Summer 8-19-2022

Generalizing the Negative Binomial-Lindley Model for Accounting Subpopulation Heterogeneity in Crash Data Analysis

A S M Mohaiminul Islam
a.s.islam@maine.edu

GENERALIZING THE NEGATIVE BINOMIAL-LINDLEY MODEL FOR ACCOUNTING SUBPOPULATION HETEROGENEITY IN CRASH DATA ANALYSIS

By

A S M Mohaiminul Islam

B.Sc. Bangladesh University of Engineering and Technology, 2018

A THESIS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

(in Civil and Environmental Engineering)

The Graduate School

The University of Maine

August 2022

Advisory Committee:

Mohammadali Shirazi, Assistant Professor of Civil and Environmental Engineering, Advisor

Per Garder, Professor of Civil and Environmental Engineering

Eric Landis, Frank M. Taylor Professor of Civil and Environmental Engineering

GENERALIZING THE NEGATIVE BINOMIAL-LINDLEY MODEL FOR ACCOUNTING SUBPOPULATION HETEROGENEITY IN CRASH DATA ANALYSIS

By A S M Mohaiminul Islam

Thesis Advisor: Dr. Mohammadali Shirazi

An Abstract of the Thesis Presented
in Partial Fulfillment of the Requirements for the
Degree of Master of Science
(in Civil and Environmental Engineering)
August 2022

Crash data are often highly dispersed; it may also include a large amount of zero observations or have a long tail. The traditional Negative Binomial (NB) model cannot model these data properly. To overcome this issue, the Negative Binomial-Lindley (NB-L) model has been proposed as an alternative to the NB to analyze data with these characteristics. Research studies have shown that the NB-L model provides a superior performance compared to the NB when data include numerous zero observations or have a long tail. In addition, crash data are often collected from sites with different spatial or temporal characteristics. Therefore, it is not unusual to assume that crash data are drawn from multiple subpopulations. Finite mixture models are powerful tools that can be used to account for underlying subpopulations and capture the population heterogeneity. This thesis first documented the derivations and characteristics of the Finite mixture NB-L model (FMNB-L) to analyze data generated from heterogeneous subpopulations with many zero observations and a long tail. The application of the model was demonstrated with a simulation study to identify subpopulations. Then the FMNB-L model was used to analyze Texas four-lane freeway crashes. These data had unique characteristics; it was highly dispersed, had many locations with very large number of crashes, as well as significant number of locations with zero

crash. Multiple goodness-of-fit metrics were used to compare the FMNB-L model with the NB, NB-L, and the finite mixture NB models. The FMNB-L identified two subpopulations in datasets. The results showed a significantly better fit by the FMNB-L compared to other analyzed models.

In addition, the differences in various temporal and spatial factors result in variations of model coefficients among different groups of observations. A grouped random parameters model is a strategy to account for such unobserved heterogeneity. In this thesis, the derivations and applications of a grouped random parameters negative binomial-Lindley model (G-RPNB-L) to account for the unobserved heterogeneity in crash data with many zero observations was proposed. First, a simulation study was designed to illustrate the proposed model. The simulation study showed the ability of the proposed model to correctly estimate the coefficients. Then, an empirical dataset in Maine was used to show the application of the proposed model. It was found that the impact of weather variables denoting “Days with precipitation greater than 1.0 inch”, and “Days with temperature less than 32°F” varied across Maine counties. The proposed model was also compared with the NB, NB-L, and grouped random-parameters NB (G-RPNB) models using different goodness-of-fit metrics. The proposed G-RPNB-L model showed a superior fit compared to the other models.

DEDICATION

I would like to dedicate this thesis to my parents, my sister, and my beautiful wife. They have been a consistent source of inspiration during the journey of my thesis. They showered me with love and affection, which motivated me to work hard every day to achieve my objectives.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and thanks to my advisor, Dr. Mohammadali (Ali) Shirazi, for his unwavering guidance and support during my MS. It was a privilege to work under his supervision and learn so much from him.

I would also like to thank my committee members, Dr. Per Garder and Dr. Eric Landis for supporting me and providing me with their insightful suggestions for my thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Every year millions of people are killed and injured in traffic crashes. Traffic crashes also cause huge monetary losses by damaging public and private properties. World Health Organization (WHO) (2018) reported that 1.35 million people are killed every year across the world because of traffic crashes. Traffic crashes have been ranked as the eighth-leading cause of death by WHO. According to the National Highway Traffic Safety Administration (NHTSA), there were 38,824 deaths on U.S. roads in 2020. The number of people injured on U.S. roads was 2,282,015 and the number of non-fatal crashes on U.S. roads was 5,215,071 (Stewart, 2022). This loss of human life in traffic crashes has irreparable consequences not only for the victims' families but also for the entire world.

Statistical models play a crucial role in predicting the frequency and severity of crashes and improving traffic safety (Lord & Mannering, 2010; Savolainen et al., 2011; Mannering & Bhat, 2014; Mannering et al., 2016). Statistical analysis builds a relationship between a response variable (usually number of crashes) and independent variables (e.g., traffic volume, geometric characteristics of the roadway, human factors, and weather variations). Some of these explanatory variables positively impact the crashes, whereas some of them have negative impacts. Collecting the explanatory variables is an extensive, expensive, and time-consuming task. Therefore, some unobserved factors influencing the occurrence of a crash cannot be collected or may remain unexplained. For example, consider gender as an observed explanatory variable in crash frequency analysis. Variations across the same gender, such as height, weight, bone structures or drinking behaviors can induce unobserved variations for this observed variable (Mannering et al, 2016).

But these unobserved variables cannot be collected. In addition, crash data often include unique characteristics such as a large number of zero responses, or a long tail. Typical models such as the Negative Binomial (NB) cannot address these issues. To overcome these limitations, recently, researchers proposed several new statistical models such as the Negative Binomial-Lindley model. This thesis proposes two new models to simultaneously account for the issue of unobserved heterogeneity and the number of zero responses.

This thesis first proposed a Finite Mixture of Negative Binomial-Lindley (FMNB-L) model to analyze crash data with heterogeneous populations. This flexible modeling approach can address subpopulations heterogeneity in crash data. It can also account for datasets with a large amount of zero crash observations or heavy tails. Then, this thesis proposed a Grouped Random Parameters NB-L (G-RPNB-L) model for addressing unobserved heterogeneity in crash data. This random-parameters modeling approach allows parameters to vary across groups of observations. Similar to the first proposed model, this model can also account for datasets that contain a large number of zero observations. This model can also account for unobserved heterogeneity in crash data due to variations across different groups (e.g., regions) for the impact of different explanatory variables.

1.1. Research Problem

The negative binomial (NB) model is the most frequently used model in analyzing crash data (Lord and Mannering, 2010; Mannering and Bhat, 2014). But this model is not without limitations. For example, it cannot address the presence of a large number of zero crash observations or heavy tails in crash datasets (Zou et al., 2015; Shirazi et al., 2016a). Researchers have developed several models over the last few years by using the mixture of NB and other distributions to account for such datasets (Geedipally et al., 2012; Vangala et al., 2015; Shirazi et al., 2016a; Khodadadi et al.,

2022a). Different extensions of NB-L have been proposed by researchers such as Random Parameters NB-L (RPNB-L) (Rusli et al., 2018; Shaon et al., 2018; Tang et al., 2020; Behara et al., 2021) and Empirical Bayes NB-L (Khodadadi et al., 2022b) because of its popularity in dealing with highly dispersed datasets containing a large number of zeros.

Unobserved heterogeneity is another unique characteristic of crash data and researchers need to address this issue (Mannering & Bhat, 2014; Mannering et al., 2016). Unobserved heterogeneity may exist in crash data because of the presence of latent subpopulations. These latent variations in crash data can be caused by different spatial, environmental, or temporal factors. Finite mixture models provide flexibility in accounting for heterogeneous subpopulations in crash data. Researchers have used finite mixture models in the past to account for heterogeneity in crash data due to latent subpopulations (Frühwirth-Schnatter, 2006; Park & Lord, 2009; Park et al., 2010; Xiong & Mannering, 2013; Zou et al., 2013). But these studies did not account for subpopulations heterogeneity in crash data with a large amount of zero observations or heavy tails. Considering this issue, this research introduced a finite mixture of NB-L model to account for heterogeneous subpopulations in crash data with excess zeros.

Unobserved heterogeneity in crash data can also be caused by regional variations due to the impacts of different explanatory variables. Different regions may have variations in traffic, geometric, human behavior, and weather characteristics. These variations may have different impacts on crashes across different regions. Grouped random parameters approach has been used by several researchers in crash frequency and crash severity studies (Cai et al., 2018; Fountas et al., 2018a; Fountas et al., 2018b; Li Jia et al., 2018; Fanyu et al., 2021).This approach allows parameters to vary across groups of observations. In addition, as noted earlier, crash data may contain a large amount of zero observations for different regions or subgroups. Keeping this in

mind, this research proposed a grouped random parameters NB-L (G-RPNB-L) model to address unobserved heterogeneity due to variations across regions in crash data with a large number of zero observations.

1.2. Thesis Objectives

The primary goal of this thesis is to generalize the NB-L model to address the subpopulations heterogeneity in crash data. To attain this purpose, the following objectives are followed:

First, the derivations and characteristics of the finite mixture of NB-L (FMNB-L) model are documented. A simulation study is demonstrated to evaluate the performance of the proposed model to identify latent subpopulations for different simulated scenarios (i.e., sample mean and different percentages of zero observations). The model performance is illustrated using three empirical datasets of Texas 4-lane freeways and compared with NB, NB-L, and FMNB models based on goodness-of-fit (GOF) measures.

Second, the characteristics and formulations of the grouped random parameters NB-L model are discussed. A simulation study is designed to illustrate the performance of the model in addressing unobserved heterogeneity due to variations across groups (e.g., regions, towns) for various scenarios with different percentages of zeros across groups. An empirical dataset of rural Interstates in Maine is used to demonstrate the performance of the model in addressing unobserved heterogeneity due to regional variations in crash data and compared with NB, NB-L, and G-RPNB models based on goodness-of-fit (GOF) measures.

1.3. Thesis Outline

The outline of this thesis is as follows:

Chapter 2 provides a systematic literature review of finite mixture models and grouped random parameter models in addressing unobserved heterogeneity in crash frequency and crash severity analysis.

Chapter 3 documents the FMNB-L model developed to identify latent subpopulations in crash data that contain a large amount of zero observations using a simulation study and three empirical datasets.

Chapter 4 documents the grouped random parameters NB-L model developed to account for unobserved heterogeneity in crash data while addressing the issue of excess zeros using a simulation study and one empirical dataset.

Chapter 5 presents the conclusions of the research and provides recommendations for future study.

CHAPTER 2

LITERATURE REVIEW

2.1. Introduction

The goal of this research is to generalize negative binomial-Lindley (NB-L) to address subpopulations heterogeneity in crash data. This chapter is structured around a discussion of peer-reviewed articles on unobserved heterogeneity, finite mixture models, and grouped random parameters model. Crash data often exhibit overdispersion. Another important issue that needs to be addressed is unobserved heterogeneity (Mannering & Bhat, 2014; Mannering et al., 2016). These topics are covered at the start of this chapter. Following that, the use of finite mixture models to account for subpopulations heterogeneity is described. Unobserved heterogeneity in crash data can also be addressed using the grouped random parameters models. Several articles on this approach have been reviewed and discussed in this chapter.

2.2. Unobserved Heterogeneity and Data With Many Zeros

Crash data often contains a large number of zeros or has a heavy tail (Zou et al., 2015; Shirazi et al., 2016a). The widely used negative binomial (NB) model cannot address the modeling limitations associated with datasets containing a large number of zeros or heavy tails. Several researchers have tried to deal with this issue by proposing different statistical models such as zero-inflated models (Shankar et al., 1997, 2003), negative binomial-Lindley (NB-L) model (Geedipally et al., 2012), negative binomial-generalized exponential (NB-GE) model (Vangala et al., 2015), negative binomial-Dirichlet process (NB-DP) model (Shirazi et al., 2016a), random parameters NB-L (RPNB-L) model (Rusli et al., 2018; Shaon et al., 2018). The NB-L model has exhibited

superior performance compared to the NB model in capturing over-dispersed data with heavy tails or excess zero observations.

Unobserved heterogeneity is caused by unobservable factors in crash data which results in inconsistent parameter estimations and erroneous interpretations of explanatory variables (Mannering & Bhat, 2014; Mannering et al., 2016). Crash data consists of various traffic, geometric, human behavior, and weather characteristics, which may have correlations between them. For instance, statistical analysis may consider traffic and weather conditions as explanatory variables when interpreting parameters, but there may be some association with human behavior as well. This unobserved correlation may affect model interpretation. Researchers have proposed random parameters multinomial logit model (Behnood & Mannering, 2015), finite mixture models (Park & Lord, 2009; Park et al., 2010; Zou et al., 2013, 2018), and latent-class models with random parameters within class (Xiong & Mannering, 2013) to account for unobserved heterogeneity in crash data.

2.3. Finite Mixture Models

According to Frühwirth-Schnatter, finite mixture models have a wide range of applications in various fields such as biology, genetics, medicine, and marketing (Frühwirth-Schnatter, 2006). They provide flexibility in modeling by accounting for latent subpopulations in heterogeneous data. The finite mixture modeling technique has been widely used in traffic safety studies (Park & Lord, 2009; Park et al., 2010; Eluru et al., 2012; Y. Xie et al., 2012; Xiong & Mannering, 2013; Zou et al., 2013; Behnood et al., 2014). Crash data are often collected from different spatial and temporal attributes. Finite mixture models are powerful tools for addressing unobserved heterogeneity in crash data because of the population heterogeneity caused by these attributes. These models account for hidden sub-groups in crash data.

Park & Lord (2009) demonstrated the application of the finite mixture modeling approach in identifying latent subpopulations in crash data. Crash data are collected from different geographic, environmental, and geometric attributes, which may create heterogeneous subpopulations in crash data. The proposed Finite Mixtures of Poisson (FMP) and Finite Mixtures of NB (FMNB) models accounted for unobserved heterogeneity in crash data due to the existence of latent subpopulations. The standard NB model cannot account for heterogeneous subpopulations, which may result in erroneous coefficients and overdispersion parameter estimation. This modeling approach allowed subpopulations or components to have varying regression coefficients and overdispersion parameters compared to traditional models. Also, this approach provided flexibility in distributional assumptions on the mixing variables.

The aforementioned study provided useful insight into capturing the unobserved heterogeneity because of the presence of heterogeneous subpopulations in crash data. But it did not account for the performance of the model in crash data analysis for a wide range of sample sizes and sample mean values. Park et al. (2010) extended the scope of their previous study by examining the bias properties of the posterior mean and median of the dispersion parameters in the two components FMNB-2 regression models. A simulation study was designed based on small mean ($\bar{y} < 1$), moderate mean ($1 < \bar{y} < 5$), and high mean ($\bar{y} > 5$) having a wide range of sample sizes. The posterior median using the non-informative prior exhibited better bias properties compared to the posterior mean for small sample sizes and small to moderate sample means. However, when sample sizes were larger, the posterior median exhibited an upward bias similar to the posterior mean. The bias in the estimates decreased when a weakly informative prior was employed for the posterior mean and median. This study recommended sample size range, suitable priors, and summary statistics for crash data analysis based on bias properties.

The prior works described above considered fixed weight parameter in formulating two components finite mixture of NB regression models. Zou et al. (2013) investigated the application of two components finite mixture of NB models with varying weight parameter for crash data analysis. The finite mixture of Poisson regression models cannot handle extra variations within components, which makes parameter interpretation unreliable. As a result, this study was conducted by comparing the model performance of two components finite mixture of NB models with fixed weight parameter (FMNB-2) and two components finite mixture of NB models with varying weight parameter (GFMNB-2). The GFMNB-2 models produced a better statistical fit and aided in the classification of high and low-risk crash sites. This model was also capable of capturing the overdispersion present in crash data. As a result, GFMNB-2 outperformed FMNB-2 in terms of capturing unobserved heterogeneity and overdispersion in crash data.

Zou et al. (2018) used finite mixture of NB models to calculate empirical Bayes (EB) in the highway safety analysis. The empirical Bayes method is widely used for hotspot identification and before-after studies in highway safety analysis. The traditional NB model is widely used for capturing overdispersion in crash data and is commonly used in the EB method. Zou et al. (2017) employed the GFMNB-2 model with varying weight parameter in their study for site rankings using EB estimates. The proposed model addressed unobserved heterogeneity in crash data due to the presence of heterogeneous population and improved crash predictions.

Finite mixture or latent class modeling approach has been widely used in crash severity analysis. Eluru et al. (2012) employed a latent class modeling approach to analyze crash severities at highway-railway grade crossings and identified various key factors influencing injury severities. Xie et al. (2012) investigated single-vehicle crash severities on rural roads using a latent class logit (LCL) model. This modeling approach allowed the coefficients of explanatory variables to vary

for different injury outcomes. This aided in understanding the impacts of various explanatory variables in crash severities. Behnood et al. (2014) examined the impacts of age, gender, and alcohol consumption on crash severities using a latent class multinomial logit modeling technique. This model accounted for heterogeneous effects across the subpopulations in this study. Xiong & Mannering (2013) implemented a finite mixture random parameters approach to study the heterogeneous effects of guardian supervision on crash severities. Thus, finite mixture models are widely used in both crash frequency and crash severity studies to account for latent subpopulations in crash data.

2.4. Grouped Random Parameters Models

Grouped random parameters models allow the mean and variance of the coefficients to vary across observations or groups (Mannering et al., 2016; Meng et al., 2017; Sarwar et al., 2017). The concept of modeling unobserved heterogeneity in crash data with Grouped Random Parameters is a powerful tool. This technique has been used in several crash frequency investigations (Cai et al., 2018; Fountas et al., 2018a; Heydari et al., 2018; Li Jia et al., 2018). Developing a reliable and efficient model for the analysis of crash occurrence on segments and intersections is necessary because they constitute a major part of the road network. To avoid omitted variable bias and inconsistent parameter estimations, appropriate explanatory variables must be used (Lord & Mannering, 2010; Mannering et al., 2016).

The influence of zonal factors on crash data modeling at segments and intersections was investigated by Cai et al. (2018). A grouped random parameters multivariate spatial model was implemented to account for zonal effects and unobserved heterogeneity in crash data modeling at segments and intersections. The addition of zonal characteristics like traffic characteristics (e.g., daily vehicle miles driven, percentage of heavy vehicles) and socio-demographic data (e.g.,

population, median household income) enhanced model estimation significantly. This study analyzed a crash dataset from Central Florida, which included 24.7 percent intersection-related crashes and 75.3 percent segment-related crashes. Integration of zonal factors (e.g., daily vehicle miles traveled, percentage of heavy vehicles, population, median family income) at segments and intersections also contributed to addressing unobserved heterogeneity in crash data and resulted in more robust parameter estimations. This modeling approach also looked at the heterogeneous and spatial correlations of zonal impacts on crash occurrences at segments and intersections and identified significant heterogeneous correlations.

