Developing A Random Parameters Negative Binomial-Lindley Model to Analyze Highly Over-Dispersed Crash Count Data

Mohammad Razaur Rahman Shaon
Ph.D. Student
Department of Civil and Environmental Engineering
University of Wisconsin-Milwaukee, Milwaukee, WI 53211, USA.
Tel.: +1-605-690-0810; E-mail: mshaon@uwm.edu

Xiao Qin, Ph.D., P.E.*
Associate Professor
Department of Civil and Environmental Engineering
University of Wisconsin-Milwaukee, P.O. Box 784, Milwaukee, WI 53201-0784, USA.
Tel.: +1- 414-229-7399; E-mail: qinx@uwm.edu

Mohammadali Shirazi
Ph.D. Candidate
Zachry Department of Civil Engineering
Texas A&M University, College Station, TX 77843, USA.
Email: alishirazi@tamu.edu

Dominique Lord, Ph.D.
Professor
Zachry Department of Civil Engineering
Texas A&M University, College Station, TX 77843, USA.
Tel.: +1- 979 458-3949; Email: d-lord@tamu.edu

Srinivas Reddy Geedipally, Ph.D., P.E.
Associate Research Engineer
Texas A&M Transportation Institute
110 N Davis Dr., Arlington, TX 76013, USA.
Tel.: +1- 817-462-0519; Email: srinivas-g@tti.tamu.edu

*: corresponding author

Abstract

The existence of preponderant zero crash sites and/or sites with large crash counts can present challenges during the statistical analysis of crash count data. Additionally, unobserved heterogeneity in crash data due to the absence of important variables could negatively impact the estimated model parameters. The traditional negative binomial (NB) model with fixed parameters might not adequately handle highly over-dispersed data or unobserved heterogeneity. Many research efforts that have involved the negative binomial–Lindley (NB-L) model or the random parameters negative binomial (RPNB) model, for example, have attempted to improve the inference of estimated coefficients by explicitly accounting for extra variation in crash data. The NB-L is a mixed modeling approach which provides flexibility to account for additional dispersion in data. The RP modeling approach accommodates the effect of unobserved variables by allowing the model parameters to vary from one observation to another. The following study proposes a combination of these models – the random parameters NB-L (RPNB-L) generalized linear model (GLM) – to account for underlying heterogeneity and address excess over-dispersion. The results show that the RPNB-L model not only provides a superior goodness-of-fit (GOF) with the sample data, but also offers a better understanding about the effects of potential contributing factors. The paper uses the Bayesian framework to provide a strategy for eliminating the potential for poor mixing in the Markov Chain Monte Carlo (MCMC) chains during the estimation of the RPNB-L model.

Keywords: excess zero observations; over-dispersion; unobserved heterogeneity; mixed model; random parameters model; negative binomial-Lindley

1. Introduction

A roadway crash is a multifaceted event involving circumstances such as highway geometry, traffic exposure, contextual factors, driver characteristics, vehicle factors, as well as the interactions among them. Identifying key crash risk factors and understanding their effects is critical to finding cost-effective strategies for the prevention and reduction of traffic crashes. Typically, a quantitative safety analysis is performed through descriptive statistics to identify patterns, and regression models are used to identify factors associated with crashes. Once the association is properly established, additional insights about the crash can be revealed and evaluated. Lastly, the mean crash count can be estimated by mathematical formulation (Mitra and Washington 2012).

Crash data are often characterized by the existence of a large sample variance compared with the sample mean[1] (Lord *et al.* 2005, Mitra and Washington 2007). Extensive research has been devoted to modeling and analyzing this type of crash dataset (Lord and Mannering 2010, Mannering and Bhat 2014, Mannering *et al.* 2016). A notable accomplishment resulting from this research is the application of the Negative Binomial (NB) model in analyzing crash frequency data. The NB model can handle data over-dispersion by assuming a gamma distribution for the exponential function of the disturbance term in the Poisson mean. However, recent studies have pointed out that with a heavy-tailed crash dataset, the NB model can produce biased parameter estimates (Zou *et al.* 2015, Shirazi *et al.* 2016). A heavy-tailed distribution is a statistical phenomenon that occurs when sample observations have a few very high crash counts with preponderant zero observations; this shifts the overall sample mean to near zero (Shirazi *et al.* 2016). Failure to account for data over-dispersion could lead to biased and inconsistent parameter estimates, which in turn causes researchers to make erroneous inferences from models and also leads to inaccurate crash prediction values.

