error terms varying across sites; R and K denote the scale matrix and the degrees of freedom, respectively, in a Wishart density. The standard multivariate Poisson-lognormal model can be defined as

$$
\begin{aligned}
& y_{ik} | \boldsymbol{X}_{ik}, \beta_{0k}, \varepsilon_{ik} \sim Poisson\ (\lambda_{ik}) \\
& log(\lambda_{ik}) = \beta_{0k} + \boldsymbol{\beta}_k \boldsymbol{X}_{ik} + \varepsilon_{ik} \\
& \varepsilon_{ik} \sim MVN(0, \Sigma)
\end{aligned}
\qquad (2\text{-}15)
$$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{kk} \end{bmatrix}$$

$$\Sigma^{-1} \sim Wishart(R, K)$$

The dependency across outcomes is captured through correlated error terms that are assumed to follow a multivariate normal density with the mean 0 and covariance matrix Σ. Therefore, a unique dependency structure is assumed here for all data points resulting in a homogeneous correlation structure. The main idea here is that after accounting for the effect of known covariates, some sources of correlation (due to omitted covariates) may still exist in the data that can be captured through the correlated error terms. The significance of omitted variables in crash modeling is highlighted in the safety literature (Mitra and Washington, 2012).

2.2 Safety Analysis of Railway Grade Crossings

Studies on grade crossing safety have been relatively limited in the crash literature. These studies are mainly concerned with accident modeling in which the aim is to identify factors that affect crash frequency or injury-severity at crossings. This includes accident-frequency analyses (Saccomanno and Lai, 2005; Oh et al., 2006; Yan et al., 2010) and accident-consequence analyses (Saccomanno et al., 2004; Eluru et al., 2012). For a summary, see Chadwick et al. (2014).

Regarding accident-frequency modeling at railway crossings, earlier accident prediction models include (i) the Peabody Dimmick Formula developed in 1941, (ii) the National Cooperative Highway Research Program Hazard Index developed in 1964, (iii) the New Hampshire Index developed in early 1970s (Austin and Carson, 2002), and (iv) the US Department of Transportation (USDOT) accident models developed in early 1980s. However, these models have different limitations. For example, the first three models allow for a limited number of independent variables (contributing factors) in the accident model. The USDOT model accommodates more contributing factors compared to other earlier accident models, but it is not flexible to allow alternative covariates in the model and is relatively complex to be employed (Oh et al., 2006). Readers are referred to Oh et al., (2006) for a comprehensive summary and discussion about earlier efforts of accident modeling for grade crossings. In Canada, Saccomanno

et al. (2003, 2004) developed a set of crossing accident models using Canadian grade crossing data, which was found to be comparable to the USDOT model (Chaudhary et al., 2011).

An important step in crossing accident modeling was made by employing the Poisson and the negative binomial models (Austin and Carson, 2002, Saccomanno et al., 2004; Park and Saccomanno, 2005a; Park and Saccomanno, 2005b; Millegan et al., 2009). In this regard, significant advances were made by Saccomanno and Lai (2005). The authors adopted a factor/cluster analysis approach to divide crossings to homogeneous groups and then developed accident models for each group separately. This framework allowed considerable improvement in the accuracy of developed accident models. Other researchers adopted alternative statistical models such as zero-inflated and gamma models to resolve issues such as the excess number of zero accidents in data and under-dispersion (Lee et al., 2004; Oh et al., 2006; Hu and Lee, 2008). However, zero-inflated models have been criticized due to their assumption of safe state, which does not seem to be realistic in the context of road safety analysis (Lord et al., 2005; Lord et al., 2007). Instead of the above-mentioned parametric approaches, there have also been a few studies that employed nonparametric methods such as hierarchical tree-based regression model for accident-frequency analyses (Yan et al., 2010; Thakali, 2016).

2.3 Safety Analysis of Active Modes (Walking and Cycling)

Many researchers have investigated factors such as built environment, socio-economic characteristics, traffic exposure, facility types (e.g., intersection, road segment, etc.) and their geometric and operational characteristics that may affect pedestrian and cyclist crash frequencies or injury-severity levels (Lyon and Persaud, 2002; Shankar et al., 2003; Noland and Quddus, 2004; Lee and Abdel-Aty, 2005; Eluru et al., 2008; Cho et al., 2009; Clifton et al., 2009; Pulugurtha and Sambhara, 2011; Tay et al., 2011; Ukkusuri et al., 2012; Mohamed et al., 2013; Wang and Kockelman, 2013; Strauss et al., 2014; Quistberg et al., 2015; Zhang et al., 2015; Aldred, 2016; Amoh-Gyimah et al., 2016; Behnood and Mannering, 2016; Jung et al., 2016; Osama and Sayed, 2016; Yasmin and Eluru, 2016).

Previous research studies have considered various empirical and methodological aspects of modeling non-motorist safety and provided valuable insights in this regard. For example, Behnood and Mannering (2016) analyzed differing injury-severity levels

sustained by pedestrians in Chicago using both latent class and random parameter logit models, which better account for unobserved heterogeneity compared to conventional models. The authors focused on addressing temporal stability in modeling pedestrian injury-severity levels while comparing the above models. Their study showed that the effect of explanatory variables on pedestrian injury-severity levels may change over time. Another recent study conducted by Yasmin and Eluru (2016) employed latent class negative binomial models to analyze pedestrian and bicyclist crashes separately at the traffic analysis zone level in the Island of Montreal and the City of Toronto. The latter study confirmed the superiority of the latent class formulation compared to the standard negative binomial model.

In the pedestrian and cyclist crash literature, while both aggregate spatial units (macro-level such as census tract) and disaggregate spatial units (micro-level such as intersections) have been considered, most studies have analyzed pedestrian and cyclist crash data separately. Only a few research efforts have jointly modeled pedestrian and cyclist safety using multivariate settings that allow the analyst to account for correlation among these two outcomes (Narayanamoorthy et al., 2013; Lee et al., 2015; Nashad et al., 2016). For instance, Narayanamoorthy et al. (2013) recast count models as a special case of generalized ordered-response models to jointly analyze pedestrian and cyclist injury-severities at the census tract level. Similarly, Nashad et al. (2016) adopted a macro-level analysis approach considering crashes at the statewide traffic analysis zone level in Florida to simultaneously analyze pedestrian and cyclist crashes.

Nevertheless, to our knowledge, previous research studies on the joint analysis of pedestrian and bicyclist injuries at a micro-level (e.g., intersections) are rare if non-existent. While a macro-level safety analysis is valuable in terms of zone level policy decision making (Hadayeghi et al., 2010; Lee et al., 2015), a micro-level modeling approach usually provides superior predictive performance, more specific high-crash location identification, and more detailed insight on factors that affect traffic safety, allowing decision makers to select safety improvement programs more properly (Huang et al., 2016). Furthermore, the quality of statistical inferences in macro-level modeling may be compromised due to the aggregation of data (Davis, 2004). For a discussion in this regard, see Osama and Sayed (2016).

2.4 Bayesian Posterior Inference

The most commonly used model estimation techniques are maximum likelihood estimation (MLE) and Bayesian inference. MLE is largely used in different applications (Washington et al., 2011). Here, we briefly discuss Bayesian inference, a viable alternative to MLE.

Bayesian inference is developed based on the Bayes theorem (Gelman et al., 2003), which essentially allows prior belief or knowledge to be utilized in analyses. For every parameter of interest, in effect, a prior distribution must be specified. The outcome is then obtained by mixing the prior and the likelihood, i.e., data. In the Bayesian context, analyses outcomes are referred to as posterior densities. That is, all model outcomes are in the form of probability density functions, whereas in classical methods (e.g., MLE) point estimates are provided (Carlin and Louis, 2009). This property of Bayesian statistics allows accounting for uncertainties in its holistic form. Under the Bayesian paradigm, the posterior density $f(\boldsymbol{\beta}|y)$ is defined as

$$f(\boldsymbol{\beta}|y) \propto f(y|\boldsymbol{\beta})f(\boldsymbol{\beta})$$

$$f(\boldsymbol{\beta}|y) = \frac{f(y|\boldsymbol{\beta})f(\boldsymbol{\beta})}{\int f(y|\boldsymbol{\beta})f(\boldsymbol{\beta})d\boldsymbol{\beta}} \quad (2\text{-}16)$$

where $\boldsymbol{\beta}$ is the vector of unknown parameters; y denotes observed data; $f(y|\boldsymbol{\beta})$ is the likelihood density; and $f(\boldsymbol{\beta})$ is the prior density.

Since inferring posterior densities involves high dimensional integrals, and thus, is highly intensive in terms of computation, the use of Bayesian methods has been limited in the past. These integrals cannot be solved analytically; instead, Markov chain Monte Carlo (MCMC) simulation techniques are commonly employed. Following advances made in computational power in 1990s, Bayesian methods had become the center of attention in many areas of research. As a powerful estimation technique, Bayesian concepts have been extensively used in the analysis of complex data in different fields such as biostatistics, economics, reliability engineering, computer science, and social sciences.

Although a number of studies have recognized the advantages of the Bayesian statistics in analyzing transportation data in the past 10 to 15 years (Miaou and Song, 2005;

Aguero-Valverde and Jovanis, 2006; Miranda-Moreno et al., 2007; Lord et al., 2008; Ma et al., 2008; El-Basyouny and Sayed, 2012; Miranda-Moreno et al., 2013), its practical use in transportation engineering has been limited as compared to other fields. Since the Bayesian framework will be mainly used in this research, here, we provide a concise summary of its four main advantages: (Interested readers are referred to Heydari et al., (2014b) for a discussion in this regard.)

First, information from different sources and in different forms, such as expert opinion and previous studies, can be incorporated into the analysis when assigning prior distributions to model parameters. This is a vital feature as it has the potential to eliminate estimation biases due to limited data (Lord and Miranda-Moreno, 2008; Miranda-Moreno et al., 2013; Daziano et al., 2013; Heydari et al., 2013). In fact, using proper priors, the sample size required to conduct a reliable road safety analysis can be reduced (Heydari et al., 2014b). Second, Bayesian statistics have the advantage of accommodating hierarchical models, which are capable of dealing with complex data structures (Gelman et al., 2003). Such data structures are often encountered in transportation research.

Third, regardless of the model complexity, an analyst can always apply the Bayes theorem to derive the posterior inference and then run MCMC simulations to obtain the posteriors without any additional effort to develop a method or algorithm for solving the problem. In contrast, in the Frequentist framework, when the standard MLE cannot solve a problem, a simulation based solution is still necessary, requiring development of problem-specific computational models and algorithms in most cases.

Lastly, a Bayesian approach allows direct interpretation of credible intervals (the Bayesian version of confidence intervals) on the probability that an estimate occurs in these intervals for a specific dataset. Such interpretation is not possible in the Frequentist paradigm; that is, the Frequentist approach cannot conclude, for an observed dataset, the probability for an estimate being in a certain interval. It can only state that a confidence interval contains the estimated value given that a substantial number of trials are repeated.

CHAPTER 3

METHODOLOGY

This chapter first provides a brief methodological background describing the main components of our approach. We then describe the proposed method in univariate, multilevel, and multivariate settings, the most common settings in transportation safety studies. This section concludes by discussing cluster detection algorithm and model selection criteria based on cross-validation predictive densities. The contents of this chapter have been published in Analytical Methods in Accident Research (see page iv for details).

3.1 Methodological Background

This thesis introduces a class of flexible statistical models that are rooted in Bayesian nonparametric literature based on Dirichlet process mixtures (Escobar and West, 1998; Walker et al., 1999; Neal, 2000; Muller and Quintana, 2004; Jain and Neal, 2004; Ohlssen et al., 2007; Hjort et el., 2010). In this regard, Escobar and West (1998) state that "Bayesian models involving Dirichlet process mixtures are at the heart of the modern nonparametric Bayesian movement". The Bayesian models used in this thesis are however semi-parametric since parametric distributional assumptions are not relaxed for all model parameters. This is mainly to retain the usual interpretation of explanatory variables.

The original ideas of Bayesian nonparametric methods were initially developed and discussed by Freedman (1963), Ferguson (1973), and Antoniak (1974); however, their applications were very limited due to computational complexities. It was mainly in the 1990s that Bayesian nonparametric models have attracted the attention of more researchers due to improvements in MCMC schemes and also substantial computational advances during those years. At that stage, several developments have been made in various aspects of Bayesian nonparametric modeling (Escobar, 1994; West and Turner, 1994; Bush and MacEachern, 1996; Mukhopadhyay and Gelfand, 1997; Kuo and Mallick, 1997; Hjort et al., 2010). Consequently, Bayesian nonparametric concepts have been used in different scientific articles mainly in biostatistics and computer science research (Ohlssen et al., 2007; Muller et al., 2007; Dhavala et al., 2010, Hannah et al., 2011; Gershman and Blei, 2012), whereas their use in transportation research, especially, transportation safety has been rare.

One of the main motivations behind the nonparametric Bayesian inference is to remove constrains associated with restrictive parametric assumptions. These constraints may affect inferences made by restrictive parametric models. Therefore, employing the Bayesian nonparametric approach enables us to circumvent restrictive distributional assumptions and make statistical models more reliable in terms of statistical inference. It is important to mention that, under the Bayesian paradigm, the term nonparametric does not mean that the model is parameter-free. In contrast, it may have an infinite number of parameters (Mallick and Walker, 1997; Muller and Quintana, 2004). In Bayesian nonparametrics, in effect, the number of parameters increases as the complexity of the data escalates. This characteristic leads to an important difference with finite mixture modeling in which the number of latent components must be decided in advance of the analysis. In Bayesian nonparametric modeling, however, the number of latent components is estimated as part of the estimation algorithm, which is more realistic, convenient, and flexible.

3.1.1 Realization of a Dirichlet process & Dirichlet process mixing

Let G, G_0 and α be an unknown density for a parameter of interest, a continuous baseline distribution (defining the location of the Dirichlet process) and a positive real precision (concentration) parameter, respectively. A Dirichlet process can be notated as

$$G \sim DP(\alpha G_0) \tag{3-1}$$

A Dirichlet process is a probability measure on the space of all measures (Mukhopadhyay and Gelfand, 1997; Escobar and West, 1998), where for any finite segment $S_1,\ldots, S_n$ of the parameter space, the vector of probabilities $(G(S_1),\ldots, G(S_n))$ follows a Dirichlet distribution with a vector of parameters $(\alpha G_0(S_1),\ldots, \alpha G_0(S_n))$ (Escobar and West, 1998; Muller and Quintana, 2004; Ohlssen et al., 2007). This can be denoted as

$$(G(S_1), \ldots, G(S_n)) \sim Dirichlet(\alpha G_0(S_1), \ldots, \alpha G_0(S_n)) \tag{3-2}$$

The concentration parameter α indicates the variability of a Dirichlet process around its baseline distribution. A low value of α indicates that G can be far from G_0, and vice versa. Therefore, the model with the above structure can be used as a diagnostic tool to verify the validity of a parametric assumption (Escobar and West, 1998; Ohlssen et al., 2007). A stick-breaking procedure (Ishwaran and James, 2002; Ohlssen et al., 2007) can be implemented to obtain random density functions drawn from a Dirichlet process. The main aim here is to have a set of random probabilities generated sequentially having a sum of one. Such restriction can be guaranteed by the stick-breaking algorithm (Sethuraman, 1994) that breaks a stick with a unit length into an infinite number of partitions. For a detailed discussion see Ishwaran and James (2002) and Muller and Quintana (2004). The stick-breaking procedure (ii) and (iii), as discussed in Ohlssen et al. (2007), is briefly described as follows:

(i) draw a set of random variables $\theta_1, \theta_2,\ldots$ from G_0;
(ii) draw a set of random variables $\xi_1, \xi_2,\ldots$ from a Beta(1, α);
(iii) allocate probabilities $p_1 = \xi_1$, $p_2 = (1 - \xi_1)\xi_2$, $p_3 = (1 - \xi_1)(1 - \xi_2)\xi_3$, ... to $\theta_1, \theta_2, \theta_3,\ldots$, respectively.

Note that the probability p and the expectation E for $\xi_1, \xi_2,\ldots$ (Beta distributed random variables) can be obtained from Eq. 3-3 and Eq. 3-4.

$$p(\xi_n) = \alpha {\xi_n}^{\kappa-1} \tag{3-3}$$

$$E(\xi_n) = (1+\alpha)^{-1} \tag{3-4}$$

An infinite mixture of points, which is the density function *f(.)* corresponding to *G*, represents realizations of the Dirichlet process (Muller and Quintana, 2004).

$$f(\cdot) = \sum_{n=1}^{\infty} p_n I_{\theta_n}, \quad \theta_n \sim G_0 \tag{3-5}$$

In Eq. 3-5, I_θ is an indicator function (measure) corresponding to θ. Note that *f(.)*, as defined in Eq. 3-5, is a discrete random probability model. As discussed in Ohlssen et al. (2007), a truncated Dirichlet process (TDP) can be used to approximate a full Dirichlet process with less computational effort, employing standard MCMC methods. To do so, it is necessary to limit the maximum number of possible clusters to *C* (i.e., substitute ∞ with *C*). Indeed, the truncation occurs at *C*; and therefore, *G* depends also on *C*, i.e., *G* ~ TDP(α, G_0, *C*). In this truncation, it is necessary to restrict the final probability p_c to be a very small value that is obtained from Eq. 3-6. The choice of *C* could in part be based on the precision parameter α and is approximately equal to $5\alpha+2$ (Ohlssen et al., 2007).

$$p_C = 1 - \sum_{n=1}^{C} p_n \tag{3-6}$$

$$C \approx 1 + log(\varepsilon)/log\left[\frac{\alpha}{1+\alpha}\right] \approx 5\alpha + 2 \tag{3-7}$$

The final form of *f*(.) collapses into a finite mixture model (Eq. 3-8) that estimates the posterior density of the number of latent clusters in data.

$$f(\cdot) = \sum_{n=1}^{\infty} p_n I_{\theta_n} \approx \sum_{n=1}^{C} p_n I_{\theta_n} \tag{3-8}$$

We discuss the specification of baseline distribution and priors for model parameters including the precision parameter in Chapter 4.

3.2 Overview of the Methodology

This section provides a schematic view (Fig. 3-1) of the methodological framework adopted in this thesis. This schematic view summarizes our approach. The method is centered on Dirichlet process mixing under the Bayesian nonparametrics paradigm. Note that the Dirichlet process mixture of points is applied to all settings while the Dirichlet process mixture of multivariate normals is applied to the multivariate setting.

The structure of the data gives rise to an applied setting for which Dirichlet process mixing allows departures from restrictive assumptions, adding flexibility to the model. Such flexibility is expected to result in a better model specification, as we show in this thesis, that in turn helps capture the underlying structure of the data more accurately.

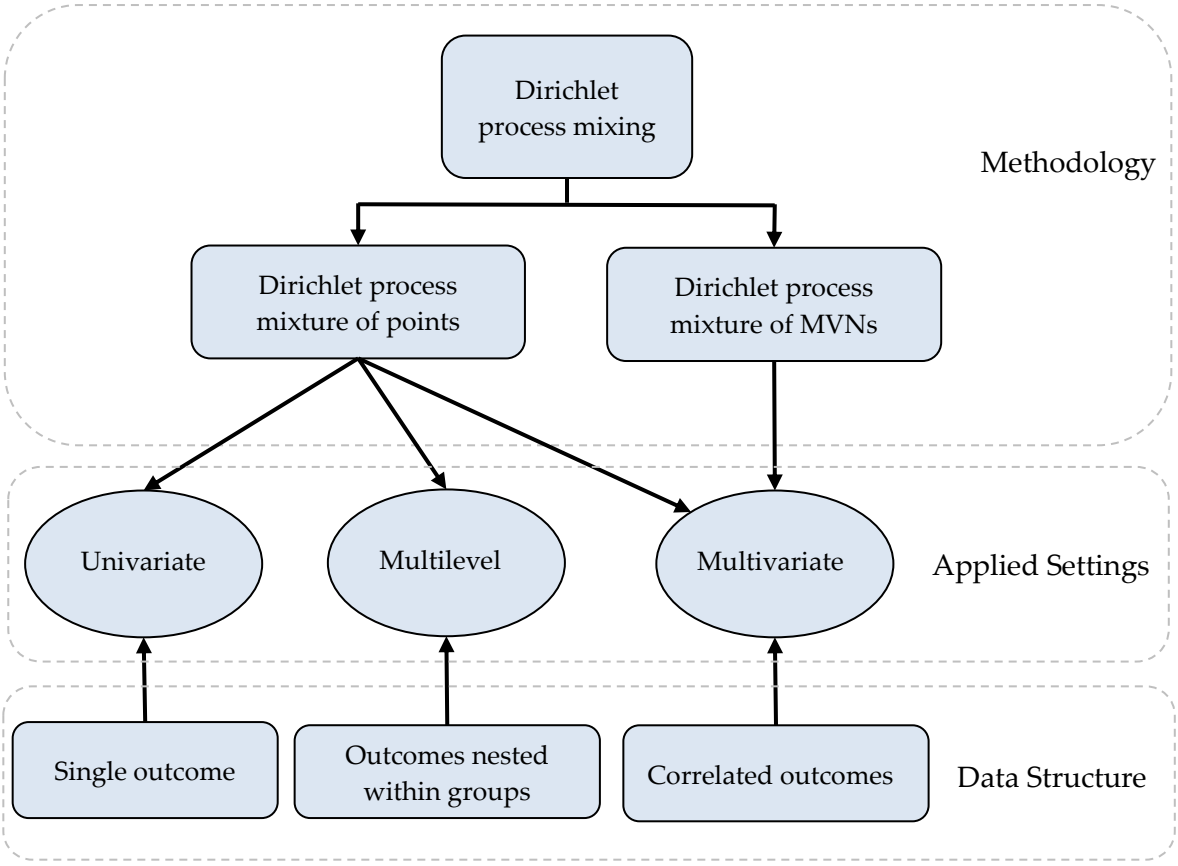


Figure 3-1 Schematic representation of the proposed approach

3.3 Univariate Settings

Univariate settings are the simplest form of crash datasets in which a single outcome (e.g., crash frequency) is modeled while there is not any grouping in the data. This means that all observations are assumed to be independent. In univariate analysis of crash data, generalized linear models are often used.

Given the modeling limitations discussed in Chapter 2 regarding the unobserved heterogeneity problem, this section introduces an alternative, a flexible Bayesian semiparametric generalized linear model (Escobar and West, 1998; Walker et al., 1999; Neal, 2000; Gelfand and Kottas, 2002; Muller and Quintana, 2004; Hjort et al., 2010). While the Bayesian nonparametric approach is used in other fields (Mukhopadhyay and Gelfand, 1997; Kleinman and Ibrahim, 1998; Ohlssen et al., 2007; Jara et al., 2007; Muller et al., 2007; Dhavala et al., 2010), its applications in transportation research or road safety studies are rare (Heydari et al., 2016b; Shirazi et al., 2016; Yu et al., 2016).

Bayesian nonparametric models are flexible in the sense that the number of parameters is not fixed and can vary according to data complexity (Gershman and Blei, 2012), taking advantage of Dirichlet process mixtures. These models relax restrictive parametric assumptions of conventional statistical models and allow the identification of latent components (Escobar and West, 1998). Interestingly, the number of latent components can be inferred from the data as part of the analysis, whereas this number must be prespecified (depending on how a priori uncertain it is) in finite mixture models. Inferring the estimated number of latent subpopulations through a systematic mathematical algorithm is more desirable and methodologically sound, assuming the data support such inferences. The following two subsections introduce extensions made to the models discussed in Section 2.1.5 based on a Dirichlet process mixing approach.

3.3.1 Generalized linear Dirichlet process mixture model

To add further flexibility to the standard random effects generalized linear model and as a surrogate to the over-dispersed generalized linear model (discussed in Section 2.1.7), a Dirichlet process mixture can be adopted to obtain the generalized linear

Dirichlet process mixture model. Including the error term in the intercept, we first write the standard random effects model presented in Section 2.1.5.1 as

$$
\begin{aligned}
&y_i|\boldsymbol{X}_i, \varepsilon_i \sim Poisson\ (\lambda_i) \\
&log(\lambda_i) = \beta_{0i} + \boldsymbol{\beta X}_i \\
&\beta_{0i}|v_0 \sim normal(m, v)
\end{aligned}
\tag{3-9}
$$

where m and v are the mean and the variance for the random intercepts. We then employ a Dirichlet process mixture over the intercepts β_{0i} to tackle heterogeneity with respect to the location of the mean by allowing multimodality as in finite mixture models (Mukhopadhyay and Gelfand, 1997). We retain the linear form for coefficients $\boldsymbol{\beta}$, which in turn retain their usual interpretations.

Given the notation in Section 2.1.5, let β_{0r} be the intercept for cluster r (1,2,…, C) and G_0 be a normally distributed baseline distribution for β_{0r} with the mean m_0 and the variance v_0. A generic form of the generalized linear Dirichlet process mixture model can be written as follows:

$$
\begin{aligned}
&y_i|\boldsymbol{X}_i, \beta_{0r} \sim Poisson\ (\lambda_i) \\
&log(\lambda_i) = \beta_{0r} + \boldsymbol{\beta X}_i \\
&\beta_{0r} \sim Dirichlet(\alpha G_0) \\
&G_0|m_0, v_0 \sim normal(m_0, v_0)
\end{aligned}
\tag{3-10}
$$

In this model, the precision parameter α follows a prior distribution $h(.)$. Therefore, its posterior density is estimated as part of the analysis. Similarly, the posterior density of the number of latent components occupied by observations in the data is inferred from the data.