Heydari et al. (2018) employed another Grouped Random Parameters approach to deal with the complex crash mechanisms at highway-railway grade crossings. Unobserved heterogeneity may exist in grade crossing crash data, resulting in erroneous parameter estimates. This study implemented a heteroskedastic grouped random parameters Poisson lognormal model with heterogeneity in mean and variance. This hierarchical Bayesian modeling approach allowed for the comparison of different geographic regions in terms of grade crossing safety. Unobserved heterogeneity was captured by modeling heterogeneity in the mean and variance of grouped random parameters as a function of explanatory variables. The study found the dispersion of crash frequencies was greater in urban areas than in rural areas because the variance is 0.134 times higher for urban areas.

Fountas et al. (2018a) developed a dynamic correlated grouped random parameters binary logit model to study the mixed effects of both non-time varying and time-varying explanatory variables on crashes and capture unobserved heterogeneity in crash data. This model used an unrestricted covariance matrix approach to estimate grouped random parameters, which allowed for parameter correlations as well as accounted for unobserved heterogeneity. The marginal effects

of stationary explanatory variables such as segment length and median width revealed that an increase in these variables resulted in an increase in the likelihood of crashes by 0.0183 and 0.0002, respectively. The marginal effect of dynamic explanatory variables such as the relative humidity indicator in t-30 minutes showed that an increase in this variable resulted in an increase in the likelihood of crashes by 0.0357.

Another crash frequency study incorporating Grouped Random Parameters approach was implemented by Li Jia et al. (2018). This study investigated the relationship between the Level of Safety (LOS) and traffic safety at signalized intersections by considering temporal attributes and different types of crashes. A grouped random parameters negative binomial model was proposed to study the LOS-safety relationship for total crashes, and a bivariate grouped random parameters negative binomial model was proposed for rear-end and left-turn crashes. The relationship varied across times for different types of crashes.

Grouped Random Parameters modeling approach has been implemented in crash severity analysis too. Fountas et al. (2018b) developed a correlated random parameters ordered probit model to analyze crash severity. This model accounted for unobserved heterogeneity and also addressed interactions among observed or unobserved characteristics. Fanyu et al. (2021) studied the effect of the presence of trucks of different classes on non-truck-related crashes by developing a correlated grouped random parameters binary logit model. This approach accounted for unobserved heterogeneity at both the observation level and space-time level. In addition to crash frequency and severity investigations, the Grouped Random Parameters technique has been used to analyze perceived and observed aggressive driving behavior (Sarwar et al., 2017) and pedestrian safety studies (Pantangi et al., 2021).

2.5. Chapter Summary

The nature of crashes, which are often acquired from different temporal and spatial attributes, may produce unobserved heterogeneity in crash data. Unobserved heterogeneity needs to be addressed in both crash frequency and crash severity studies (Mannering & Bhat, 2014; Mannering et al., 2016). The presence of latent subpopulations may cause unobserved heterogeneity in crash data. Finite mixture models provide a flexible modeling approach to address subpopulations heterogeneity in crash data. Different traffic (e.g., traffic volume, speed, driver behavior), geometric (e.g., skid number, lane width, curve presence), and weather (e.g., rainfall, snowfall, visibility, temperature) characteristics have different impacts on crashes and these effects may vary across different regions. As a result, unobserved heterogeneity may exist in crash data due to variations across regions too. Grouped Random Parameters modeling approach has become popular nowadays because it allows parameters to vary across groups of observations. It also allows accounting for zonal factors on crash occurrences. This chapter discussed various peer-reviewed articles to have a better understanding of these topics.

CHAPTER 3

FINITE MIXTURE NEGATIVE BINOMIAL-LINDLEY

3.1. Introduction

Statistical models play a crucial role in improving safety. Over the last decade, research studies have proposed various statistical models to analyze crash data (Lord & Mannering, 2010; Mannering & Bhat, 2014; Mannering et al., 2016; Lord et al., 2021). These models attempt to address unique characteristics in crash data that are not typically found in other research fields. As such, crash data are often highly dispersed and characterized by many zero observations or a long (or heavy) tail (Zou et al., 2015; Shirazi et al., 2016a). Several researchers have proposed models to analyze these data. Initially, zero-inflated models were introduced to account for excess zero observations. However, zero-inflated models have important limitations. Research studies have documented multiple limitations of these models (Lord et al., 2005; Lord et al., 2007; Lord et al., 2021), such as the strict dual state process or a state with a long-term mean that is equal to zero. Recently, using the mixture of NB and other distributions has received significant attention from researchers to account for such data characteristics (Geedipally et al., 2012; Vangala et al., 2015; Shirazi et al., 2016a; Khodadadi et al., 2022a). Negative binomial-Lindley (NB-L) is one of the most popular models in this line of modeling. The NB-L distribution was first introduced by Zamani & Ismail (2010). Lord & Geedipally (2011) later demonstrated its application to model crash data with many zero observations. Previous research studies also developed NB-L generalized linear model (GLM) (Geedipally et al., 2012), Random Parameters NB-L (RPNB-L) (Rusli et al., 2018; Shaon et al., 2018; Tang et al., 2020; Behara et al., 2021) and Empirical Bayes NB- (Khodadadi et al., 2022b), and showed its superior performance compared to the NB.

Addressing unobserved heterogeneity is another important challenge in modeling (Mannering et al., 2016). Crash data are often collected from groups with various geographical, environmental, behavioral, or other spatial or temporal attributes. The simple NB model cannot account for population heterogeneity in modeling when data comes from heterogeneous sources. Finite mixture models (Frühwirth-Schnatter, 2006; Park & Lord, 2009; Park et al., 2010; Xiong & Mannering, 2013; Zou et al., 2013) are a class of models that address the heterogeneity in population by accounting for latent subpopulations (or groups or classes) in the data. For example, one subpopulation may include data with high mean and variations, but another with low mean and variations. Park & Lord (2009) demonstrated the application of the finite mixture negative binomial GLM (FMNB GLM) in modeling heterogeneous crash data drawn from different subpopulations and documented its superior performance to the simple NB model using several datasets and multiple goodness-of-fit (GOFs) statistics.

This chapter documents the derivations and applications of the finite mixture NB-L GLM (FMNB-L GLM) in modeling crash data with many zero observations and a long tail. This research was motivated by two concepts or ideas. The first idea can be explained by taking a closer look at the structure of the NB-L and FMNB models and the flexibility they provide in modeling. The NB-L GLM provides additional flexibility in modeling by mixing the NB with the Lindley distribution. This additional flexibility allows the model to account for excess zero observations or a long tail (Shirazi et al., 2016a). The FMNB models also provide very flexible models by accounting for subpopulations in the data. We are deriving the FMNB-L model to benefit from the strength of both strategies. It is hypothesized that the FMNB-L will provide significantly flexible models that account for both heterogeneity in population and numerous zero observations. Secondly, the research was inspired by taking a closer look at zero-inflated models. Zero-inflated

models assume a dual state, with two distinctive components, one with a mean that is always zero (Lord et al., 2005; Hilbe, 2011). As noted above, zero-inflated models have been criticized for having a stage with a long-term mean of zero, which is theoretically impossible for numerous cases or scenarios (see, e.g., Lord et al., 2005; Allison, 2012; H. Xie et al., 2013; Fisher et al., 2017; Lord et al., 2019; Lord et al. 2021). Finite mixture models, however, assume that each observation can belong to all subpopulations with certain probabilities where none of the subpopulations has a long-term mean equal to zero. Using the FMNB-L model, we can account for the excess number of zero observations without assuming a subgroup with a long-term mean of zero.

In this chapter, first, we document the characteristics of the FMNB-L model. Then, we document the results of a simulation study to evaluate the performance of the FMNB-L model in identifying subpopulations for data with different characteristics (i.e., the population mean and the percentage of zero observations). In the end, we demonstrate the performance of the FMNB-L model using three empirical datasets and compare the results with the NB, NB-L, and FMNB models based on multiple GOF measures.

3.2. Background

Before documenting the derivations and characteristics of the FMNB-L model, let us first briefly review the NB, NB-L, and FMNB GLMs. The NB is the most common model used to analyze over-dispersed crash count data (Lord and Mannering, 2010; Mannering and Bhat, 2014). The probability mass function (pdf) of the negative binomial distribution is defined using the following equation (Hilbe, 2011):

$$\mathrm{NB}(\mu_i, \varphi) \equiv \mathrm{P}(y_i \mid \mu_i, \varphi) = \frac{\Gamma(y_i + \varphi)}{\Gamma(y_i + 1) \times \Gamma(\varphi)} \left(\frac{\mu_i}{\mu_i + \varphi}\right)^{y_i} \left(\frac{\varphi}{\mu_i + \varphi}\right)^{\varphi}; \ \varphi \text{ and } \mu_i > 0 \qquad (1)$$

Where y_i is the crash observation at site i, and $\Gamma(.)$ is the gamma function. The parameters μ_i and φ respectively denote the mean response of observations at site i, and the inverse dispersion parameter. Often, it is assumed that the mean response of the observations has a log-linear relationship with regression coefficients (denoted by βs) and a set of m-dimensional covariates (denoted by X) as follows:

$$\ln(\mu_i|\beta_0, \beta_1, \dots \beta_m) = \beta_0 + \sum_{j=1}^{m} \beta_j X_{ij} \quad (2)$$

The NB-L model is a mixture of the NB and Lindley distribution and can be written as follows (Geedipally et al., 2012):

$$\text{NB-L}(\mu_i, \varphi, \theta) \equiv P(Y = y_i|\, \mu_i, \varphi, \theta) = \int \text{NB}(y_i|\ \varepsilon_i\mu_i, \varphi)\text{Lindley}(\varepsilon_i|\theta)\, d\varepsilon_i \quad (3)$$

Note that the Lindley distribution can be written as a mixture of the following two gamma distributions (Zamani & Ismail, 2010):

$$\varepsilon_i|\theta \sim \frac{1}{1+\theta}\ \text{gamma}(2, \theta) + \frac{\theta}{1+\theta}\ \text{gamma}(1, \theta) \quad (4)$$

This expression is equal to the following hierarchical representation:

$$\varepsilon_i|z_i, \theta\ \sim \text{gamma}(1 + z_i, \theta) \quad (5\text{-}1)$$

$$z_i|\theta \sim \text{Bernoulli}\left(\frac{1}{1+\theta}\right) \quad (5\text{-}2)$$

Hence, the NB-L model can be presented as the following hierarchical representation (Zamani & Ismail, 2010):

$$y_i|\ \varepsilon_i\mu_i, \varphi \sim \text{NB}(\varepsilon_i\mu_i, \varphi) \quad (6\text{-}1)$$

$$\varepsilon_i|z_i, \theta\ \sim \text{gamma}(1 + z_i, \theta) \quad (6\text{-}2)$$

$$z_i|\theta \sim \text{Bernoulli}\left(\frac{1}{1+\theta}\right) \quad (6\text{-}3)$$

$$\ln(\mu_i|\beta_0, \beta_1, \ldots \beta_m) = \beta_0 + \sum_{j=1}^{m} \beta_j X_{ij} \quad (6\text{-}4)$$

The expectation and variance of the variable $y_i \sim$NB-L(μ_i, φ, θ) is given as follows (Geedipally et al., 2012):

$$E(y_i|\mu_i, \theta) = \mu_i \times \frac{\theta + 2}{\theta(\theta + 1)} \quad (7\text{-}1)$$

$$var(y_i|\mu_i, \varphi, \theta) = \mu_i \times \frac{\theta + 2}{\theta(\theta + 1)} + \left(\mu_i{}^2 \times \frac{1 + \varphi}{\varphi}\right) \times \frac{2(\theta + 3)}{\theta^2(\theta + 1)} - \left(\mu_i \times \frac{\theta + 2}{\theta(\theta + 1)}\right)^2 \quad (7\text{-}2)$$

The NB-L model provides greater flexibility to account for excess zero observations or data characterized by a long tail or large skewness (Shirazi et al., 2016a , Shirazi et al., 2017a).

Finite mixture models are another class of models that provide flexibility in modeling especially when data are originated from heterogeneous populations (Park & Lord, 2009; Park et al., 2014; Zou et al., 2018). In finite mixture models, each observation belongs to the finite mixture of distributions with certain probabilities. The general form of finite mixture models with K components is defined in Eq. (8) as follows [note that vectors are shown in bold fonts]:

$$p(y_i|\boldsymbol{\Theta}) = \sum_{k=1}^{K} w_k f_k(y_i|\boldsymbol{\Theta_k}) \quad (8\text{-}1)$$

$$\sum_{k=1}^{K} w_k = 1 \quad (8\text{-}2)$$

where w_k and $f_k(.|\boldsymbol{\Theta_k})$ represent the mixing weight, and the distribution of the k-th component respectively. The vector $\boldsymbol{\Theta_k}$ indicates the parameters of the k-th distribution. The mean and variance of finite mixture models (based on the above general form) are given as follows (Park & Lord, 2009):

$$E(y_i|\boldsymbol{\Theta}) = \sum_{k=1}^{K} w_k E(y_i|\boldsymbol{\Theta_k}) \tag{9-1}$$

$$var(y_i|\boldsymbol{\Theta}) = \sum_{k=1}^{K} (E^2(y_i|\boldsymbol{\Theta_k}) + var(y_i|\boldsymbol{\Theta_k}) - E^2(y_i|\boldsymbol{\Theta}) \tag{9-2}$$

Given the definition in Eq. (8), the general form of the finite mixture of negative binomial (FMNB) model with k components, and parameters $\boldsymbol{\mu_i} = \{\mu_{i1}, \mu_{i2}, \ldots, \mu_{ik}\}$, $\mathbf{w} = \{w_1, w_2, \ldots, w_k\}$, and $\boldsymbol{\varphi} = \{\varphi_1, \varphi_2, \ldots, \varphi_k\}$ is defined in Eq. (10) (Park & Lord, 2009):

$$p(y_i|\mathbf{w}, \boldsymbol{\mu_i}, \boldsymbol{\varphi}) = \sum_{k=1}^{K} w_k NB(\mu_{ik}, \varphi_k)$$

$$= \sum_{k=1}^{K} w_k \left[\frac{\Gamma(y_i + \varphi_k)}{\Gamma(y_i + 1)\Gamma(\varphi_k)} \left(\frac{\varphi_k}{\mu_{ik} + \varphi_k} \right)^{\varphi_k} \left(\frac{\mu_{ik}}{\mu_{ik} + \varphi_k} \right)^{y_i} \right] \tag{10}$$

The mean and variance of the FMNB model is given as follows:

$$E(y_i|\mathbf{w}, \boldsymbol{\mu_i}) = \sum_{k=1}^{K} w_k \mu_{ik} \tag{11-1}$$

$$var(y_i|\mathbf{w}, \boldsymbol{\mu_i}, \boldsymbol{\varphi}) = \sum_{k=1}^{K} \left(w_k \mu_{ik} + \mu_{ik}^2 \left(\frac{1 + \varphi_k}{\varphi_k} \right) \right) - E^2(y_i|\boldsymbol{\Theta}) \tag{11-2}$$

Where μ_k and φ_k show the mean and inverse dispersion parameter of the k-th NB component.

3.3. Finite Mixture Negative Binomial-Lindley

This section describes the characteristics of the FMNB-L GLM. Let us assume the model includes K latent NB-L subpopulations. Therefore, the FMNB-L model with k components and the vector

of parameters denoted by $\boldsymbol{\mu_i} = \{\mu_{i1}, \mu_{i2}, \ldots, \mu_{ik}\}$, $\mathbf{w} = \{w_1, w_2, \ldots, w_k\}$, $\boldsymbol{\varphi} = \{\varphi_1, \varphi_2, \ldots, \varphi_k\}$ and $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_k\}$ is defined by the following closed form:

$$p(y_i|\mathbf{w}, \boldsymbol{\mu_i}, \boldsymbol{\varphi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k \mathrm{NBL}(\mu_{ik}, \varphi_k, \theta_k)$$

$$= \sum_{k=1}^{K} w_k \int \mathrm{NB}(y|\, \varepsilon_{ik}\mu_{ik}, \varphi_k)\mathrm{Lindley}(\varepsilon_{ik}|\theta_k)\, d\varepsilon_{ik} \tag{12}$$

Eq. (12) can also be rewritten as follows:

$$p(y_i|\mathbf{w}, \boldsymbol{\mu_i}, \boldsymbol{\varphi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k \mathrm{NB}(\varepsilon_{ik}\mu_{ik}, \varphi_k)\,;\ \varepsilon_{ik} \sim \mathrm{Lindley}(\theta_k) \tag{13}$$

Given Eq. (7) and Eq. (11), the mean and the variance of the FMNB-L model can be written as follows:

$$\mathrm{E}(y_i|\mathbf{w}, \boldsymbol{\mu_i}, \boldsymbol{\varphi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} w_k \left(\mu_{ik} \times \frac{\theta_k + 2}{\theta_k(\theta_k + 1)}\right) \tag{14-1}$$

$$\mathrm{var}(y_i|\mathbf{w}, \boldsymbol{\mu_i}, \boldsymbol{\varphi}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \left(w_k \mu_{ik} \frac{\theta_k + 2}{\theta_k(\theta_k + 1)} + \mu_{ik}^2 \left(\frac{1 + \varphi_k}{\varphi_k}\right)\left(\frac{2(\theta_k + 3)}{\theta_k^2(\theta_k + 1)}\right)\right) - \mathrm{E}^2(y_i|\mathbf{w}, \boldsymbol{\mu_i}, \boldsymbol{\varphi}, \boldsymbol{\theta}) \tag{14-2}$$

Given the hierarchical representation of the NB-L model described in Eq. (6), we can write Eq. (13) in the following hierarchical Bayesian representation:

$$y_i|\, \varepsilon_{ik}, \mu_{ik}, \varphi_k \sim \textstyle\sum_{k=1}^{K} w_k \mathrm{NB}(\varepsilon_{ik}\mu_{ik}, \varphi_k) \tag{15-1}$$

$$\varepsilon_{ik}|z_{ik}\,, \theta_k \sim \mathrm{gamma}(1 + z_{ik}, \theta_k) \tag{15-2}$$

$$z_{ik}|\theta_k \sim \text{Bernoulli}\left(\frac{1}{1+\theta_k}\right) \tag{15-3}$$

$$\ln(\mu_{ik}|\beta_{0k}, \beta_{1k}, \ldots \beta_{mk}) = \beta_{0k} + \sum_{j=1}^{m} \beta_{jk} X_{ij} \tag{15-4}$$

The Markov Chain Monte Carlo (MCMC) simulation can be used to estimate the parameters of the hierarchical model described in Eq. (15). In addition, given that all distributions in Eq. (15) have standard distributions if suitable prior distributions are used, Eq. (15) can be implemented in statistical software programs such as WinBUGS (Spiegelhalter et al., 2003) for MCMC analysis.