The mixed model is a well-known methodology used to incorporate heterogeneity into statistical analysis. Safety literature shows that mixed distribution NB models expanded the linear mixed model for continuous responses to discrete responses (e.g., crash count) by incorporating correlated non-normally distributed outcomes. Several mixed NB models have been proposed, including the NB-Lindley (NB-L), NB-Generalized Exponential (NB-GE), and NB-Dirichlet process (NB-DP) generalized linear models (GLMs) (Geedipally *et al.* 2012, Vangala *et al.* 2015, Rahman Shaon and Qin 2016, Shirazi *et al.* 2016). The advantage of using a mixed model is that it adds a mixed distribution to account for extra variance in the crash data which is caused by preponderant zero crash responses and/or a heavy-tail of crash counts (Shirazi et al., 2016). The underlying hypothesis is that the crash datasets are comprised of distinct subpopulations which have different probabilistic distributions. Accessing all data items associated with the likelihood of crash occurrence and/or injury severity is nearly impossible, but omitting important variables causes data heterogeneity which adds extra variation in the effects of explanatory variables. Random parameters (RP) models can account for unobserved heterogeneity by allowing the parameter of variables to vary from one observation to the next and by estimating the unbiased mean effect of explanatory variables (Mannering *et al.* 2016). Therefore, incorporating both random parameters and mixed probabilistic distributions within a single model can be a viable alternative for handling crash data with high over-dispersion and unobserved heterogeneity.

The objective of this study was to develop and document an RPNB model with Lindley mixed effect for heterogeneous count data that features an excess number of zero responses and/or a heavy-tail. The proposed RPNB-L model was developed in a Bayesian hierarchical framework that is expanded from fixed-coefficients NB-L GLM (Geedipally *et al.* 2012, Rahman Shaon and Qin 2016). The study utilized two crash datasets, one from Indiana and one from South Dakota, to calibrate the parameters in RPNB-L GLM.

[1] In a statistical term, the sample data is over-dispersed when the variance is greater than the mean. Data over-dispersion is often caused by unobserved data heterogeneity due to unobserved, unavailable, or unmeasurable variables that are important to explain model responses.

The datasets were characterized by over-dispersion with a very high percentage of zero responses and a heavy-tail. The model fitting and the modeling results were compared with the traditional NB, RPNB and NB-L models.

2. Literature Review

The existence of preponderant zero crash sites with a heavy tail can create highly over-dispersed data. The NB distribution has been used to model crash frequencies for decades because it can handle data over-dispersion. However, some studies have noted that the NB distribution cannot adequately handle over-dispersion caused by a heavy tail in the crash data (Guo and Trivedi 2002, Park *et al.* 2010, Zou *et al.* 2015, Shirazi *et al.* 2016). Guo and Trivedi (2002) noted that a negligible probability is usually assigned to higher crash counts in the NB model during the modeling of highly over-dispersed data with a heavy tail. Lord *et al.* (2005) pointed out that over-dispersion arises from the actual nature of the crash process. One limitation of the NB distribution is that it assumes that only one underlying process affects the likelihood of crash frequency (Shankar *et al.* 1997).