3.3.2 Over-dispersed generalized Dirichlet process mixture model

After accounting for heterogeneity in the data through the generalized linear Dirichlet process mixture model, some extra variability may still exist in some datasets. To

account for extra variability, it is possible to use a Dirichlet process mixture over the over-dispersed generalized linear model discussed in Section 2.1.5.1. Doing so, besides accounting for over-dispersion by allowing for a flexible model resulting in a mixture of points (in contrast to the parametric unimodal distribution), the remaining variability is accounted for by the error term. As in the generalized linear Dirichlet process mixture model (discussed in previous section), we adopt the method suggested by Mukhopadhyay and Gelfand (1997) in which the authors use a Dirichlet process mixture over the intercept. As discussed previously, this allows maintaining the convenient form of the conventional over-dispersed generalized linear models for the covariates. Given the above notation, the over-dispersed generalized linear Dirichlet process mixture model can be specified as

$$
\begin{aligned}
& y_i|\boldsymbol{X}_i, \beta_{0r}, \varepsilon_i \sim Poisson\ (\lambda_i) \\
& log(\lambda_i) = \eta_r + \boldsymbol{\beta X}_i + \varepsilon_i \\
& \beta_{0r} \sim Dirichlet(\alpha G_0) \\
& G_0|m_0, v_0 \sim normal(m_0, v_0)
\end{aligned}
\qquad \text{(3-11)}
$$

To circumvent identifiability issues, the mean of the error term ε_i is fixed to be equal to zero; i.e., $\varepsilon_i \sim$ normal(0, v_ε).

3.4 Multilevel (Hierarchical) Settings

In this section, we discuss multilevel models that accommodate the hierarchical structure of crash data. Such hierarchical structure, which occurs often in transportation safety studies (Dupont et al., 2013), requires allowing one or more model parameters to vary across groups of observations (e.g., regions). As discussed in Section 2.1.9, random effects and/or random parameter models are often employed in analyzing multilevel data. Due to the higher computational complexity involved in random parameter models, the majority of those studies involving multilevel analyses have used random effects models (Vanlaar, 2005, Lenguerrand et al., 2006; Kim et al., 2007; Helai et al., 2008; Park et al., 2010; Yannis et al., 2010; Jovanis et al., 2011; Papadimitriou et al., 2014). In

this thesis, among other reasons, we therefore focus on the use of random effects models (in particular, random intercepts models) in multilevel settings. We first discuss the limitations associated with random effects models and provide a flexible latent class model to address such limitations.

To clarify one problem that may arise when adopting standard random effects models, suppose a multilevel scenario in which the analyst is only interested in potential variations in the intercepts (random intercepts model) among groups (e.g., geographical areas). A simplistic approach is to assume that all groups have exactly the same intercept and that there is no extra variability due to grouping in data. Obviously, this assumption does not take into consideration the fact that there might be some unknown and/or unmeasured attributes that vary between groups. Basically, this approach ignores the hierarchical structure of the data.

In the aforementioned scenario, two major approaches have been proposed in literature to accounting for group-specific effects, tackling unobserved heterogeneity through the random intercepts. The first approach is estimating the intercepts for the individual groups separately based on the belief that they differ completely, the assumption of complete independence (Ohlssen et al., 2007). This assumption is not realistic since groups of observations (e.g., intersections or municipalities) are not totally dissimilar and they certainly share some similar features. A more appropriate approach, which is also the most commonly applied, is to assume that the intercepts vary between groups but are generated from a single population. Thus, the intercepts are assumed to share a common distribution being usually a unimodal normal density. Depending on the extent to which standard distributional assumptions are capable of capturing heterogeneity in a dataset, say, in the form of random intercepts, the results would be biased by various degrees. It should be noted that standard parametric assumptions in traditional random effects models—such as normally distributed random effects—usually do not accommodate skewness, kurtosis, and multimodality (Xiong and Mannering, 2013).

One limitation of the model described above is that different groups in data may be generated from widely differing subpopulations instead of a single population. As an example of a problem that may arise following standard parametric assumptions in random effects models, let's consider a multilevel dataset in which observations are

nested in different groups such as geographical regions. When there are outlier regions (extreme cases) in data, large outlier regions affect other regions excessively. Consequently, estimates relating to smaller outlier regions erroneously tend to approach the overall mean. In these circumstances, a more flexible modeling approach is necessary (Ohlssen et al., 2007). The flexible model must satisfy two requirements: (i) it should be able to avoid the complete independence assumption of groups described above; and (ii) it should be able to relax the distributional assumption while adapting itself to the complexity of the observed data.

Following our discussion above related to the presence of subpopulations, interest may lie in clustering groups (e.g., regions), the higher level of the hierarchy. For example, such clustering allows identifying regions that perform similarly when analyzing sites nested within regions. The model presented in Section 3.4.1 helps achieve this goal.

3.4.1 Flexible Dirichlet process mixture multilevel model

Based on the observations above, we extend the random intercepts multilevel model presented in Section 2.1.8.1 as follows. The flexible Dirichlet process mixture multilevel model relaxes the distributional assumption (normal density on η_r) of the standard random intercepts multilevel model by estimating a flexible mixture of points model. Doing so, it also provides further insights by identifying latent clusters and outliers. It should be underlined that the flexible Dirichlet process mixture multilevel model allows accommodating outliers in the analysis using a more flexible distribution. Therefore, there is no need to diagnose and remove outliers from a dataset in advance of the analysis (Ohlssen e al., 2007). The flexible Dirichlet process mixture multilevel model can be defined as

$$
\begin{aligned}
&y_{ri}|\boldsymbol{X}_{ri}, \varepsilon_{ri}, \eta_r \sim Poisson\,(\lambda_{ri}) \\
&log(\lambda_{ri}) = \eta_r + \boldsymbol{\beta X}_{ri} + \varepsilon_{ri} \\
&\varepsilon_{ri}|v_\varepsilon \sim normal(0, v_\varepsilon) \\
&\eta_r = \eta_{DP} \sim Dirichlet(\alpha\eta_0) \;\&\; r = 1,2,\dots,C
\end{aligned}
\qquad (3\text{-}12)
$$

$$\eta_0 | m_0, v_0 \sim normal(m_0, v_0)$$

where η_0 (with unknown parameters, the mean m_0 and the variance v_0) is the baseline distribution for η_r and α is the precision parameter as explained earlier in Section 3.2. Recall that r denotes latent clusters (at grouping or subject level) and C stands for the maximum possible (allowed) number of latent clusters or mass points (see Section 3.2.). In the previous model presented in Section 2.1.9.1, the random intercepts η_r were normally distributed, whereas under the flexible Dirichlet process mixture multilevel model the intercepts are modeled non-parametrically using a Dirichlet process mixture. Doing so, we remove the restriction of the standard distributional assumption and allow the observed dataset to decide its proper form of the random intercepts. One should also take into account that the parameters of the baseline distribution, η_0, are estimated here as part of the modeling process allowing us to account for uncertainties associated with the baseline distribution for the random intercepts.

It can be seen in the flexible Dirichlet process mixture multilevel model that the vector of coefficients $\boldsymbol{\beta}$ associated with the known covariates vector $\boldsymbol{X}$ (site characteristics) does not follow a Dirichlet process. Therefore, the convenient linear relationship between covariates and the response (crash frequency) is maintained, retaining their usual interpretations. Other extensions are obviously possible; for example, one might allow the effect of one or more covariates to vary across different regions. Note also that a Dirichlet process mixture over the intercepts (as in our suggested extension) allows us to deal with heterogeneity in data with respect to the location of the mean (Mukhopadhyay and Gelfand, 1997); that is, average crash frequency of each grouping. In the study in context, thus, such model enables the identification of latent clusters among different regions, after adjusting for the effect of covariates.

3.4.2 Schematic representation of the model

To highlight the differences between the standard random intercepts multilevel Poisson-lognormal model and the flexible Dirichlet process mixture multilevel model, a directed acyclic graph of both models is shown in Fig. 3-2. In Fig. 3-2, solid arrows indicate stochastic relationships, dashed arrows indicate deterministic relationships, circles stand for model parameters, small rectangles are deterministic data points (X and

y), and large rectangles stand for loops. It can be clearly seen in Fig. 3-2 that the flexible Dirichlet process mixture multilevel model involves a higher number of parameters that allow the data to decide the form of the random intercepts η_r. This in turn adds more flexibility to the model. Recall that the flexible Dirichlet process mixture multilevel model is developed based on Bayesian nonparametric literature in which the number of parameters is not fixed and is inferred from the model based on the data. Obviously, this may come with a higher cost: more parameters to deal with, longer time of execution, and higher computational complexities.

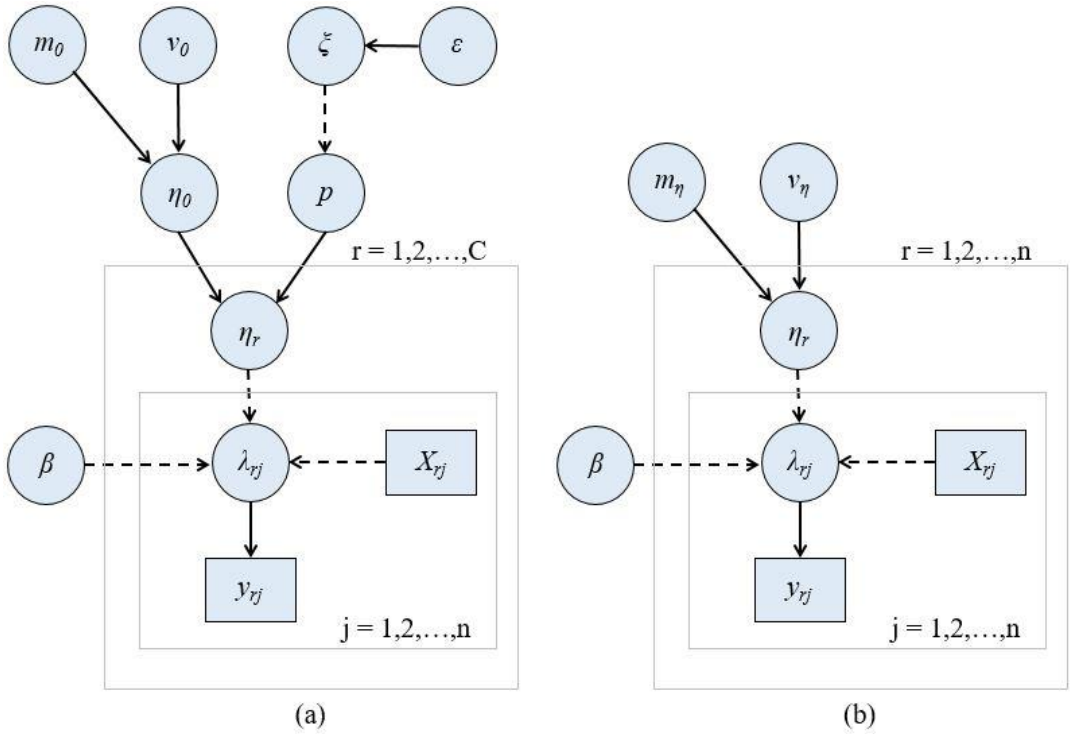


Figure 3-2 Directed acyclic graph of (a) flexible Dirichlet process mixture multilevel model and (b) random intercepts multilevel model

3.5 Multivariate Settings

In multivariate settings, two or more correlated outcomes (e.g., differing injury-severity levels) are modelled simultaneously. As described in Chapter 2, multivariate models capture the existing correlation through their dependence structure. Recall that such

dependence between outcomes is mainly due to the presence of unobserved or unmeasured factors that affect correlated outcomes simultaneously. Obviously, neglecting the correlation among correlated outcomes may result in misleading statistical inferences.

While providing valuable insights that help our understanding of crash mechanisms in the presence of correlated outcomes, most previous multivariate traffic safety studies (discussed in Section 2.1.9) have not considered whether their assumptions relating to the dependence structure reflects the true structure of the data. In effect, in terms of the methodological framework, despite the general consensus that restrictive assumptions (e.g., homogeneity) in dependence structure may have an adverse effect on the accuracy of estimates, studies addressing the sensitivity of the results to these assumptions in multivariate settings are surprisingly rare, especially, in transportation safety studies. Examples of flexible multivariate models can more easily be found in other fields such as econometrics and biostatistics (Müller et al., 1996; Cameron et al., 2004; Jara et al., 2007). Note that the degree of correlation between crash types may vary significantly from one site (intersection, neighborhood, etc.) to another, for example, due to variation in intersection geometric and operational characteristics or some other unknown contributing factors. Therefore, representing correlation through a homogeneous density such as the multivariate normal distribution may lead to misleading statistical inferences. In the safety literature, Nashad et al. (2016) discussed the above issue and suggested a copula based approach to formulate a flexible multivariate model that accounts for heterogeneity in the dependence structure.

Building on previous research, this section introduces two flexible Bayesian latent class multivariate models to jointly analyze correlated crash outcomes. Both models are in the form of "infinite" mixture multivariate models that can be developed with Dirichlet process mixing (Müller et al., 1996; Müller et al., 2007). Here, the term "infinite" is used since the number of components can theoretically go to infinity under the Bayesian nonparametric approach. In this regard, most studies have used a Dirichlet process mixing over the conventional model parameters (random intercepts or covariate coefficients) in univariate or multilevel settings (Mukhopadhyay and Gelfand, 1997; Kleinman and Ibrahim, 1998; Ohlssen et al., 2007; Dhavala et al., 2010; Heydari et al., 2016a; Heydari et al., 2016b).

In this thesis, we also use a Dirichlet process mixture for building a flexible correlation structure in multivariate settings that accommodates multimodality; and consequently, allows for heterogeneity in the dependence structure. Our model not only conveniently infers the number of latent subpopulations as part of its estimation algorithm, but it also allows this number to be large (Ohlssen et al., 2007; Gershman and Blei, 2012). Recall that classical latent class models usually employ a limited number of prespecified components (Mannering and Bhat, 2014).

In this section, we discuss our proposed extensions relaxing the homogeneity in the dependence structure with respect to the location of the correlation structure, and then relaxing the homogeneity with respect to both the location and the covariance matrix. The proposed extensions allow for departures from restrictive parametric assumptions in multivariate modeling of crash datasets, adding flexibility to the multivariate framework.

3.5.1 Multivariate mixture of points model

This model relaxes the homogeneity assumption of the dependence structure with respect to its location. To extend the standard multivariate model to the Dirichlet process mixture multivariate model, the error term, ε_{ik}, can be included in the intercepts to allow variation across observations with respect to the intercepts. Given the notation presented in Section 2.1.10, let m_k denote the mean of the correlated random intercepts. We can thus write

$$log(\lambda_{ik}) = \beta_{0ik} + \boldsymbol{\beta}_k \boldsymbol{X}_{ik}$$
$$\beta_{0ik} \sim MVN(m_k, \Sigma) \quad (3\text{-}13)$$

The Bayesian nonparametric allows relaxing of the parametric assumption for the jointly distributed error terms (here correlated random intercepts). The multivariate mixture of points model uses a parametric density that is usually a multivariate normal distribution as its baseline density G_0 and then allows departures from this parametric assumption. Note that the same analogy can be used in simultaneous equation modeling to relax restrictive parametric assumptions. The model can thus be defined as

$$y_{ik}|\boldsymbol{X}_{ik}, \beta_{0L_i k} \sim Poisson\ (\lambda_{ik})$$

$$log(\lambda_{ik}) = \beta_{0L_i k} + \boldsymbol{\beta}_k \boldsymbol{X}_{ik} \tag{3-14}$$

$$\beta_{0L_i k} \sim DP(\kappa G_0)$$

$$G_0 \sim MVN(m_{0k}, \Sigma)$$

where m_{0k} is the mean of the outcome k for the multivariate normal baseline G_0, and L_i denotes an allocation variable indicating latent clusters. The correlated parameters (random intercepts) are modeled as a mixture of points. While one can allocate a Dirichlet prior on the error term without involving the intercept, this results in further complexity as the mean of the Dirichlet cannot be equal to 0.

3.5.2 Mixtures of normal densities and multivariate normal densities

As discussed in Section 3.2, suppose $f(.)$ is the density relating to G, the unknown distribution function of interest; θ are random draws from the baseline G_0; and p denotes the probability for infinite mixtures of mass points, being a discrete distribution function. We can then write

$$f(.) = \sum_{j=1}^{\infty} p_j I_{\theta_j}\,; \quad \theta_j \sim G_0 \tag{3-15}$$

To build a mixture of normal (continuous) densities instead of a mixture of points represented in Eq. 3-15, we substitute the indicator function I_θ with a continuous density $h(.|\theta_j)$ while truncating at C.

$$f(.) = \sum_{j=1}^{C} p_j h(.|\theta_j) \approx \sum_{j=1}^{\infty} p_j h(.|\theta_j), \quad \theta_j \sim G_0 \tag{3-16}$$

In particular, as in Ohlssen et al. (2007), we need to present each component of the mixture with a mean and a variance when considering a mixture of normal densities:

$$\delta_i \sim Normal(\theta_j, v_j) \quad (3\text{-}17)$$

where θ and v are, respectively, the mean and the variance of the j^{th} component, and δ is a generic model parameter of interest for observation i. An extension to a mixture of multivariate normal (MVN) densities for k correlated outcomes can then readily follow for a set of dependent parameters δ_{ik} as

$$\delta_{ik} \sim MVN(\theta_j, \sum_j) \quad (3\text{-}18)$$

where θ and Σ are, respectively, the mean and the covariance matrix associated with the j^{th} component of the mixture of multivariate normals. In the next section, details on the extension to a mixture of multivariate normal densities are provided in the study in context.

3.5.3 Mixture of multivariate normal densities

To add further flexibility, one can let the association structure vary with respect to both the location and the covariance matrix. Similar to the model presented in Section 3.4.1, such an approach induces a latent class construct into the model to account for unobserved heterogeneity in addition to addressing dependence across correlated outcomes. To this end, one can include intercepts in the dependence structure to represent its mean hypothesizing that not only error terms but also intercepts can affect both outcomes simultaneously. In fact, there may be unknown or unmeasured covariates that are common to sites (e.g., intersections) and that can be incorporated into the model by pooling their effects into the intercept terms. Then, whatever common feature that is left over (i.e., not modeled) can go into the covariance matrix.

Therefore, we first use a joint distribution $p(\beta_{0i1}, \beta_{0i2}, \ldots, \beta_{0ik})$, for k possibly correlated outcomes for each site i, over the dependent random intercepts $\beta_0 = (\beta_{0i1}, \beta_{0i2}, \ldots, \beta_{0ik})$ to account for correlation among different crash types or injury-severity levels. Note that β_0 varies across sites for each crash type k under the flexible multivariate model; however, it is fixed across sites under the standard multivariate model presented previously. Such an approach conveniently allows us to introduce a mixture of multivariate normal densities, based on Dirichlet process mixing, into the dependence

element of the correlated outcomes. The main idea here is that n data points are generated form J latent subpopulations. Thus, instead of a single restrictive multivariate normal distribution, the model can accommodate J mixtures of multivariate normal densities with non-zero mean values, MVN (β_{0jk}, $\sum_j$). In addition to the above notation, suppose μ_{0k} and σ_{0k} are the baseline mean and its respective standard deviation. The proposed flexible Dirichlet process mixture of multivariate normals model can then be written as follows. For computational convenience, we used an allocation variable L_i instead of j.

$$
\begin{aligned}
&y_{ik}|\boldsymbol{X}_{ik}, \beta_{0ik} \sim Poisson\ (\lambda_{ik}) \\
&log(\lambda_{ik}) = \beta_{0ik} + \boldsymbol{\beta}_k X_{ik} \\
&\beta_{0ik} \sim MVN(\mu_{L_i k}, \Sigma_{L_i}) \\
&\mu_{L_i k} \sim Normal(\mu_{0k}, \sigma_{0k}) \\
&\Sigma_{L_i} = \begin{bmatrix} \sigma_{L_i 11} & \cdots & \sigma_{L_i 1k} \\ \vdots & \ddots & \vdots \\ \sigma_{L_i k1} & \cdots & \sigma_{L_i kk} \end{bmatrix} \\
&\Sigma_{L_i}^{-1} \sim Wishart(R_{L_i}, K)
\end{aligned}
\qquad (3\text{-}19)
$$

It can be seen in the latter formulation that the correlation structure varies with respect to both the location and the covariance matrix. In other words, each identified latent subpopulation has its own correlation resulting in a correlation mixture from which a correlation summary can be created. This adds considerable flexibility to the model, so that it could handle skewness and multimodality in the correlation structure. Note that, in our model, the correlation is in the form of a mixture and the location of the correlation mixtures (components) is not fixed. Therefore, these mixtures can move between the range -1 and +1 accommodating skewed correlation densities to some extent. This is shown in Section 4.3.2.2.

3.6 Cluster Detection Algorithm

In the previous sections of Chapter 3, we explained our method in univariate, multilevel, and multivariate settings, and showed how to relax restrictive assumptions. Another important advantage of our proposed Dirichlet process models is the possibility to identify latent clusters. In fact, it is possible to compute the probability of similarities between pairs of observations. This is implemented through indicator variables I_{ab} being an NxN matrix, where N indicates the number of data points based on which clustering is to be done. Let L_i and L_j be the components of interest (allocation variables). Then we can write

$$I_{ij} = \begin{cases} 1 & if \quad L_i = L_j \\ 0 & if \quad L_i \neq L_j \end{cases} \tag{3-20}$$

This is obtained at each iteration of the MCMC algorithm. Then, averaging over the total number of iterations gives the probability of i and j being in the same cluster. The total number of observations sharing the same cluster is $\sum_{i \neq j} I_{ij}$. The latter statistic can be used to detect outliers. The above clustering and outlier detection approach are further discussed in terms of their applications in Section 4.2.5.

3.7 Model Selection and Performance Measures

As discussed in Section 3.2, Dirichlet process mixture models can be used to check how closely a parametric assumption might hold. If a parametric assumption (for example, a normality assumption for random intercepts) seems far from true, there is justification to avoid using that parametric model. At this point, there is no further need for other model selection methods for that model, which has been ruled out. Nevertheless, we do discuss some model fitting criteria here as an extra piece of information; for example, to show that how the predictive capability of a model can be affected by an assumption that does not hold.

The deviance information criterion (DIC) is usually used for model selection in Bayesian crash data analysis. However, the DIC is sensitive to different parameterizations (Geedipally et al., 2014) and of questionable use in case of multimodal posteriors

(Washington et al., 2011). A discussion about some of the limitations associated with the DIC can be found in Carlin and Louis (2008). We used cross-validation predictive densities (Gelfand, 1996; Mukhopadhyay and Gelfand, 1997; Vehtari and Lampinen, 2002; Ntzoufras, 2009) to compute conditional predictive ordinates (CPOs) that in turn allow estimating the log pseudo marginal likelihood (LPML) and the pseudo Bayes factor (PBF). The cross-validation method compares alternative models in terms of their predictive abilities. The main idea behind cross-validation methods constitutes the base for the estimation of the CPOs. In cross-validation, a given dataset is divided into two groups. One is used to make the posterior inference, whereas the second group is used to validate the previously estimated model. The problem here is the sensitivity of the results to how these groups are selected. The CPO circumvents this problem by leaving out only one observation each time. A relatively detailed discussion in this regard is provided in Ntzoufras (2009) and Carlin and Louis (2008). Here, we briefly discuss the main components of this method.

Suppose Y_i is the i^{th} observation, T stands for the total number of iterations in an MCMC simulation, ψ represents the estimated model parameters, and $f(.)$ is the density function evaluated at Y_i. For each observation, the CPO can be estimated as

$$CPO_i = \left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{f(Y_i|\psi^{(t)})}\right)^{-1} \tag{3-21}$$

The product of CPOs across all observations gives the pseudo marginal likelihood (PML), from which the PBF of comparing model 1 against model 2 can be obtained

$$PBF = PML_{model\ 1}/PML_{model\ 2} \tag{3-22}$$

Alternatively, the LPML (Gelfand et al., 1992), given in Eq. (3-23), is easier to calculate.

$$LPML = log\{\prod_{i=1}^{l} CPO_i\} = \sum_{i=1}^{l} log(CPO_i) \tag{3-23}$$

Finally, log pseudo Bayes factors (LPBF) (Basu and Chib, 2003; Ntzoufras, 2009), especially useful in the presence of mixture models, can be obtained as follows:

$$LPBF = LPML_{MODEL\,1} - LPML_{MODEL\,2} \tag{3-24}$$

Table 3-1 reports the interpretation of log Bayes factors according to Kass and Raftery (1995) and Ntzoufras (2009). (The interpretation is also valid for log pseudo Bayes factors.)