Eq. (15) presented a FMNB-L model with intercept terms (β_{0k}); however, as noted in previous studies (Geedipally et al., 2012; Shirazi et al., 2016a), there are strong correlations between the intercept (β_{0k}) and the site frailty terms (ε_k). To overcome this issue, it is recommended to either use an informative prior for ε_k that ensures $E(\varepsilon_k) = 1$ or drop the intercept initially from the model, and then once the model converged, approximately estimate the intercept using the following equation.

$$\beta_{0k} = E\left(\log\left(E(\varepsilon_k)\right)\right) = E\left(\log\left(\frac{\theta_k + 2}{\theta_k(\theta_k + 1)}\right)\right) \tag{16}$$

Eq. (16) can easily be estimated using MCMC. For this purpose, the value of $\log\left(\frac{\theta_k+2}{\theta_k(\theta_k+1)}\right)$ needs to be recorded in each iteration of MCMC. After the completion of MCMC, an average is taken over all simulated samples. The average values will be presented as the intercepts. We used the later approach in our analysis.

It is worth pointing out that the MCMC outputs can be used to determine the association probabilities of subpopulations for each observation. For this purpose, it is more convenient to revise Eq. (15) using a subpopulation allocation parameter for each site “i" denoted by δ_i (Ohlssen et al., 2007). Let us assume a categorical distribution with probabilities of w_k on δ_i. Then, Eq. (15) is revised as follows:

$$y_i|\, \varepsilon_{ik}, \mu_{ik}, \varphi_k\, , (\delta_i = k) \sim NB(\varepsilon_{ik}\mu_{ik}, \varphi_k) \tag{17-1}$$

$$\varepsilon_{ik}|z_{ik}\, , \theta_k, (\delta_i = k)\ \sim \text{gamma}(1 + z_{ik}, \theta_k) \tag{17-2}$$

$$z_{ik}|\theta_k, (\delta_i = k)\ \sim \text{Bernoulli}\left(\frac{1}{1+\theta_k}\right) \tag{17-3}$$

$$\ln(\mu_{ik}|\beta_{0k}, \beta_{1k}, \dots \beta_{mk}, (\delta_i = k)\ ) = \beta_{0k} + \sum_{j=1}^{m} \beta_{jk} X_{ij} \tag{17-4}$$

If we run the MCMC for N iterations. The association probability of subpopulation k for the i-th observation (p_{ik}) is derived as:

$$p_{ik} = \frac{\left(\sum_{n=1}^{N} I_{i,n}(\delta_i = k)\right)}{N} \tag{18}$$

Where, for each i-th observation, the indicator parameter $I_{i,n}(\delta_i = k)$ denotes a sample from the posterior of association probabilities at iteration n of MCMC, which is equal to one if $\delta_i = k$, and zero otherwise. In the next section, we document the results of a simulation study to evaluate the performance of the FMNB-L model to estimate the coefficients of subpopulations given a range of mean and percentage of zero observations.

3.4. Simulation Analysis

As noted earlier, we proposed the FMNB-L GLM to identify latent subpopulations in data characterized by many zero observations or a long tail. In this section, the results of a simulation study to evaluate the performance of the FMNB-L model in estimating the coefficients of subpopulations are presented and discussed for a range of scenarios. For this purpose, we simulated data for a range of characteristics (i.e., different crash means and percentages of zero responses), and then used the FMNB-L to find the subpopulations. This section is divided into two parts. The first part documents the simulation protocol used in this study to simulate data for different scenarios (or different characteristics, to be exact). The second part illustrates the results of applying the simulation protocol.

3.4.1. Simulation Protocol

Simulation has been used by various studies to demonstrate an idea, draw conclusions about the advantages and limitations of a methodology, or provide guidelines (Lord, 2006; Shirazi et al., 2016b; Shirazi et al., 2017b; Shirazi et al., 2021). Simulation is a powerful method due to its ability to create controlled scenarios when known input variables are available. We use simulation to understand the FMNB-L strength in identifying the mixing components for a range of sample mean and zero responses. We designed a simulation study similar to the Park et al. (2010) work. Without loss of generality, we considered a two component FMNB-L model with mixing weights of 0.5 ($w_1 = w_2 = 0.5$). For simplicity, we assumed a common inverse dispersion parameter (φ) for the two components. We first simulated two covariates (X_1 and X_2) from a normal distribution with a mean of zero and variance of 1. Then, given two sets of coefficients (i.e., β_{11}, β_{21} and β_{12}, β_{22}), we constructed the subpopulations means (μ_{i1}, μ_{i2}) using Eq. (2). We then simulated two sets of unobserved site frailty terms ($\varepsilon_{i1}, \varepsilon_{i2}$) from two different Lindley distributions. We controlled over

the range of mean and percentage of zeros by modifying the values of Lindley parameters (θ_1, θ_2). Then, observations were simulated from a mixture of two negative binomial distributions as follows:

$$y_i \sim \sum_{k=1}^{2} w_k NB(\varepsilon_{ik}\mu_{ik}, \varphi) \quad (19)$$

To accomplish the latest step, we randomly sampled crash observations from the mixture model in Eq. (19) with a probability of w_k using a categorical distribution.

The detailed steps of the simulation protocol are described below:

Step 1. Initializations.

1.1. Set the regression coefficients for the two subpopulations. β_{11} and β_{21} represent the two coefficients of the first component, and β_{12} and β_{22} the coefficients of the two coefficients of the second component.

1.2. Set the value of Lindley parameters (θ_1 and θ_2) for the two subpopulations.

1.3. Set the value for the inverse dispersion parameter (φ).

Step 2. Simulate covariates and find the mean.

2.1 Simulate 10,000 draws for variables X_{i1} and X_{i2} from a standard normal distributions as follows:

$$X_{i1} \sim N(0, 1); \quad i = 1, \dots, 10{,}000$$

$$X_{i2} \sim N(0, 1); \quad i = 1, \dots, 10{,}000$$

2.2. Calculate the mean of components (or subpopulations) using the following equations:

$$\mu_{i1} = \exp(\beta_{11} X_{i1} + \beta_{12} X_{i2})$$

$$\mu_{i2} = \exp(\beta_{12} X_{i1} + \beta_{22} X_{i2})$$

Step 3. Simulate site-specific frailty terms (ε_1 and ε_2)

3.1. Simulate 10,000 site-specific frailty terms for component 1 (ε_1) and component 2 (ε_2) from Lindley distributions with parameters θ_1 and θ_2 respectively.

$$\varepsilon_{i1} \sim \text{Lindley}(\theta_1); \qquad i = 1, \ldots, 10{,}000$$

$$\varepsilon_{i2} \sim \text{Lindley}(\theta_2); \qquad i = 1, \ldots, 10{,}000$$

Step 4. Simulate crash observations.

4.1 Simulate crash observations (y_i) from a mixture of two negative binomial distributions as follows:

$$y_i \sim \sum_{k=1}^{2} w_k \text{NB}(\varepsilon_{ik}\mu_{ik}, \varphi)$$

where $w_1 = w_2 = 0.5$.

Step 5. Fit the Model.

5.1. Use the FMNB-L model (Eq. 15) to fit the model and estimate the coefficients.

3.4.2. Simulation Results

We ran the simulation protocol for a range of data that include scenarios with low mean ($\bar{y} < 1$), moderate mean ($1 < \bar{y} < 2$), and high mean ($\bar{y} > 5$). In addition, we controlled the percentages of zeros by simulating datasets that include approximately 60%, 70%, and 80% of zero observations. An additional dataset containing approximately 90% of zero observations was also considered for the low mean ($\bar{y} < 1$) category. For other mean categories, having more than 80% zero observations is nearly impossible; hence, they were not considered in simulation analysis. Although data with high mean ($\bar{y} > 5$) and 70% or 80% zero observations are rarely observed in

real crash data, we considered them in the analysis for the sake of simulation completeness. Table 1 the mean and percentage of zeros for simulated datasets. As shown in this table, the mean and percentages of zeros were recorded for both subpopulations and the population. We ensured that there are distinct subpopulations in simulated datasets. Therefore, often, the simulated data include one component with a smaller mean, and another with a larger mean, or components with different percentages of zeros.

Table 1: Summary Statistics of Simulated Data

Data Mean	**Zeros in Data**	**~ 60%**		**~ 70%**		**~ 80%**		**~ 90%**	
	Components	**Mean**	**Zeros**	**Mean**	**Zeros**	**Mean**	**Zeros**	**Mean**	**Zeros**
Low ($\bar{y} < 1$)	**Combined**	0.9	61%	0.8	69%	0.4	80%	0.2	90%
	Component 1	1.2	51%	0.6	75%	0.3	86%	0.2	87%
	Component 2	0.5	70%	0.9	64%	0.5	73%	0.1	93%
Moderate ($1 < \bar{y} < 2$)	**Combined**	1.7	60%	1.8	70%	1.2	79%		
	Component 1	0.5	77%	0.3	84%	1.0	82%		
	Component 2	2.9	42%	3.1	57%	1.4	76%		
High ($\bar{y} > 5$)	**Combined**	8.3	59%	7.1	70%	5.3	79%		
	Component 1	0.2	87%	1.3	81%	13.0	79%		
	Component 2	19.0	31%	12.0	58%	0.5	80%		

We considered both extreme cases related to having data with numerous zero observations (80% or 90%) as well as data with a long tail in our analysis. The former was considered by simulating data with low mean ($\bar{y} < 1$) and high percentages of zero (e.g., ~80% or ~90%). The latter was considered in the simulation study by considering data with high mean ($\bar{y} > 5$) with two clear and distinct components, one component with small mean and high percentages of zeros and the other with very high mean and small percentages of zeros. This consideration is clearly observed in the case of a population with high mean ($\bar{y} > 5$) and 60% zero observations. For this scenario, we simulated two subpopulations, one with a mean of 0.2 and 87% zero observations and the other with a mean of 19.0 and about 31% zero observations.

We implemented the model in WinBUGS (Spiegelhalter et al., 2003). Priors were specified to estimate the unknown parameters. Priors for regression coefficients (βs), inverse dispersion parameter (φ), and weighs (w) were assumed to have a normal, gamma, and uniform distributions respectively. We also considered a uniform prior on $1/(1 + \theta)$ parameter. To overcome limitations related to the correlations between the site frailty terms and intercepts, we dropped the intercepts from the model first. But after MCMC convergence, we calculated the intercepts using Eq. (16). We conducted the MCMC analysis for 3 chains and 30,000 iterations. The first 5,000 samples were considered as burn-in samples and excluded from the analysis. The remaining 25,000 samples were used for estimating the posterior means and standard deviations. We used the thinning method to ensure the generated samples are random. For this purpose, only every 10-th sample was kept. Reviewing the convergence, auto correlation, kernel density, and tracing plots, the results showed excellent convergence and mixing for MCMC for all parameters in all simulation scenarios. No label switching or multi-modality was found in the posterior distribution which shows the stability of the model.

Table 2: Modeling Results for the Simulated Data

Percentage of Zeros	Parameters		Low Mean ($\bar{y} < 1$)		Moderate Mean ($1 < \bar{y} < 2$)		High Mean ($\bar{y} > 5$)	
			True Value	Estimated[1] Values	True Value	Estimated Values	True Value	Estimated Values
~ 60%	Component 1	β_1	-0.5	-0.534 (0.035)	1	0.981 (0.054)	0.5	0.454 (0.059)
		β_2	-0.5	-0.506 (0.038)	1	0.949 (0.054)	1	0.919 (0.062)
		θ_1	1.5	1.574 (0.056)	6	5.812 (0.469)	11	9.900 (0.836)
		W_1	0.5	0.521 (0.027)	0.5	0.509 (0.012)	0.5	0.492 (0.009)
	Component 2	β_1	0.5	0.445 (0.057)	-0.5	-0.511 (0.031)	1.5	1.521 (0.021)
		β_2	0.5	0.508 (0.058)	-1	-0.990 (0.033)	-1	-1.010 (0.021)
		θ_2	3	2.885 (0.209)	1	0.955 (0.028)	0.5	0.499 (0.010)
		W_2	0.5	0.479 (0.027)	0.5	0.491 (0.012)	0.5	0.508 (0.009)
		ϕ	5	5.612 (1.205)	5	4.327 (0.810)	5	5.340 (0.777)
~ 70%	Component 1	β_1	-1	-0.981 (0.056)	-0.5	-0.567 (0.058)	-1	-0.936 (0.048)
		β_2	1	1.008 (0.049)	1	1.130 (0.072)	2	1.997 (0.057)
		θ_1	5	4.842 (0.353)	8	8.748 (0.906)	11	10.660 (0.858)
		W_1	0.5	0.489 (0.015)	0.5	0.471 (0.012)	0.5	0.506 (0.009)
	Component 2	β_1	1	0.954 (0.039)	1	1.036 (0.029)	2	1.962 (0.034)
		β_2	-0.5	-0.498 (0.034)	-1.5	-1.527 (0.036)	-1.5	-1.478 (0.033)
		θ_2	2.5	2.545 (0.101)	2	2.145 (0.079)	2.5	2.421 (0.098)
		W_2	0.5	0.489 (0.015)	0.5	0.529 (0.012)	0.5	0.494 (0.009)
		ϕ	5	5.395 (1.206)	5	5.002 (1.057)	5	5.139 (0.926)
~ 80%	Component 1	β_1	-1.5	-1.443 (0.093)	-1	-0.928 (0.047)	3	2.942 (0.065)
		β_2	0.5	0.598 (0.053)	2	1.974 (0.058)	1.5	1.434 (0.053)
		θ_1	12.5	11.810 (1.808)	14	13.370 (1.131)	20	19.070 (1.946)
		W_1	0.5	0.508 (0.020)	0.5	0.509 (0.012)	0.5	0.517 (0.022)
	Component 2	β_1	0.5	0.498 (0.063)	1.5	1.496 (0.044)	1.5	1.558 (0.054)
		β_2	-0.5	-0.561 (0.043)	-1.5	-1.478 (0.047)	-0.5	-0.403 (0.052)
		θ_2	3.5	3.690 (0.221)	8	7.572 (0.516)	8	7.714 (0.528)
		W_2	0.5	0.492 (0.020)	0.5	0.491 (0.012)	0.5	0.483 (0.022)
		ϕ	5	5.155 (1.535)	5	5.742 (1.643)	5	4.226 (1.011)
~ 90%	Component 1	β_1	-0.5	-0.509 (0.055)	-[2]	-	-	-
		β_2	1	0.991 (0.061)	-	-	-	-
		θ_1	10	9.466 (0.909)	-	-	-	-
		W_1	0.5	0.501 (0.037)	-	-	-	-
	Component 2	β_1	1	0.940 (0.103)	-	-	-	-
		β_2	-0.5	-0.614 (0.101)	-	-	-	-
		θ_2	25	27.010 (5.365)	-	-	-	-
		W_2	0.5	0.499 (0.037)	-	-	-	-
		ϕ	5	5.452 (3.498)	-	-	-	-

[1]Standard deviation is shown in parenthesis.
[2]Datasets cannot adequately be simulated when both samples means, and percentages of zeros are large

Table 2 summarizes the modeling results for the simulated datasets. The estimated parameters from the fitted FMNB-L model were compared with the true parameters. For all simulation scenarios, both estimated and true values are very close. Also, the estimated values did not include any zero in the 95% confidence interval which indicated the values to be statistically significant. As shown in Table 2, The FMNB-L model estimated the subpopulations for two extreme cases of having numerous zero observations (e.g., $\bar{y} < 1$ and 80% of zero observations or $\bar{y} < 1$ and 90% of zero observations) and has a long tail (e.g., $\bar{y} > 5$ and ~60% zero observations) with good accuracy. Note that small deviations from the true value are unavoidable given that there are almost always some similarities between subpopulations; as subpopulations become more and more distinct, better estimations are expected. In addition, as noted in previous studies (Lord, 2006; Lord & Miranda-Moreno, 2008), the estimation of the inverse dispersion parameter (φ) for the NB distribution is often biased especially for data characterized by a small sample mean. Although the φ has a different interpretation in NB-L or FMNB-L models compared to NB or FMNB due to the existence of Lindley terms, similar inaccuracy in estimation is expected in our simulation study as well.

3.5. Application to Empirical Data

We used an empirical dataset (separated by severity levels) with a long tail (and different crash means, variances, and percentages of zero responses) to demonstrate the application of the FMNB-L to model real crash data and compare the results with other competitive models (i.e.: NB, NB-L, and FMNB). This section is divided into two parts. The first part describes the characteristics of datasets used in this study while the second part presents the modeling results.

Table 3: Characteristics of Texas Four-Lane Freeway Data

Variables	Min	Max	Avg.	S.D.
Multi-Vehicle Fatal Injury Crashes (5 years)	0	135	3.07	7.59
Multi-Vehicle Property Damage Only Crashes (5 years)	0	343	6.94	16.72
Number of Multi-Vehicle Crashes (5 years)	0	478	10.01	23.80
Annual average daily traffic in 5 years (AADT)	1,651	267,131	43,935	27,353
Segment length (in miles) (L)	0.001	5.192	0.304	0.454

3.5.1. Data Description

For this research, we used the multi-vehicle property damage only (PDO) and fatal-injury (FI) crash data collected in 5 years on 4,192 segments of Texas four-lane freeways. This dataset has unique characteristics that are not found in other typical crash datasets. Crash data are highly dispersed and include segments with very large number of crashes while still a significant number of segments did not experience any crashes. Table 3 indicates the summary statistics of the data. We divided the data into three datasets (i.e., multi-vehicle FI crashes, multi-vehicle PDO crashes, and Total PDO and FI multi-vehicle crashes) to evaluate the performance of FMNB-L for a range of crash mean, dispersion, the maximum number of crashes, and percentage of zero observations. The FI crash dataset has a mean of 3.07, and standard deviation of 7.59. It includes many segments with large number of crashes (note the maximum number of 135 crash); yet it includes about 50% of zero observations. The PDO crash dataset has a mean of 6.94 and standard deviation of 16.72. This dataset also includes many segments with very large number of crashes (note the maximum number of 343 crash) while still around 39% of segments did not include any crash. Finally, we considered the combined FI and PDO (i.e., Total) multi-vehicle crashes for the analysis. The total FI and PDO crash dataset has a mean of 10.01, standard deviation of 23.80. In addition, the dataset includes many segments with very large number of crashes (note the maximum number of 478

crash) while still around 33% of segments did not include any crash. The original data only included segment length and traffic flow as variables; considering only these variables simplifies delineating the boundary between potential classes or components for the purpose of this work. Given that all models in the results section are estimated using the same data, the omitted variable bias would not be an issue. We used segment length as an offset and Annual Average Daily Traffic (AADT) as a variable in the model. The segment length varied from 0.001 to 5.192 miles, with an average of 0.304 miles. The AADT data varied from 1,651 to 267,131 with an average of 43,935.