Table 3-1 Bayesian model selection via Bayes factor

Bayes factor	**Log Bayes factor**	**Degree of support for the model of interest**
1-3	0-1	No evidence of support
3-20	1-3	Support
20-150	3-5	Strong support
>150	>5	Very strong support

CHAPTER 4

ANALYSIS & RESULTS

This chapter discusses the applications of our proposed methods in univariate, multilevel, and multivariate settings for transportation safety studies. We adopted two simulated and six real datasets to demonstrate the performance of the proposed models and their merits considering various crash data types, scenarios, and settings. For each setting, we describe the adopted datasets, followed by discussing prior specification, model computation, and the results. We then provide a policy example that assists to draw transportation engineering insights from the proposed methods. Lastly, we summarize our findings for each setting separately. Note that the contents of this chapter (except Section 4.1.6) have been published in Analytical Methods in Accident Research (see page iv for details).

4.1 Univariate Modeling

As discussed in Section 3.2, there is only one outcome of interest in univariate modeling and no hierarchical structure in data exists or is considered. To clarify our method, we first show how it works devising a simulation exercise in Section 4.1.1, then two real datasets with distinct characteristics are used to apply the method. The first dataset is characterized by relatively high mean values, being highly overdispersed. The second dataset is characterized by low mean values and excess zero counts. The basics of the

Dirichlet process mixing approach used in the univariate settings are then extended in the multilevel and multivariate settings that have more complex data and model structures.

4.1.1 Simulated data

One important advantage of simulated data is that the true parameters and the underlying structure of the data are known, so that one can evaluate the accuracy of posterior inferences from any model. In this section, two simulated data are used: (1) a dataset with bimodal intercepts concentrated at two distinct values; and (2) a dataset with intercepts concentrated at a single value, creating a unimodal density. For the first data simulation scenario, we generated two crash datasets, both with 100 observations, and varying only in their intercepts. The total number of generated observations is 200, which is sufficient here since the simulated data are only intended as an example to illustrate how Dirichlet process mixture models work. If we were instead aiming to provide detailed properties of a new model via a simulation study, then a larger sample might have been indicated. Data were generated from a Poisson distribution with expected crash frequency, a function of a single hypothetical covariate, say, traffic exposure. Given the notation in Section 3.3, in particular, we generated the data as follows:

$$y_i|X_i \sim Poisson\ (\lambda_i)$$
$$log(\lambda_i) = \beta_0 + \beta X_i \quad (4\text{-}1)$$

To create the above scenario, we randomly selected 100 observations from a railway grade crossing dataset, described in Section 4.1.3, where traffic exposure, *X*, was known. Since we assumed that covariates and their effect are identical, we used the same set of observations selected above to build the second subset containing 100 observations. We set the value of β (in Eq. 4-1) to be 0.492. To generate crashes based on the model structure defined above, we set the intercept to be equal to -4 for the first subset and 3 for the second subset. Both subsets were then combined to create a single dataset with 200 observations. Doing so, both subsets were identical except in their two distinct intercepts. We then analyzed the simulated data using the proposed Dirichlet process

mixture model, the finite-mixture Poisson-gamma model with two and three latent components, the standard Poisson-gamma (negative binomial) model, and the random intercepts (random effects) Poisson model. Readers are referred to Chapter 2 for details relating to the above models.

The Dirichlet process mixture model correctly identifies the two clusters in the simulated data (see Fig. 4-1a). It also accurately estimates other model parameters. The results are reported in Table 4-1. The conventional finite mixture model with two components also performs well. The intercept and beta coefficient are estimated accurately, and over-dispersion parameters for each component are estimated to be very large indicating that the distribution of crash frequency in each subset is Poisson. Note that a large value of over-dispersion is expected here since we generated each subset from a simple Poisson distribution. The finite mixture model with three components (wrong number of components) works less well, with biased estimates, and similarly biased results are obtained from the Poisson-gamma and the random intercepts models. The Poisson-gamma model assumes that the intercept is fixed, while the random intercepts model allows the intercept to vary, but following a normal density. With neither assumption holding, it is not surprising that these models do not work well. Conversely, the Dirichlet process mixture model works well when these assumptions do not hold.

For the second data simulation scenario, we randomly selected 200 observations from a grade crossing dataset. Similar to the first scenario, this simulated dataset was generated using Eq. 4-1 with only one covariate; i.e., traffic exposure. Model parameters β and β_0 were set to be 0.492 and -4, respectively. The data were generated from a simple Poisson distribution with fixed parameters, so that there is no multimodality in any component of the data. We first analyzed this simulated dataset using the standard negative binomial model. This model estimated β and β_0 to be 0.488 and -4.06, respectively. The over-dispersion parameter was estimated to be 65.73 indicating that the distribution of the data is close to the simple Poisson.

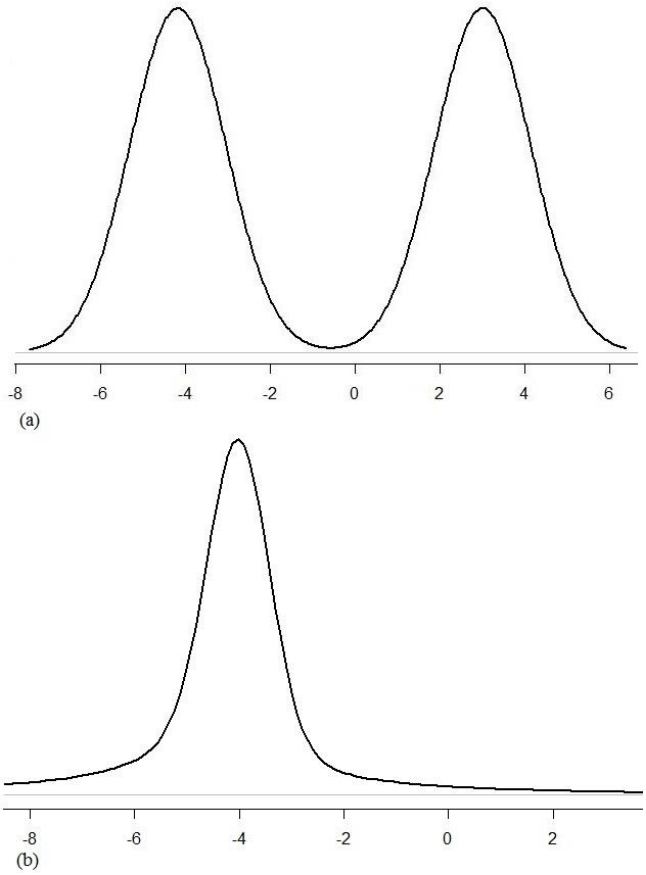


Figure 4-1 Kernel density plot of the posterior density for intercepts: (a) scenario 1; and (b) scenario 2.

We then analyzed the later simulated dataset using the Dirichlet process model; the results were found to be very similar to those obtained from the negative binomial model. In particular, the Dirichlet process model estimated β and $\beta 0$ to be 0.486 and -4.066, respectively. A kernel density plot of the posterior density for the intercept, obtained from the Dirichlet process model, is illustrated in Fig. 4-1b showing a unimodal density concentrated at -4, as expected. Similar to the first scenario, the Dirichlet process model accurately estimated the model parameters and the structure of the data (here, the form of the intercepts). In both scenarios, the Dirichlet process model performed well. This is a valuable property of the Dirichlet process mixture models that can adjust themselves to the complexity of any data (Gershman and Blei, 2012).

Table 4-1 Posterior inference for the simulated data

	Posterior Mean	Std. Dev.	Credible intervals 2.50%	97.50%
Dirichlet process mixture Poisson model				
Intercept mean	-0.601	0.270	-1.143	-0.085
Intercept variance	13.160	1.630	12.020	16.820
Covariate coefficient	0.491	0.001	0.489	0.494
Baseline mean	-1.423	3.930	-9.457	6.559
Baseline Std. Dev.	5.999	2.107	2.428	9.742
Precision parameter α	0.706	0.382	0.312	1.698
Finite-mixture Poisson-gamma model with 2 components				
Intercept (component 1)	-4.148	0.452	-5.049	-3.283
Intercept (component 2)	2.998	0.018	2.963	3.035
Covariate coefficient (component 1)	0.492	0.045	0.404	0.580
Covariate coefficient (component 2)	0.492	0.002	0.488	0.496
Over-dispersion (component 1)	47.330	55.510	4.844	207.900
Over-dispersion (component 2)	794.600	207.900	460.100	1,269.000
Finite-mixture Poisson-gamma model with 3 components				
Intercept (component 1)	-2.557	0.615	-4.202	-1.868
Intercept (component 2)	2.998	0.019	2.962	3.035
Intercept (component 3)	1.953	1.472	-0.850	4.999
Covariate coefficient (component 1)	0.368	0.063	0.279	0.495
Covariate coefficient (component 2)	0.492	0.002	0.488	0.497
Covariate coefficient (component 3)	0.527	0.744	-0.126	2.605
Over-dispersion (component 1)	18.160	41.040	0.046	134.700
Over-dispersion (component 2)	798.600	210.000	459.100	1,275.000
Over-dispersion (component 3)	0.022	0.009	0.012	0.041
Standard Poisson-gamma model				
Intercept	2.758	0.518	1.841	3.842
Covariate coefficient	0.442	0.057	0.328	0.545
Over-dispersion	0.119	0.011	0.098	0.142
Random intercepts Poisson model				
Intercept mean	-4.413	0.469	-5.293	-3.442
Intercept variance	22.030	2.891	17.040	28.340
Covariate coefficient	1.010	0.043	0.902	1.071

4.1.2 Vehicle injury data

This dataset contains vehicle-injury counts for 647 signalized intersections in Montreal from 2003 to 2008. The dataset is highly over-dispersed and characterized by a relatively high mean value. The vehicle-injury data were provided by ambulance services. Other information such as geometric characteristics (number of lanes, presence of median, etc.), built environment characteristics (population, land use, presence of bus and subway stations, etc.), and traffic control characteristics (signal type, etc.) were obtained from various sources. Summary statistics of this dataset are reported in Table 4-2.

The vehicle-injury dataset has an average mean value of 4.6 injuries in a six-year period. Among 647 signalized intersections, 143 (22.10%) were three-leg intersections, 458 (70.79%) were in the proximity of bus stops, and 364 (56.26%) were in a distance of less than 400 meters from a school. The number of intersections with at least one raised median was 290 (44.82%). For further discussion relating to this dataset, see Strauss et al. (2014).

Table 4-2 Summary statistics for the vehicle-injury data

Variable	Mean	Std. Dev.	Min	Max
Through AADT	19,467.96	11,084.39	1,790.00	76,525.00
Left-turning AADT	2,602.72	2,641.86	0	23,843.00
Right-turning AADT	2,668.01	2,697.45	0	23,792.00
Ratio of pedestrians & bikes over total AADT	0.226	0.467	0.003	7.574
Total number of lanes for all approaches	6.90	2.60	2.00	16.00
Number of subway stations in 400 m	0.44	0.70	0.00	4.00
Three-leg (1 if three-leg intersection; 0 otherwise)	0.22	0.42	0.00	1.00
Bus stop (1 if present in 50 m; 0 otherwise)	0.71	0.46	0.00	1.00
Raised median (1 if present; 0 otherwise)	0.47	0.50	0	1.00
School (1 if present in 400 m; 0 otherwise)	0.56	0.50	0.00	1.00
Vehicle-injury frequency	4.60	6.37	0.00	58.00

4.1.3 Railway grade crossing data

This dataset is characterized by a very low mean value and excess zero counts. The dataset records crash frequencies at 6,617 automated railway grade crossings in Canada. Automated crossings are equipped with flashing lights, bells and/or gates to inform road users about approaching trains. The data were provided by Transportation Safety Board of Canada covering a six-year period from 2008 to 2013. A host of independent variables (including geometric and operational attributes) were available in the database, the most important shown in Table 4-3, where summary statistics of the crossing data are reported.

Table 4-3 Summary statistics for the grade crossing data

Variable	Mean	Std. Dev.	Min	Max
Train flow (number of trains daily)	11.071	12.976	0.100	162.000
Vehicle flow (AADT)	3,082.396	5,636.744	1.000	71,500.000
Exposure (product of train flow and vehicle flow)	29,695.710	94,428.280	0.270	3,000,000.000
Train ratio (ratio of train flow to vehicle flow)	0.170	1.854	0.000	54.000
Number of rail tracks	1.292	0.612	1.000	7.000
Number of lanes	2.164	0.671	1.000	7.000
Road speed (speed limit in km/h)	62.333	17.879	5.000	110.000
Train speed (maximum train speed in km/h)	63.910	36.446	1.608	160.800
ln(road speed)*ln(ratio of train flow to vehicle flow)	-20.323	9.471	-56.350	17.314
Track angle (deviation from 90 degrees)	19.496	19.709	0.000	87.000
Gate (1 if gate is present; 0 otherwise)	0.364	0.481	0.000	1.000
Whistle prohibition (1 if prohibited; 0 otherwise)	0.130	0.336	0.000	1.000
Urban (1 if located in urban area; 0 otherwise)	0.354	0.478	0.000	1.000
Ont./Qc. (1 if located in Ontario or Quebec)	0.578	0.494	0.000	1.000
Pac./Atl. (1 if located in Pacific or Atlantic region)	0.154	0.361	0.000	1.000
Crash frequency	0.080	0.317	0.000	4.000

We also created three dummy variables to reflect spatial effects to some extent based on similarities observed in an exploratory data analysis phase. The prairie region, consisting of the provinces of Manitoba, Saskatchewan, and Alberta, was selected as the reference group. Ontario and Quebec formed another group. Finally, the Pacific region (British Columbia) and the Atlantic region (New Brunswick, Newfoundland and Labrador, Nova Scotia, and Prince Edward Island) formed the Pacific/Atlantic region. The crossing dataset had a very low mean crash frequency with almost 90% of crossings experiencing no crash over the aforementioned period. Interested readers may refer to Heydari and Fu (2015) for details relating to the Canadian grade crossing data.

4.1.4 Prior specification and model computation

Bayesian analysis requires the elicitation of priors for parameters of interest. Non-informative priors were set for regression coefficients and the mean of the baseline distribution. The model parameters are described in Section 3.3. In particular, we used normally distributed priors with mean zero and a large variance. We used gamma(0.01,0.01) priors for the inverse variance. For a detailed discussion related to prior specification in road safety studies in univariate settings, see Heydari et al. (2014b). With respect to the baseline distribution, note that we did not fixed the baseline parameters (mean and variance). Instead, we used vague hyper priors on these parameters and let the model estimate them. It is important to consider that if a baseline does not support the range of a dataset, the model would not be able to make proper posterior inference. Using vague hyper priors for the baseline helps prevent such condition.

One also needs to select a prior for the Dirichlet precision parameter α, an important choice, since its posterior density is critical in deciding how closely a parametric distributional assumption holds. A gamma or uniform prior is usually selected for this prior. For example, Ohlssen et el. (2007) chose a uniform prior with lower and upper limits of 0.3 and 10, respectively, while Ishwaran and James (2002) suggested a gamma(2, 2) prior that supports both small and large values of α. For a detailed discussion in this regard, see Ishwaran (2000), Ohlssen et al. (2007), Dorazio (2009), and Murugiah and Sweeting (2012). The prior for α can be related to the maximum number of possible clusters C, discussed in Section 3.2. For the vehicle injury dataset, we

considered C to be equal to 52 (based on $5\alpha + 2$, where the upper limit of α is set to 10), a relatively large number, so that we were able to approximate an infinite mixture of points.

For the grade crossing dataset, we first used a uniform prior with lower and upper limits of 0.3 and 10, respectively, for the Dirichlet precision parameter. However, the dataset, being limited as it contains very large proportion of zero crashes, couldn't provide much information about the Dirichlet precision parameter. The estimated interval around this parameter varied from 0.7 to 8.4 that is quite a large interval, which is similar in range to the specified prior. We then used a gamma prior with the shape and scale parameters set to one, a somewhat more informative prior. While the selected gamma prior is inclined to result in small values of the precision parameter, it has a relatively heavy tail that also allows larger values although the probability for such values is small.

In addition, note that almost 90% of the crossings in this dataset did not experience any crash over a 6-year period, so lower values of the precision parameter were more plausible, justifying that prior choice. In the grade crossings dataset, the density of observed crashes bunches up at zero with around 90% of observations concentrated at this peak. It is doubtful that the random intercepts follow a normal distribution, but we rather expect these to concentrate near zero, and with a limited number of latent components. The gamma(1, 1) supported these expectations and resulted in a better and quicker mixing of chains in the MCMC algorithm.

WinBUGS (Lunn et al., 2000) was used to generate MCMC samples for Bayesian posterior inference. For the Vehicle-injury dataset, two chains with 80,000 iterations were considered among which the first 20,000 were discarded for burn-in and model convergence, so 120,000 samples were utilized for inference. This was sufficient for low Monte Carlo errors. History plots, trace plots, and the Gelman-Rubin statistic (Gelman and Rubin, 1992; Brooks and Gelman, 1998) were used to ensure that convergence was reached. See Spiegelhalter et al., (2002) for Bayesian measures of complexity and fit. For the crossings dataset, we used two chains with 100,000 iterations each, using WinBUGS. The first 20,000 iterations were considered as burn-in and convergence requirements. Posterior estimates were thus obtained using 80,000 iterations or 160,000 samples.

4.1.5 Results and discussions

We first analyzed the vehicle-injury data using the standard Poisson-gamma model. We then allowed the intercept to vary across observations (random intercepts or random effects model) following a normal distribution. We also analyzed the data using a finite-mixture Poisson-gamma model with two and three components, but since the 3-component model did not improve the fit, only the results for the 2-component model are reported.

We compared these standard models to the Dirichlet process mixture on intercept. The results from the Dirichlet process mixture model showed that the Dirichlet precision parameter α is concentrated at some point close to the lower limit of 0 (Fig. 4-2). To verify the sensitivity to the initial prior choice for α (a uniform distribution with lower and upper bounds of 0.3, 10, respectively), we analyzed the data using a different prior, $\alpha \sim$ uniform(0.3, 20), and obtained similar results. Recall that a low value of α indicates that G is far from G_0, as discussed in Section 3.2. Therefore, the normal assumption for intercept is unlikely to hold. That is the 647 random intercepts are not normally distributed, with evidence of multimodality that can be captured in the form of latent clusters. In fact, the Dirichlet process model estimates the posterior median of the number of clusters to be 8 (3, 25). A histogram of the posterior number of clusters is shown in Fig. 4-3. It should be noted that the number of clusters is a stochastic parameter under the Bayesian nonparametric approach. As discussed in Section 3.1, the number of clusters is estimated as part of the analysis algorithm; and therefore, a posterior density can be obtained for this parameter.

We also verified the sensitivity to hyper prior choice for the baseline density although this was not a priori of a major concern as the initial hyper priors were selected to be vague. For example, when we changed the variance of the specified hyper prior from 100 to 400, only a minor difference was observed, with the point estimate of α remaining stable changing from 1.771 to 1.784, without any particular change in the form of the posterior. The log pseudo marginal likelihoods suggest that the random intercept model does not provide a better fit compared to the Poisson-gamma model. This is similar to the case discussed in Ohlssen et al. (2007).

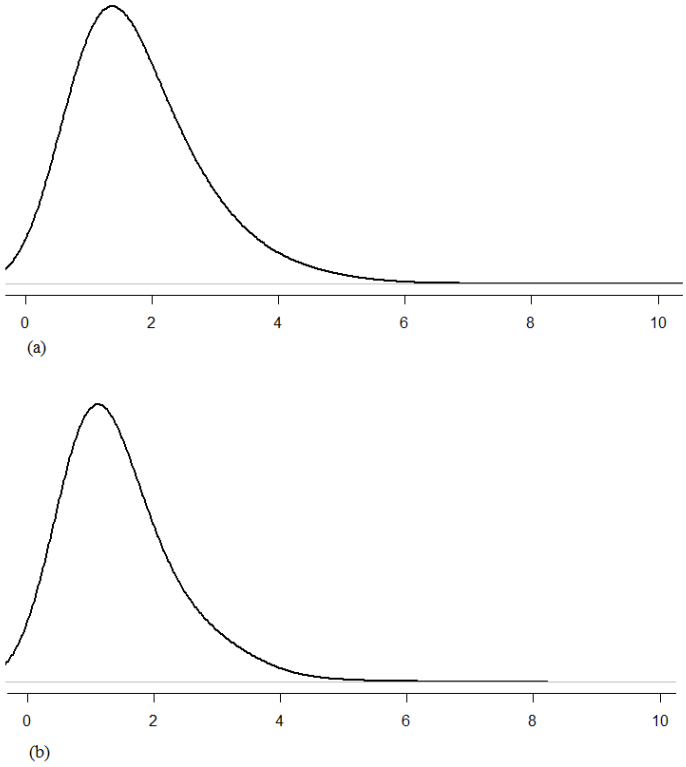


Figure 4-2 Kernel density plots of Dirichlet precision parameter for the vehicle-injury dataset: (a) α ~ uniform(0.3, 10); and (b) α ~ uniform(0.3, 20)

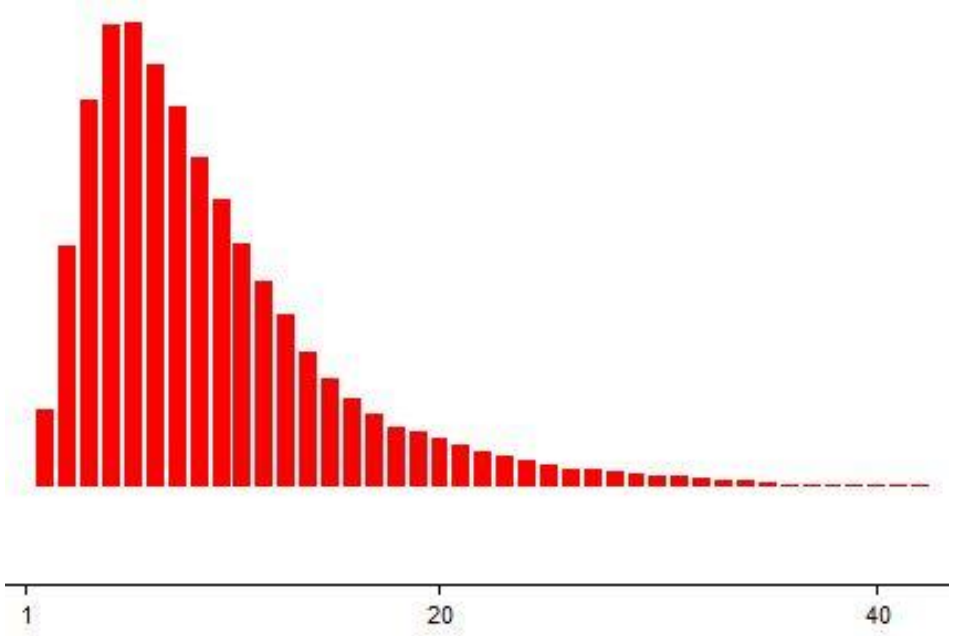


Figure 4-3 Histogram of the posterior number of latent clusters - vehicle-injury data

Table 4-4 Posterior inference for the vehicle-injury dataset

	Posterior Mean	Std. Dev.	Credible intervals 2.50%	97.50%
Over-dispersed Dirichlet process mixture Poisson model				
Intercept mean	-8.746	1.189	-11.290	-6.743
Intercept variance	14.370	15.830	2.796	58.410
ln(through AADT)	0.484	0.099	0.293	0.738
ln(right-turning AADT)	0.240	0.050	0.149	0.354
ln(left-turning AADT)	0.177	0.041	0.091	0.252
ln(ratio of pedestrians & bikes over total AADT)	-0.112	0.040	-0.189	-0.031
Presence of bus stop	0.298	0.131	0.048	0.561
Presence of subway station	0.199	0.118	0.001	0.423
Dirichlet baseline mean	-11.090	3.441	-19.010	-5.224
Dirichlet Baseline Std. Dev.	4.984	2.268	1.673	9.552
Dirichlet precision parameter α	1.771	1.372	0.384	5.628
Variance v_ε (for extra variation)	0.487	0.141	0.125	0.705
Log pseudo marginal likelihood	-1,462.670	-	-	-
Finite-mixture Poisson-gamma model with 2 components				
Component 1				
Intercept	-8.205	0.627	-9.593	-7.075
ln(through AADT)	0.645	0.044	0.556	0.722
ln(right-turning AADT)	0.433	0.095	0.268	0.626
ln(left-turning AADT)	-0.129	0.053	-0.248	-0.060
ln(ratio of pedestrians & bikes over total AADT)	0.338	0.092	0.178	0.529
Presence of bus stop	1.782	0.356	1.186	2.517
Presence of subway station	-0.153	0.229	-0.613	0.299
Over-dispersion	0.583	0.095	0.003	0.415
Component 2				
Intercept	-0.153	0.229	-0.613	0.299
ln(through AADT)	0.209	0.149	0.020	0.409
ln(right-turning AADT)	0.095	0.038	0.033	0.171
ln(left-turning AADT)	0.401	0.056	0.293	0.515
ln(ratio of pedestrians & bikes over total AADT)	-0.165	0.045	-0.269	-0.087
Presence of bus stop	-0.120	0.131	-0.395	0.119
Presence of subway station	0.124	0.146	-0.166	0.430
Over-dispersion	4.230	1.277	2.542	7.300
Log pseudo marginal likelihood	-1,549.690	-	-	-

Table 4-4 (continued)

	Posterior Mean	Std. Dev.	Credible intervals 2.50%	97.50%
Standard Poisson-gamma model				
Intercept	-6.983	0.921	-8.792	-5.180
ln(through AADT)	0.486	0.092	0.306	0.668
ln(right-turning AADT)	0.201	0.035	0.132	0.271
ln(left-turning AADT)	0.189	0.034	0.121	0.256
ln(ratio of pedestrians & bikes over total AADT)	-0.092	0.040	-0.172	-0.014
Presence of bus stop	0.462	0.123	0.218	0.703
Presence of subway station	0.348	0.126	0.101	0.597
Over-dispersion	0.763	0.060	0.654	0.886
Log pseudo marginal likelihood	-1,496.86	-	-	-
Random intercepts over-dispersed Poisson model				
Intercept mean	-7.649	0.768	-9.066	-6.078
Intercept variance	0.886	0.489	0.047	1.540
ln(through AADT)	0.499	0.083	0.331	0.656
ln(right-turning AADT)	0.214	0.044	0.128	0.302
ln(left-turning AADT)	0.174	0.044	0.089	0.261
ln(ratio of pedestrians & bikes over total AADT)	-0.013	0.048	-0.107	0.080
Presence of bus stop	0.718	0.135	0.455	0.985
Presence of subway station	0.329	0.137	0.059	0.600
Variance v_ε (for extra variation)	0.490	0.480	0.010	1.408
Log pseudo marginal likelihood	-1,512.18	-	-	-

The vehicle-injury dataset is highly over-dispersed with a relatively high mean value, so that it is not surprising to find models accommodating these features supported by the criteria reported in Table 4-4. It can be implied from Table 4-4 that log pseudo marginal likelihoods significantly differ from one model to another. The over-dispersed generalized linear Dirichlet process mixture model provides the highest log pseudo marginal likelihood of -1,462.67 followed by the standard Poisson-gamma model, the random intercept model, and then the finite mixture Poisson-gamma model. In comparison to the Poisson-gamma model, for example, we obtain a log pseudo Bayes factor of 34.01 (1,496.86-1,462.67) that provides support for the over-dispersed Dirichlet process mixture model. Based on the results, through AADT, left-turning AADT, right-turning AADT, the presence of bus stop, and the presence of subway station are

positively associated with vehicle-injury counts. However, the ratio of the number of pedestrians and cyclists to motorized traffic is negatively associated with vehicle-injury counts. This indicates that as pedestrians' and cyclists' activities increase, vehicle-injury frequencies decrease likely due to an increase in drivers' level of concentration and a decrease in operating speed.