3.5.2. Modeling Results

This section presents the results of application of the FMNB-L to the Texas four-lane datasets described in the previous section. As discussed in greater details below, we selected two components for FMNB-L model due to the existence of the Lindley terms. We also compared the results with the NB, NB-L, and two-components FMNB models. As noted earlier, we used only AADT as a variable. Segment length was considered as an offset. So, we assumed that the mean response of crashes increases linearly as the segment length increases. We used the method explained earlier to overcome the correlation between the intercept and site frailty terms. We dropped the intercept initially from the model. After convergence, we calculated the intercept using Eq. (16). We implemented the models in WinBUGS (Spiegelhalter et al., 2003) and used MCMC for parameter estimations. We ran the MCMC for 3 chains and 30,000 iterations. Similar to the simulation analysis, all models converged well. No label switching or bimodality was observed in the MCMC. After the MCMC, we treated the first 5,000 samples as burn-in and discarded them from posterior estimations. We also used thinning and only considered every 10-th sample in posterior estimations. We used four GOF metrics to evaluate the fit and select the best model. These measures include Log-likelihood (LL), Deviance Information Criterion (DIC) (Geedipally

et al., 2014), Widely Applicable Information Criterion (WAIC), and Leave-One-Out Cross-Validation Information Criterion (LOOIC) (Vehtari et al., 2017; Ahmed et al., 2020; Khodadadi et al., 2021). While the log-likelihood metric does not consider complexity in its estimation and often favors a model with a complex structure, the other three metrics consider complexity in their estimations. Hence, given that the complexities of the NB, NB-L, FMNB, and FMNB-L are not the same, they are more reliable metrics for model comparison in this study.

Tables 4-6 respectively show the modeling results for multi-vehicle FI and PDO and Total multi-vehicle crashes. All estimated AADT coefficients and model parameters are significant at 95% confidence interval. However, there are clear distinctions between the estimated coefficients by different models. Both FMNB and FMNB-L found two subpopulations in the data. It is worth pointing out that, although we assumed only two components in this example, without loss of generality of both FMNB and FMNB-L, finite mixture models can include more than 2 components in modeling. However, due to the Lindley terms in FMNB-L models, it is expected that unlike FMNB, more components are not needed to classify data. By the same token, note that large estimation for inverse dispersion parameters is also expected for NB-Lindley models, as a significant portion of dispersion in these models is captured by Lindley terms.

Table 4: Modeling Results for Texas Four-Lane Freeway Multi-Vehicle FI Crashes

Parameters		NB	NB-L	FMNB-2	FMNB-L-2
		Mean (S.D.)	Mean (S.D.)	Mean (S.D.)	Mean (S.D.)
Intercept (β_0)	Component 1	-13.87 (0.430)	-14.03 (0.470)	-16.21 (0.595)	-15.23 (0.638)
	Component 2	-	-	-10.22 (1.242)	-9.536 (2.464)
Ln (AADT) (β_1)	Component 1	1.516 (0.040)	1.533 (0.044)	1.702 (0.054)	1.625 (0.058)
	Component 2	-	-	1.279 (0.116)	1.244 (0.239)
Inverse Over Dispersion (ϕ)	Component 1	1.068 (0.045)	4.668 (0.715)	1.951 (0.178)	30.53 (12.65)
	Component 2	-	-	1.101 (0.366)	3.696 (3.788)
Lindley Parameter (θ)	Component 1	-	1.754 (0.040)	-	2.135 (0.084)
	Component 2	-	-	-	0.510 (0.120)
Component Weight (W)	Component 1	-	-	0.846 (0.043)	0.923 (0.030)
	Component 2	-	-	0.154 (0.043)	0.077 (0.030)
Log-Likelihood		-6852.7	-5104.3	-6061.8	**-4405.0**[1]
DIC		13714.6	12611.5	13596.5	**12130.0**
WAIC		13712.9	12756.5	13369.7	**12037.6**
LOOIC		13712.9	13211.0	13416.7	**12727.1**

[1]Bold values indicate a better fit.

Table 5: Modeling Results for Texas Four-Lane Freeway Multi-Vehicle PDO Crashes

Parameters		NB	NB-L	FMNB-2	FMNB-L-2
		Mean (S.D.)	Mean (S.D.)	Mean (S.D.)	Mean (S.D.)
Intercept (β_0)	Component 1	-12.46 (0.393)	-12.81 (0.418)	-14.73 (0.481)	-14.2 (0.496)
	Component 2	-	-	-9.379 (1.157)	-8.872 (1.602)
Ln (AADT) (β_1)	Component 1	1.469 (0.037)	1.503 (0.039)	1.640 (0.044)	1.601 (0.046)
	Component 2	-	-	1.291 (0.108)	1.273 (0.152)
Inverse Over Dispersion (ϕ)	Component 1	0.860 (0.028)	2.291 (0.171)	1.726 (0.130)	20.05 (9.437)
	Component 2	-	-	0.778 (0.133)	1.800 (0.888)
Lindley Parameter (θ)	Component 1	-	0.809 (0.017)	-	1.098 (0.038)
	Component 2	-	-	-	0.211 (0.037)
Component Weight (W)	Component 1	-	-	0.827 (0.031)	0.897 (0.025)
	Component 2	-	-	0.173 (0.031)	0.104 (0.025)
Log-Likelihood		-9382.5	-7669.1	-8396.0	**-6180.0**[1]
DIC		18774.9	17321.8	18108.0	**16020.0**
WAIC		18773.6	17765.8	18064.9	**16154.0**
LOOIC		18773.7	18112.9	18140.1	**17038.1**

[1]Bold values indicate a better fit.

Table 6: Modeling Results for Texas Four-Lane Freeway Total Multi-Vehicle Crashes

Parameters		NB	NB-L	FMNB-2	FMNB-L-2
		Mean (S.D.)	Mean (S.D.)	Mean (S.D.)	Mean (S.D.)
Intercept (β_0)	**Component 1**	-12.15 (0.368)	-12.50 (0.385)	-14.23 (0.436)	-13.78 (0.462)
	Component 2	-	-	-9.55 (1.167)	-9.17 (1.558)
Ln (AADT) (β_1)	**Component 1**	1.476 (0.035)	1.509 (0.036)	1.631 (0.040)	1.597 (0.042)
	Component 2	-	-	1.347 (0.111)	1.339 (0.150)
Inverse Over Dispersion (ϕ)	**Component 1**	0.888 (0.027)	2.276 (0.156)	1.689 (0.108)	21.34 (10.37)
	Component 2	-	-	0.740 (0.122)	1.658 (0.756)
Lindley Parameter (θ)	**Component 1**	-	0.583 (0.012)	-	0.796 (0.027)
	Component 2	-	-	-	0.146 (0.025)
Component Weight (W)	**Component 1**	-	-	0.842 (0.026)	0.895 (0.023)
	Component 2	-	-	0.158 (0.026)	0.105 (0.023)
Log-Likelihood		-10584.7	-8782.1	-9560.3	**-7126.1**[1]
DIC		21170.6	19555.9	20479.4	**17747.9**
WAIC		21177.9	20102.1	20391.9	**18209.8**
LOOIC		21177.9	20453.3	20464.4	**19179.4**

[1]Bold values indicate a better fit.

For all estimated models, we calculated the log-likelihood, DIC, WAIC and LOOIC metrics for model selection. The FMNB-L has the best log-likelihood. The NB-L, FMNB and NB are ranked after FMNB-L, respectively. As noted earlier, the log-likelihood metric does not consider complexity in modeling. In terms of complexity, FMNB-L is the most complex model, with NB-L and FMNB coming right after that. The NB model has the least complexity among the four. Therefore, it is not unexpected to observe that the FMNB-L has the best log-likelihood. We reported log-likelihood just for information purposes. However, it is important to compare models using measures that consider complexity. We used three measures to address this issue. First, we considered DIC. The DIC is widely used for comparing models in Bayesian Statistics or crash data modeling (Geedipally et al., 2014). DIC is derived using the following equations:

$$\mathrm{DIC} = \overline{\mathrm{D}(\Theta)} + \rho_{\mathrm{D}}$$

$$\rho_{\mathrm{D}} = \overline{\mathrm{D}(\Theta)} - \mathrm{D}(\overline{\Theta})$$

where Θ and $\bar{\Theta}$ respectively denote model parameters, and posterior estimations of parameters. $D(\bar{\Theta}) = -2LL(\bar{\Theta})$ is the deviance calculated using the posterior estimates, $\overline{D(\Theta)} = E(-2LL(\Theta))$ is expectation of deviance, and LL(.) is log-likelihood. Models with less $\overline{D(\Theta)}$ show a better fit. However, the ρ_D term is used as a penalty term to advocate for models with less complexity.

As shown in Tables 4-6, the FMNB-L shows a clear superiority over other models in terms of DIC values. For the FI crash dataset (with mean=3.07, S.D.=7.59, max crash=135, and percentage of zeros=50%), the DIC value for FMNB-L model is 12130 which is ranked the best. The NB-L (with DIC= 12611.5), FMNB (with DIC=13596.5), and NB (with DIC=13714.6) models are ranked in sequence after the FMNB-L model. For the PDO crash dataset (with mean=6.94, S.D.=16.72, max crash=343, and percentage of zeros=39%), the DIC value for FMNB-L model is 16020 which is ranked as the best model with a clear superiority. The NB-L (with DIC=17321.8), FMNB (with DIC=18108), and NB (with DIC=18774.9) models are ranked in sequence after the FMNB-L model. Lastly, for the Total (PDO+FI) multi-vehicle dataset (with mean=10.01, S.D.= 23.80, max crash=478, and percentage of zeros=33%), the FMNB-L model with a DIC of 17747.9 is ranked as the best model with significantly better DIC. For this dataset, the NB-L (with DIC=19555.9), FMNB (with DIC=20479.4), and NB (with DIC=21170.6) models are ranked in sequence after the FMNB-L model.

One notable observation is that the NB-L model also shows a better fit compared to the FMNB model. However, although the NB-L fits the data better than the FMNB, the results clearly show that two subpopulations exist. Hence, if we consider the concept of "goodness-of-logic" as a selection criterion (Shirazi et al., 2017a; Shirazi & Lord, 2019; Lord et al., 2021), the FMNB should be selected over the NB-L. In other words, we know that data are drawn from a heterogeneous population given that two classes of subpopulations are identified by the FMNB

and FMNB-L models; hence, given the "goodness of logic" concept, we should only consider finite mixture models (i.e., FMNB-L and FMNB) for model comparisons.

We also used two new GOF statistics (WAIC and LOOIC) to evaluate the proposed models. Both measures consider the model complexity in their estimations as well. Vehtari et al. (2017) noted that these measures are more reliable that the DIC. In fact, WAIC and LOOIC can be considered an improvement over the DIC metric. For example, one limitation of the DIC is producing negative outcomes for the number of parameters in some situations. WAIC overcomes this limitation. As noted in (Geedipally et al., 2014), the DIC is also sensitive to parametrizations. WAIC, on the other hand, is invariant to parametrizations. LOOIC is a robust version of the WAIC; it works better for models with weak prior information. The detailed steps for derivations and calculations of the WAIC and LOOIC can be found in the work of Vehtari et al. (2017).

The results show the same trends as observed for the DIC measure. The FMNB-L model consistently shows the best fit among the four models. For the Multi-vehicle FI crash dataset, the values of the WAIC and LOOIC measures for the FMNB-L mode are 12037.6 and 12727.1 respectively. In sequence, after the FMNB-L model, the NB-L (with WAIC=12756.5, and LOOIC=13211.0), FMNB (with WAIC=13369.7 and LOOIC=13416.7), and NB (with WAIC=13712.9 and LOOIC=13712.9) models show the best goodness of fit. For the Multi-vehicle PDO crashes, the values of WAIC and LOOIC are 16154.0 and 17038.1 respectively. The NB-L (with WAIC=17765.8 and LOOIC=18122.9) and FMNB (with WAIC=18064.9 and LOOIC=18140.1) and NB (with WAIC=18773.6 and LOOIC=18773.7) are ranked after FMNB-L in sequence. Lastly, for the Total (PDO+FI) multi-vehicle crash dataset, the WAIC of 18209.8 and LOOIC of 19179.4 selected the FMNB-L as the best model. The NB-L (with WAIC=20102.1

and LOOIC=20453.2), FMNB (with WAIC=20391.9 and LOOIC=20464.4), and NB (with WAIC=21177.9 and LOOIC=2177.9) models were ranked after FMNB-L subsequently.

As noted earlier, it is possible to find the association probabilities of each component for each observation in the dataset, using the MCMC information. We used the procedure described in Eq. (17) and Eq. (18) to find the association probabilities. Table 7 shows the association probabilities for the 15 sites with the highest number of crashes for the three datasets analyzed in this section. For example, in the multi-vehicle FI crash model, site 3 belongs to components 1 and 2 with 91.8% and 8.2% probabilities respectively. As another example, in the multi-vehicle PDO crashes model, site 13 with 147 Multi-Vehicle PDO crashes belongs to component 1 with 27.6%, and component 2 with 72.4% probability.

Table 7: Probabilities of Components for 15 Observations with Highest Number of Total Multi-Vehicles Crashes for the FMNB-L Model

Data	Models								
	Multi-Vehicle FI			Multi-Vehicle PDO			Total Multi-Vehicle		
Site	Number of Crashes	Components Probability		Number of Crashes	Components Probability		Number of Crashes	Components Probability	
		p_1	p_2		p_1	p_2		p_1	p_2
1	135	29.2%	70.8%	343	7.4%	92.6%	478	7.6%	92.4%
2	102	91.2%	8.8%	181	91.4%	8.6%	283	90.2%	9.8%
3	77	91.8%	8.2%	202	84.6%	15.4%	279	88.2%	11.8%
4	68	92.6%	7.4%	200	85.1%	14.9%	268	87.1%	12.9%
5	90	0.9%	99.1%	170	1.2%	98.8%	260	0.6%	99.4%
6	70	96.0%	4.0%	159	94.4%	5.6%	229	94.4%	5.6%
7	43	94.4%	5.6%	179	87.1%	12.9%	222	89.6%	10.4%
8	47	92.0%	8.0%	164	83.4%	16.6%	211	85.5%	14.5%
9	65	94.0%	6.0%	138	93.4%	6.6%	203	94.6%	5.4%
10	84	87.9%	12.1%	119	91.9%	8.1%	203	89.2%	10.8%
11	92	90.6%	9.4%	103	93.6%	6.4%	195	91.5%	8.5%
12	67	93.9%	6.1%	125	93.0%	7.0%	192	93.0%	7.0%
13	36	85.1%	14.9%	147	27.6%	72.4%	183	42.3%	57.7%
14	76	95.0%	5.0%	100	94.2%	5.8%	176	95.4%	4.6%
15	46	95.8%	4.2%	129	93.6%	6.4%	175	94.3%	5.7%

Recently the Random Parameters Negative Binomial-Lindley (RPNB-L) model has also been proposed to account for unobserved heterogeneity and many zero observations (Rusli et al., 2018; Shaon et al., 2018). We decided to compare this model with the finite mixture model proposed in this study. Table 8 shows the modeling results for the RPNB-L and the FMNB-L (previously documented in Tables 4-6). As noted above, given that the modeling results indicated two subpopulations in data, FMNB and FMNB-L are recommended to be used instead of the single component NB, NB-L, or RPNB-L models regardless of the GOF measures. This supports the discussions found in Miaou & Lord (2003), Shirazi et al. (2017a), Shirazi & Lord (2019), and Lord et al., (2021) that the selection of the models should also be based on the data generation process. However, as shown in Table 8, the FMNB-L model shows a better statistical performance compared to the RPNB-L as well.

Table 8: Comparison between the FMNB-L and RPNB-L

Parameters	Component	Models					
		Multi-Vehicle FI		Multi-Vehicle PDO		Total Multi-Vehicle	
		FMNB-L	RPNB-L	FMNB-L	RPNB-L	FMNB-L	RPNB-L
Mean of Parameters							
Intercept (β_0)	Component 1	-15.23 (0.638)[1]	-13.97 (0.486)	-14.2 (0.496)	-12.8 (0.431)	-13.78 (0.462)	-12.61 (0.397)
	Component 2	-9.536 (2.464)	-	-8.872 (1.602)	-	-9.17 (1.558)	-
Ln (AADT) (β_1)	Component 1	1.625 (0.058)	1.527 (0.045)	1.601 (0.046)	1.501 (0.040)_	1.597 (0.042)	1.519 (0.037)
	Component 2	1.244 (0.239)	-	1.273 (0.152)	-	1.339 (0.150)	-
Standard Deviation of Random Parameters							
Ln (AADT) (β_1)	Component 1	-	0.238 (0.064)	-	0.247 (0.045)	-	0.258 (0.070)
	Component 2	-	-	-	-	-	-
Goodness of Fits							
Log-Likelihood		**-4405.0**[2]	-5076.3	**-6180.0**	-7620.6	**-7126.1**	-8718.5
DIC		**12130.0**	12513.8	**16020.0**	17278.7	**17747.9**	19523.1
WAIC		**12037.6**	12714.3	**16154.0**	17724.5	**18209.8**	20056.6
LOOIC		**12727.1**	13194.2	**17038.1**	18090.7	**19179.4**	20436.3

[1]The number in parenthesis is the standard deviation (S.D.) of the estimate.
[2]Bold values indicate a better fit.

As a closing note to this section, it is worth pointing out that we analyzed 5 years of aggregated data. If disaggregated data (e.g., yearly or monthly observations) were used in modeling, greater unobserved heterogeneity could exist in the dataset. This larger unobserved heterogeneity could affect the mixture probabilities or number of latent classes in finite mixture models. In addition, the disaggregated dataset could also include larger number of zero observations; the excess number of zeros could better be modeled with the FMNB-L than with the FMNB.

3.6. Summary and Conclusions

Crash data are often drawn from heterogeneous locations, with different populations, environments, and geographic patterns. Furthermore, crash data may also include many zero observations or have a long tail. The typical statistical models (e.g., the NB model) cannot model these data properly. In this research, we proposed the finite mixture NB-L model to account for unobserved heterogeneity due to latent subpopulations in data with many zero observations or long tails. We designed and used a simulation analysis to evaluate the performance of the FMNB-L model in identifying subpopulations under different ranges of sample means and zero percentages. The results show that the FMNB-L can reasonably identify the subpopulations and account for large percentages of zero observations. We also used the FMNB-L to model crash data for three datasets collected for four-lane freeways in Texas (all characterized by high dispersion and a long tail) and compared the results with other models (i.e., NB, NB-L, and FMNB). We used the DIC, WAIC, and LOOIC as model selection metrics to compare the GOF of the models. All these metrics consider the complexity of models in their estimations. The GOF statistics show that the FMNB-L model fits the data significantly better than the NB, NB-L, and FMNB models. As, discussed in previous work (Miaou & Lord, 2003; Shirazi et al., 2017a; Shirazi & Lord, 2019),

GOF should not be the sole factor in selecting a model. In the datasets used for this work, there is also clear evidence that subpopulations exist. Overall, the modeling results show the robustness of the proposed model in addressing the issues of subpopulation heterogeneity and excess number of zero observations in crash data analysis. To simplify the application analysis, we used flow-only datasets to demonstrate the application of the model. Hence, further research is recommended to explore the FMNB-L model with more independent variables. In addition, future research should explore the application of the FMNB-L model with varying weight parameters, or explore the performance of the model in simulated scenarios with different skewness, or different numbers of independent variables in each component.

CHAPTER 4

GROUPED RANDOM PARAMETERS NEGATIVE BINOMIAL-LINDLEY

4.1. Introduction

According to World Health Organization (WHO) (2018), the worldwide death toll due to traffic crashes has reached 1.35 million people per year, equating to 3,700 deaths every day. Traffic crashes have risen from ninth to the eighth position on the list of the world's top leading causes of death according to statistical data. Over the previous century, around 3.8 million Americans have died in traffic crashes. The National Highway Traffic Safety Administration (NHTSA) reported that 38,824 people were killed in traffic crashes in the United States in 2020 (Stewart, 2022). Maine Department of Transportation (MaineDOT) (2020) stated that there had been 28,746 reported traffic crashes in Maine in 2020, with 150 of those being fatal. This devastating cause of death has claimed the lives of people all over the world. As a result, the improvement of traffic safety has become a major concern for transportation safety analysts all over the world. Predicting crashes and identifying the key explanatory variables behind these crashes are of utmost importance to improve highway safety. Researchers have spent a significant amount of time in the past developing robust statistical models to analyze crash data (Lord & Mannering, 2010; Savolainen et al., 2011; Mannering & Bhat, 2014; Mannering et al., 2016). These statistical models can be used to estimate the number of crashes, identify crash contributing factors, or locate high-risk crash locations.

Overdispersion (i.e., variance greater than mean) is a common feature often found in crash data. The negative binomial (NB) model is the most popular model to address overdispersion in crash data (Lord & Mannering, 2010). Data with many zero observations is another characteristic

found in crash data. When crash datasets contain a large amount of zero observations, the NB model cannot be estimated properly. To overcome this limitation, a mixture of the NB with other distributions has been proposed by several studies to provide flexibility in capturing the large number of zero observations in crash data (Shirazi et al., 2016a). The NB-Lindley (NB-L) generalized linear model (Geedipally et al., 2012) is one of the most popular models in this category. Recently a few more advanced variations of this model have also been proposed and found superior to the NB model for datasets containing excess zero observations (Shaon et al., 2018; Rusli et al., 2018; Tang et al., 2020; Behara et al., 2021; Islam et al., 2022; Khodadadi et al., 2022a; Khodadadi et al., 2022b).

Unobserved heterogeneity may also exist in the data or model, especially due to variations in temporal and spatial characteristics among groups of observations. As a result, the explanatory variables may not have the same effect on all segments or regions in the network. In fact, the effect of various variables such as skid number (for pavement friction), driver behavior, climate, surface condition, and weather characteristics may vary substantially across different groups of observations (e.g., regions). For example, the impact of weather factors such as rainfall or snow, as one of the key contributing factors in lane departure crashes, can vary significantly from one location to another due to variations in terrain, climate, or other location characteristics. Researchers have found that weather variables such as rainfall, precipitation, fog, visibility, wind speed, snow, temperature, etc. have mixed effects on crash occurrence (Qiu & Nixon, 2008; El-Basyouny et al., 2014; Theofilatos & Yannis, 2014; Zhao et al., 2019; Sawtelle et al., 2022). Some of these variables positively interact with crashes, whereas some of them have negative interactions with crashes and the effect is not similar even in two nearby regions or locations in the network. Several researchers have proposed Grouped Random Parameters models to address

the unobserved heterogeneity due to spatial or temporal variations in different groups of observations in data (Cai et al., 2018; Heydari et al., 2018), for example, across different regions of a state. These models showed better capabilities to account for unobserved heterogeneity and consequently better estimation of the model coefficients.

Several studies used the grouped random parameters model to analyze crash frequency. As such, Cai et al. (2018) proposed a grouped random parameters multivariate spatial model to study the observed zonal effects and unobserved heterogeneity at the zonal level on crash count data. This study considered traffic data and socio-demographic information as zonal factors, which have significant effects on crashes. Heydari et al. (2018) introduced a Grouped Random Parameters approach to benchmark different geographic regions based on crash frequency. This study implemented a heteroskedastic grouped random parameters Poisson lognormal model with heterogeneity in mean and variance to address unobserved heterogeneity and provided important guidelines to grade crossing safety analysis in Canada. Another crash frequency study incorporating Grouped Random Parameters approach was implemented by Li et al. (2018). This study proposed a grouped random parameters negative binomial model to study the relationship between the level of service (LOS) and traffic safety. They also developed a bivariate grouped random parameters negative binomial model to analyze rear-end and left-turn crashes. In another study, Fountas et al. (2018a) implemented a dynamic correlated grouped random parameters binary logit model to study the mixed effects of both non-time varying and time-varying explanatory variables and address unobserved heterogeneity in crash data.

Grouped Random parameters modeling approach has also been implemented in crash severity analysis. As such, the effect of the presence of trucks of different classes on non-truck-related crash severity was studied by Fanyu et al. (2021). This study proposed a correlated grouped

random parameters binary logit model to account for unobserved heterogeneity at both observation level and space-time level. This study accounted for temporal instability because driver's behavior, risk perceptions, weather conditions, vehicle technology, and socio-economic conditions vary with time. Fountas et al. (2018b) implemented a correlated random parameters ordered probit model to address unobserved heterogeneity and account for interaction among observed or unobserved characteristics. Grouped Random parameters approach has also been implemented to study aggressive driver behavior by Sarwar et al. (2017). This study implemented a grouped random parameters bivariate probit model to investigate perceived and observed aggressive driving behavior based on surveys and driving simulation experiments. This study also addressed cross-equation error correlation among the dependent variables, panel effects, and other unobserved factors that may vary systematically across the participants. Grouped Random parameters approach has been used in Pedestrian safety studies too (Pantangi et al., 2021). Pedestrians are one of the most vulnerable road users. Pedestrian-involved accidents lead to fatal accidents most of the time compared to other motorist-involved accidents. High Visibility Crosswalks (HVC) play a vital role in improving pedestrian safety. The study was designed to evaluate the efficacy of HVCs in improving pedestrian safety and evaluate their potential to modify driving behavior.

This research was motivated to overcome two modeling limitations in crash data analysis, first addressing data with many zero observations, and second, accounting for unobserved heterogeneity in crash data across group of observations, using the grouped RP paradigm. In this chapter, we propose the derivations and characteristics of the grouped random parameters negative binomial-Lindley (G-RPNB-L) model, to account for unobserved heterogeneity in groups of observations in crash data while addressing the issue of excess zero observations. The model is first illustrated using a simulation study. Then, the application of the model is demonstrated using

an empirical dataset collected on rural Interstates in Maine. This dataset includes over 90% of zeros. Weather in Maine varies significantly from region to region, county to county, and even within towns, mainly due to vast variations in terrain or climate in Maine. Using data collected in Maine, we illustrated the variations in the effect of monthly weather variations on crashes in different counties in Maine. The proposed model was compared with the NB, the grouped random parameters NB (G-RPNB), and the NB-L models based on different Goodness-of-Fit (GOF) metrics, and results were discussed. The next section briefly documents the characteristics of the NB and Grouped Random Parameters NB models.

4.2. Grouped Random Parameters Negative Binomial

This research documents the derivations of the grouped random parameters negative binomial-Lindley (G-RPNB-L), and its characteristics to model crash data. To better explain the formulation process, however, let us start with a brief discussion on the formulations of the typical NB and the grouped RPNB models. The NB model is a widely used model to address overdispersion in crash data. The NB generalized linear model (NB-GLM) is defined as follows (Hilbe, 2011; Cameron & Trivedi, 2013):

$$y_i|\ \mu_i, \varphi \sim \mathrm{NB}\ (\mu_i, \varphi) \equiv \frac{\Gamma(y_i + \varphi)}{\Gamma(y_i + 1) \times \Gamma(\varphi)} \left(\frac{\mu_i}{\mu_i + \ \varphi}\right)^{y_i} \left(1 - \frac{\mu_i}{\mu_i + \ \varphi}\right)^{\varphi};\ \varphi > 0, \mu_i > 0 \qquad (20\text{-}1)$$

$$\ln(\mu_i|\beta_0, \beta_1, \ldots, \beta_M) = \ \beta_0 + \sum_{j=1}^{M} \beta_j X_{ij} \qquad (20\text{-}2)$$

Where φ denotes the inverse overdispersion parameter, and μ_i denotes the long-term mean for the i-th site; Eq. (20-2) defines the natural log of the long-term mean as a linear function of "M" covariates (denoted by X), and coefficients (denoted by β).

The NB model, however, does not account for unobserved heterogeneity adequately (Anastasopoulos & Mannering, 2009; Mannering et al., 2016; Behnood & Mannering, 2017; Zamenian et al., 2017; Shaon et al., 2018). Random Parameters (RP) models are a class of models that account for unobserved heterogeneity by allowing parameters to vary from one observation to another. The random parameters negative binomial (RPNB) model is one of the typical models proposed by researchers to address unobserved heterogeneity. This model allows the coefficients of the NB model to vary among different observations. The grouped RP models (G-RP) are a special case of RP models where the model coefficients vary from one group of observations to another (e.g., from one region in the network to another) (Mannering et al., 2016; Meng et al., 2017; Sarwar et al., 2017). Let us assume the model includes M covariates with fixed parameters (denoted by X) and M' covariates with varying parameters that change from one group of observations to another (denoted by Z). Likewise, let us assume "K" groups of observations in data. In addition, let the symbol k(i) denotes the group of observations that the i-th site is associated with (e.g., the i-th segment is part of the k-th region). The G-RPNB model, with an intercept that also varies among groups of observations, can be formulated as the following hierarchical Bayesian model:

$$y_i|\ \mu_i, \varphi \sim \mathrm{NB}\ (\mu_i, \varphi) \tag{21-1}$$

$$\ln\left(\mu_i|\beta_1, \dots, \beta_M, \Upsilon_{0,k(i)}, \Upsilon_{1,k(i)}, \dots, \Upsilon_{M',k(i)}\right) = \Upsilon_{0,k(i)} + \sum_{j=1}^{M} \beta_j X_{ij} + \sum_{j=1}^{M'} \Upsilon_{j,k(i)} Z_{ij} \tag{21-2}$$

$$\Upsilon_{0,k(i)}|\ \mu_0, \sigma_0 \sim \mathrm{N}(\mu_0, \sigma_0) \tag{21-3}$$

$$\Upsilon_{j,k(i)}|\ \mu_j, \sigma_j \sim \mathrm{N}(\mu_j, \sigma_j);\ \forall j \in \{1, \dots, M'\} \text{ and } \forall k \in \{1, \dots, K\} \tag{21-4}$$

where,

β_j = The fixed coefficient for the j-th fixed-parameters covariate.

X_{ij} = The value of the j-th fixed-parameters covariate at the i-th site.

$\Upsilon_{0,k(i)}$ = The grouped random intercept for the k-th group of observations.

$\Upsilon_{j,k(i)}$ = The grouped random coefficient for the k-th group of the j-th covariate.

Z_{ij} = The value of the j-th random-parameters covariate at the i-th site.

μ_0 = The intercept mean.

σ_0 = The intercept standard deviation.

μ_j = The mean of random parameters for the j-th random-parameters covariate.

σ_j = The standard deviation of random parameters for the j-th random-parameters covariate.

4.3. Grouped Random Parameters Negative Binomial-Lindley

In this section, the derivations and characteristics of the grouped random parameters Negative Binomial- Lindley model is documented. Let us first introduce the NB-L model. The NB-L model is written as the following mixture model (Geedipally et al., 2012):

$$\text{NB-L}(\mu_i, \varphi, \theta) \equiv P(Y = y_i | \mu_i, \varphi, \theta) = \int \text{NB}(y_i | \varepsilon_i \mu_i, \varphi) \text{Lindley}(\varepsilon_i | \theta) \, d\varepsilon_i \quad (22)$$

The above mixture model can be revised as a hierarchical Bayesian model. Note that the Lindley distribution with parameter "θ" can be written as the following hierarchical structure:

$$\varepsilon_i|z_i,\theta \sim \text{gamma}(1+z_i,\theta) \tag{23-1}$$

$$z_i|\theta \sim \text{Bernoulli}\left(\frac{1}{1+\theta}\right) \tag{23-2}$$

Therefore, the multi-level hierarchical representation of the NB-L model can be given as follows (Geedipally et al., 2012):

$$y_i|\ \varepsilon_i\mu_i,\ \varphi \sim \text{NB}\ (\varepsilon_i\mu_i,\varphi) \tag{24-1}$$

$$\varepsilon_i|z_i,\theta \sim \text{gamma}(1+z_i,\theta) \tag{24-2}$$

$$z_i|\theta \sim \text{Bernoulli}\left(\frac{1}{1+\theta}\right) \tag{24-3}$$

$$\ln(\mu_i|\beta_0,\beta_1,\dots,\beta_M) = \beta_0 + \sum_{j=1}^{M}\beta_j X_{ij} \tag{24-4}$$

Researchers have shown that the NB-L model provides additional flexibility to address the issue of the excess number of zero responses. However, the coefficients of the covariates may vary from one group of observations to another, due to unobserved heterogeneity, as described above. In addition, the Lindley term may also vary among groups of observations. For example, different regions may include different percentages of zero observations. Hence, different Lindley terms might also be needed, to account for the number of zeros in different regions, instead of one unique or universal term. Keeping that in mind, the G-RPNB-L model with varying coefficients and Lindley terms across groups of observations can be defined as the following hierarchical model.

$$y_i|\ \varepsilon_i\mu_i,\ \varphi \sim \text{NB}\ (\varepsilon_i\mu_i,\varphi) \tag{25-1}$$

$$\varepsilon_i|z_{k(i)},\theta_{k(i)} \sim \text{gamma}\left(1+z_{k(i)},\ \theta_{k(i)}\right) \tag{25-2}$$

$$z_{k(i)}|\theta_{k(i)} \sim \text{Bernoulli}\left(\frac{1}{1+\theta_{k(i)}}\right) \tag{25-3}$$

$$\ln\left(\mu_i|\beta_1,\dots,\beta_M,\gamma_{0,k(i)},\gamma_{1,k(i)},\dots,\gamma_{M',k(i)}\right) = \gamma_{0,k(i)} + \sum_{j=1}^{M}\beta_j X_{ij} + \sum_{j=1}^{M'}\gamma_{j,k(i)}Z_{ij} \tag{25-4}$$

$$\Upsilon_{0,k(i)}|\,\mu_0,\sigma_0 \sim N(\mu_0,\sigma_0) \tag{25-5}$$

$$\Upsilon_{j,k(i)}|\,\mu_j,\sigma_j \sim N\left(\mu_j,\sigma_j\right);\ \forall j \in \{1,\dots,M'\} \text{ and } \forall k \in \{1,\dots,K\} \tag{25-6}$$

where,

β_j = The fixed coefficient for the j-th fixed-parameters covariate.

X_{ij} = The value of the j-th fixed-parameters covariate at the i-th site.

$\Upsilon_{0,k(i)}$ = The grouped random intercept for the k-th group of observations.

$\Upsilon_{j,k(i)}$ = The grouped random coefficient for the k-th group of the j-th covariate.

Z_{ij} = The value of the j-th random-parameters covariate at the i-th site.

μ_0 = The intercept mean.

σ_0 = The intercept standard deviation.

μ_j = The mean of random parameters for the j-th random-parameters covariate.

σ_j = The standard deviation of random parameters for the j-th random-parameters covariate.

$\theta_{k(i)}$= The Lindley parameter for the k-th group of observations.

Let us assume a normal prior on fixed parameters coefficients (β), a gamma prior on inverse dispersion parameter (φ), and a uniform prior on parameters $1/1+\theta_{k(i)}$. In addition, let us assume a normal prior on μ_0 and μ_j and a gamma prior on $1/\sigma_0$ and $1/\sigma_j$. Then, the above hierarchical Bayesian model can be implemented in WinBUGS (Spiegelhalter et al., 2003) for inference of parameters using the Monte Carlo Markov Chain (MCMC) approach.

As a closing note to this section, it is worth noting that in the above formulation, there are correlations between the grouped random intercepts ($\gamma_{0,k(i)}$) and the Lindley terms (ε_i) which could result in poor MCMC convergence or mixing. However, there are two ways to overcome this limitation. First, it is possible to use informative priors on Lindley terms in a way that ensures $E(\varepsilon_i) = 1$ (Geedipally et al., 2012; Shaon et al., 2018). Second, it is also possible to drop the intercepts from the model and then calculate the grouped intercepts from Lindley terms after convergence using Eq. (26) (Shirazi et al., 2016a; Islam et al., 2022).

$$\Upsilon_{0,k(i)} = E\left(\log\left(E\left(\varepsilon_{i,k}\right)\right)\right) = E\left(\log\left(\frac{\theta_k + 2}{\theta_k\,(\theta_k + 1)}\right)\right) \tag{26}$$

Therefore, $\Upsilon_{0,k(i)}$ can be calculated using the MCMC samples at no additional computational expenses. For this purpose, a sample is also drawn from the posterior of the $\log\left(\frac{\theta_k+2}{\theta_k\,(\theta_k+1)}\right)$ in each MCMC iteration. The average of these samples, then, can be reported as the group intercept.

4.4. Simulation Study

This section documents a simulation study to evaluate the accuracy of the proposed G-RPNB-L model in estimating the grouped random parameters. This section is divided into two parts. The

first section describes the simulation protocol to generate scenarios. The second section illustrates the results of the simulation study.

4.4.1. Simulation Protocol

Several studies have used simulation to demonstrate the applicability of a theory, document the strength or weaknesses of a model, or provide recommendations and guidelines (Lord, 2006; Shirazi et al., 2016b; Shirazi et al., 2017a; Shirazi et al., 2021; Bhowmik et al., 2021; Islam et al., 2022; Khodadadi et al., 2022a). We designed a simulation study to evaluate the performance of the proposed model in estimating grouped random parameters for data with excess zero observations. We simulated several scenarios with different percentages of zero observations ranging from 50% to 90%. Without loss of generality, we assumed three groups of observations in our study, each also with different percentages of zero observations. Let us assume the simulated dataset includes 9,000 records. We first simulated two fixed-parameters covariates (X_1 and X_2), and two random-parameters covariates (Z_1 and Z_2) from standard normal distributions with a mean of zero and standard deviation of one. Let us assume β_1 and β_2 are the regression coefficients for the fixed-parameters covariates. Likewise, let us assume $\gamma_{1,1}$, $\gamma_{1,2}$, and $\gamma_{1,3}$ are the grouped random coefficients for random-parameters covariate denoted as Z_1, and $\gamma_{2,1}$, $\gamma_{2,2}$, and $\gamma_{2,3}$ are the grouped random coefficients for the random-parameters covariate denoted as Z_2. The Lindley terms (ε_i) were also simulated from three Lindley distributions with parameters of θ_1, θ_2, and θ_3 representing the three groups of observations (e.g., regions) in the data. Then, the sample mean (μ_i) was calculated using Eq. (25-4), and data were simulated from the NB distribution using Eq. (25-1). Finally, the simulated data were used to fit the model and estimate the coefficients. The estimated coefficients were compared with their true value. The step-by-step instruction of the simulation protocol is described in the following:

Step 1. Initializations.

1.1. Set the value of β_1 and β_2 to represent the coefficients of the first and second fixed-parameters covariates.

1.2. Set the value of $\gamma_{1,1}, \gamma_{1,2}$, and $\gamma_{1,3}$ to represent the random coefficients for the first grouped random-parameters covariate, and $\gamma_{2,1}, \gamma_{2,2}$, and $\gamma_{2,3}$ to represent the random coefficients for the second grouped random-parameters covariate.