Similar to the vehicle-injury dataset, the Dirichlet precision parameter α is close to 0 in the grade crossing dataset, again suggesting that the underlying random intercept distribution is not normal. The posterior density of α based on both gamma and uniform priors is shown in Fig. 4-4. We have support for the specified gamma prior based on two model-fitting measures: the cross-validation predictive density and the predictive ability of the model in replicating excess zero values. The generalized linear Dirichlet process mixture model (the simple Poisson model with a Dirichlet mixture over the intercepts) identifies around 8 (3, 18) latent components for the crossing dataset. Note that, since the grade crossing dataset is not highly over-dispersed, this is not an over-dispersed model in contrast to that used for the vehicle-injury dataset. The variance of the error term v_ε was estimated to be very close to 0 when we analyzed the crossing dataset using the over-dispersed Dirichlet process mixture model, and so was dropped from further consideration.

It can be implied from Table 4-5 that the regression coefficients estimates obtained from different models are similar. Traffic exposure (the product of train flow and vehicle flow), train speed, interaction between the logarithm of road speed and the logarithm of the train flow to vehicle flow ratio were found to be positively associated with crash frequencies. In contrast, the presence of a gate in addition to the flashing lights and bells was found to reduce crash frequency. Finally, grade crossings located in Ontario, Quebec, Pacific region, and Atlantic region was found to have a lower chance of crash frequency compared to those located in the Prairie region. In terms of goodness-of-fit, the generalized linear Dirichlet process mixture model provides the highest log pseudo marginal likelihood; that is, -1687.65. This results in a log pseudo Bayes factor of 43.6 when comparing this model with the commonly used Poisson-gamma model (negative binomial), the conventional over-dispersed generalized linear model. When comparing the random intercept model with the Poisson-gamma model, a log pseudo Bayes factor of 20.96 provides support for the random intercept model in the grade crossing dataset.

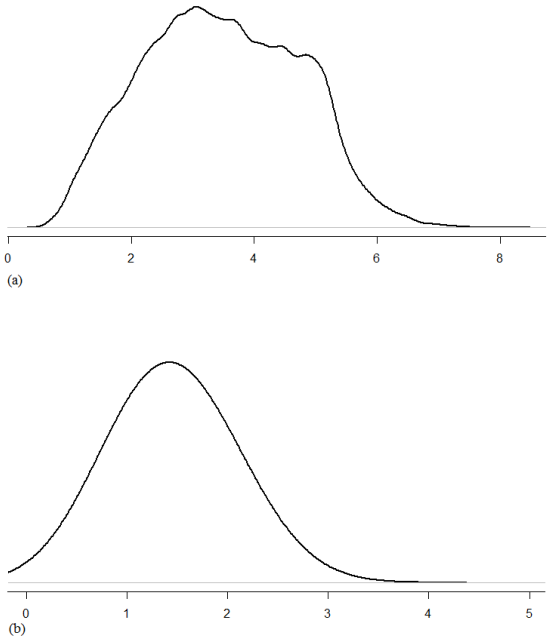


Figure 4-4 Kernel density plots of Dirichlet precision parameter for the grade crossing dataset: (a) α ~ uniform(0.3, 10); and (b) α ~ gamma(1, 1)

We also examined the performance of the proposed model in terms of its ability to replicate a high proportion of zero crashes as in the grade crossing dataset. This posterior predictive check is based on a selected statistic of interest as discussed in Rubin (1984). To implement, we first replicated crash observations based on estimated expected crash frequencies inside the MCMC algorithm. A Bayesian p-value (Gelman et al., 1996) then compares the proportions of zeros in replicated and observed data. A p-value of 0.5 indicates a perfect similarity between the observed and replicated data with respect to the proportion of zero crashes. To obtain the above p-value, at each iteration of the MCMC simulations, we examined whether replicated observations based on the developed model are equal to zero. We then obtained the proportion of zero crashes in the replicated data at each iteration and compared this proportion to that of the observed data. Summarizing the results of this comparison over all iterations, we calculated the p-value of interest. The results of the posterior predictive check in terms of the proportion of zero counts estimated a Bayesian p-value of 0.529, which is very close to the value 0.5 indicating a very good match between observed and replicated zero counts. Therefore, the Dirichlet process mixture model is excellent in this regard as well.

Table 4-5 Posterior inference for the grade crossing crash dataset

	Posterior Mean	Std. Dev.	Credible intervals 2.50%	97.50%
Dirichlet process mixture Poisson model				
Intercept mean	-7.693	1.248	-11.250	-6.223
Intercept variance	5.383	14.200	0.139	41.440
ln(traffic exposure)	0.488	0.035	0.420	0.557
ln(train speed)	0.226	0.098	0.036	0.423
ln(road speed)*ln(train ratio)	0.014	0.008	0.001	0.029
Presence of gate	-0.686	0.126	-0.934	-0.437
Ontario/Quebec[1]	-0.913	0.105	-1.120	-0.705
Pacific/Atlantic region[1]	-0.575	0.145	-0.860	-0.292
Dirichlet baseline mean	-8.045	2.785	-15.600	-4.393
Dirichlet Baseline Std. Dev.	3.374	2.286	0.847	9.168
Dirichlet precision parameter α	0.922	0.509	0.323	2.232
Log pseudo marginal likelihood	-1,687.65	-	-	-
Standard Poisson-gamma model				
Intercept	-6.631	0.523	-7.651	-5.525
ln(traffic exposure)	0.490	0.037	0.420	0.566
ln(train speed)	0.181	0.101	0.012	0.386
ln(road speed)*ln(train ratio)	0.016	0.008	0.001	0.032
Presence of gate	-0.676	0.126	-0.925	-0.434
Ontario/Quebec	-0.911	0.105	-1.117	-0.711
Pacific/Atlantic region	-0.584	0.147	-0.867	-0.289
Over-dispersion	0.912	0.225	0.599	1.467
Log pseudo marginal likelihood	-1,731.720	-	-	-
Random intercepts over-dispersed Poisson model				
Intercept mean	-7.364	0.545	-8.595	-6.577
Intercept variance	0.871	0.161	0.586	1.218
ln(traffic exposure)	0.491	0.036	0.422	0.562
ln(train speed)	0.237	0.128	-0.940	-0.439
ln(road speed)*ln(train ratio)	0.013	0.102	0.062	0.456
Presence of gate	-0.688	0.008	0.001	0.029
Ontario/Quebec	-0.914	0.105	-1.121	-0.707
Pacific/Atlantic region	-0.572	0.147	-0.866	-0.290
Variance v_ε (for extra variation)	0.871	0.161	0.586	1.218
Log pseudo marginal likelihood	-1,710.760	-	-	-

[1] The Prairie region is the reference region.

4.1.6 An example of policy implications

In terms of engineering insights, this section provides a policy example to show how the proposed model specification affects high-crash location identification procedures, a critical component of the transportation safety management process. For this purpose, we use the railway grade crossing data described above. Crossing data are usually analyzed employing a Poisson-gamma (negative binomial) model. In the previous section, however, we showed that the Poisson-gamma model is not an appropriate choice for this dataset. Consequently, we showed that the Dirichlet process mixture Poisson model performs better because of an enhanced model specification.

Suppose the aim is to prioritize the funding of grade separation projects. Specifically, we are interested in selecting the top 20 (most hazardous) sites for grade separation, which is a costly countermeasure, eliminating the risk of vehicle-train collision. Different ranking criteria and strategies exist in road safety literature to conduct a prioritization process (Washington and Oh, 2006). We used the posterior expected crash frequency–which is among the most valid high-crash location identification methods– to rank grade crossings, using the aforementioned models. A list of high-crash locations reporting the top 20 hazardous sites is reported in Table 4-6.

As it can be seen in Table 4-6, sites 1404, 4721, 3925, and 715 are selected by the Dirichlet process mixture model in the list of the most 20 hazardous crossings, whereas these sites are not among high-crash locations according to the Poisson-gamma model, the model with a spurious statistical assumption. Therefore, using a simplistic statistical model that cannot capture the underlying structure of the data may cause the false negative problem: a site is not selected as hazardous while it belongs to the list of high-crash locations based on a specified criterion. The presence of such problem will obviously lead to an ineffective allocation of funds, reducing the effectiveness of an implemented countermeasure and jeopardizing people's safety.

To illustrate the sensitivity to the size of the hazardous site's list, Fig. 4-5 displays how the number of false negative cases varies as the number of sites in the list increases. For the dataset adopted here with 6,617 observations, the number of false negatives increases for lists including up to 500 sites, then it gradually decreases as expected intuitively. Note that this pattern is observed in this dataset and may not be generalized

to other data. Nevertheless, this example clearly highlights that such cases can exist. The use of our proposed model is therefore justified since it provides more reliable engineering insights due to its superior model specification that better captures the hidden structure of the data.

Table 4-6 Comparison of high-crash location lists

Dirichlet process mixture Poisson model			**Poisson-gamma model**		
Crossing ID	**Posterior mean**	**Rank**	**Crossing ID**	**Posterior mean**	**Rank**
5826	2.144	1	5826	1.71	1
4673	1.653	2	2701	1.516	2
5048	1.575	3	5705	1.331	3
2701	1.412	4	2030	1.33	4
3914	1.364	5	3914	1.321	5
5577	1.349	6	3793	1.188	6
3793	1.299	7	4673	1.076	7
4988	1.11	8	5048	1.041	8
5705	1.092	9	725	1.021	9
2030	1.076	10	3326	0.9906	10
1404	1.055	11	2699	0.9901	11
725	0.8448	12	3507	0.9807	12
3324	0.8399	13	3324	0.8762	13
3507	0.8264	14	4988	0.845	14
2699	0.8214	15	5577	0.8396	15
3326	0.8177	16	4418	0.8123	16
4721	0.7501	17	3481	0.7677	17
3925	0.7327	18	3436	0.7334	18
715	0.7146	19	3784	0.7249	19
4418	0.7134	20	2284	0.7242	20

Note: italic underlined crossings are not in the list selected by the Poisson-gamma model

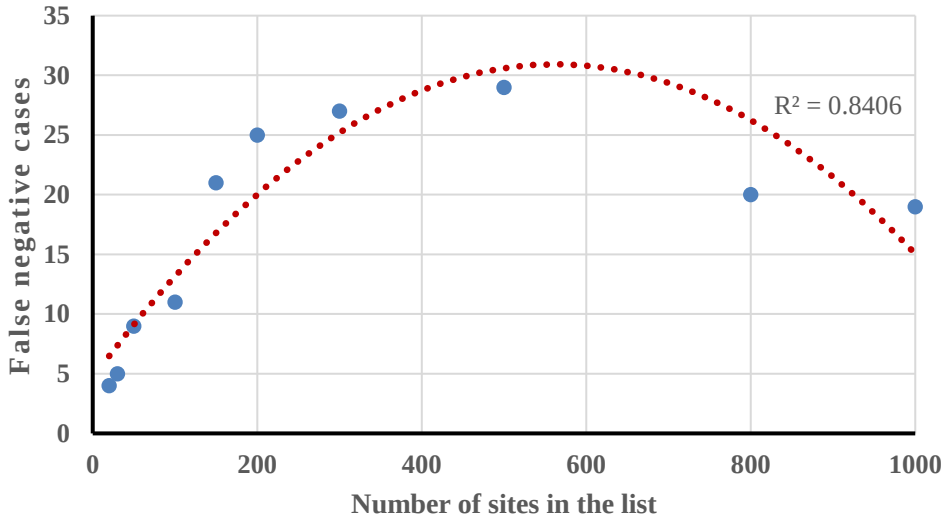


Figure 4-5 Variation in the number of false negatives as a function of list's size

4.1.7 Summary of univariate modeling

Section 4.1 introduced Dirichlet process mixture models to analyze crash data in univariate settings. The proposed technique derives from the Bayesian nonparametric literature, and presents a semiparametric model based on Dirichlet process priors. We followed Mukhopadhyay and Gelfand (1997) and Ohlssen et al. (2007) to refine the model to one which is not computationally cumbersome. The nonparametric part of the model manifests in the intercepts in the univariate settings. Modeling intercepts nonparametrically allows us to conveniently retain the linear form of the vector of coefficients in relation to log-transformed responses (e.g., crash frequencies or differing injury-severity levels). This in turn retains usual interpretations made by conventional generalized linear models.

Using two simulated data, we first highlighted how the proposed model works and compares to conventional models used in road safety literature. We then adopted two real datasets for our univariate setting: (1) a vehicle-injury count dataset from signalized intersections that is somewhat highly over-dispersed and is characterized by a relatively large mean value; and (2) a railway grade crossing crash dataset that is characterized by the low mean value problem and excess zero counts. The proposed model allowed us to examine the sensitivity to parametric assumptions, providing a better fit to both

datasets compared to other conventional models commonly used in road safety studies. The results showed that the proposed model performs well on different data with dissimilar different characteristics. We also provided a policy exercise to show the effect of model specification on the identification of high-crash locations for engineering safety improvements. The results indicated that model misspecification may result in erroneously selected sites for safety improvement programs, reducing the effectiveness of implemented countermeasures.

4.2 Multilevel Modeling

In this section, we account for the hierarchical structure of data. Specifically, we used two grade crossing datasets in which crossings are nested within different geographical areas. Doing so, we account for spatial dependency among crossings located in the same regions. Firstly, we analyze both datasets using the simple Poisson-lognormal model (Section 2.1.3). Recall that this model neglects the multilevel structure of the data. Secondly, we use the standard random intercepts multilevel model described in Section 2.1.7.1, Eq. 2-13, assuming that all random intercepts are generated from a unique normal density. Thirdly, we employ our flexible model proposed for the multilevel settings as discussed in Section 3.4, Eq. 3-12, without imposing restrictions on the form of the random intercepts. Different model comparisons and a policy example then follow.

4.2.1 Province level grade crossing data

Across Canada, 4,213 crossings (with flashing lights and bells) were selected for the province-level data. A total of 303 crashes were observed at these crossings during a six-year period, 2008-2013. The grade crossings are located in eight different Canadian provinces: British Columbia, Alberta, Saskatchewan, Manitoba, Ontario, Quebec, New Brunswick, and Nova Scotia. Note that Canada has a total of ten provinces and three territories; the two provinces (Prince Edward Island and Newfoundland and Labrador) and the tree territories were not included because there are a few crossings with flashing lights and bells available in the data in these regions. The number of crossings in each

province is reported in Table 4-7. In the data preparation process, an id was assigned to each province and observations nested within each province.

Table 4-7 Spatial distribution of crossings in various provinces

Province	Frequency	Percent
Alberta	461	10.94
British Columbia	373	8.85
Manitoba	305	7.24
New Brunswick	187	4.44
Nova Scotia	177	4.20
Ontario	1,242	29.48
Quebec	1,070	25.40
Saskatchewan	398	9.45

Table 4-8 Summary statistics of the province-level data

Variable	Mean	Std. Dev.	Min	Max
Train flow (average annual daily)	5.83	6.61	0.01	46.00
Vehicle flow (average annual daily)	2,567.95	5,085.48	1.00	71,500.00
Log of exposure (product of train and vehicle flows)	7.86	1.80	-1.39	12.98
Number of tracks	1.13	0.42	1.00	6.00
Number of lanes	2.14	0.62	1.00	6.00
Track angle (deviation from 90°)	20.94	20.41	0.00	87.00
Road speed (km/h)	63.19	18.41	5.00	110.00
Train speed (km/h)	48.94	24.09	1.61	110.86
Whistle prohibition (1 if present, 0 otherwise)	0.07	0.25	0.00	1.00
Urban area (1 if urban area, 0 otherwise)	0.33	0.47	0.00	1.00
Urban/Whistle prohibition interaction	0.04	0.20	0.00	1.00
Crash frequency (6-year period)	0.07	0.31	0.00	4.00

A large number of explanatory variables were considered; however, most of them were not found to have an important effect. Summary statistics of the province-level dataset are provided in Table 4-8. The interaction between the dummy variables urban (1 if crossing is in an urban area) and whistle prohibition (1 if whistle prohibition applies) was considered in the analysis as it was found to provide a better fit to the data. Among

4,213 crossings in the province-level dataset, 185 (4.39%) were those crossings for which whistle prohibition was in effect and were situated in urban areas. Whistle prohibition were applied to 6.74% of the crossings and 32.68% of the crossings were located in urban areas.

4.2.2 Municipality-level grade crossing data

To prepare this dataset, municipalities with at least 10 crossings in their boundary were considered. The final municipality-level data included 1,513 crossings located in 81 municipalities, which come from 8 major Canadian provinces: British Columbia, Alberta, Saskatchewan, Manitoba, Ontario, Quebec, New Brunswick, and Nova Scotia. A total of 135 crashes were observed in the municipality-level data. This dataset includes all major Canadian cities such as Toronto, Montreal, Winnipeg, Edmonton, Vancouver, etc. It should be underscored that a number of factors (e.g., driver behavior, climate, regulations, etc.) might differ between different municipalities, so that one scope of this research was to verify the existence of dependencies (similarities) among grade crossings situated within the same municipality. More importantly, we aimed at examining the standard parametric assumption for the data while accounting for its multilevel form. Table 4-9 provides descriptive statistics of the data for the most important variables.

Table 4-9 Summary statistics of the municipality-level data

Variable	Mean	Std. Dev.	Min	Max
Train flow (average annual daily)	5.54	6.07	0.01	28.86
Vehicle flow (average annual daily)	4,034.12	7,206.42	1.00	71,500.00
Log of exposure (product of train and vehicle flows)	8.15	1.97	-1.31	12.98
Number of tracks	1.12	0.45	1.00	6.00
Number of lanes	2.26	0.79	1.00	6.00
Track angle (deviation from 90°)	21.34	20.55	0.00	80.00
Road speed (km/h)	61.44	16.17	5.00	100.00
Train speed (km/h)	41.10	24.51	1.61	100.58
Whistle prohibition (1 if present, 0 otherwise)	0.14	0.35	0.00	1.00
Urban area (1 if urban area, 0 otherwise)	0.37	0.48	0.00	1.00
Crash frequency (6-year period)	0.09	0.36	0.00	4.00

We were also interested in identifying outlier municipalities (those that perform differently from the rest of the data), and municipalities that manifest similar patterns (latent subpopulations) in terms of crash frequency at crossings equipped with flashing lights and bells. Among 1,513 grade crossings in this dataset, 36.55% were in urban areas, and whistle prohibition were applied to 14.28% of them. A host of explanatory variables were available, but many of them did not have any important effect on crash frequencies or were removed from the model due to collinearity.

4.2.3 Prior specification and model computation

Given the parameters presented in Section 3.4.1, we used non-informative normal priors with mean zero for $\boldsymbol{\beta}$ and m_0. For the inverse of variances v_ε, v_η, and v_0, we used a gamma prior with shape and rate (inverse-scale) parameters being equal to 0.01. It is also necessary to define a prior distribution for the precision parameter α for which different priors are possible such as gamma, exponential, and uniform. This prior could agree with the maximum number of allowed clusters C (see Section 3.2).

For the province-level data, since observations were nested in 8 provinces and we were interested to cluster over provinces (random intercepts at province-level), we used a C value of 8. A uniform prior with an upper bound of 1 was selected. This results in a maximum of 8 clusters. Therefore, we assume $\alpha \sim$ uniform(0.2, 1) for the latter dataset. A lower bound of 0.2 was selected here to allow smaller values of α and also to circumvent problems associated with the estimation of p_n (see Section 3.2). Since the above prior is informative, we also used a uniform(0.2, 10), which allows larger values of α. We discuss this further in the next section.

For the municipality-level dataset, we mainly used similar priors as for the province-level data. We set the maximum number of cluster C to be 50 given the number of municipalities being 81. We chose a uniform prior with an upper bound of 10 that corresponds to approximately 50 clusters based on Eq. 3-7. Therefore, we assume a uniform(0.2, 10) for α.

A total of 20,000 MCMC iterations, in addition to 5,000 burn-in iterations, with 2 chains were utilized to obtain posterior inferences. All three models ran smoothly and converged relatively quickly. For example, the flexible Dirichlet process mixture

multilevel model converged at around 4,000 iterations. This is an indication of well-defined models and priors. The MCMC convergence was verified through history plots, trace plots, and Gelman-Rubin diagram, being available in WinBUGS. Note that the Gelman-Rubin diagram is a visual demonstration of the Gelman-Rubin statistic, which is a quantitative measure of convergence (Gelman and Rubin, 1992; Brooks and Gelman, 1998). The readers are referred to the WinBUGS manual (Spiegelhalter et al., 2003) for a detailed treatment of convergence verification techniques in WinBUGS.

In addition to checking convergence, other methods are available to examine the accuracy of the posteriors. For every parameter of interest in WinBUGS, for instance, a Markov chain error estimate is provided among other statistics. For every stochastic parameter of interest, as a rule of thumb, the value of the Markov chain error should be smaller than 5% of the estimated standard deviations.

4.2.4 Results and discussions

Regarding the province level data, estimation results for the simple Poisson-lognormal model, the random intercepts multilevel Poisson-lognormal model, and the flexible Dirichlet process mixture multilevel model are presented in Table 4-10. The number of lanes was found to have an important effect on crash frequencies among railway crossings under the simple Poisson-lognormal model, the single-level model. Interestingly, in the multilevel models, however, this variable did not have an important effect. This is in accordance with previous research (Kim et al., 2007; Jovanis et al., 2011; Dupont et al., 2013). As discussed by Dupont et al. (2013), single-level models—such as the simple Poisson-lognormal model employed here—assume that all observations are generated from a unique homogeneous population. This, in turn, implies that the residuals are independent resulting in underestimated standard errors; and consequently, erroneous confidence intervals.

With respect to the random intercepts multilevel Poisson-lognormal model, the estimated variance of the random intercepts clearly confirms the need for the multilevel approach. The estimated intra-province correlations (Eq. 2-14) are 0.54 and 0.64 according to the standard and the flexible multilevel models, respectively. Therefore, the simple Poisson-lognormal model is not an appropriate choice. Recall that the latter

does not account for the hierarchical structure of the data; thus, neglecting similarities between crossings located within the same region. One can also notice a considerable improvement in the model fitting (comparing log pseudo marginal likelihoods in Table 4-10) when using the random intercepts multilevel Poisson-lognormal model instead of the simple Poisson-lognormal model. Based on this multilevel model, exposure, train speed, and urban/whistle prohibition interaction were found to be associated with crash frequencies considering a 5% level of confidence.