1.3. Set three Lindley parameters θ_1, θ_2, and θ_3 to represent three groups of observations.

1.4. Set the value of inverse dispersion parameter (φ).

Step 2. Simulate Covariates.

2.1. Simulate the fixed-parameters covariates (X_{i1}, X_{i2}) from standard normal distributions as follows (with 9,000 samples):

$$X_{i1} \sim N(0,1); \quad i = 1, \dots, 9000$$

$$X_{i2} \sim N(0,1); \quad i = 1, \dots, 9000$$

2.2. Simulate the random-parameters covariates (Z_{i1}, Z_{i2}) from standard normal distributions as follows (with 9,000 samples):

$$Z_{i1} \sim N(0,1); \quad i = 1, \dots, 9000$$

$$Z_{i2} \sim N(0,1); \quad i = 1, \dots, 9000$$

Let the first 3,000 samples from Z_{i1}, and Z_{i2} belong to the first group, the second 3,000 samples belong to the second group and the third 3,000 samples belong to the third group of observations in the data.

Step 3. Simulate the Lindley terms and calculate the mean.

3.1. Simulate the Lindley terms for three groups of observations from Lindley distributions with parameters θ_1, θ_2, and θ_3 as follows:

$$\varepsilon_{i1} \sim \text{Lindley}(\theta_1); \qquad i = 1, \ldots, 3000$$

$$\varepsilon_{i2} \sim \text{Lindley}(\theta_2); \qquad i = 3001, \ldots, 6000$$

$$\varepsilon_{i3} \sim \text{Lindley}(\theta_3); \qquad i = 6001, \ldots, 9000$$

3.2. Calculate the mean from the regression coefficients and simulated covariates using the following equations:

$$\mu_{i1} = \exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_{1,1} Z_{i1} + \gamma_{2,1} Z_{i2}); \qquad i = 1, \ldots, 3000$$

$$\mu_{i2} = \exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_{1,2} Z_{i1} + \gamma_{2,2} Z_{i2}); \qquad i = 3001, \ldots, 6000$$

$$\mu_{i3} = \exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_{1,3} Z_{i1} + \gamma_{2,3} Z_{i2}); \qquad i = 6001, \ldots, 9000$$

Step 4. Simulate Crash observations.

4.1. Simulate 3,000 observations for each group of observations from the NB distributions as follows:

$$y_{i1} \sim \text{NB}(\varepsilon_{i1}\mu_{i1}, \varphi); \qquad i = 1, \ldots, 3000$$

$$y_{i2} \sim \text{NB}(\varepsilon_{i2}\mu_{i2}, \varphi); \qquad i = 3001, \ldots, 6000$$

$$y_{i3} \sim \text{NB}(\varepsilon_{i3}\mu_{i3}, \varphi); \qquad i = 6001, \ldots, 9000$$

4.2. Combine the simulated data to create a population dataset with 9,000 records.

Step 5. Fit the Model.

5.1. Use the G-RPNB-L model (Eq. 25) to fit the model to the simulated dataset and estimate the coefficients.

4.4.2. Simulation Results

This section illustrates the results of the simulation study. As noted earlier, we designed the simulation for a range of data with different percentages of zero observations. We controlled over the regression coefficients, and Lindley parameters to ensure simulating highly dispersed datasets with approximately 50%, 60%, 70%, 80%, and 90% of zero observations. We also tried to ensure that different regions constituted different percentages of zero observations. Table 9 indicates the characteristics of the simulated dataset.

The mean and standard deviation of the simulated datasets varied from 0.3 to 6.4 and from 2.3 to 37.1 respectively. To ensure simulating scenarios where the percentage of zeros varies across different groups of observations, we controlled the percentage of zeros across the three groups of observations. As shown in Table 9, the dataset with approximately 50% zero observations respectively had 29%, 56%, and 69% zero observations in groups 1-3. For the dataset with 60% of zero observations, the three groups had 44%, 60%, and 74% of zero observations respectively. The dataset with approximately 70% zero observations respectively had 57%, 72%, and 81% zero observations in groups 1-3. In the case of 80% of zero responses, groups 1 to 3 had 73%, 79%, and 85% zero responses respectively. For the dataset with 90% of zero responses, groups 1 to 3 had 88%, 87%, and 93% zeros respectively. The values of the regression coefficients and inverse dispersion parameter used in each simulation scenario are shown in Table 9, denoted as “true” values.

Table 9: Characteristics of Simulated Data

Total Percentage of Zeros	Mean	Standard Deviation	Percentage of Zeros in Groups		
			Group 1	Group 2	Group 3
~ 50%	6.4	37.1	29%	56%	69%
~ 60%	3.2	15.3	44%	60%	74%
~ 70%	1.5	8.5	57%	72%	81%
~ 80%	0.7	3.6	73%	79%	87%
~ 90%	0.3	2.3	88%	87%	93%

We implemented the G-RPNB-L model in WinBUGS software (Spiegelhalter et al., 2003) and estimated the coefficients using the MCMC approach. To ensure proper convergence, we considered 30,000 MCMC iterations with 3 chains. The results of the first 5,000 posterior samples were considered as burn-in samples and discarded from the analysis. We used various diagnostic tools to evaluate the MCMC; all metrics showed excellent convergence for the model. However, to ensure adequate mixing in the MCMC experiment, and remove any autocorrelation between simulated samples, we only considered every 10-th sample in the analysis (i.e., we considered a thinning of 10). We then estimated the posterior mean of the parameters using the remaining samples. Table 10 shows the modeling results for the simulated datasets. As shown in this table, all estimated parameters were significant at 95% confidence level and close to the true value. Most importantly, both group coefficients as well as Lindley parameters (θ_1, θ_2, and θ_3) were estimated with high precision using the G-RPNB-L model.

Table 10: Modeling Results for the Simulated Data

Parameters		Percentage of Zeros									
		~ 50%		~ 60%		~70%		~ 80%		~ 90%	
		True Value	Est. Value[1]	True Value	Est. Value	True Value	Est. Value	True Value	Est. Value	True Value	Est. Value
Fixed Parameters											
β_1		-0.5	-0.495 (0.015)[2]	-0.5	-0.490 (0.018)	-0.5	-0.511 (0.020)	-0.5	-0.466 (0.025)	-0.5	-0.461 (0.033)
β_2		1	1.006 (0.016)	1	1.004 (0.018)	1	1.004 (0.021)	1	0.980 (0.026)	1	1.036 (0.035)
Grouped Random Parameters											
γ_1	**Group 1**	0.5	0.496 (0.020)	0.5	0.519 (0.025)	0.5	0.501 (0.029)	0.5	0.469 (0.038)	0.5	0.604 (0.056)
	Group 2	1	0.942 (0.029)	1	1.003 (0.032)	1	0.973 (0.038)	1	1.003 (0.044)	1	1.002 (0.057)
	Group 3	1.5	1.516 (0.038)	1.5	1.490 (0.042)	1.5	1.440 (0.049)	1.5	1.508 (0.053)	1.5	1.531 (0.079)
γ_2	**Group 1**	1	1.005 (0.022)	1	1.017 (0.025)	1	1.027 (0.031)	1	0.994 (0.039)	1	0.976 (0.059)
	Group 2	1.5	1.409 (0.033)	1.5	1.500 (0.036)	1.5	1.499 (0.043)	1.5	1.368 (0.048)	1.5	1.455 (0.058)
	Group 3	0.5	0.507 (0.034)	0.5	0.456 (0.039)	0.5	0.465 (0.043)	0.5	0.475 (0.048)	0.5	0.509 (0.072)
θ	**Group 1**	0.5	0.505 (0.010)	1	1.018 (0.023)	2	2.026 (0.058)	5	4.835 (0.199)	18	16.900 (1.238)
	Group 2	2	1.980 (0.058)	2.5	2.548 (0.085)	6	6.093 (0.283)	10	9.639 (0.579)	25	24.460 (2.117)
	Group 3	4	3.931 (0.151)	6	6.031 (0.280)	10	9.835 (0.562)	15	14.590 (0.996)	45	45.680 (5.320)
Inverse Over dispersion Parameter											
φ		10	11.460 (2.679)	10	10.170 (2.375)	10	11.230 (3.953)	10	11.960 (5.576)	10	11.110 (6.220)

[1]Estimated value.

[2]Standard deviation of the estimate is shown in parenthesis

4.5. Application to Empirical Data

This section illustrates the application of the proposed G-RPNB-L model. This section consists of two parts. The first part describes the characteristics of the empirical dataset used in this study. The second part documents the application of the model and compares the results with NB, G-RPNB, and NB-L models.

4.5.1. Data Description

We used lane departure crashes data of rural Interstates in Maine from the years 2015 to 2019 during the winter months of November to April (when often considered the winter period in Maine) to demonstrate the application of the proposed model and compare the results with existing models. Lane departure crashes are the leading type of crash in Maine, comprising over 70% of fatal crashes on Maine roadways (Sawtelle et al., 2022). The rural Interstates in Maine pass through eight counties in the state, Androscoggin, Aroostook, Cumberland, Kennebec, Penobscot, Sagadahoc, Somerset, and York. Table 11 presents the summary statistics of the traffic and geometric characteristics of rural Interstates in Maine.

The dataset contains information about 1236 roadway segments. This dataset includes monthly AADT, speed, shoulder width, presence of curve, and segment length. All rural Interstates segments in the data have a lane width of 12 feet. Shoulder width varies from 12 to 20 feet. The speed limit varies from 50 to 75 mph. This dataset has a very low mean of 0.1 and a standard deviation of 0.27 and includes 94.5% of zero crash observations. The number of zero observations in different counties also varies from 90% to 97%. The space scale (with short lengths) cannot be changed, since the characteristics of the adjacent segments were different. Aggregating data over a time scale will also lead to the loss of information (Shirazi et al., 2021; Lord et al., 2021). Hence,

as discussed in Lord & Geedipally (2018), the use of alternative models, such as the NB-L and its variations, is justified.

Table 11: Characteristics of Rural Interstates Roadways in Maine

Variables		Min	Max	Avg.	S.D.
Number of crashes		0	5	0.1	0.27
Speed limit (in mph)		50	75	69.2	4.47
Shoulder width (in feet)		12	20	14.3	0.85
Presence of curve (1 if present, 0 if absent)		0	1	0.29	0.45
Segment length (in miles)		0.01	4.9	0.5	0.61
Monthly average daily traffic (MADT)	**November**	229	40,366	12,386.9	9,050.2
	December	200	36,659	11,428.3	8,512.4
	January	178	33,467	10,181.4	7,550.1
	February	185	33,261	9,981.1	7,234.1
	March	188	35,526	10,675.7	7,827.1
	April	210	38,224	11,823.9	8,710.1

Table 12: Mean and Standard Deviation of Weather Variables for Maine Counties

Weather Variables	Winter Months					
	Nov	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**
Days with precipitation > 1.0 inch [1]	7.75 (3.27)[2]	8.42 (2.26)	7.09 (2.32)	7.43 (1.62)	5.99 (2.15)	8.74 (2.56)
Days with precipitation > 0.1 inch	12.16 (3.15)	12.76 (2.89)	10.91 (2.64)	12.20 (1.78)	10.57 (2.33)	14.27 (3.51)
Days with temperature < 32°F [1]	2.88 (3.25)	13.37 (5.56)	18.01 (5.33)	14.95 (6.44)	8.11 (3.93)	0.69 (0.83)
Days with snowfall > 1.0 inch	1.62 (2.03)	4.49 (2.03)	5.09 (2.25)	6.08 (2.11)	3.20 (1.82)	1.42 (1.01)
Average monthly temperature (°F)	34.90 (6.34)	25.24 (5.44)	19.86 (4.71)	20.41 (7.35)	28.22 (4.15)	41.32 (2.46)
Maximum monthly temperature (°F)	44.44 (4.37)	34.05 (4.88)	29.42 (3.84)	31.19 (6.70)	38.01 (3.35)	51.54 (2.47)
Minimum monthly temperature (°F)	26.39 (5.08)	16.42 (6.21)	10.30 (5.77)	9.64 (8.23)	18.45 (5.37)	31.08 (2.99)
Total monthly precipitation (inch)	4.15 (1.99)	4.47 (1.11)	4.18 (1.22)	3.40 (1.19)	2.75 (0.97)	4.20 (1.33)

[1]The weather variables used to demonstrate the model.
[2]Standard deviation is shown in parenthesis.

We combined the above data with monthly weather data collected at a weather station located in each county during the same period. While Maine experiences adverse weather conditions during the winter months of November to April, the weather variables often vary from county to county, and even from town to town, mainly due to vast variations in terrain and geography. We hypothesize that the impact of the weather variables on crashes could also be different from one region (or county) to another. We will use the G-RPNB-L to explore this hypothesis. Table 12 shows the summary statistics of the weather variables considered in this study (Sawtelle et al., 2022). Given that weather variables are generally correlated, they cannot all be included in the model simultaneously. After careful consideration, and test of significance, we chose two weather variables for the analysis. These variables are "Days with precipitation greater than 1.0 inch", and "Days with temperature less than 32°F". For the variable denoting the "Days with precipitation greater than 1.0 inch", the lowest average value is 5.99, which happened in March, and the highest average value is 8.74 which happened in April. For the "Days with temperatures less than 32°F" variable, the lowest mean is 0.69 which occurred in April, and the highest mean is 18.01 which occurred in January.

4.5.2. Modeling Results

This section presents the application of the G-RPNB-L model to the empirical dataset explained in the previous section. The results were also compared with the NB, NB-L, and G-RPNB models. To examine the goodness of fit (GOF), we used three commonly used criteria, Deviance Information Criterion (DIC), Widely Applicable Information Criterion (WAIC), and Leave-One-Out Cross-Validation Information Criterion (LOOIC) (Geedipally et al., 2014; Vehtari et al., 2017; Khodadadi et al., 2021; Islam et al., 2022). As noted earlier, the segment length was considered as an offset in our study. Without loss of generality, we considered the MADT, shoulder

width, speed limit, and the presence of the curve as fixed-parameters variables, and the two weather variables as grouped random-parameters variables in the model. This is mainly because we were interested to examine the impact of different weather variables across different counties.

We implemented the models in WinBUGS and used Bayesian inference and MCMC to estimate the parameters of the model (Spiegelhalter et al., 2003). As discussed earlier, to remove the correlation between the intercepts and the Lindley terms, we dropped the intercepts from the model, and calculated the intercepts based on Lindley terms using Eq. (26). We ran 30,000 MCMC iterations in WinBUGS with 3 chains. We discarded the first 5,000 samples as burn-in samples and estimated parameters from the remaining 25,000 posterior samples. To ensure removing any autocorrelation between simulated samples, we considered every 10-th sample from the posterior to compute the posterior mean of the parameters. The MCMC results showed excellent convergence and mixing. The density plots also showed clear unimodality for all the parameters. Table 13 shows the modeling results. As shown in this table, the coefficients of all the traffic and geometric variables were significant at 95% confidence level. The variables MADT, speed limit, and presence of the curve had positive interactions with crashes, which indicated an increase in these variables resulted in a higher number of crashes. Shoulder width had a negative effect on crashes; this means that an increase in shoulder width results in the reduction of lane departure crashes. Although with different coefficient values, these variables are also significant in other models.

Table 13: Modeling Results of Rural Interstates Data in Maine

Parameters	NB		G-RPNB		NB-L		G-RPNB-L	
	Mean	**S.D.[1]**	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**
Mean of Parameters								
Intercept	-9.193	1.284	-10.070	1.477	-9.163	1.321	-10.350	1.476
Ln (MADT)	0.668	0.047	0.800	0.060	0.667	0.049	0.812	0.063
Speed Limit	0.035	0.009	0.026	0.011	0.035	0.009	0.028	0.012
Shoulder Width	-0.168	0.055	-0.161	0.057	-0.167	0.055	-0.152	0.058
Presence of Curve (1 if present, 0 if absent)	0.275	0.053	0.240	0.054	0.275	0.055	0.243	0.056
Days with precipitation > 1.0 inch	0.061	0.009	*0.076*[2]	0.073	0.061	0.009	*0.076*[2]	0.075
Days with temperature < 32°F	0.043	0.003	*0.043*[2]	0.070	0.044	0.003	*0.044*[2]	0.070
Standard Deviation of Random Parameters								
Days with precipitation > 1.0 inch	-	-	0.197	0.061	-	-	0.198	0.063
Days with temperature < 32°F	-	-	0.187	0.058	-	-	0.188	0.060
Inverse Over dispersion (φ)	1.461	0.246	1.752	0.331	21.050	10.490	24.480	11.160
Model Performance								
DIC	15236.4		15101.1		14605.9		**14491.3**[3]	
WAIC	15236.8		15101.7		14964.9		**14831.5**[3]	
LOOIC	15236.8		15101.8		15169.6		**15038.6**[3]	

[1]Standard deviation.
[2]Italic font shows insignificant at 95% confidence level.
[3]Bold values indicate a better fit.

Table 14: Regional Estimates and Standard Deviations of the G-RPNB and G-RPNB-L Models for Weather Variables

Maine counties	Lindley Parameter	Intercept		Weather Variables			
				Days with precipitation > 1.0 inch		Days with temperature < 32°F	
	G-RPNB-L	G-RPNB	G-RPNB-L	G-RPNB	G-RPNB-L	G-RPNB	G-RPNB-L
Androscoggin	29.06 (3.543)	-9.003 (1.446)[1]	-9.319 (1.497)[1]	*-0.009*[2] *(0.040)*	*-0.009*[2] *(0.040)*	*0.023*[2] *(0.015)*	*0.022*[2] *(0.014)*
Aroostook	39.41 (6.304)	-10.920 (1.505)	-11.220 (1.540)	0.169 (0.044)	0.165 (0.044)	0.051 (0.012)	0.050 (0.012)
Cumberland	35.98 (2.514)	-9.870 (1.451)	-10.240 (1.501)	0.058 (0.018)	0.057 (0.019)	0.041 (0.007)	0.042 (0.007)
Kennebec	35.91 (2.349)	-9.735 (1.452)	-10.100 (1.497)	*0.033*[2] *(0.023)*	*0.033*[2] *(0.023)*	0.046 (0.008)	0.047 (0.008)
Penobscot	26.20 (1.333)	-9.813 (1.452)	-10.190 (1.500)	0.082 (0.019)	0.082 (0.020)	0.051 (0.005)	0.051 (0.005)
Sagadahoc	72.82 (9.309)	-10.780 (1.505)	-11.270 (1.551)	0.081 (0.036)	0.088 (0.037)	0.049 (0.016)	0.051 (0.017)
Somerset	19.88 (1.595)	-9.901 (1.453)	-10.180 (1.512)	0.117 (0.032)	0.112 (0.033)	0.057 (0.010)	0.056 (0.010)
York	47.68 (3.768)	-10.280 (1.460)	-10.670 (1.508)	0.083 (0.021)	0.084 (0.021)	0.035 (0.007)	0.036 (0.008)

[1]Standard deviations are shown in parenthesis.
[2]Italic font shows insignificant at 95% confidence level.

As noted earlier, we treated the two weather variables in the dataset as grouped random-parameters variables, each stratified at the county level. Although unlike the simpler models (i.e., NB and NB-L), the results of the G-RPNB-L model showed that these two variables were not significant at 95% confidence interval, the standard deviations of the random parameters are significant. This means that these variables are still important and impact lane departure crashes and should be kept in the model. A closer look at these two variables shows that both variables are in fact significant for several counties in Maine. Table 14 shows the coefficient of these variables for each county in Maine. As shown in Table 14, the variable denoting "Days with precipitation greater than 1.0 inch" was significant at 95% confidence level for all counties in Maine, except for Androscoggin and Kennebec counties. In other words, although the mean of the random

parameters for this variable was insignificant at 95% confidence level, the regional variation is significant; there are significant variations regarding the impact of this variable across different counties in Maine. For the variable denoting the "Days with temperature less than 32°F", the estimated coefficients for all counties, except Androscoggin County, were significant at 95% confidence level. For this variable, also, the parameter mean was insignificant at 95% confidence level. But the standard deviation of the random parameters was significant at 95% confidence level. In other words, although this variable is insignificant for one county, it is significant for the rest, but with different values. Furthermore, as expected, both of these variables had a positive interaction with crashes, which is similar to the findings from previous studies (Qiu & Nixon, 2008; El-Basyouny et al., 2014).

In the simple NB-L model, a single Lindley term is considered in the model to address the issue of excess zero observations. As noted earlier, this may not be an ideal strategy due to the differences in the number of zero observations in different regions or unobserved heterogeneity. In the G-RPNB-L model, instead, a random Lindley term is considered for each group of observations (here, each county). Therefore, the model can better address the issue of the number of zeros or unobserved heterogeneity, and subsequently, result in a better fit. Figure 1 shows the value of the Lindley parameter (θ) for each group of observations (here, counties). As shown in this figure, the value of the Lindley parameter substantially varies from one county to another. In particular, the value of this parameter varies from 19.88 (for Somerset County) to 72.82 (for Sagadahoc County). Allowing the Lindley parameter to vary among groups of observations allows the G-RPNB-L model to better account for variations in the number of zeros or address the unobserved heterogeneity in different counties.

Figure 2 shows variations of the coefficient of the variable denoting the "Days with precipitation greater than 1 inch" across different counties in Maine. The results show that the effect of this variable is the highest for Aroostook County. Aroostook County is in the northern region of Maine, which is less developed than the southern regions of Maine. These results presumably could be due to low traffic volume, less frequent winter maintenance, older roads, high speeds, and more adverse pavement conditions, in this region compared to southern Maine. As a result, adverse weather such as precipitation has a higher impact on crashes in this county compared to others. The impact of this variable was found to be the lowest for Cumberland County. Cumberland County is in the southwest region of Maine. The precipitation in this county is generally lower than in other counties (located on the east coast) which may lead to fewer crashes compared to other counties.

Figure 3 shows the coefficients of the variable denoting the "Days with temperature less than 32°F" for different counties in Maine. The impact of this variable on lane departure crashes also varies across different counties. The impact of this variable is highest for Somerset County. This county is also located in the northern region of Maine; hence, it experiences a colder climate compared to other counties. In addition, most of the mountains in Maine are in this county resulting in a more mountainous terrain compared to other counties. Presumably, low temperature and mountainous terrain led to a higher coefficient for this county.

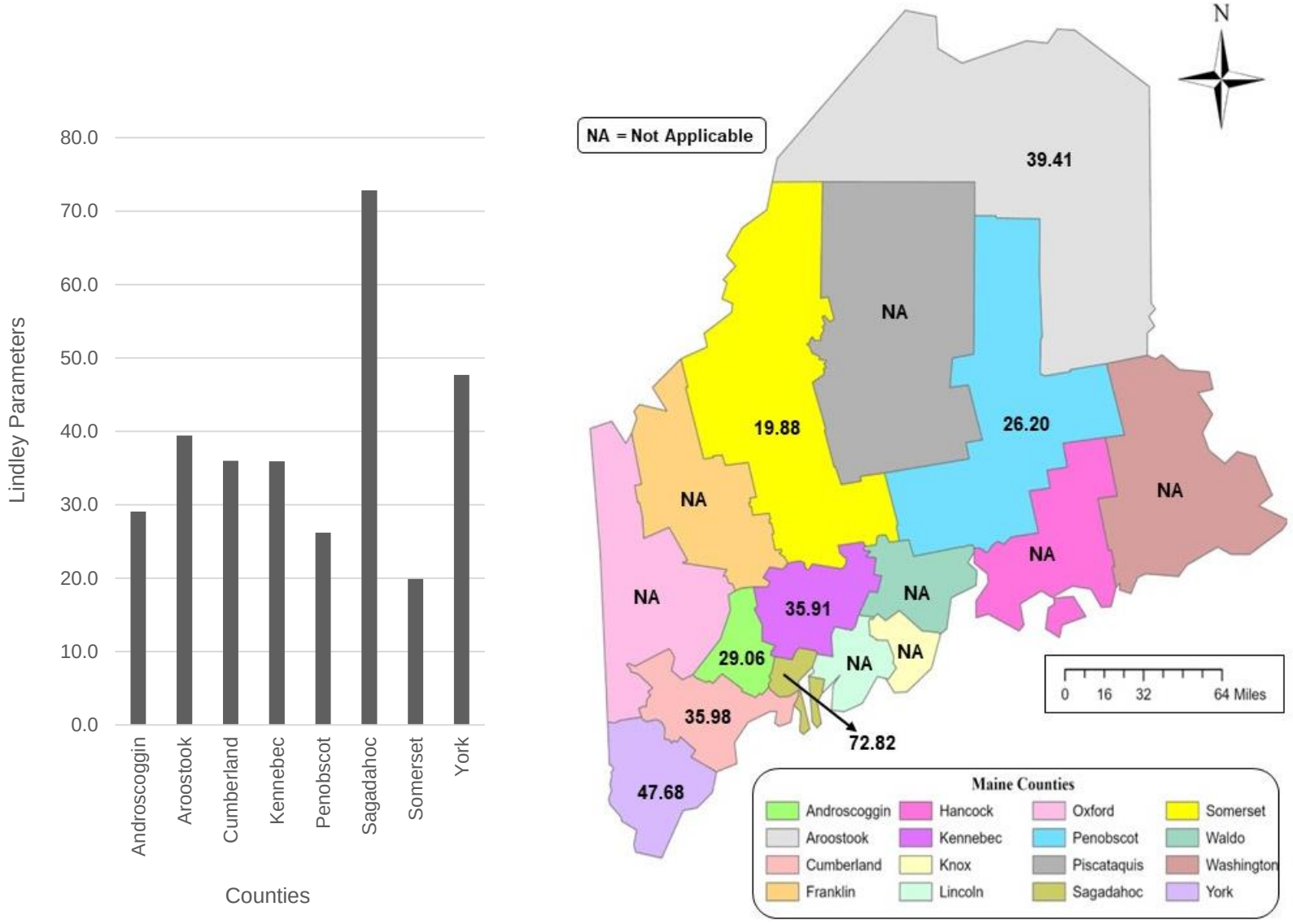

Figure 1: Variations of Lindley parameters across different Maine counties.

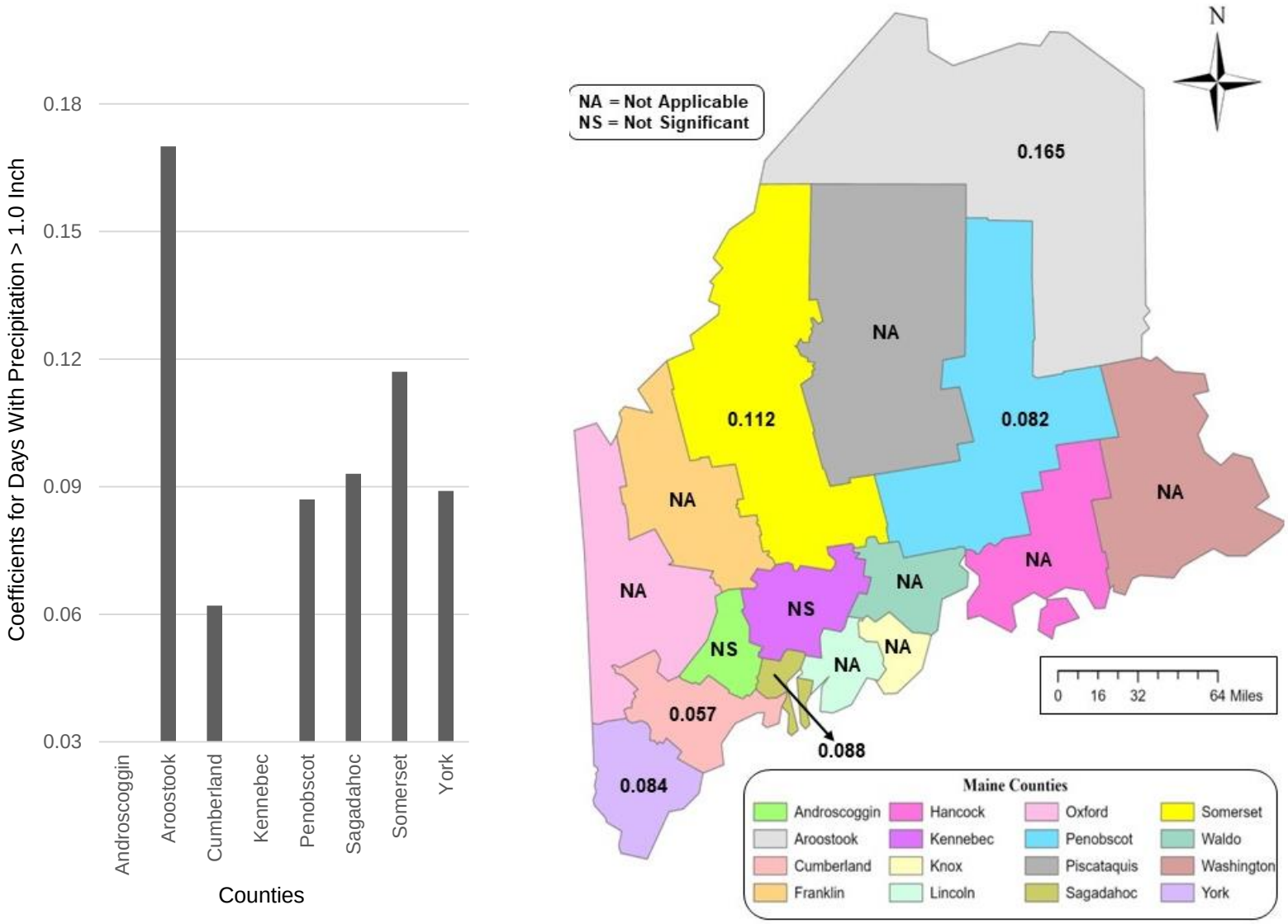

Figure 2: Variations of the coefficients of the "Days with precipitation greater than 1.0-inch" variable across different Maine counties.

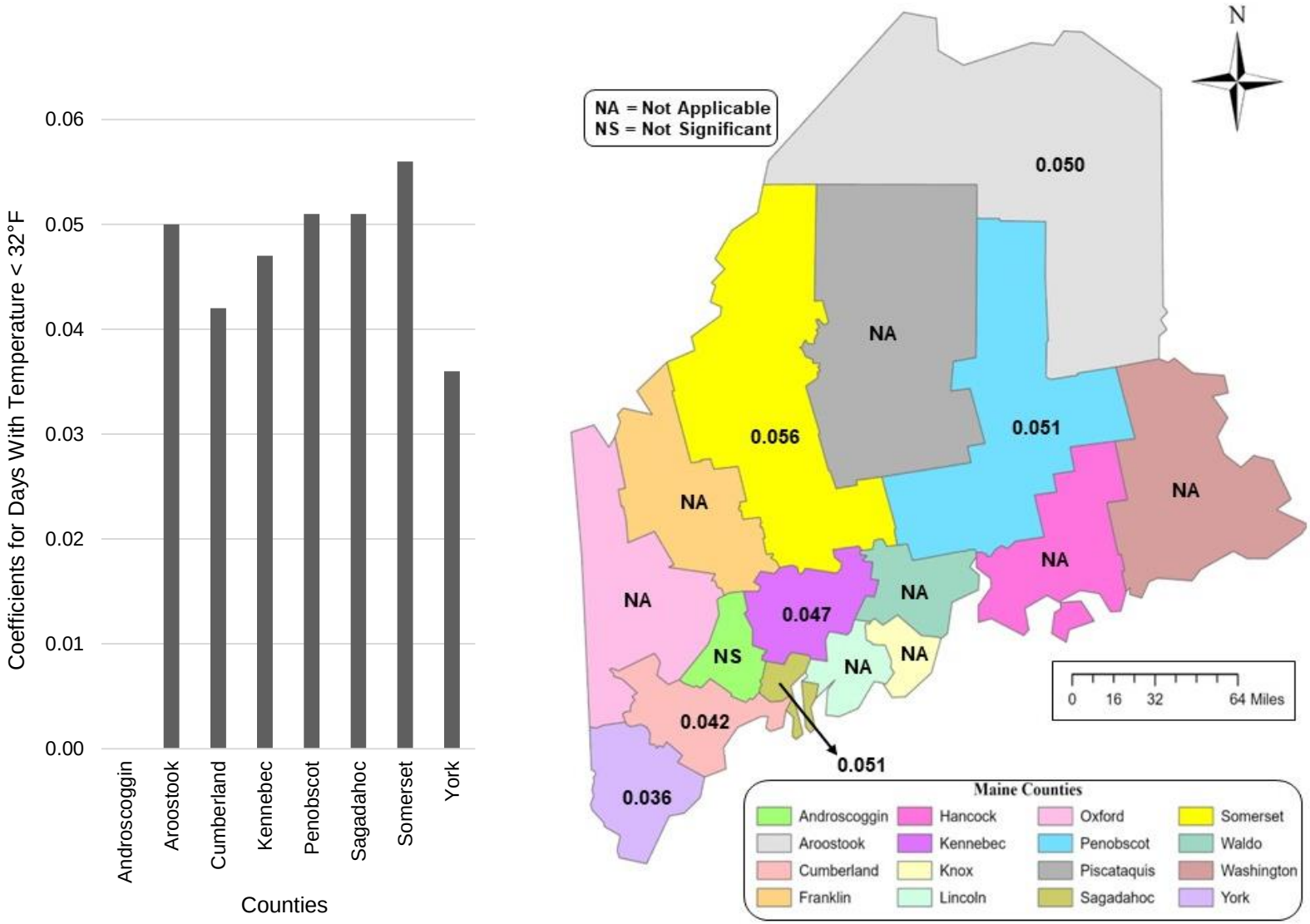

Figure 3: Variations of the coefficients of the "Days with temperature less than 32°F" variable across different Maine counties.

We used three goodness-of-fit metrics to compare the models. DIC is a popular goodness-of-fit metric to compare models with different complexities in Bayesian statistics (Geedipally et al., 2014); We recorded DIC for NB, NB-L, G-RPNB, and G-RPNB-L models. Based on DIC, the G-RPNB-L model had a superior fit with the DIC of 14491.3. The NB-L model had the second-best fit with a DIC of 14605.9; the G-RPNB was third with a DIC of 15101.1, and the NB was fourth in the order with a DIC of 15236.4. DIC however is sensitive to parameterizations (Geedipally et al., 2014). Therefore, we also used two more robust goodness-of-fit metrics known as WAIC and LOOIC (Vehtari et al., 2017). These two metrics also consider model complexity in their assessment metrics. Several researchers have used WAIC and LOOIC metrics for model comparison (Ahmed et al., 2020; Khodadadi et al., 2021; Mertens et al., 2021; Khodadadi et al., 2022b). For the WAIC metric, the G-RPNB-L exhibited the best fit with the WAIC of 14831.5. The NB-L with WAIC of 14964.9, G-RPNB with WAIC of 15101.7, and NB with WAIC of 15236.8 were ranked after the G-RPNB-L model. Using the LOOIC metric, the G-RPNB-L (LOOIC=15038.6) was the best model following the G-RPNB (LOOIC=15101.8), NB-L (LOOIC=15169.6), and NB (LOOIC=15236.8).

As a closing note to this section, it is worth pointing out that the NB-L model also had better goodness of fit (DIC and WAIC) compared to the G-RPNB, possibly due to excess zero observations in the data. Note that the dataset had more than 90% of zero observations. Several previous studies also showed that the NB-L model performs well when datasets are abundant with zero crash observations. However, the G-RPNB model can account for unobserved heterogeneity. Particularly, if the analyst is interested in a better understanding of the variations among different groups of observations (here, counties), the G-RPNB is preferred to the NB-L model, although the

NB-L may have a better fit. This concept is referred to as “goodness-of-logic” as illustrated in the work of Miaou & Lord (2003), Shirazi et al. (2017b), and Shirazi & Lord (2019).

4.6. Summary and Conclusions

Most often crash data are highly dispersed. Crash data may also contain a large amount of zero observations. The NB-L model can provide additional flexibility to the NB model to address the issue of the excess number of zero observations. The effect of different explanatory variables may also vary across different groups of observations, such as counties, regions, or cities, due to unobserved heterogeneity. In addition, different subgroups or regions may also have different percentages of zeros. To overcome these limitations, in this chapter, we proposed and documented the derivations and characteristics of the grouped random parameters negative binomial-Lindley (G-RPNB-L) model to address the regional heterogeneity in crash datasets with an excess number of zeros. We illustrated the feasibility of the model with a simulation study. The simulation study examined several scenarios with different percentages of zero observations. Then, we showed the application of the proposed model using an empirical dataset. We explored the effect of weather variations on crashes across different counties of Maine. This dataset had a large amount of zero crash observations. Our results showed that the coefficient of the weather variables varied across different counties. We also compared the proposed G-RPNB-L model with NB, NB-L, and G-RPNB models. We used three goodness-of-fit metrics for model comparison. The goodness-of-fit statistics showed the superiority of the G-RPNB-L model over the NB, NB-L, and G-RPNB models.

CHAPTER 5

SUMMARY AND RECOMMENDATIONS

This chapter summarizes the results and discussions of the two proposed models described in this thesis and provides recommendations for future research. This chapter is divided into two sections. The summary of the findings of the FMNB-L and G-RPNB-L models is discussed in the first section. The second section recommends the scope for future research.

5.1. Summary

Chapter 3 documented the derivations and characteristics of the finite mixture NB-L GLM to analyze crash data. This model was developed to account for unobserved heterogeneity due to latent subpopulations in data with many zero observations or long tails. The performance of the FMNB-L model in identifying subpopulations was evaluated using a simulated study with various sample means and zero percentages. The simulation results suggested that the FMNB-L can accurately identify subpopulations and account for a large percentage of zero observations. The application of the FMNB-L model in the empirical analysis was demonstrated using three highly dispersed and long-tailed datasets collected from Texas 4-lane Freeways. Then the results from the FMNB-L model were compared with NB, NB-L, and FMNB models based on several model selection metrics such as DIC, WAIC, and LOOIC. These goodness-of-fit metrics were used because they considered the complexity of models in their estimations. The FMNB-L model fitted the data significantly better than other models, according to the GOF statistics. The presence of subpopulations in these datasets was evident from the analysis.