One key scope of this research was to examine the adequacy of the standard parametric assumption for the intercepts in the random intercepts multilevel Poisson-lognormal model, using the flexible Dirichlet process mixture multilevel model. The results for the latter model are also represented in Table 4-10. In terms of covariates, similar site characteristics, as in the random intercepts multilevel Poisson-lognormal model, are found to be important in the model. With respect to model-fitting measures, a log pseudo Bayes factor of 2.7 provides support (according to Kass and Raftery (1995), Table 3-1) for the conventional random intercepts multilevel model. Therefore, our flexible model does not provide a superior fit to the province-level data. Also, the posterior distribution of the Dirichlet precision parameter is similar to its prior when using uniform(0.2, 1) or uniform(0.2, 10), indicating that the data do not provide enough information about this parameter. Based on the above observations, one can postulate that the flexible model may not be needed for the province-level data.

Table 4-11 presents the analyses results related to the municipality-level data. Similar to the province-level data, the simple Poisson-lognormal model provided a poor fit compared to other two models that account for the hierarchical structure of the data, crossings nested within municipalities. The estimated intra-municipality correlations are 0.57 and 0.84 according to the standard and the flexible multilevel models, respectively. The results highlighted that traffic exposure, urban area, whistle prohibition, and train speed increase crash frequencies at crossings. The variance of the intercepts in the multilevel framework indicates that crossings nested in the same municipalities are somehow dependent. Therefore, the simple Poisson-lognormal model is not a proper choice.

Table 4-10 Posterior inference for province-level data

Variable	Posterior mean	Std. dev.	Bayesian intervals 2.50%	97.50%
Simple Poisson-lognormal model				
Log of exposure	0.485	0.046	0.395	0.574
Train speed	0.010	0.003	0.005	0.015
Number of lanes	0.206	0.081	0.043	0.362
Urban/Whistle prohibition interaction	0.738	0.210	0.322	1.151
Intercept	-8.245	0.455	-9.115	-7.332
Variance v_ε	0.795	0.255	0.367	1.326
Log pseudo marginal likelihood	-1,001.31	-	-	-
Random intercepts multilevel Poisson-lognormal model				
Log of exposure	0.505	0.042	0.423	0.586
Train speed	0.008	0.002	0.004	0.013
Urban/Whistle prohibition interaction	0.696	0.208	0.282	1.103
Intercept mean	-7.852	0.504	-8.852	-6.866
Intercept variance	0.660	0.817	0.113	2.465
Variance v_ε	0.560	0.249	0.106	1.036
Log pseudo marginal likelihood	-979.928	-	-	-
Flexible Dirichlet process mixture multilevel model				
Log of exposure	0.510	0.041	0.430	0.590
Train speed	0.009	0.003	0.004	0.014
Urban/Whistle prohibition interaction	0.721	0.212	0.299	1.131
Intercept mean	-8.001	0.577	-9.154	-7.026
Intercept variance	1.162	7.826	0.083	5.076
Intercept's baseline mean m_0	-8.051	1.485	-10.610	-5.353
Intercept's baseline variance v_0	8.482	96.200	0.103	43.880
Variance v_ε	0.662	0.210	0.328	1.139
Dirichlet precision parameter α	0.752	0.202	0.275	0.990
Log pseudo marginal likelihood	-982.634	-	-	-

Whistle prohibition is significant at a level of confidence of 0.05 in the simple Poisson-lognormal model, but this variable is only significant at a 10% level of confidence in other two models. Similar to the province-level data, this again confirms the fact that standard errors and intervals around the mean may be estimated erroneously in the single-level model. In contrast to the province-level data, this time our flexible model provided the best fit to the municipality-level data. The log marginal likelihood of the flexible Dirichlet process mixture multilevel model is the highest (see Table 4-11).

Table 4-11 Posterior inference for municipality-level data

Variable	Posterior mean	Std. dev.	Bayesian intervals 2.5%	97.5%
Simple Poisson-lognormal model				
Log of exposure	0.488	0.067	0.358	0.618
Urban area	0.605	0.219	0.181	1.036
Whistle prohibition	0.522	0.246	0.039	1.010
Train speed	0.012	0.004	0.003	0.020
Intercept	-8.204	0.673	-9.553	-6.946
Variance v_ε	0.963	0.393	0.177	1.817
Log pseudo marginal likelihood	-414.610	-	-	-
Random intercepts multilevel Poisson-lognormal model				
Log of exposure	0.504	0.063	0.379	0.628
Urban area	0.498	0.245	0.021	0.977
Whistle prohibition	0.452	0.253	0.044	0.873
Train speed	0.016	0.005	0.006	0.026
Intercept mean	-8.558	0.666	-9.934	-7.282
Intercept variance	0.714	0.379	0.162	1.612
Variance v_ε	0.534	0.321	0.047	1.259
Log pseudo marginal likelihood	-404.837	-	-	-
Flexible Dirichlet process mixture multilevel model				
Log of exposure	0.496	0.065	0.370	0.626
Urban area	0.541	0.230	0.084	0.986
Whistle prohibition	0.443	0.240	0.051	0.838
Train speed	0.017	0.005	0.007	0.027
Intercept mean	-8.768	0.798	-10.360	-7.361
Intercept variance	2.356	8.844	0.340	9.670
Intercept's baseline mean m_0	-9.078	1.495	-12.230	-6.793
Intercept's baseline variance v_0	7.854	48.560	0.314	42.250
Variance v_ε	0.463	0.344	0.032	1.256
Dirichlet precision parameter α	3.700	2.610	0.415	9.546
Log pseudo marginal likelihood	-401.347	-	-	-

Note: Whistle prohibition is significant at a 10% level of significance in multilevel models.

When comparing the flexible Dirichlet process mixture multilevel model with the random intercepts multilevel Poisson-lognormal model, a log pseudo Bayes factor of 3.5 indicates support for the flexible model. Also, the posterior density of the precision parameter supports smaller values of α. Therefore, the adequacy of the standard

parametric assumption on the random intercepts can be questioned. In other words, assuming a single distribution for all 81 municipalities does not seem to be appropriate.

4.2.5 An example of policy implications

Employing Dirichlet process mixture models to analyze multilevel crash data in which observations are nested within different geographical areas appears to be immensely useful in terms of regional or national safety policy and benchmarking. Specifically, our method allows the estimation of pairwise probabilities of similarities between regions; and consequently, helps identify (i) clusters of regions that perform similarly, and (ii) outliers, those performing very different from other regions. Such analyses stimulate further investigations to find reasons for inter-regional variations in safety performances, a task that can be achieved by an in-depth research. Here, we demonstrate how the above analysis can provide practical engineering insights using both datasets analyzed in Section 4.2. It should be mentioned that the Dirichlet process mixing is in general more useful for a larger number of groupings (here, provinces); for instance, the municipality-level data.

In this study, although we did not find a strong evidence to rule out the parametric assumption for the province level data (as discussed earlier), for the sake of demonstration, we identify latent clusters and outliers among various Canadian provinces using the flexible Dirichlet process mixture multilevel model. To this end, we employed the cluster detection algorithm described in Section 3.6. The results are shown in Fig. 4-6 in which provinces in the same cluster are filled by identical colors. Table 4-12 reports estimation results using the cluster detection algorithm, allowing for the detection of the most similar and dissimilar provinces.

For example, the most similar provinces in terms of total crash frequencies were Ontario and Quebec with an expected probability of 0.82, followed by Alberta and Saskatchewan with an expected probability of 0.71. It can be inferred from Table 4-12 that Nova Scotia is an outlier province since it has only one province (i.e., itself) in its cluster. Nova Scotia's size of cluster has an expected value of one. It is important to mention that Table 4-12 uses a threshold probability of 60% to define clusters (and outliers) among different provinces. Obviously, alternative threshold values result in

different clusters. Larger probabilities will result in higher number of clusters. In other words, the number of remaining provinces that share the same cluster with province *i* approaches 0 as the threshold probability approaches 1.

Figure 4-6 Latent clusters among the 8 Canadian provinces.

For the municipality level data, it can be seen in Table 4-11 that the expected number of non-empty clusters (mass points) is 11.66. The list of municipalities and their corresponding ID are provided in Appendix I. Fig. 4-7 provides a grey-scale plot of probabilities of similarities or clustering between pairs of municipalities, as in Ghosh et al. (2010). In this plot, darker squares indicate larger pairwise probabilities of similarities. As an example, we found that the following municipalities share the same cluster with a probability greater than 0.60: Calgary, Edmonton, Regina, Saskatoon,

Winnipeg, Grand Prairie, and Nanaimo. We monitored the total number of municipalities that share the same clusters as described in Section 3.6; however, the results indicate that there is no outlier municipality in the data. Note that the above clustering is obtained after adjusting for the effect of covariates.

Table 4-12 Cluster and outlier identification results - province-level data

Province	Size of cluster (95% interval)	Similar provinces with probability > 0.60
Alberta	2 (1, 5)	Saskatchewan
British Columbia	4 (1, 5)	Manitoba
Manitoba	3 (1, 5)	British Columbia
New Brunswick	3 (1, 5)	Ontario & Quebec
Nova Scotia	1 (1, 5)	None
Ontario	3 (1, 5)	New Brunswick & Quebec
Quebec	3 (1, 5)	New Brunswick & Ontario
Saskatchewan	3 (1, 5)	Alberta

Note: size of cluster is the median of the number of provinces in the same cluster

4.2.6 Summary of multilevel modeling

To overcome unobserved cross-group heterogeneity in multilevel crash data, random effects (including random intercepts) models are often used in transportation safety studies. Standard distributional assumptions are an intrinsic part of random effects models. Since sensitivity to such assumptions might be of a major concern in some datasets or applications, Section 4.2 proposes a class of flexible statistical models that allows us to investigate the adequacy of parametric assumptions. Our approach has some other advantages such as the ability to identify outliers and latent subpopulations in data at the higher level of the hierarchy. Our model collapses into a form of finite mixture models. In classical finite mixture models, the number of latent components should be prespecified while in most applications there is not any sound justification for selecting the number of components prior to the analysis. By contrast, our model treats the number of latent components as an unknown stochastic parameter and estimates the expected number of clusters among groupings (here, regions) as part of its mathematical algorithm. Note that studies using finite mixture models to identify

clusters among the higher level of hierarchies or groupings of data are rare, if non-existent, in the crash literature.

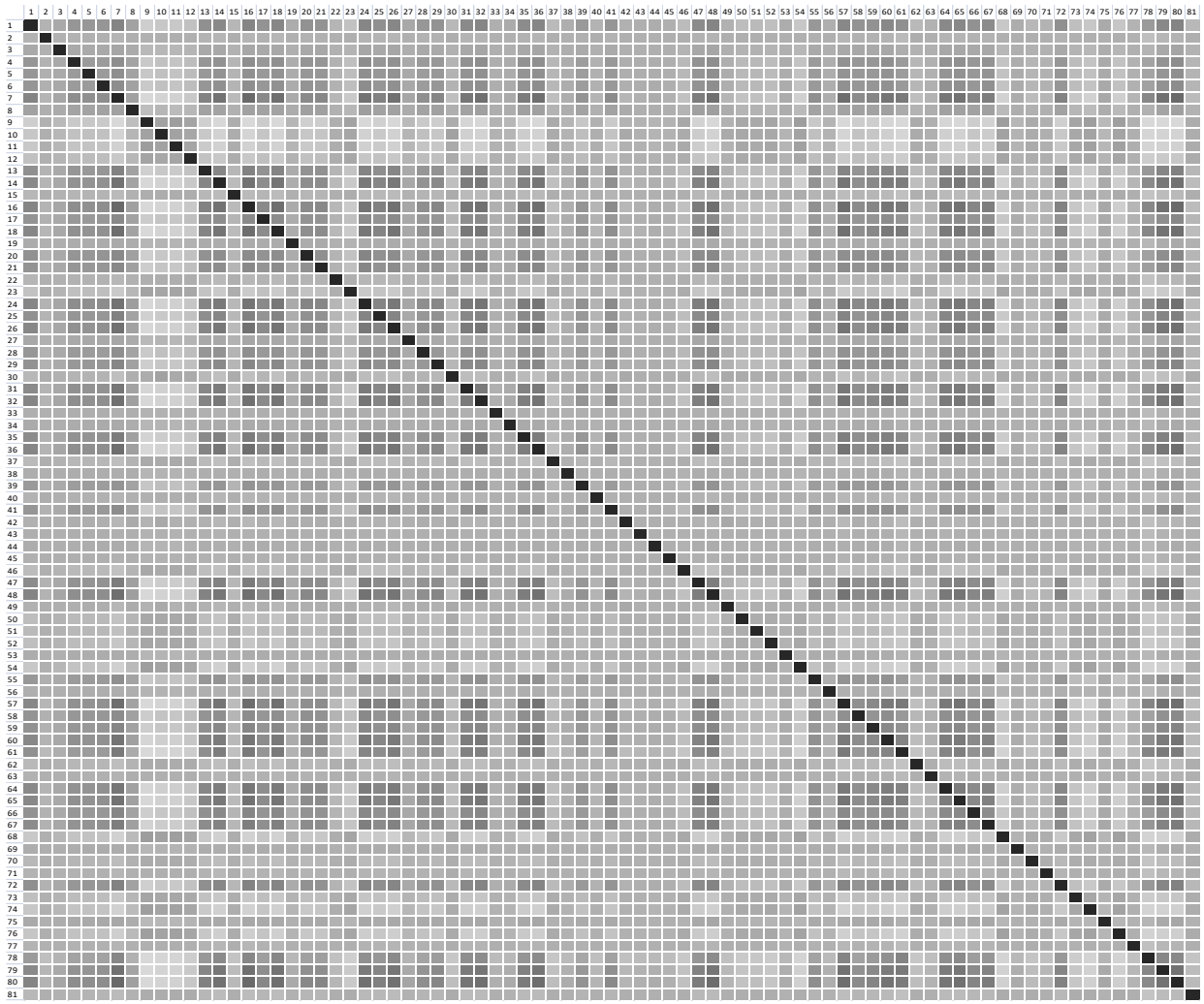


Figure 4-7 Grey-scale plot of pairwise probabilities of similarities of the 81 municipalities. Note: Darker squares indicate larger probabilities of similarities (clustering).

In this dissertation, we present the flexible Dirichlet process mixture multilevel model as an alternative to the random effects and random parameter models to accounting for unobserved heterogeneity in multilevel settings. Recall that random parameter models allow some or all model covariates to vary across observations as way of overcoming unobserved heterogeneity in data; however, due to computational complexities, especially in large datasets, only a limited number of parameters can often be allowed as random parameters.

We adopted two multilevel datasets—containing crash frequencies for grade crossings equipped with flashing lights and bells in Canada—to show the feasibility of our flexible model. Log pseudo marginal likelihood and log pseudo Bayes factors (computed from conditional predictive ordinates) were utilized for model comparison. The results confirmed the need for a multilevel modeling approach. We found that non-multilevel models (simple Poisson-lognormal models) underestimated standard errors of the coefficients associated with the number of lanes and whistle prohibition in province-level and municipality-level data, respectively. Traffic exposure, the location of crossing (urban vs. non-urban), train speed, whistle prohibition, and the interaction between whistle prohibition and urban area were positively associated with crash frequencies. Based on the evidence provided by the two datasets, the results illustrated that the standard distributional assumption for the random intercepts could not be ruled out for the province-level data, whereas this assumption was found to be under question for the municipality-level data.

In our policy example, we identified latent subpopulations among Canadian provinces and municipalities. In terms of outlier regions, the results indicated that the province of Nova Scotia is an outlier province in the study in context and that there is not any outlier municipality among those analyzed in this research. It should be noted that identifying latent clusters among various regions has a significant interpretative value. This is an indicator of common unmeasured/unknown factors among those regions that are in the same clusters. Based on the identified clusters, further investigations can be conducted to detect or hypothesize the presence (or extent) of such unidentified attributes. Note that latent similarities and dissimilarities are expected among different regions due to variations in different regional policies, population demography, driver behavior, climate, traffic regulations, etc.

4.3 Multivariate Modeling

This section presents the data and analyses results relating to multivariate models discussed previously in Section 2.1.8.1 and Section 3.5. The first dataset is used to demonstrate the application of the multivariate mixture of points model discussed in Section 3.5.1. The second dataset is used to demonstrate the application of the mixture

of multivariate normal density model explained in Section 3.5.3. The first dataset contains correlated injury-severity levels. The second dataset contains correlated non-motorist crash types: pedestrian and cyclist injury counts. For both dataset, we compare the proposed models with the standard multilevel Poisson lognormal model presented in Eq. 2-15. We investigate departures from restrictive assumptions and provide a policy example to show how restrictive dependence structures in multivariate settings can affect the interpretation of the explanatory variables.

4.3.1 Highway segment injury-severity data

This dataset was provided by the Ontario Ministry of Transportation. We employed the multivariate mixture of points model (See Section 3.5.1) to analyze this dataset, which consists of crash data from 418 highway segments in Ontario (highway 401) collected over a 3-year period (2006 to 2008). Descriptive statistics of the highway 401 dataset are provided in Table 4-13.

Highway 401 connects eastern Ontario (the Quebec boarder) to south west Ontario (the Michigan boarder). This highway is a major roadway in Ontario with a very large number of vehicles passing through it on a daily basis. The crash data are divided into three categories of severities: fatal, injury, and property damage only crashes. Due to limited number of fatal crashes, we divided the crash data into two categories of injury-fatal and property damage only crashes, which are modeled here simultaneously. The dataset does not distinguish between various levels of injury such as incapacitating injury. In addition to the crash history, major roadway-segment attributes were available.

We noticed a higher rate of crashes among segments with a median shoulder width of smaller than 1.80 meters during an exploratory data analysis phase. Based on the median (inside) shoulder width, we created a dummy independent variable, here named narrow median shoulder. No information relating to the vertical alignment of segments was available, but we were able to obtain the average horizontal curve degree per kilometer of highway segment.

Table 4-13 Summary statistics for the highway 401 data

Variable	Mean	Std. Dev.	Min	Max
AADT all vehicles	80,369.420	95,760.440	14,499.940	44,2900.300
AADT commercial vehicles	14,383.640	6,890.880	4,864.000	42,075.500
Percentage of commercial vehicles	29.027	12.300	3.100	49.100
Segment length (km)	1.952	2.061	0.206	12.703
Number of lanes	5.445	2.428	4.000	12.000
Median (inside) shoulder width (m)	1.598	1.194	0.000	5.190
Median width (m)	11.106	6.147	0.600	30.500
Outside shoulder width (m)	3.135	0.285	2.600	4.000
Lane width (m)	3.707	0.301	1.830	5.625
Average horizontal curve degree curvature per km	0.945	1.864	0	16.592
Paved outside shoulder (1 if paved; 0 otherwise)	0.586	0.493	0.000	1.000
Surface type (1 if HCB[1]; 0 otherwise)	0.526	0.500	0.000	1.000
Narrow median shoulder (1 if < 1.8 m; 0 otherwise)	0.629	0.493	0.000	1.000
Property-damage-only crash frequency	18.715	38.257	0.000	336.000
Injury-fatal crash frequency	4.530	9.334	0.000	96.000

[1] HCB stands for high class bituminous pavement.

4.3.2 Pedestrian/Cyclist data

The data used in this section are derived from a 6-year period (2003-2008) of pedestrian and cyclist injury counts for 647 signalized intersections in Montreal. The spatial location of the intersections is shown in Fig. 4-8. One limitation of the data is that differing injury-severity levels (e.g., minor, major, fatal) are not reported. Had such information been available, we would have been able to conduct a more comprehensive study providing detailed insights on factors that affect each injury severity level. In this dataset, the sustained injuries were mostly the consequence of crash with motorized vehicles, but very limited injuries resulted from cyclist with cyclist or pedestrian with

cyclist crashes. The availability of both crash types provides a valuable opportunity to simultaneously study crash correlates of walking and cycling in an urban area, while implementing a flexible modeling approach. In modeling pedestrian and cyclist injury counts, for instance, intersections' proximity to alcohol dispensing locations (e.g., bars) or nightlife activities may influence drivers, pedestrians, and cyclist's behavior affecting both pedestrian and bicyclist injuries in a similar way. However, such potentially important variables are often omitted from crash models due to data limitations. In these circumstances, a joint analysis helps improve the quality of estimates.

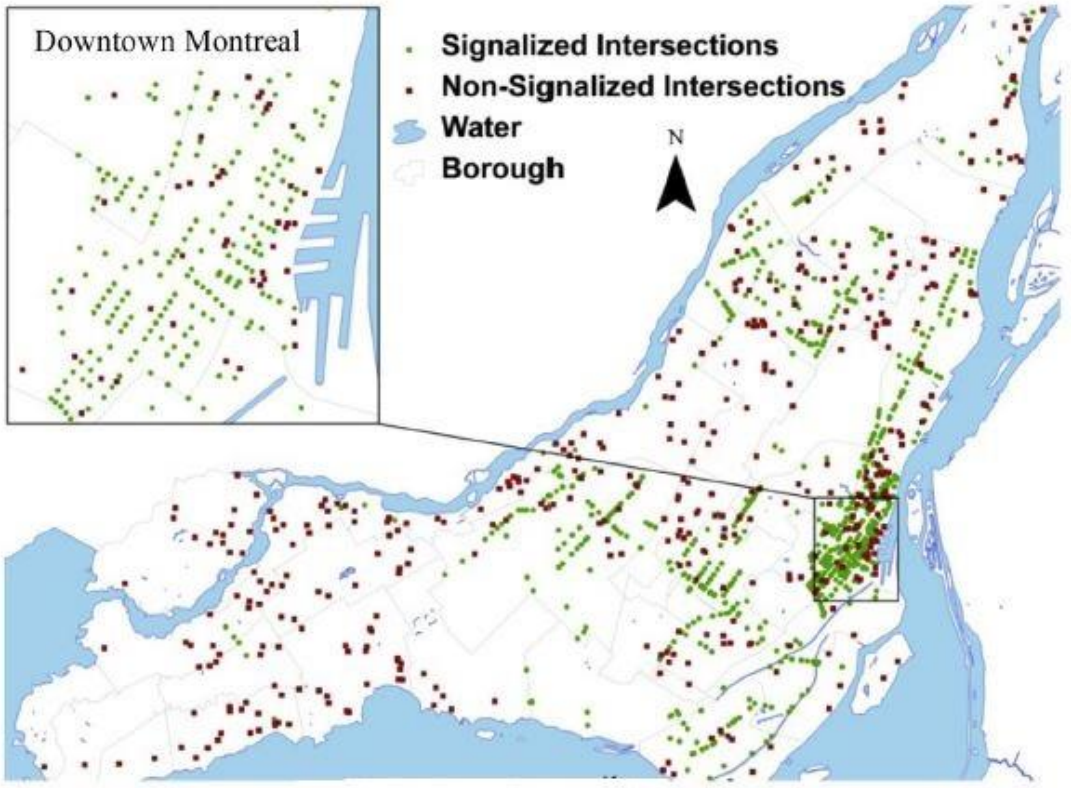


Figure 4-8 Spatial distribution of intersections (from Strauss et al. (2014)).
See the electronic version for a color view.

Several data sources (the City of Montreal, police records, ambulance services, Statistics Canada, etc.) were used to obtain motorized and non-motorized traffic flow, built environment characteristics, intersection geometric/operational characteristics, and

pedestrian/cyclist crash history. Here we explain main variables used in this research. Interested readers are referred to Strauss et al. (2014) for further details on the data. A disaggregate vehicle volume information, on the basis of turning directions, was available at each intersection. This information is useful to examine how right-turning, left-turning, and non-turning traffic differentiate in their association with different crash types. In addition, pedestrian and cyclist counts were also available, but not at a disaggregate level. Intersection geometric attributes such as the form (e.g., three-leg), crosswalk length, number of lanes, presence of median (raised or not) were available in the data. Intersection operational characteristics such as dedicated pedestrian crossing light, which were found to be important in our exploratory data analysis, were also considered. Note that the dedicated pedestrian light gives an exclusive right of way to pedestrians to cross. This was operated at some intersections by dedicating a full phase or part of a phase to pedestrian crossing while keeping all other traffic lights red. Dedicated pedestrian lights aim at providing a safer crossing experience for pedestrians.

Built environment variables such as the length of cycling facilities, the presence of bus stops, subway stations, and schools in the proximity (all within a range of 50 m, 400 m, and 800 m) of intersections, employment, land use mix, and area of commercial land use were also available in the data. Although these variables may not seem to be directly related to pedestrian/cyclist safety, they can be used as a proxy to other intersection features, for example, indicating the level of motorized and non-motorized activity around intersections. A summary of the dataset is given in Table 4-14. Fig. 4-9 provides a histogram of injury counts by crash type. Fig. 4-10 summarizes the distribution of the motorized and non-motorized (active mode) traffic.

We estimated the effects of total non-motorized volume and the ratio of non-motorized volume, based on the hypothesis that these variables may have a bearing on injury counts among vulnerable road users. In this regard, some earlier studies suggest that higher pedestrian and cyclist activity may help enhance safety of active modes of transport (Leden, 2002; Pucher and Buehler, 2008; Jacobsen, 2015; Stoker et al., 2015). However, the non-motorized intensity was not found to be useful in our models. Among other variables, land use mix that may influence the type of vehicles circulating at an intersection (and consequently, driver behavior) was examined too. Moreover, high co-linearity among some covariates meant that we could not include some subsets

of covariates in the models at the same time. Lastly, we verified co-linearity between disaggregate motorized traffic volumes to avoid including highly correlated volumes in the models.

Table 4-14 Summary statistics of the pedestrian/cyclist data

Variable types	Variables	Mean	Std. Dev.	Min	Max
Crash type	Cyclist injury counts	0.628	1.324	0.000	20.000
	Pedestrian injury counts	1.151	1.880	0.000	16.000
Exposure measure	Cyclist counts	444.915	717.616	1.662	6,433.217
	Pedestrian counts	1,578.071	3,531.822	1.000	40,958.300
	Total non-motorized volume	2,022.985	3,792.451	2.963	41,541.050
	Left-turning motorized volume	2,602.724	2,641.855	0.000	23,843.000
	Right-turning motorized volume	2,668.011	2,697.447	0.000	23,792.000
	Non-turning motorized volume	19,467.960	11,084.390	1,790.000	76,525.000
	Total motorized volume (AADT)	24,738.650	12,526.060	3,751.271	84,389.650
	Ratio of non-motorized to motorized	0.129	0.304	0.000	4.006
Built environment	Employment (800 m) (in 10000)	0.580	0.304	0.026	1.492
	Commercial area (800 m) (in 10000 m^2)	1.212	1.343	0.000	8.695
	Land use mix (400 m)	0.514	0.199	0.000	0.920
	Land use mix (800 m)	0.666	0.149	0.000	0.920
	Length of cycling facilities (400 m) (km)	0.536	0.606	0.000	2.959
	Number of schools (400 m)	1.045	1.238	0.000	6.000
	Number of subway stations (400 m)	0.439	0.701	0.000	4.000
	Presence of bus stop (50 m)	0.708	0.455	0.000	1.000
	Presence of School (400 m)	0.563	0.496	0.000	1.000
	Presence of subway stations (400 m)	0.342	0.475	0.000	1.000
Geometric & operational	Maximum speed (km/h)	61.824	9.879	50.000	100.000
	Dedicated traffic light for pedestrians	0.247	0.432	0.000	1.000
	Three-leg intersection	0.221	0.415	0.000	1.000
	Presence of raised median	0.478	0.500	0.000	1.000
	Total number of lanes	6.870	2.631	3.000	16.000

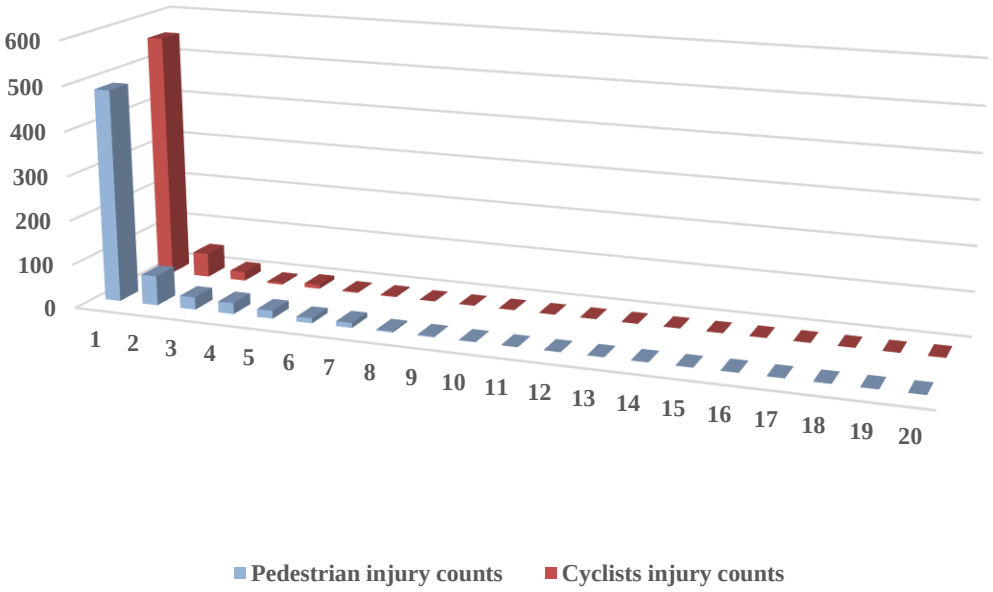


Figure 4-9 Histogram of injury counts for pedestrians and cyclists

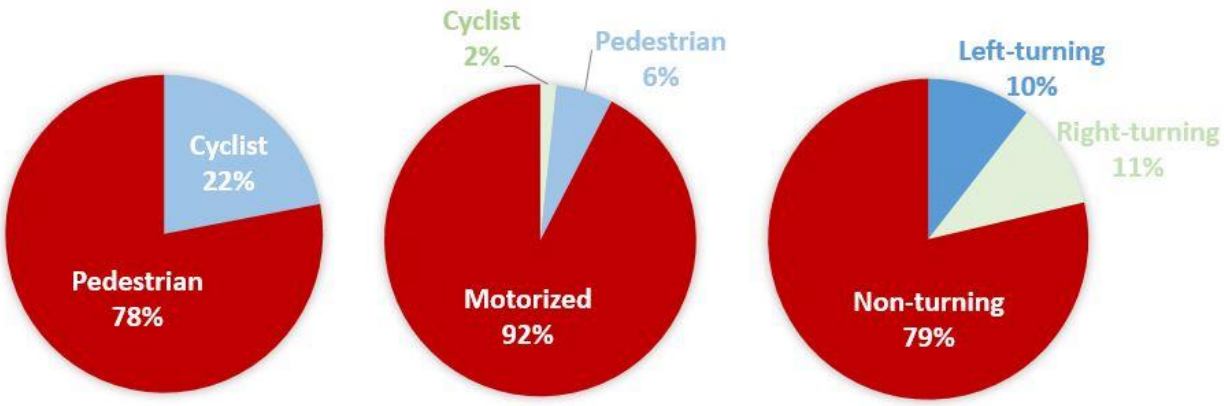


Figure 4-10 Distribution of motorized and non-motorized traffic by type.

4.3.3 Prior specification and model computation – multivariate settings

For the covariates coefficients $\boldsymbol{\beta}$ and the mean of the baseline distribution μ_0, we specified normally distributed vague priors, normal(0, 100). For σ_0, the standard deviation of the baseline density, uniform(0, 10) priors were specified, a relatively vague specification. As it is common in multivariate settings, we set a Wishart distribution for the inverse of covariance matrix Σ^{-1} (Tunaru, 2002) with K=2 (for two correlated outcomes) and a 2x2 scale matrix R with $R[1,1]=R[2,2]=0.01$ and $R[1,2]=R[2,1]=0$, resulting in a non-informative specification.

Based on the maximum number of components (i.e., 50) and discussion provided in Section 3.2, we initially used a uniform prior for α with a lower limit of 0.3 and an upper limit of 10 for the highway 401 dataset. Such values allow small and large values of α while avoiding problems relating to the calculation of p_n as discussed previously. We also verified the sensitivity to the prior choice for the Dirichlet precision parameter by choosing an upper limit of 100 for the highway 401 dataset. The results in this regard are reported in Section 4.3.3.

We used a gamma(0.05, 0.05) prior with a truncation at 0.3 for α, the precision parameter of the Dirichlet process. The truncation is applied to prevent potential difficulties in the computation due to small probability values in the stick-breaking algorithm (Ohlssen et al., 2007). As indicated in Section 3.2, the Dirichlet precision parameter controls the level of similarity between the mixing density G and its prior G_0, referred to as the baseline density. Note that a sensitivity analysis was carried out to verify the robustness of the results with respect to the assumed priors for the precision parameter and the baseline distribution. Since the selected priors have relatively large variances, we did not observe any reportable variation in the results. In fact, the ratio of posterior estimates to prior variances were smaller than 0.05.

WinBUGS was used to generate MCMC samples for the Bayesian posterior inference. For the highway segment data, two chains with 80,000 iterations were considered among which the first 20,000 were discarded for burn-in and model convergence, so 120,000 samples were utilized for inference. This was sufficient for low Monte Carlo errors. History plots, trace plots, and the Gelman-Rubin statistic were used to ensure that convergence was reached.

One key advantage of the proposed flexible multivariate model is its computational simplicity in WinBUGS. While Bayesian nonparametric models having an infinite number of parameters are often intractable computationally, our model is formulated in a way that can be estimated employing standard MCMC algorithms. We simplify the Bayesian nonparametric model to a regular finite mixture model. For the pedestrian/cyclist dataset, two chains were used in our MCMC simulations each containing 100,000 iterations. We discarded the first 40,000 iterations to meet convergence requirements, so that posterior inferences were drawn from 120,000 samples. This was sufficient for low Monte Carlo errors and for verifying the Gelman-Rubin convergence statistic.

4.3.4 Results and discussions – multivariate settings

For the highway injury-severity dataset, the results indicate that the Dirichlet precision parameter has a posterior mean that is away from the lower limit of 0 (Fig. 4-11a). Therefore, the Dirichlet process mixture model does not appear to provide strong evidence for an underlying non-normal multivariate density for this dataset. To examine the sensitivity to the prior choice of a uniform(0.3, 10) distribution for the precision parameter, we also analyzed the data using a uniform(0.3, 100) distribution. A kernel density plot of the precision parameter with different priors is shown in Fig. 4-11. Although the value of α varies, it remains bounded away from zero.

The log pseudo marginal likelihoods and coefficient estimates obtained from the Dirichlet process mixture multivariate model are similar to those from the standard multivariate Poisson-lognormal model (Table 4-15). This indicates that employing the flexible model, although it is not needed, does not penalize the model in terms of predictive performance. In other words, the flexible model approximates the standard model. The results show that as traffic flow and segment length increase, crash frequencies of both type of severity increase. In contrast, an increase in median shoulder width or median width results in decreased injury-fatal crashes. The chance of property-damage-only crashes is higher among segments with a narrow median shoulder while the chance of injury-fatal crashes is lower among segments with paved outside shoulders. We also found that the degree of horizontal curve per km is negatively

associated with both injury-fatal and property damage only crashes, and that two crash outcomes are highly correlated with a correlation of 0.876.

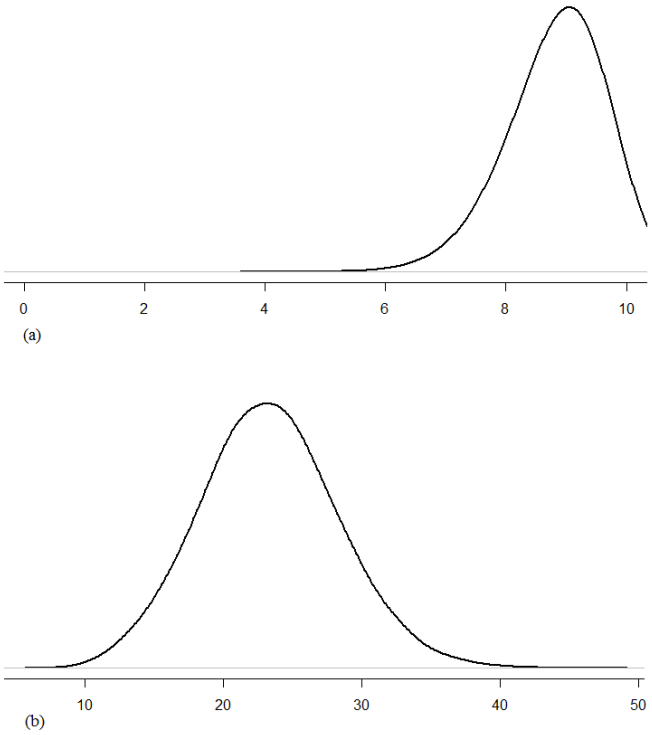


Figure 4-11 Kernel density plot of the precision parameter, highway 401 dataset: (a) α ~ uniform(0.3, 10); and (b) α ~ uniform(0.3, 100).

For the pedestrian/cyclist data, with respect to the dependence structure, Fig. 4-12 shows a kernel density plot of the correlation obtained from both the standard and the flexible model. Recall that the flexible model is obtained by allowing the correlation structure to vary across latent subpopulations in data. This can be seen in the form of the correlation density that is the distribution of the correlation amongst intersections while accounting for probabilities (weights) of different latent components. It can be implied from Fig. 4-12b that the density of correlation is quite spread in the range -1 to +1 while being mainly concentrated on the positive side of the graph. Note that a similar pattern in the correlation structure was observed by Jara et al. (2007) in modeling correlated binary outcomes.

Table 4-15 Posterior inference for the highway 401 dataset

	Posterior Mean	Std. Dev.	Credible intervals 2.50%	97.50%
Multivariate mixture of points Poisson-lognormal model				
Property damage only crashes				
Intercept mean	-10.950	0.291	-11.500	-10.460
Intercept variance	0.703	0.127	0.513	1.006
ln(AADT)	1.267	0.026	1.223	1.316
ln(length)	0.754	0.028	0.701	0.810
Average horizontal curve degree curvature per km	-0.146	0.016	-0.176	-0.113
Narrow median (inside) shoulder	0.160	0.042	0.080	0.249
Baseline mean	-10.980	0.390	-11.760	-10.270
Injury-fatal crashes				
Intercept mean	-12.050	0.415	-12.800	-11.21
Intercept variance	0.322	0.057	0.235	0.457
ln(AADT)	1.291	0.034	1.220	1.354
ln(length)	0.803	0.033	0.739	0.869
Average horizontal curve degree curvature per km	-0.072	0.019	-0.108	-0.034
Median (inside) shoulder width	-0.079	0.019	-0.116	-0.042
ln(median width)	-0.077	0.034	-0.145	-0.011
Paved outside shoulder	-0.245	0.068	-0.377	-0.113
Baseline mean	-12.090	0.448	-12.910	-11.200
Dirichlet precision parameter α	8.752	1.035	6.184	9.961
Correlation between outcomes	0.943	0.030	0.866	0.983
Log pseudo marginal likelihood	-2,022.710	-	-	-
Standard multivariate Poisson-lognormal model				
Property damage only crashes				
Intercept mean	-10.720	0.403	-11.290	-9.870
ln(AADT)	1.247	0.036	1.172	1.298
ln(length)	0.748	0.049	0.650	0.842
Average horizontal curve degree curvature per km	-0.146	0.024	-0.193	-0.100
Narrow median shoulder	0.157	0.073	0.014	0.298
Injury-fatal crashes				
Intercept mean	-11.280	0.506	-12.420	-10.400
ln(AADT)	1.232	0.043	1.158	1.327
ln(length)	0.793	0.045	0.706	0.881
Average horizontal curve degree curvature per km	-0.073	0.023	-0.117	-0.028
Median (inside) shoulder width	-0.084	0.024	-0.131	-0.037
ln(median width)	-0.132	0.039	-0.209	-0.054
Paved outside shoulder	-0.238	0.072	-0.380	-0.097
Correlation between outcomes	0.876	0.022	0.829	0.914
Log pseudo marginal likelihood	-2,021.390	-	-	-

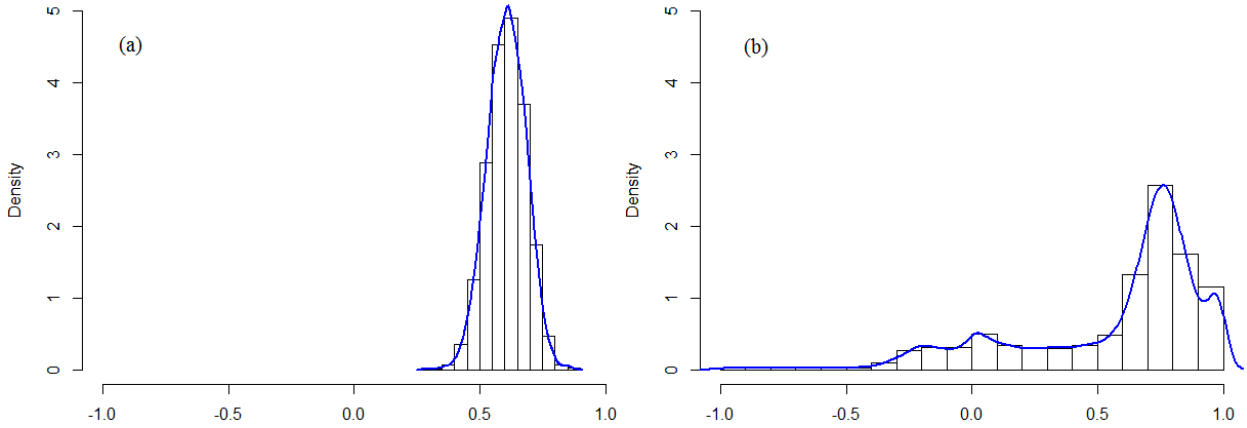


Figure 4-12 Histogram and kernel density plot for the estimated correlation between pedestrian and cyclist injury counts: (a) standard multivariate model and (b) flexible multivariate model.

The posterior estimate of the mean of the correlation for the entire dataset is 0.57 (0.09, 0.94) based on the flexible multivariate model. The posterior estimate of the correlation under the standard multivariate model with a unimodal density is 0.61 (0.38, 0.81). Based on the standard multivariate model, Fig. 4-12a implies an erroneously narrower interval (compared to the flexible model) around the correlation mean estimate, which does not reflect the reality of the data. It can also be inferred from Fig. 4-12 that the correlation is slightly overestimated under the standard multivariate model. As discussed in Jara et al. (2007), this could be associated with the fact that, although the standard model adjusts for the effect of observed factors, it does not accommodate unobserved confounders that lead to the formation of latent subpopulations in the data.

In addition, the flexible mixture of multivariate normals yields a right skewed posterior distribution for the Dirichlet precision parameter with a peak at 0.936 (0.314, 2.53), being away from zero (Fig. 4-13). The above findings relating to both the estimated precision parameter and the shape of the correlation density suggest that the true underlying dependence structure is away from the homogeneous multivariate normal density. In fact, the flexible model finds 5 clusters in the data resulting in a superior model-fitting compared to the standard multivariate model, a log pseudo marginal likelihood of -1,437 versus -1,457. This results in a log pseudo Bayes factor of 20, which provides strong support for the flexible multivariate model.

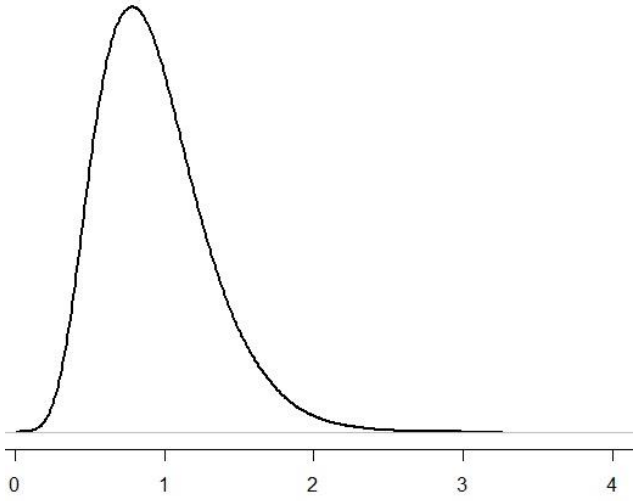


Figure 4-13 Kernel posterior density plot of the Dirichlet precision parameter, pedestrian/cyclist data

The lack of homogeneity in our dataset is intuitively plausible. This is because safety mechanisms that influence the magnitude of correlation between pedestrian and cyclist injury frequencies may vary from one intersection to another. Such variation may, for example, be caused by differences in geometric and operational attributes of intersections and variations in urban patterns relating to motorist and non-motorist activity and behavior amongst intersections. This study therefore points out the importance of considering the presence of subpopulations in data and heterogeneity in correlation structure when modeling correlated crash types or injury severities.

One aspect of this research was to identify contributing factors that affect injury frequencies of pedestrians and cyclists simultaneously at signalized intersections in urban settings. Tables 4-16 and 4-17 provide a summary of the results. While significant differences were observed in the correlation structure including the intercept structure, covariate estimates including their Bayesian intervals were found to be similar. However, marginal effects displayed real differences in covariate estimates.

Based on the covariate estimates, we found that as the numbers of pedestrians and cyclists increase, their injury counts increase. Both left-turning and right-turning motorized flow were positively associated with cyclists' injury counts, whereas only left-turning flow was found to have a positive effect on pedestrians' injury counts. Therefore, if the major scope is to improve pedestrian safety, for example, particular

attention should be given to intersections with left-turning flow, implementing countermeasures that could help protect pedestrians against left-turning motorized traffic. Non-turning flow was not found to have an important effect on neither pedestrian safety nor cyclist safety. The effect of right-turning flow was greater than that of the left-turning flow for cyclists.

We also found that the presence of bus stop within a range of 50 m is highly correlated with pedestrian and cyclist injury counts at signalized intersections, again suggesting attention to improving safety at these intersections. The area of commercial land use and employment are also positively associated with pedestrian injury counts, whereas having a dedicated pedestrian crossing light is negatively correlated with pedestrian safety. Employment and the length of cycling facilities within a range of 400 m around intersections were positively correlated with cyclist injury counts. The data used in this study do not distinguish between separated and non-separated cycling facilities precluding a more detailed and informative examination of their possibly distinct effects.

Table 4-16 Posterior inference for active modes, standard multivariate model

Crash type	Variables	Mean	Std. Dev.	2.50%	97.50%
Pedestrian	ln(pedestrian counts)	0.299	0.034	0.232	0.366
	ln(left-turning motorized volume)	0.271	0.056	0.166	0.385
	Presence of bus stop	0.771	0.149	0.481	1.065
	Employment	0.710	0.219	0.281	1.138
	Commercial area	0.013	0.004	0.005	0.022
	Dedicated traffic light for pedestrians	-0.343	0.143	-0.626	-0.063
	Intercept	-5.377	0.523	-6.461	-4.407
Cyclist	ln(cyclist counts)	0.383	0.063	0.261	0.511
	ln(left-turning motorized volume)	0.174	0.063	0.055	0.304
	ln(right-turning motorized volume)	0.247	0.074	0.110	0.398
	Presence of bus stop	0.694	0.169	0.367	1.032
	Employment	0.671	0.239	0.200	1.143
	Length of cycling facilities (km)	0.365	0.112	0.145	0.586
	Intercept	-7.343	0.732	-8.833	-5.950

Table 4-17 Posterior inference for active modes, mixture of multivariate normals

Crash type	Variables	Mean	Std. Dev.	2.50%	97.50%
Pedestrian	ln(pedestrian counts)	0.295	0.036	0.230	0.370
	ln(left-turning motorized volume)	0.301	0.046	0.211	0.390
	Presence of bus stop	0.760	0.149	0.477	1.059
	Employment	0.755	0.212	0.332	1.161
	Commercial area	0.014	0.004	0.005	0.022
	Dedicated traffic light for pedestrians	-0.326	0.150	-0.617	-0.030
	Intercept mean	-5.674	0.421	-6.444	-4.746
	Baseline mean	-6.178	1.756	-9.657	-2.225
	Baseline standard deviation	2.522	1.905	0.564	8.138
Cyclist	ln(cyclist counts)	0.408	0.065	0.276	0.530
	ln(left-turning motorized volume)	0.192	0.065	0.073	0.324
	ln(right-turning motorized volume)	0.305	0.077	0.160	0.459
	Presence of bus stop	0.661	0.170	0.330	0.999
	Employment	0.721	0.228	0.285	1.177
	Length of cycling facilities (km)	0.412	0.109	0.198	0.628
	Intercept mean	-8.731	0.809	-9.794	-6.954
	Baseline mean	-10.410	2.918	-15.940	-4.169
	Baseline standard deviation	4.935	2.061	1.740	9.404

4.3.5 An example of policy implications

We used marginal effects to highlight an important advantage of our model from a practical engineering perspective. Specifically, to interpret the impact of variables on non-motorist injury counts, we computed marginal effects that allow estimating the effect of one unit change in l^{th} independent variable on the outcome of interest (Washington et al., 2011). Given the notation in previous sections, marginal effects can be obtained as in Eq. 4-2 (in its simplest form for the Poisson regression).

$$\frac{\partial E[Y_i|x_i]}{x_{il}} = \beta_l EXP(\beta x_i) \quad (4\text{-}2)$$

The average marginal effects of the covariates over all observations are reported in Table 4-18 for both models. These are computed in WinBUGS given the structure of the models. As described in Section 4.2.1, the standard multivariate model is not an

appropriate model to analyze the dataset used in this study. Nevertheless, we report its marginal effects for comparison purposes. From the estimated marginal effects, we infer that the influence of exploratory variables (built environment, etc.), particularly, motorized and non-motorized traffic is significantly larger on pedestrian safety compared to cyclist safety. For example, based on the flexible mixture of multivariate normals, we infer from Table 4-18 that intersections that are equipped with dedicated traffic lights for pedestrian crossing have an average expected pedestrian injury frequency that is 0.154 lower than other intersections. Intersections that are in proximity to bus stops have an average expected pedestrian and cyclist injury counts that are, respectively, 0.358 and 0.101 higher. Marginal effects also indicate that an increase in the length of cycling facilities, on average, leads to 0.063 increase per kilometer in the expected bicyclist injury counts. The marginal effect of the log of cyclist counts is 0.063 that translates to a 0.142 increase in cyclist injury counts for every thousand increase in cyclist counts.