Chapter 4 documented the derivations and characteristics of the grouped random parameters NB-L to analyze crash data. This model was developed to account for unobserved

heterogeneity due to variations across groups in crash data with many zero observations. A simulation study with varying zero percentages was used to assess the grouped random parameters NB-L model's performance in detecting variations across groups of observations. These simulated groups had different percentages of zero observations too. In the simulated scenarios, the proposed model efficiently addressed unobserved heterogeneity due to variations across groups. Then the proposed model was applied to an empirical dataset of rural Interstates in Maine that contained a large amount of zero observations across different regions. The G-RPNB-L model accounted for unobserved heterogeneity in the data due to the variations in the impacts of different weather characteristics across different regions. Then the model performance was compared with NB, NB-L, and G-RPNB models for various GOF metrics such as DIC, WAIC, and LOOIC. The proposed model was found superior to fit the data based on these GOF statistics.

5.2. Recommendations

The following recommendations are proposed based on the outcomes of this research. These recommendations include both methodological and practical aspects.

5.2.1. Methodological Recommendations

When datasets with a large amount of zero crash observations or heavy tails are suspected to include heterogeneous subpopulations, the finite mixture NB-L model should be used instead of NB or NB-L to identify latent subpopulations. There are a few suggestions for future study that should be considered. The application of FMNB-L model with varying weight parameters should be investigated further in the future. Future research on the finite mixture of random parameters NB-L is also recommended in crash data analysis.

When datasets with a large amount of zero crash observations across different groups (e.g., regions, towns) are suspected to have unobserved heterogeneity due to variations across groups, the grouped random parameters NB-L model should be employed instead of NB or NB-L to address unobserved heterogeneity. The effect of temporal attributes should be explored in future studies for grouped random parameters NB-L modeling approach.

5.2.2. Practical Recommendations

Apart from crash prediction, finite mixture NB-L models can be used for hotspot identifications. The applications of finite mixture NB-L models in empirical Bayes (EB) estimations for hotspot identifications and before-after studies can be explored in the future. Also, sample size guidelines for different sample means should be investigated in the future.

Skid number has a significant effect on crashes. Skid number can also vary across different regions. The impact of skid number on crashes should be investigated. Grouped random parameters NB-L model can be applied to datasets containing varying skid number across regions to capture unobserved heterogeneity due to variations across regions.

BIBLIOGRAPHY

Ahmed, I. U., Gaweesh, S. M., & Ahmed, M. M. (2020). Exploration of Hazardous Material Truck Crashes on Wyoming's Interstate Roads using a Novel Hamiltonian Monte Carlo Markov Chain Bayesian Inference. *Transportation Research Record*, *2674*(9), 661–675. https://doi.org/10.1177/0361198120931103

Allison, P. D. (2012). *Logistic regression using SAS: Theory and application*. SAS institute.

Anastasopoulos, P. C., & Mannering, F. L. (2009). A note on modeling vehicle accident frequencies with random-parameters count models. *Accident Analysis & Prevention*, *41*(1), 153–159

Behara, K. N., Paz, A., Arndt, O., & Baker, D. (2021). A random parameters with heterogeneity in means and Lindley approach to analyze crash data with excessive zeros: A case study of head-on heavy vehicle crashes in Queensland. *Accident Analysis & Prevention*, *160*, 106308.

Behnood, A., & Mannering, F. (2017). Determinants of bicyclist injury severities in bicycle-vehicle crashes: A random parameters approach with heterogeneity in means and variances. *Analytic Methods in Accident Research*, *16*, 35–47.

Behnood, A., & Mannering, F. L. (2015). The temporal stability of factors affecting driver-injury severities in single-vehicle crashes: Some empirical evidence. *Analytic Methods in Accident Research*, *8*, 7–32. https://doi.org/10.1016/j.amar.2015.08.001

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*, 56–91.

Cai, Q., Abdel-Aty, M., Lee, J., Wang, L., & Wang, X. (2018). Developing a grouped random parameters multivariate spatial model to explore zonal effects for segment and intersection crash modeling. *Analytic Methods in Accident Research*, *19*, 1–15. https://doi.org/10.1016/j.amar.2018.05.001

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (Vol. 53). Cambridge university press.

El-Basyouny, K., Barua, S., & Islam, M. T. (2014). Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson lognormal models. *Accident Analysis & Prevention*, *73*, 91–99. https://doi.org/10.1016/j.aap.2014.08.014

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 & Prevention*, *47*, 119–127.

Fanyu, M., Sze, N., Cancan, S., Tiantian, C., & Yiping, Z. (2021). Temporal instability of truck volume composition on non-truck-involved crash severity using uncorrelated and correlated grouped random parameters binary logit models with space-time variations. *Analytic Methods in Accident Research*, *31*, 100168.

Fisher, W. H., Hartwell, S. W., & Deng, X. (2017). Managing inflation: On the use and potential misuse of zero-inflated count regression models. *Crime & Delinquency*, *63*(1), 77–87.

Fountas, G., Sarwar, M. T., Anastasopoulos, P. Ch., Blatt, A., & Majka, K. (2018a). Analysis of stationary and dynamic factors affecting highway accident occurrence: A dynamic correlated grouped random parameters binary logit approach. *Accident Analysis & Prevention*, *113*, 330–340. https://doi.org/10.1016/j.aap.2017.05.018

Fountas, G., Anastasopoulos, P. C., & Abdel-Aty, M. (2018b). Analysis of accident injury-severities using a correlated random parameters ordered probit approach with time variant covariates. *Analytic Methods in Accident Research*, *18*, 57–68.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models* (Vol. 425). Springer.

Geedipally, S. R., Lord, D., & Dhavala, S. S. (2012). The negative binomial-Lindley generalized linear model: Characteristics and application using crash data. *Accident Analysis & Prevention*, *45*, 258–265. https://doi.org/10.1016/j.aap.2011.07.012

Geedipally, S. R., Lord, D., & Dhavala, S. S. (2014). A caution about using deviance information criterion while modeling traffic crashes. *Safety Science*, *62*, 495–498. https://doi.org/10.1016/j.ssci.2013.10.007

Heydari, S., Fu, L., Thakali, L., & Joseph, L. (2018). Benchmarking regions using a heteroskedastic grouped random parameters model with heterogeneity in mean and variance: Applications to grade crossing safety analysis. *Analytic Methods in Accident Research*, *19*, 33–48. https://doi.org/10.1016/j.amar.2018.06.003

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Islam, A. S. M. M., Shirazi, M., & Lord, D. (2022). Finite mixture Negative Binomial-Lindley for modeling heterogeneous crash data with many zero observations. *Accident Analysis & Prevention*, *175*, 106765. https://doi.org/10.1016/j.aap.2022.106765

Khodadadi, A., Tsapakis, I., Das, S., Lord, D., & Li, Y. (2021). Application of different negative binomial parameterizations to develop safety performance functions for non-federal aid system roads. *Accident Analysis & Prevention*, *156*, 106103. https://doi.org/10.1016/j.aap.2021.106103

Khodadadi, A., Shirazi, M., Geedipally, S., & Lord, D. (2022a). Evaluating alternative variations of Negative Binomial–Lindley distribution for modelling crash data. *Transportmetrica A: Transport Science*, 1–22. https://doi.org/10.1080/23249935.2022.2062480

Khodadadi, A., Tsapakis, I., Shirazi, M., Das, S., & Lord, D. (2022b). Derivation of the Empirical Bayesian method for the Negative Binomial-Lindley generalized linear model with application in traffic safety. *Accident Analysis & Prevention*, *170*, 106638. https://doi.org/10.1016/j.aap.2022.106638

Li, J., Wang, X., Yu, R., & Tremont, P. J. (2018). Relationship between Level of Service and Traffic Safety at Signalized Intersections: Grouped Random Parameter Method. *Journal of Transportation Engineering, Part A: Systems*, *144*(6), 04018022. https://doi.org/10.1061/JTEPBS.0000142.

Lord, D. (2006). Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis & Prevention*, *38*(4), 751–766. https://doi.org/10.1016/j.aap.2006.02.001

Lord, D., & Geedipally, S. R. (2011). The negative binomial–Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. *Accident Analysis & Prevention*, *43*(5), 1738–1742.

Lord, D., Geedipally, S. R., Guo, F., Jahangiri, A., Shirazi, M., Mao, H., & Deng, X. (2019). *Analyzing Highway Safety Datasets: Simplifying Statistical Analyses from Sparse to Big Data*.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, *44*(5), 291–305. https://doi.org/10.1016/j.tra.2010.02.001

Lord, D., & Miranda-Moreno, L. F. (2008). Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: A Bayesian perspective. *Safety Science*, *46*(5), 751–770.

Lord, D., Qin, X., & Geedipally, S. (2021). *Highway safety analytics and modeling*. Elsevier.

Lord, D., Washington, S., & Ivan, J. N. (2007). Further notes on the application of zero-inflated models in highway safety. *Accident Analysis & Prevention*, *39*(1), 53–57.

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 & Prevention*, *37*(1), 35–46.

Lord, D., & Geedipally, S. R. (2018). Safety prediction with datasets characterised with excess zero responses and long tails. In *Safe Mobility: Challenges, Methodology and Solutions* (Vol. 11, pp. 297-323). Emerald Publishing Limited.

Maine Department of Transportation (MaineDOT). (2020). *Crash and Highway Facts 2020*. https://www.maine.gov/mdot/safety/docs/2021/Statewide%20Crash%20Publication%202020.pdf

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. https://doi.org/10.1016/j.amar.2013.09.001

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. https://doi.org/10.1016/j.amar.2016.04.001

Meng, F., Wong, W., Wong, S., Pei, X., Li, Y., & Huang, H. (2017). Gas dynamic analogous exposure approach to interaction intensity in multiple-vehicle crash analysis: Case study of crashes involving taxis. *Analytic Methods in Accident Research*, *16*, 90–103.

Miaou, S.-P., & Lord, D. (2003). Modeling traffic crash-flow relationships for intersections: Dispersion parameter, functional form, and Bayes versus empirical Bayes methods. *Transportation Research Record*, *1840*(1), 31–40.

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.

Pantangi, S. S., Ahmed, S. S., Fountas, G., Majka, K., & Anastasopoulos, P. Ch. (2021). Do high visibility crosswalks improve pedestrian safety? A correlated grouped random parameters approach using naturalistic driving study data. *Analytic Methods in Accident Research, 30*, 100155. https://doi.org/10.1016/j.amar.2020.100155

Park, B.-J., & Lord, D. (2009). Application of finite mixture models for vehicle crash data analysis. *Accident Analysis & Prevention, 41*(4), 683–691. https://doi.org/10.1016/j.aap.2009.03.007

Park, B.-J., Lord, D., & Hart, J. D. (2010). Bias properties of Bayesian statistics in finite mixture of negative binomial regression models in crash data analysis. *Accident Analysis & Prevention*, *42*(2), 741–749. https://doi.org/10.1016/j.aap.2009.11.002

Park, B.-J., Lord, D., & Lee, C. (2014). Finite mixture modeling for vehicle crash data with application to hotspot identification. *Accident Analysis & Prevention*, *71*, 319–326. https://doi.org/10.1016/j.aap.2014.05.030

Qiu, L., & Nixon, W. A. (2008). Effects of Adverse Weather on Traffic Crashes: Systematic Review and Meta-Analysis. *Transportation Research Record*, *2055*(1), 139–146. https://doi.org/10.3141/2055-16

Rusli, R., Haque, Md. M., Afghari, A. P., & King, M. (2018). Applying a random parameters Negative Binomial Lindley model to examine multi-vehicle crashes along rural mountainous highways in Malaysia. *Accident Analysis & Prevention*, *119*, 80–90. https://doi.org/10.1016/j.aap.2018.07.006

Sarwar, M. T., Anastasopoulos, P. Ch., Golshani, N., & Hulme, K. F. (2017). Grouped random parameters bivariate probit analysis of perceived and observed aggressive driving behavior: A driving simulation study. *Analytic Methods in Accident Research*, *13*, 52–64. https://doi.org/10.1016/j.amar.2016.12.001

Savolainen, P. T., Mannering, F. L., Lord, D., & Quddus, M. A. (2011). The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis & Prevention*, *43*(5), 1666–1676. https://doi.org/10.1016/j.aap.2011.03.025

Sawtelle, A., Shirazi, M., Garder, P. E., & Rubin, J. (2022). Exploring the impact of seasonal weather factors on frequency of lane-departure crashes in Maine. *Journal of Transportation Safety & Security*, 1–22. https://doi.org/10.1080/19439962.2022.2086952

Shankar, Milton, J., & Mannering, F. (1997). Modeling accident frequencies as zero-altered probability processes: An empirical inquiry. *Accident Analysis & Prevention*, *29*(6), 829–837.

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

Shaon, M. R. R., Qin, X., Shirazi, M., Lord, D., & Geedipally, S. R. (2018). Developing a Random Parameters Negative Binomial-Lindley Model to analyze highly over-dispersed crash count data. *Analytic Methods in Accident Research*, *18*, 33–44. https://doi.org/10.1016/j.amar.2018.04.002

Shirazi, M., Dhavala, S. S., Lord, D., & Geedipally, S. R. (2017a). A methodology to design heuristics for model selection based on the characteristics of data: Application to investigate when the Negative Binomial Lindley (NB-L) is preferred over the Negative Binomial (NB). *Accident Analysis & Prevention*, *107*, 186–194. https://doi.org/10.1016/j.aap.2017.07.002

Shirazi, M., Geedipally, S. R., & Lord, D. (2021). A simulation analysis to study the temporal and spatial aggregations of safety datasets with excess zero observations. *Transportmetrica A: Transport Science*, *17*(4), 1305–1317. https://doi.org/10.1080/23249935.2020.1858993

Shirazi, M., & Lord, D. (2019). Characteristics-based heuristics to select a logical distribution between the Poisson-gamma and the Poisson-lognormal for crash data modelling. *Transportmetrica A: Transport Science*, *15*(2), 1791–1803.

Shirazi, M., Lord, D., Dhavala, S. S., & Geedipally, S. R. (2016a). 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 & Prevention*, *91*, 10–18. https://doi.org/10.1016/j.aap.2016.02.020

Shirazi, M., Lord, D., & Geedipally, S. R. (2016b). Sample-size guidelines for recalibrating crash prediction models: Recommendations for the highway safety manual. *Accident Analysis & Prevention*, *93*, 160-168.

Shirazi, M., Reddy Geedipally, S., & Lord, D. (2017b). A Monte-Carlo simulation analysis for evaluating the severity distribution functions (SDFs) calibration methodology and

determining the minimum sample-size requirements. *Accident Analysis & Prevention*, *98*, 303–311. https://doi.org/10.1016/j.aap.2016.10.004

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS user manual*.

Stewart, T. (2022). *Overview of Motor Vehicle Crashes in 2020*. Retrieved from, https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813266.

Tang, F., Fu, X., Cai, M., Lu, Y., Zhong, S., & Lu, C. (2020). Applying a correlated random parameters negative binomial lindley model to examine crash frequency along highway tunnels in china. *Ieee Access*, *8*, 213473–213488.

Theofilatos, A., & Yannis, G. (2014). A review of the effect of traffic and weather characteristics on road safety. *Accident Analysis & Prevention*, *72*, 244–256. https://doi.org/10.1016/j.aap.2014.06.017

Vangala, P., Lord, D., & Geedipally, S. R. (2015). Exploring the application of the Negative Binomial–Generalized Exponential model for analyzing traffic crash data with excess zeros. *Analytic Methods in Accident Research*, *7*, 29–36. https://doi.org/10.1016/j.amar.2015.06.001

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

World Health Organization (WHO). (2018). *Global status report on road safety 2018. Geneva: World Health Organization; 2018*.

Xie, H., Tao, J., McHugo, G. J., & Drake, R. E. (2013). Comparing statistical methods for analyzing skewed longitudinal count data with many zeros: An example of smoking cessation. *Journal of Substance Abuse Treatment*, *45*(1), 99–108.

Xie, Y., Zhao, K., & Huynh, N. (2012). Analysis of driver injury severity in rural single-vehicle crashes. *Accident Analysis & Prevention*, *47*, 36–44.

Xiong, Y., & Mannering, F. L. (2013). The heterogeneous effects of guardian supervision on adolescent driver-injury severities: A finite-mixture random-parameters approach. *Transportation Research Part B: Methodological*, *49*, 39–54. https://doi.org/10.1016/j.trb.2013.01.002

Zamani, H., & Ismail, N. (2010). Negative binomial-Lindley distribution and its application. *Journal of Mathematics and Statistics, 6*(1), 4–9.

Zamenian, H., Mannering, F. L., Abraham, D. M., & Iseley, T. (2017). Modeling the frequency of water main breaks in water distribution systems: Random-parameters negative-binomial approach. *Journal of Infrastructure Systems, 23*(2), 04016035.

Zhao, S., Wang, K., Liu, C., & Jackson, E. (2019). Investigating the effects of monthly weather variations on Connecticut freeway crashes from 2011 to 2015. *Journal of Safety Research, 71*, 153–162. https://doi.org/10.1016/j.jsr.2019.09.011

Zou, Y., Ash, J. E., Park, B.-J., Lord, D., & Wu, L. (2018). Empirical Bayes estimates of finite mixture of negative binomial regression models and its application to highway safety. *Journal of Applied Statistics, 45*(9), 1652–1669. https://doi.org/10.1080/02664763.2017.1389863

Zou, Y., Wu, L., & Lord, D. (2015). Modeling over-dispersed crash data with a long tail: Examining the accuracy of the dispersion parameter in negative binomial models. *Analytic Methods in Accident Research, 5*, 1–16.

Zou, Y., Zhang, Y., & Lord, D. (2013). Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis. *Accident Analysis & Prevention, 50*, 1042–1051. https://doi.org/10.1016/j.aap.2012.08.004

BIOGRAPHY OF THE AUTHOR

A S M Mohaiminul Islam was born and raised in Dhaka, Bangladesh. He attended Monipur High School in Dhaka, Bangladesh, for his Secondary School education, and Notre Dame College in Dhaka, Bangladesh, for his Higher Secondary School education. In October 2018, he graduated from the Bangladesh University of Engineering and Technology (BUET) in Dhaka, Bangladesh, with a Bachelor of Science in Civil Engineering. Then he began his Master of Science in Civil and Environmental Engineering at the University of Maine in the Fall 2020 session. His research is focused on Transportation Safety. At the 70th Annual Maine Better Transportation Association (MBTA) Meeting (2020), he was awarded second place in the Student Paper Presentation category. Mohaiminul also attended the Transportation Research Board (TRB) 101st Annual Meeting in 2022 and presented his research in the Poster Session. After graduation, he plans to join the industry to serve as a Traffic Engineer. Mohaiminul is a candidate for the Master of Science degree in Civil and Environmental Engineering from the University of Maine in August 2022.