Table 4-18 Average marginal effects for pedestrian/cyclist data

Crash type	**Variables**	**Flexible multivariate model**	**Standard multivariate model**
Pedestrian	ln(pedestrian counts)	0.139	0.149
	ln(left-turning motorized volume)	0.142	0.152
	Presence of bus stop	0.358	0.384
	Employment	0.356	0.381
	Commercial area	0.007	0.007
	Dedicated traffic light for pedestrians	-0.154	-0.165
Cyclist	ln(cyclist counts)	0.063	0.123
	ln(left-turning motorized volume)	0.029	0.058
	ln(right-turning motorized volume)	0.047	0.092
	Presence of bus stop	0.101	0.199
	Employment	0.111	0.217
	Length of cycling facilities (km)	0.063	0.124

While the differences in coefficient estimates between the two models seem to be small (see Table 4-16 and Table 4-17), after computing their marginal effects, differences

become more apparent. And, we see that the marginal effects obtained from the standard model are almost twice those obtained from the proposed flexible model for cyclists. Therefore, when modeling correlated outcomes, at least, the marginal effects of some outcomes (here, bicyclists injuries) may be poorly estimated under the standard multivariate model that erroneously assumes homogeneity in the correlation structure. This may affect the accuracy of information provided to decision makers, and consequently, the countermeasure selection process. A biased interpretation of the impact of covariates on safety may result in an ineffective allocation of funds since an expected improvement in safety conditions may not be achieved.

4.3.6 Summary of multivariate modeling

Section 4.3 contributes to the crash literature in presenting two flexible Bayesian multivariate mixture models based on a Dirichlet process. The models allow for a heterogeneous correlation structure with respect to the location (mean) and/or the covariance matrix of the dependence component. The proposed models are in the form of multivariate latent class models that account for unobserved heterogeneity while accommodating correlation among outcomes. As indicated in Mannering et al. (2016), this is an appealing way to account for unobserved heterogeneity in multivariate settings.

In our models, the number of latent subpopulations is itself a stochastic parameter to be inferred from the data using a rigorous mathematical algorithm, whereas this number must be prespecified – usually without any sound justification – in conventional latent class models. The models allow the number of parameters, in terms of the hidden or latent structure of the data, to grow according to data complexity. In addition, while a few sup-populations are often assumed in traditional latent class models, our models can accommodate a large number of clusters.

The models' high flexibility better captures complex data structures when modeling correlated outcomes such as crash types or injury-severity levels. Thus, it helps avoid inconsistencies with real crash data generation mechanisms. In this study, we applied the proposed models to correlated counts, but it can be also employed to model correlated outcomes of different types such as binary, continuous, etc. Another

advantage of the proposed model is its ability to accommodate outliers without compromising the results (Jara et al., 2007).

We first used a highway segment data from Ontario to jointly model different crash types by severity. We showed how to extend the standard multivariate Poisson-lognormal model to a more flexible multivariate mixture of points model, thereby accounting for dependence nonparametrically. This model relaxes the homogeneity assumption of the correlation structure with respect to the mean. We investigated the multivariate normal distribution assumption, and found it is reasonable, at least for the highway 401 data. To add further flexibility to the model, we relaxed the homogeneity assumption in both the mean and the covariance structure of the dependence structure.

We applied the latter model to a pedestrian/cyclist crash dataset including 647 signalized intersections in Montreal. We modeled pedestrian and cyclist injury counts simultaneously using a multivariate modeling framework that captures the effects of unobserved factors in addition to the effects of the exogenous variables. A non-restrictive joint modeling of pedestrian and bicyclist injuries, thus, improves our collective understanding of vulnerable road users' safety. This understanding is further improved by using a micro-level (intersection) analysis that takes advantage of direct exposure measures. In modeling non-motorist safety, many studies lacked detailed motorist and non-motorist exposure information and used proxy measures instead. In contrast, the current study takes advantage of detailed exposure measures: the pedestrian and cyclist counts together with vehicle flow at a disaggregate level.

We show how a limiting multivariate model structure, the standard model with multivariate normal density for modeling correlation, compromises goodness-of-fit and leads to spurious interpretation of variables in the model. Indeed, we found that the proposed flexible mixture of multivariate normals substantially improves the predictive performance of the model. It also prevented overestimating the impact of variables influencing non-motorist injury frequencies, especially for bicyclists. Providing more accurate estimates, the empirical findings of this study can be useful for policy decisions such as planning safety improvement programs for pedestrians and cyclists at signalized intersections, particularly, in urban areas. An enhanced safety condition would then help foster active modes of transport.

CHAPTER 5

CONCLUSIONS

Crash data are often heterogeneous due to various unobserved factors that have a bearing on crash frequencies and injury-severities. Therefore, many road safety studies have focused on addressing the unobserved heterogeneity problem. Based on the major limitations to the conventional models (random effects, random parameter, and latent class models) commonly used in addressing unobserved heterogeneity in the crash literature, this thesis introduces a class of flexible latent class models that are rooted in the Bayesian nonparametric literature. As we use a truncated Dirichlet process, our models reduce to the form of finite mixture models that can be estimated using standard MCMC algorithms. The proposed models infer the number of latent clusters from crash data as part of their estimation procedures. Our approach is extremely rich, offering a number of advantages that we have clearly shown in this dissertation using different univariate, multilevel, and multivariate crash datasets and policy examples. A summary of our main contributions is provided as follows.

5.1 Major Contributions

This thesis contributes to the road safety literature methodologically and empirically:

- To add flexibility to the standard generalized linear models, this thesis employed a Dirichlet process mixture over the vector of intercepts to tackle heterogeneity with respect to the location of the mean by allowing multimodality as in finite mixture models. We retain the linear form for model coefficients, which in turn retain their usual interpretations.
    - The resulting model (Dirichlet process mixture of generalized linear models) offers considerable promise in addressing unobserved heterogeneity and over-dispersion in analyzing crash datasets, including those characterized by the low mean value problem and excess zero counts such as railway grade crossing data. We also showed that our model is more reliable in identifying high-crash locations.
- To better circumvent the cross-group unobserved heterogeneity in multilevel settings, this thesis incorporated a latent structure into multilevel models allowing the analyst to detect latent subpopulations among groups of observations. This approach is mainly appealing when apparent clustering exists among groups, and separation between groups matters (e.g., sites nested within different geographic areas). The method also appears to be particularly useful in the presence of outlier groupings.
    - This study indicates that outliers and latent clusters could exist among different Canadian regions in terms of railway grade crossing safety measures. This indicates that Canadian regions may widely differ according to their safety performances. The method therefore allows for monitoring the performance of different regions in terms of specific safety measures, assisting to draw Canada-wide safety policy insights.
    - This research confirms the presence of spatial dependencies among railway grade crossings nested within the same province or municipality using a latent class multilevel model. Such dependencies are expected due to unmeasured or unknown regional similarities, for example, in climate and traffic regulations.

- In analyzing the railway grade crossing crash data, this thesis addressed the omitted variables problem, a major concern in road safety studies, especially grade crossings (Jovanis et al., 2011; Mannering and Bhat, 2014; Wu et al., 2015). We have dealt with this problem as follows. First, we have attempted to include all known important variables that may significantly affect crash frequencies at grade crossings based on literature. Second, it is presumed that, since there could be other unknown or unmeasured spatial factors that may have a bearing on crash frequencies, our multilevel approach at the regional level is expected to capture some of these unknown/unmeasured variables. Note that spatially related attributes are recognized as an important source of omission as discussed in Mitra and Washington (2012). Lastly, we have employed a Dirichlet process mixture of generalized linear model that includes multiple latent components to handle unobserved heterogeneity. This is expected to help minimize undesirable consequences of the omitted variables problem. Note that, as discussed in road safety literature, some statistical models such as random parameter models can minimize the bias caused by the omission of variables as they account for unobserved heterogeneity (Anastasopoulos and Mannering, 2009; Mitra and Washington, 2012; Chen and Tarko, 2014).

- In the joint analysis of correlated outcomes, this research allowed departures from restrictive homogeneous dependency structures such as that of the multivariate normal density. We added flexibility to the standard multivariate model by allowing the location or both the location and the covariance matrix to vary across observations. Doing so, we derived two flexible multivariate models: the multivariate mixture of points model and the multivariate mixture of normal densities model. Both models are in the form of multivariate latent class models that, according to the crash literature, appear to be immensely appealing in overcoming unobserved heterogeneity in modeling correlated outcomes.

- We analyzed pedestrian and cyclist injuries simultaneously based on the hypothesis that crash mechanisms of pedestrians and cyclists have some similarities, being non-motorized modes of transport. We showed that correlation exists between the two outcomes. This correlation is better captured when employing the flexible multivariate mixture of multivariate normal densities due to a better model specification. Therefore, a more accurate estimation of safety was obtained for vulnerable road users at intersections in an urban setting.

5.2 Future Research

We identified the following future research directions considering both methodological and empirical aspects:

- To retain the interpretability of the Dirichlet process mixture of generalized linear model, this study selects a Dirichlet process prior for the random intercepts. The method could be applied to extend the flexibility to other model coefficients associated with explanatory variables. Obviously, a number of challenges may emerge; for instance, such extension could be computationally intensive.
- In this research, we used a mixture of points approach to add flexibility in modeling univariate and multilevel settings. It would be interesting to investigate the use of mixture of normal densities as well.
- The proposed flexible multilevel model, while flexible in accounting for spatial dependencies for sites located in similar regions, does not account for the neighborhood effects among sites, for example, as in conditional autoregressive models. Similarly, the conditional autoregressive model cannot discover latent clusters in data as our model does. Therefore, it would be interesting to explore the feasibility of incorporating spatial autocorrelation into our flexible multilevel model, thus addressing this limitation of conditional autoregressive models.

- In terms of empirical aspects, it would be an interesting line of future research to employ the proposed models for estimating the effectiveness of safety improvement programs. As discussed previously, we expect more reliable estimates when applying our approach due to its enhanced model specification that better captures the underlying structure of crash data.

BIBLIOGRAPHY

Aguero-Valverde, J., Jovanis, P., 2006. Spatial analysis of fatal and injury crashes in Pennsylvania. Accident Analysis and Prevention 38, 618–625.

Aguero-Valverde, J., Jovanis, P., 2008. Analysis of road crash frequency with spatial models. Transportation Research Record 2061, 55–63.

Aguero-Valverde, J., Jovanis, P., 2009. Bayesian multivariate Poisson lognormal models for crash severity modeling and site ranking. Transportation Research Record: Journal of the Transportation Research Board 2136, 82-91.

Aguero-Valverde, J., 2013. Multivariate spatial models of excess crash frequency at area level: case of Costa Rica. Accident Analysis and Prevention 59, 365-373.

Aldred, R., 2016. Cycling near misses: their frequency, impact, and prevention. Transportation Research Part A 90, 69-83.

Amoh-Gyimah, R., Saberi, M., Sarvi, M., 2016. Macroscopic modeling of pedestrian and bicycle crashes: a cross-comparison of estimation methods. Accident Analysis and Prevention 93, 147-159.

Anastasopoulos, P.C., 2016. Random parameters multivariate tobit and zero-inflated count data models: addressing unobserved and zero-state heterogeneity in accident injury-severity rate and frequency analysis. Analytic Methods in Accident Research 11, 17-32.

Anastasopoulos, P., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41 (1), 153-9.

Anastasopoulos, P.C., Mannering, F.L., 2016. The effect of speed limits on drivers' choice of speed: a random parameters seemingly unrelated equations approach. Analytic Methods in Accident Research 10, 1-11.

Anastasopoulos, P.C., Shankar, V.N., Haddock, J.E., Mannering, F.L., 2012. A multivariate tobit analysis of highway accident-injury-severity rates. Accident Analysis and Prevention 45, 110-119.

Antoniak, C.E., 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics 2 (6), 1152-1174.

Anwaar, A., Anastasopoulos, P., Ong, G.P., Labi, S., Islam, M.B., 2012. Factors affecting highway safety, health care services, and motorization—an exploratory empirical analysis using aggregate data. Journal of Transportation Safety & Security 4 (2), 94-115.

Austin, R., Carson, J., 2002. An alternative accident prediction model for highway-rail interfaces. Accident Analysis and Prevention 34, 31–42.

Barua, S., El-Basyouny, K., Islam, M.T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. Analytic Methods in Accident Research 9, 1-15.

Basu, S., Chib, S., 2003. Marginal likelihood and Bayes factors for Dirichlet process mixture models. Journal of the American Statistical Association 98 (461), 224-235.

Behnood, A., Mannering, F.L., 2016. An empirical assessment of the effects of economic recessions on pedestrian-injury crashes using mixed and latent-class models. Analytic Methods in Accident Research 12, 1-17.

Behnood, A., Roshandeh, A.M., Mannering, F.L., 2014. Latent class analysis of the effects of age, gender, and alcohol consumption on driver-injury severities. Analytic Methods in Accident Research 3-4, 56-91.

Bhat, C.R., Dubey, S.K., 2014. A new estimation approach to integrate latent psychological constructs in choice modeling. Transportation Research Part B 67 (1), 68-85.

Braun, L.M., Rodriguez, D.A., Cole-Hunter, T., Ambros, A., Donaire-Gonzalez, D., Jerrett, M., Mendez, M.A., Nieuwenhuijsen, M.J., De Nazelle, A., 2016. Short-term planning and policy interventions to promote cycling in urban centers: findings from a commute mode choice analysis in Barcelona, Spain. Transportation Research Part A 89 (11), 164-183.

Brooks, S.P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7 (4), 434-455.

Buddhavarapu, P., Scott, J.G., Prozzi, J.A., 2016. Modeling unobserved heterogeneity using finite mixture random parameters for spatially correlated discrete count data. Transportation Research Part B 91, 492-510.

Bush, C.A., MacEachern, S.N., 1996. A semi-parametric Bayesian model for randomized block designs. Biometrika 83 (2), 275-285.

Canadian Automobile Association, 2016. http://bikesafety.caa.ca/cyclists/bicycle-statistics.php (Accessed July 2016).

Cameron, A.C., Li, T., Trivedi, P.K., Zimmer, D.M., 2004. Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. The Econometrics Journal 7 (2), 566-584.

Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. New York, Cambridge University Press.

Carlin, B.P., Louis, T.A., 2008. Bayesian Methods for Data Analysis. Chapman & Hall/CRC. Taylor & Francis Group. Boca Raton, Florida.

Cerwick, D.M., Gkritza, K., Shaheed, M.S., Hans, Z., 2014. A comparison of the mixed logit and latent class methods for crash severity analysis. Analytic Methods in Accident Research 3-4, 11-27.

Chadwick, S.G., Zhou, N., Saat, M.R., Highway-rail grade crossing safety challenges for shared operations of high-speed passenger and heavy freight rail in U.S. safety science 68, 128-137.

Chataway, E.S., Kaplan, S., Nielsen, T.A.S., Prato, C.G., 2014. Safety perceptions and reported behavior related to cycling in mixed traffic: a comparison between Brisbane and Copenhagen. Transportation Research Part F 23, 32-43.

Chaudhary, M., Hellman, A., Ngamdung, T., John, A., 2011. Volpe National Transportation Systems Center (U.S.). Railroad Right-of-Way Incident Analysis Research. Federal Railroad Administration Office of Railroad Policy and Development, Washington, DC.

Chen, E., Tarko, A., 2014. Modeling safety of highway work zones with random parameters and random effects models. Analytic Methods in Accident Research 1, 86-95.

Cho, G., Rodríguez, D.A., Khattak, A.J., 2009. The role of the built environment in explaining relationships between perceived and actual pedestrian and bicyclist safety. Accident Analysis and Prevention 41 (4), 692-702.

Clifton, K.J., Burnier, C.V., Akar, G., 2009. Severity of injury resulting from pedestrian–vehicle crashes: what can we learn from examining the built environment? Transportation Research Part D 14 (6), 425-436.

Coruh, E., Bilgic, A., Tortum, A., 2015. Accident analysis with aggregated data: the random parameters negative binomial panel count data model. Analytic Methods in Accident Research 7, 37-49.

Cruzado, I.U., Donnell, E.., 2010. Factors affecting driver speed choice along two-lane rural highway transition zones. Journal of Transportation Engineering 136 (8), 755-764.

Davis, G.A., 2004. Possible aggregation biases in road safety research and a mechanism approach to accident modeling. Accident Analysis and Prevention 36 (6), 1119-1127.

Daziano, R.A., Miranda-Moreno, L.F., Heydari, S., 2013. Computational Bayesian statistics in transportation modeling: from road safety analysis to discrete choice. Transport Reviews 33, 570–592.

de Hartog, J.J., Boogaard, H., Nijland, H., Hoek, G., 2010. Do the health benefits of cycling outweigh the risks? Environmental Health Perspectives 118 (8), 1109-1116.

Dhavala, S.S., Mallick, B.K., Carroll, R.J., Datta, S., Khare, S., Lawhon, S.D., Adams, L.G., 2010. Bayesian modeling of MPSS data: gene expression analysis of bovine salmonella infection. Journal of the American Statistical Association 105 (491), 956-967.

Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2014. Multivariate random-parameters zero-inflated negative binomial regression model: an application to estimate crash frequencies at intersections. Accident Analysis and Prevention 70, 320-329.

Dorazio, R.M., 2009. On selecting a prior for the precision parameter of Dirichlet process mixture models. Journal of Statistical Planning and Inference 139 (9), 3384-3390.

Dupont, E., Papadimitriou, E., Martensen, H., Yannis, G., 2013. Multilevel analysis in road safety research. Accident Analysis and Prevention 60, 402-411.

El-Basyouny, K., Sayed, T., 2006. Comparison of two negative binomial regression techniques in developing. Accident Prediction Models. Transportation Research Record: Journal of the Transportation Research Board 1950, 9-16.

El-Basyouny, K., Sayed, T., 2009a. Collision prediction models using multivariate Poisson-lognormal regression. Accident Analysis and Prevention 41 (4), 820-828.

El-Basyouny, K., Sayed, T., 2009b. Accident predicyion models with random corridor parameters. Accident Analysis and Prevention 41 (5), 1118-1123.

El-Basyouny, K., Sayed, T., 2012. Measuring safety treatment effects using full Bayes non-linear safety performance intervention functions. Accident Analysis and Prevention 45, 152–163.

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

Eluru, N., Bhat, C.R., Hensher, D.A., 2008. A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. Accident Analysis and Prevention 40 (3), 1033-1054.

Escobar, M.D., 1994. Estimating normal means with a Dirichlet process prior. Journal of the American Statistical Association 89 (425), 268-277.

Escobar, M.D., West, M., 1998. Computing nonparametric hierarchical models. In: Dey, D., Müller, P., Sinha, D. eds. Practical nonparametric and semiparametric Bayesian statistics. Springer New York, New York, NY, 1-22.

Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1 (2), 209-230.

Forsyth, A., Oakes, J.M., 2015. Cycling, the built environment, and health: results of a midwestern study. International Journal of Sustainable Transportation 9 (1), 49-58.

Freedman, D., 1963. On the asymptotic behavior of Bayes estimates in the discrete case. Annals of Mathematical Statistics 34 (4), 1386-1403.

Fuzhong, L., Fisher, K.J., Ross, C.B., Mark, B., 2005. Multilevel modelling of built environment characteristics related to neighbourhood walking activity in older adults. Journal of Epidemiology & Community Health 59 (7), 558-564.

Geedipally, S.R., Lord, D., Dhavala, S.S., 2014. A caution about using deviance information criterion while modelling traffic crashes. Safety Science 62, 495-498.

Gelfand, A., 1996. Model determination using sampling-based methods, in W. Gilks, S. Richardson, and D. Spiegelhalter, eds., Markov Chain Monte Carlo in Practice, Chapman & Hall, Suffolk.

Gelfand, A.E., Dey, D.K., Chang, H., 1992. "Model determination using predictive distributions with implementation via sampling-based methods (with discussion)", In Bayesian Statistics 4, 147-169. Oxford: Clarendon.

Gelfand, A., Kottas, A., 2002. A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. Journal of Computational and Graphical Statistics 11 (2), 289-305.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2003. Bayesian data analysis, 2nd ed., Chapman & Hall/CRC Press, New York, USA.

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

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Statistical Science 7 (4), 457-472.

Gershman, S.J., Blei, D.M., 2012. A tutorial on Bayesian nonparametric models. Journal of Mathematical Psychology 56 (1), 1-12.

Ghosh, P., Gill, P., Muthukumarana, S., Swarts, T., 2010. A semiparametric Bayesian approach to network modelling using Dirichlet process prior distributions. Australian & New Zealand Journal of Statistics 52 (3), 289-302.

Greene, W.H., Hensher, D.A., 2003. A latent class model for discrete choice analysis: contrasts with mixed logit. Transportation Research Part B 37 (8), 681-698.

Gustavson, J., Svensson, A., 1976. A Poisson regression model applied to classes of road accidents with small frequencies. Scandinavian Journal of Statistics 3, 49-60.

Hadayeghi, A., Shalaby, A.S., Persaud, B.N., 2010. Development of planning level transportation safety tools using geographically weighted Poisson regression. Accident Analysis and Prevention 42 (2), 676-688.

Hannah, L.A., Blei, D.M., Powell, W.B., 2011. Dirichlet process mixtures of generalized linear models. Journal of Machine Learning Research 1, 1-33.

Hauer, E., 1977. Observational before-after studies in road safety. Elsevier Science Ltd. Oxford, United Kingdom.

Helai, H., Chor, C.H., Haque, M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis and Prevention 40 (1), 45–54.

Heydari, S., Fu, L., 2015. Developing safety performance functions for railway grade crossings: a case study of Canada. Joint Rail Conference, San Jose, CA, USA.

Heydari, S., Miranda-Moreno, L.F., Amador, L., 2013. Does prior specification matter in hotspot identification and before-after studies? Transportation Research Record 2392, 31-39.

Heydari, S., Miranda-Moreno, L.F., Liping, F., 2014a. Speed limit reduction in urban areas: a before-after study using Bayesian generalized mixed linear models. Accident Analysis and Prevention 73, 252-261.

Heydari, S., Miranda-Moreno, L.F., Lord, D., Fu, L., 2014b. Bayesian methodology to estimate and update safety performance functions under limited data conditions: a sensitivity analysis. Accident Analysis and Prevention 64, 41-51.

Heydari, S., Fu, L., Joseph, L., Miranda-Moreno, L.F., 2016a. Bayesian nonparametric modeling in transportation safety studies: applications in univariate and multivariate settings. Analytic Methods in Accident Research 12, 18-34.

Heydari, S., Fu, L., Lord, D., Mallick, B.K., 2016b. Multilevel Dirichlet process mixture analysis of railway grade crossing crash data. Analytic Methods in Accident Research 9, 27-43.

Heydecker, B.G., Wu, J., 2001. Identification of sites for road accident remedial work by Bayesian statistical methods: an example of uncertain inference. Advances in Engineering Software 32, 859–869.

Hjort, N., Holmes, C., Müller, P., Walker, S.G., 2010. Bayesian nonparametrics: principles and practice. Cambridge University Press.

Hu, S.R., Lee, C.K., 2008. Analysis of accident risk at railroad grade crossing. In: Transportation Research Board 87th Annual Meeting, Washington DC, USA.

Huang, H., Abdel-Aty, M., 2010. Multilevel data and Bayesian analysis in traffic safety. Accident Analysis and Prevention 42 (6), 1556–1565.

Huang, H., Chin, H.C., Haque, M.M., 2008. Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis. Accident Analysis and Prevention 40 (1), 45–54.

Huang, H., Song, B., Xu, P., Zeng, Q., Lee, J., Abdel-Aty, M., 2016. Macro and micro models for zonal crash prediction with application in hot zones identification. Journal of Transport Geography 54, 248-256.

Ishwaran, H., 2000. Inference for the random effects in Bayesian generalized linear mixed models. ASA Proceedings of the Bayesian Statistical Science Section, 1–10.

Ishwaran, H., James, L.F., 2002. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association 96 (453), 161-173.

Islam, M.T., El-Basyouny, K., 2015. Multilevel models to analyze before-after speed data. Analytic Methods in Accident Research 8, 33-44.

Jacobsen, P.L., 2015. Safety in numbers: more walkers and bicyclists, safer walking and bicycling. Injury Prevention 21 (4), 271-275.

Jain, S., Neal, R.M., 2004. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. Journal of Computational and Graphical Statistics 13 (1), 158-182.

Jara, A., José García-Zattera, M., Lesaffre, E., 2007. A Dirichlet process mixture model for the analysis of correlated binary responses. Computational Statistics and Data Analysis 51 (11), 5402-5415.

Jones, A, Jørgensen, S., 2003. The use of multilevel models for the prediction of road accident outcomes. Accident Analysis and Prevention 35(1),59–69.

Joshua, S.C., Garber, N.J., 1990. Estimating truck accident rate and involvement using linear and Poisson regression models. Transportation Planning and Technology 15, 41-58.

Jovanis, P.P., Aguero-Valverde, J., Wu, K-F, Shankar, V., 2011. Analysis of naturalistic driving event data: omitted-variable bias and multilevel modeling approaches. Transportation Research Record: Journal of the Transportation Research Board 2236, 49-57.

Jung, S., Qin, X., Oh, C., 2016. Improving strategic policies for pedestrian safety enhancement using classification tree modeling. Transportation Research Part A 85, 53-64.

Karlaftis, M., Tarko, A., 1998. Heterogeneity considerations in accident modeling. Accident Analysis and Prevention 30, 425–433.

Kass, R.E., Raftery, A.E., 1995. Bayes factors. Journal of the American Statistical Association 90 (430), 773-795.

Khattak, A.J., Rodriguez, D., 2005. Travel behavior in neo-traditional neighborhood developments: a case study in USA. Transportation Research Part A 39 (6), 481-500.

Kim, D.G., Lee, Y., Washington, S., Choi, K., 2007. Modeling crash outcome probabilities at rural intersections: application of hierarchical binomial logistic models. Accident Analysis and Prevention 39 (1), 125-134.

Kleinman, K.P., Ibrahim, J.G., 1998. A semiparametric Bayesian approach to the random effects model. Biometrics 54 (3), 921-38.

Krizek, K.J., Johnson, P.J., 2006. Proximity to trails and retail: effects on urban cycling and walking. Journal of the American Planning Association 72 (1), 33-42.

Kuo, L. and Mallick, B.K., 1997. Bayesian semiparametric inference for the accelerated failure-time model. Canadian Journal of Statistics 25 (4), 457-472.

Leden, L., 2002. Pedestrian risk decrease with pedestrian flow. A case study based on data from signalized intersections in Hamilton, Ontario. Accident Analysis and Prevention 34 (4), 457-464.

Lee, C., Abdel-Aty, M., 2005. Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida. Accident Analysis and Prevention 37 (4), 775-786.

Lee, J., Abdel-Aty, M., Jiang, X., 2015. Multivariate crash modeling for motor vehicle and non-motorized modes at the macroscopic level. Accident Analysis and Prevention 78, 146-154.

Lee, J., Nam, D., Moon, D., 2004. A zero-inflated accident frequency model of highway-rail grade crossing. In: Proceedings of the Transportation Research Board Annual Meeting. Washington, DC.

Lenguerrand, E., Martin, J.L., Laumon, B., 2006. Modelling the hierarchical structure of road crash data—Application to severity analysis. Accident Analysis and Prevention 38 (1), 43-53.

Li, F., Fisher, K., Brownson, R., Bosworth, M., 2005. Multilevel modelling of built environment characteristics related to neighbourhood walking activity in older adults. Journal of Epidemiology and Community Health 59 (7), 558-564.

Lord, D., Bonneson, J.A., 2007. Development of accident modification factors for rural frontage road segments in Texas. Transportation Research Record 2023, 20–27.

Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A 44 (5), 291-305.

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

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis and Prevention 37, 35–46.

Lord, D., Washington, S.P., Ivan, J.N., 2007. Further notes on the application of zero-inflated models in highway safety. Accident Analysis and Prevention 39, 53-57.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing 10 (4), 325-337.

Lyon, C., Persaud, B., 2002. Pedestrian collision prediction models for urban intersections. Transportation Research Record: Journal of the Transportation Research Board 1818, 102-107.

Ma, J., Kockelman, K.M., Damien, P., 2008. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. Accident Analysis and Prevention 40 (3), 964-975.

Maher, M.J., Summersgill, I.A., 1996. Comprehensive methodology for the fitting of predictive accident models. Accident Analysis and Prevention 28, 281-296.

Mallick, B.K., Walker, S., 1997. Combining information from several experiments with nonparametric priors. Biometrika 84 (3), 697-706.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: methodological frontier and future directions. Analytic Methods in Accident Research 1, 1-22.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research 11, 1-16.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models (2nd edition). Chapman & Hall. London, England.

McMillan, T.E., 2007. The relative influence of urban form on a child's travel mode to school. Transportation Research Part A 41 (1), 69-79.

Miaou, S-P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus Negative Binomial regressions. Proceedings of the 73rd Annual Meeting of the Transportation Research Board, Washington D.C.

Miaou, S.P., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvement: decision parameter, treatability concept, statistical criterion and spatial dependence. Accident Analysis and Prevention 37, 699-720.

Millegan, H., Yan, X., Richards, S., Han, L., 2009. Evaluation of effectiveness of stop sign treatment at highway-railroad grade crossings. Transportation Safety and Security 1, 46–60.

Milton, J.C., Mannering, F.L., 1998. The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. Transportation 25, 395-413.

Miranda-Moreno, L.F., Heydari, S., Lord, D., Fu, L., 2013. Bayesian road safety analysis: incorporation of past evidence and effect of hyper-prior choice. Safety Research 46, 31-40.

Miranda-Moreno, L.F., Labbe, A., Fu, L., 2007. Multiple Bayesian testing procedures for selecting hazardous sites. Accident Analysis and Prevention 39, 1192-1201.

Mitra, S., Washington, S., 2007. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis and Prevention 39, 459–468.

Mitra, S., Washington, S., 2012. On the significance of omitted variables in intersection crash modeling. Accident Analysis and Prevention 49 (1), 439-448.

Mohamed, M.G., Saunier, N., Miranda-Moreno, L.F., Ukkusuri, S.V., 2013. A clustering regression approach: a comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada. Safety Science 54, 27-37.

Mothafer, G.I.M.A., Yamamoto, T., Shankar, V.N., 2016. Evaluating crash type covariances and roadway geometric marginal effects using the multivariate Poisson gamma mixture model. Analytic Methods in Accident Research 9, 16-26.

Moudon, A.V., Lee, C., Cheadle, A.D., Collier, C.W., Johnson, D., Schmid, T.L., Weather, R.D., 2005. Cycling and the built environment, a US perspective. Transportation Research Part D 10 (3), 245-261.

Mukhopadhyay, S., Gelfand, A.E., 1997. Dirichlet process mixed generalized linear models. Journal of the American Statistical Association 92 (438), 633-639.

Müller, P., Erkanli, A., West, M., 1996. Bayesian curve fitting using multivariate normal mixtures. Biometrika 83 (1), 67-79.

Müller, P., Quintana, F.A., 2004. Nonparametric Bayesian data analysis. Statistical Science 19 (1), 95–110.

Müller, P., Quintana, F.A., Rosner, G.L., 2007. Semiparametric Bayesian inference for multilevel repeated measurement data. Biometrics 63 (1), 280-289.

Murugiah, S., Sweeting, T., 2012. Selecting the precision parameter prior in Dirichlet process mixture models. Journal of Statistical Planning and Inference 142 (7), 1947-1959.

Narayanamoorthy, S., Paleti, R., Bhat, C.R., 2013. On accommodating spatial dependence in bicycle and pedestrian injury counts by severity level. Transportation Research Part B 55, 245-264.

Nashad, T., Yasmin, S., Eluru, N., Lee, J., Abdel-Aty, M.A., 2016. Joint modeling of pedestrian and bicycle crashes: a copula based approach. Transportation Research Board 95th Annual Meeting, Washington DC, USA.

Neal, R.M., 2000. Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics 9 (2), 249-265.

Noland, R., Quddus, M., 2004. Analysis of pedestrian and bicycle casualties with regional panel data. Transportation Research Record: Journal of the Transportation Research Board 1897, 28-33.

Ntzoufras, I., 2009. Bayesian Modeling Using WinBUGS. Wiley Series in Computational Statistics, Hoboken, USA.

Ohlssen, D.I., Sharples, L.D., Spiegelhalter, D.J., 2007. Flexible random-effects models using Bayesian semi-parametric models: applications to institutional comparisons. Statistics in Medicine 26 (9), 2088-2112.

Osama, A., Sayed, T., 2016. Evaluating the impact of bike network indicators on cyclist safety using macro-level collision prediction models. Accident Analysis and Prevention 97, 28-37.

Papadimitriou, E., Theofilatos, A., Yannis, G., Cestac, J., Kraïem, S., 2014. Motorcycle riding under the influence of alcohol: results from the SARTRE-4 survey. Accident Analysis and Prevention 70, 121-130.

Park, B.J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accident Analysis and Prevention 41 (4), 683-91.

Park, E., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity. Transportation Research Record: Journal of the Transportation Research Board 2019, 1-6.

Park, Y.-J., Saccomanno, F.F., 2005a. Evaluating factors affecting safety at highway–railway grade crossings. Transportation Research Record 1918, 1–9.

Park, Y.-J., Saccomanno, F.F., 2005b. Collision Frequency analysis using tree-based stratification. Transportation Research Record 1908, 119–121.

Persaud, B.P., 1994. Accident prediction models for rural roads. Canadian Journal of Civil Engineering 21, 547–554.

Pucher, J., Buehler, R., 2008. Making cycling irresistible: lessons from the Netherlands, Denmark and Germany. Transport Reviews 28 (4), 495-528.

Pulugurtha, S.S., Sambhara, V.R., 2011. Pedestrian crash estimation models for signalized intersections. accident Analysis and Prevention 43 (1), 439-446.

Oh, J., Washington, S.P., Nam, D., 2006. Accident prediction model for railway-highway interfaces. Accident Analysis and Prevention 38, 346-356.

Quistberg, D.A., Howard, E.J., Ebel, B.E., Moudon, A.V., Saelens, B.E., Hurvitz, P.M., Curtin, J.E., Rivara, F.P., 2015. Multilevel models for evaluating the risk of pedestrian–motor vehicle collisions at intersections and mid-blocks. Accident Analysis and Prevention 84, 99-111.

Railway Association of Canada, 2012. Rail Trends. Ottawa, Canada. http://www.railcan.ca/assets/images/publications/2012_Rail_Trends/2012_RAC_TrendsE_Jan10a.pdf (Accessed July 2016).

Rubin, D.B., 1984. Bayesian justifiable and relevant frequency calculations for the applied statistician. The Annals of Statistics 12 (4), 1151-1172.

Russo, B.J., Savolainen, P.T., Schneider, W.H., Anastasopoulos, P.C., 2014. Comparison of factors affecting injury severity in angle collisions by fault status using a

random parameters bivariate ordered Probit model. Analytic Methods in Accident Research 2, 21-29.

Saccomanno, F., Fu, L., Miranda-Moreno, L.F., 2004. Risk-based model for identifying highway-rail grade crossing blackspots. Transportation Research Record 1862, 127–135.

Saccomanno, F., Fu, L., Ren, C., Miranda-Moreno, L.F., 2003. Identifying highway-railway grade crossing black spots: phase 1. Montreal, Canada: Transport Canada Development Centre.

Saccomanno, F., Lai, X., 2005. A model for evaluating countermeasures at highway-railway grade crossings. Transportation Research Record 1918, 18–25.

Saelens, B.E., Handy, S.L., 2008. Built environment correlates of walking: a review. Medicine and Science in Sports and Exercise 40 (7), S550-S566.

Serhiyenko, V., Mamun, S.A., Ivan, J.N., Ravishanker, N., 2016. Fast Bayesian inference for modeling multivariate crash counts. Analytic Methods in Accident Research 9, 44-53.

Sethuraman, J., 1994. A constructive definition of Dirichlet priors. Statistica Sinica 4, 639-650.

Shaheed, M., Grikitza, K., 2014. A latent class analysis of single-vehicle motorcycle crash severity outcomes. Analytic Methods in Accident Research 2, 30-38.

Shankar, V.N., Ulfarsson, G.F., Pendyala, R.M., Nebergall, M.B., 2003. Modeling crashes involving pedestrians and motorized traffic. Safety Science 41 (7), 627-640.

Shirazi, M., Lord, D., Dhaval, S.S., Geedipally, S.R., 2016. A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: characteristics and applications to crash data. Accident Analysis and Prevention 91, 10-18.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of complexity and fit (with discussion). Journal of the Royal Statistics Society, Series B 64 (4), 1–34.

Spiegelhalter, D.J., Thomas, A., Best, N.G., 2003. WinBUGS 1.4 User Manual. MRC Biostatistics Unit and Imperial College. Available from http://www.mrc-bsu.cam.ac.uk/bugs

Stoker, P., Garfinkel-Castro, A., Khayesi, M., Odero, W., Mwangi, M.N., Peden, M., Ewing, R., 2015. Pedestrian safety and the built environment: a review of the risk factors. Journal of Planning Literature 30 (4), 377-392.

Strauss, J., Miranda-Moreno, L.F., Morency, P., 2014. Multimodal injury risk analysis of road users at signalized and non-signalized intersections. Accident Analysis and Prevention 71, 201-209.

Tay, R., Choi, J., Kattan, L., Khan, A., 2011. A multinomial logit model of pedestrian–vehicle crash severity. International Journal of Sustainable Transportation 5 (4), 233-249.

Thakali, L., 2016. Nonparametric methods for road safety analysis. PhD Dissertation, Department of Civil & Environmental Engineering, University of Waterloo.

Transport Canada, 2014. Canadian motor vehicle collison statistics. https://www.tc.gc.ca/media/documents/roadsafety/cmvtcs2014_eng.pdf (Accessed May 2017)

Transportation Safety Board of Canada, 2015. Statiostical summary, railway occurences. http://www.tsb.gc.ca/eng/stats/rail/2015/sser-ssro-2015.pdf (Accessed May 2017)

Transportation Safety Board of Canada, 2014. Statiostical summary, railway occurences. http://www.tsb.gc.ca/eng/stats/rail/2014/sser-ssro-2014.pdf (Accessed May 2017)

Tunaru, R., 2002. Hierarchical Bayesian models for multiple count data. Austrian Journal of Statistics 31 (3), 221-229.

Ukkusuri, S., Miranda-Moreno, L.F., Ramadurai, G., Isa-Tavarez, J., 2012. The role of built environment on pedestrian crash frequency. Safety Science 50 (4), 1141-1151.

Vanlaar, W., 2005. Multilevel modelling in traffic safety research: two empirical examples illustrating the consequences of ignoring hierarchies. Traffic Injury Prevention 6 (4), 311–316.

Vehtari, A., Lampinen, J., 2002. Bayesian model assessment and comparison using cross-validation predictive densities. Neural computation 14 (10), 2439-68.

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

Walker, S.G., Adrian, F.M.S., Damien, P., Laud, P.W., 1999. Bayesian nonparametric inference for random distributions and related functions. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 61 (3), 485-527.

Wang, Y., Kockelman, K.M., 2013. A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. Accident Analysis and Prevention 60 (3), 71-84.

Washington, S., Karlaftis, M., Mannering, F., 2011. Statistical and Econometric Methods for Transportation Data Analysis, second edition. Chapman and Hall/CRC. Boca Raton, Florida.

Washington, S., Oh, J., 2006. Bayesian methodology incorporating expert judgment for ranking countermeasure effectiveness under uncertainty: example applied to at grade railroad crossings in Korea. Accident Analysis and Prevention, 38, 234-247.

West, M., Turner, D.A., 1994. Deconvolution of mixtures in analysis of neural synaptic transmission. The Statistician 43 (1), 31-43.

Winkelmann, R., 2008. Econometric Analysis of Count Data, 5th edition. Springer-Verlag Berline Heidelberg.

Winters, M., Davidson, G., Kao, D., Teschke, K., 2011. Motivators and deterrents of bicycling: comparing influences on decisions to ride. Transportation 38 (1), 153-168.

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

Wu, L., Lord, D., Zou, Y., 2015. Validation of crash modification factors derived from cross-sectional studies with regression models. Transportation Research Record: Journal of the Transportation Research Board 2514, 88-96.

Wu, Z., Sharma, A., Mannering, F., Wang, S., 2013. Safety impacts of signal-warning flashers and speed control at high-speed signalized intersections. Accident Analysis and Prevention 54, 90–98.

Xiong, Y., Mannering, F.L., 2013. The heterogeneous effects of guardian supervision on adolescent driver-injury severities: a finite-mixture random-parameters approach. Transportation Research Part B 49, 39-54.

Yan, X., Richards, S., Su, X., 2010. Using hierarchical tree-based regression model to predict train–vehicle crashes at passive highway-rail grade crossings. Accident Analysis and Prevention, 42, 64-74.

Yannis, G., Papadimitriou, E., Antoniou, C., 2007. Multilevel modelling for the regional effect of enforcement on road accidents. Accident Analysis and Prevention 39 (4), 818-825.

Yannis, G., Papadimitriou, E., Antoniou, C., 2008. Impact of enforcement on traffic accidents and fatalities: a multivariate multilevel analysis. Safety Science 46 (5), 738–750.

Yannis, G., Papadimitriou, E., Dupont, E., Martensen, H., 2010. Estimation of fatality and injury risk by means of in-depth fatal accident investigation data. Traffic Injury Prevention 11 (5), 492–502.

Yasmin, S., Eluru, N., 2016. Latent segmentation based count models: analysis of bicycle safety in Montreal and Toronto. Accident Analysis and Prevention 95, Part A, 157-171.

Ye, X., Pendyala, R.M., Washington, S.P., Konduri, K., Oh, J., 2009. A simultaneous equations model of crash frequency by collision type for rural intersections. Safety Science 47 (3), 443-452.

Yu, R., Wang, X., Yang, K., Abdel-Aty, M., 2016. Crash risk analysis for Shanghai urban expressways: a Bayesian semi-parametric modeling approach. Accident Analysis and Prevention 95, 495, 502.

Zeger S.L., Karim, M.R., 1992. Generalized linear models with random effects; a Gibbs sampling approach. Journal of the American Statistical Association 86 (413), 79-86.

Zhan, X., Aziz, H.M.A., Ukkusuri, S.V., 2015. An efficient parallel sampling technique for multivariate Poisson-lognormal model: analysis with two crash count datasets. Analytic Methods in Accident Research 8, 45-60.

Zhang, Y., Bigham, J., Ragland, D., Chen, X., 2015. Investigating the associations between road network structure and non-motorist accidents. Journal of Transport Geography 42, 34-47.

Zou, Y., Zhang, Y., Lord, D., 2012. Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis. Accident Analysis and Prevention 50, 1042-1051.

Zou, Y., Zhang, Y., Lord, D., 2014. Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models. Analytic Methods in Accident Research 1, 39-52.

APPENDIX I
List of Municipalities Analyzed in Section 4.2

Municipality	ID	Municipality	ID	Municipality	ID
ALGOMA	1	KITCHENER	28	SALABERRY-DE-VALLEYFIELD	55
ANTIGONISH COUNTY	2	LACOMBE COUNTY	29	SARNIA	56
BECANCOUR	3	LAKESHORE	30	SASKATOON	57
BRAMPTON	4	LEVIS	31	SAULT STE. MARIE	58
BRANDON	5	LONDON	32	SEGUIN	59
CALEDON	6	MIRABEL	33	SHERBROOKE	60
CALGARY	7	MIRAMICHI	34	SHERWOOD	61
CAMBRIDGE	8	MONTREAL	35	ST. CLAIRE	62
CAPE BRETON	9	NANAIMO (City)	36	ST. THOMAS	63
CHATHAM-KENT	10	NANAIMO (Reg Dist)	37	STRATFORD	64
CLARINGTON	11	NORFOLK COUNTY	38	STRATHCONA COUNTY	65
COLCHESTER COUNTY	12	NORTH COWICHAN DM	39	STURGEON COUNTY	66
COOKSHIRE	13	ORO - MEDONTE	40	SUDBURY	67
CORMAN PARK	14	OTTAWA	41	THAMES CENTRE	68
CUMBERLAND COUNTY	15	PETERBOROUGH	42	THOROLD	69
EDMONTON	16	PICKERING	43	THUNDER BAY	70
FRASER-FORT GEORGE	17	PICTOU COUNTY	44	THUNDER BAY	71
GRANDE PRAIRIE	18	PORT COLBORNE	45	TILLSONBURG	72
GRAVENHURST	19	PRINCE ALBERT	46	TORONTO	73
GREATER SUDBURY	20	RED DEER COUNTY	47	VANCOUVER	74
GUELPH	21	REGINA	48	WEST HANTS MD	75
HALDIMAND COUNTY	22	RICHMOND DM	49	WEST LINCOLN TWP	76
HALIFAX	23	ROCKY VIEW MD	50	WEST NIPISSING	77
HAMILTON	24	ROUYN-NORANDA	51	WEST VANCOUVER	78
HUNTSVILLE	25	SAINT JOHN	52	WINDSOR	79
INGERSOLL	26	SAINT-HYACINTHE	53	WINNIPEG	80
INVERNESS COUNTY	27	SAINT-JEAN-SUR-RICHELIEU	54	YELLOWHEAD COUNTY	81

APPENDIX II
WinBUGS Code (Vehicle-Injury Data), Poisson-gamma model

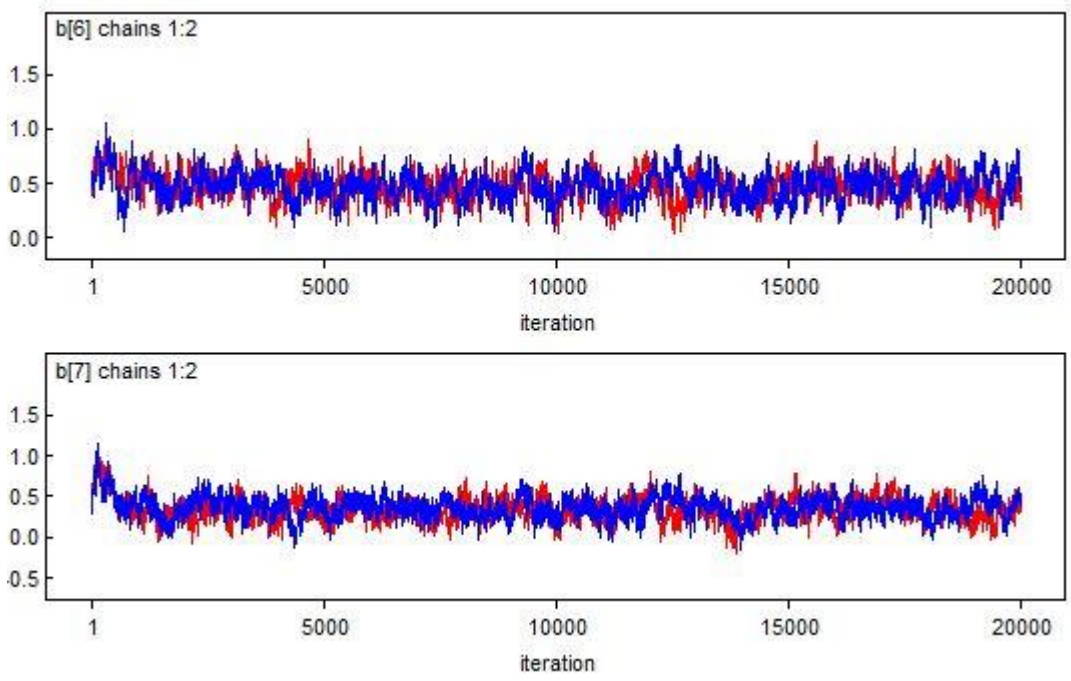
```
Model {   for( i in 1 : 647) {

y[i] ~ dpois(mu[i])

log(mu[i]) <- b[1] + b[2]*throughaadt[i] + b[3]*rightaadt[i] + b[4]*leftaadt[i] +
        b[5]*lnratiovul[i] + b[6]*bus50[i] + b[7]*metro[i] + log(r[i])

r[i] ~ dgamma(phi,phi)

}

for (k in 1:7) { b[k] ~ dnorm(0,0.01) }

phi ~ dgamma(0.001,0.001)

}
```

APPENDIX III

WinBUGS Code (Vehicle-Injury Data), an example of a Dirichlet process model

```
Model { for( i in 1: M ) {
log(mu[i]) <- theta[Z[i]] +  b[1]*throughaadt[i] + b[2]*rightaadt[i] + b[3]*leftaadt[i] +
b[4]*lnratiovul[i] + b[5]*bus50[i] + b[6]*metro[i] + e[i]
e[i] ~ dnorm(0,tau.e)
y[i] ~ dpois(mu[i])
Z[i] ~ dcat(p[])
}
p[1] <- r[1]
for (j in 2:N-1) {p[j] <- r[j]*(1-r[j-1])*p[j-1]/r[j-1]}
for (k in 1:N-1){ r[k] ~ dbeta(1,alpha)}
ps <- sum(p[1:N-1])
for(k in N:N){p[k]<-1-ps}
for(k in 1:N){theta[k] ~ dnorm(basemu,basetau) }
basemu~dnorm(0,0.01)
basetau <- pow(sigmaF0,-2)
sigmaF0 ~ dunif(0,10)
tau.e ~ dgamma(0.01,0.01); var.e <- 1/tau.e
alpha ~ dunif(0.3,10)
for (k in 1:6){ b[k] ~ dnorm(0,0.01) }
}
```

APPENDIX IV
An Example of History Plots (Mixing of Chains in MCMC)



APPENDIX V
An Example of BGR Diagrams (Convergence Check)