A mixture model is a very popular statistical modeling technique that is often used to account for data over-dispersion because it is flexible and extensible (Shankar *et al.* 1997, Aguero-Valverde and Jovanis 2008, Lord *et al.* 2008, Lord and Geedipally 2011, Geedipally *et al.* 2012, Cheng *et al.* 2013, Mannering and Bhat 2014, Rahman Shaon and Qin 2016, Shirazi *et al.* 2016). The mixture model is comprised of a convex combination of a finite number of different distributions. The NB-L GLM is a mixture of the NB and Lindley distribution in which the Lindley distribution itself is a mixture of two gamma distributions (Lindley 1958). The NB-L GLM was recently introduced to model crash frequency data (Geedipally *et al.* 2012, Rahman Shaon and Qin 2016). The count data mixture model works well when the dataset contains a large number of zero responses, is skewed, or is highly dispersed. Zamani and Ismail showed that the NB-L distribution provides a better fit compared to the Poisson and NB models when there is a large probability of crash frequency at zero (Zamani and Ismail 2010). Lord and Geedipally (2011) applied the NB-L distribution to estimate the predicted probability and frequency of crashes using both simulated and observed crash data. The authors concluded that the NB-L distribution can handle crash datasets with preponderant zero crash observations. Recently, Rahman Shaon and Qin (2016) evaluated the effect of lane and shoulder width on over-dispersed crash data using the NB-L model. The authors found that the NB-L GLM performed better than a traditional NB model when working with crash data characterized by preponderant zero responses, and that the core strength of the NB model was maintained. Overwhelmingly positive results have been reported from applying the NB-L model with many different data sources (Zamani and Ismail 2010, Lord and Geedipally 2011, Geedipally *et al.* 2012, Hallmark *et al.* 2013, Xu and Sun 2015, Rahman Shaon and Qin 2016). Although the Lindley distribution has a closed form (Zamani and Ismail 2010), the Lindley distribution cannot be mixed with the NB distribution in the context of GLM because it is not available in any standard statistical software (e.g. R, SAS, SPSS). Researchers have used the Bayesian method to create the hierarchical structure that is needed to estimate the parameters of NB-L in the context of GLM (Geedipally *et al.* 2012, Rahman Shaon and Qin 2016).

The existing crash dataset contains only a fraction of the potential variables that can significantly affect the likelihood of crash occurrence (Mannering *et al.* 2016). Unobserved heterogeneity in a regression model occurs when important covariates have been omitted during the data collection process. The influence of these variables is therefore not accounted for in the analysis. Unobserved heterogeneity in traditional NB models is usually considered to be random errors because the effect of each covariate is restricted to be the same across all observations; this causes even more dispersion problems. Such modeling strategies can cause serious model specification problems and may result in a variation of the estimated effect of observed covariates (Mannering *et al.* 2016). An overview of the potential for heterogeneity in driver behavior was highlighted by Mannering *et al.* (2016). The research found that varying lane and shoulder widths may have an impact on the likelihood of a crash event, but that these effects can vary among observations due to timevarying traffic, weather conditions, and/or the driver's reactions, all of which are not available for model development. Ignoring heterogeneous effects in explanatory variables leads to biased parameter estimates which can result in inaccurate conclusions (Mannering *et al.* 2016).

Research studies have been devoted to the task of obtaining unavailable but necessary data by utilizing statistical and econometric models[2] to account for unobserved heterogeneity. The RP modeling approach (Mannering *et al.* 2016)[3] has gained considerable attention for its use with crash count data. RP modeling addresses data heterogeneity by allowing the model parameters to vary from observation to observation. The parameter is treated as a random variable whose probability distribution is usually defined by the modelers. Anastasopoulos and Mannering (2009) introduced the RPNB model to account for data heterogeneity caused by explanatory variables and other unobserved factors. Crash data studies that have applied the RPNB model have found a significant improvement in the statistical model fit (El-Basyouny and Sayed 2009, Garnowski and Manner 2011, Venkataraman *et al.* 2011, Chen and Tarko 2014, Buddhavarapu *et al.* 2016).

In summary, the RP model incorporates the effect of unobserved variables by allowing model parameters to vary from observation to observation, but this method is susceptible to observations generated from different data sources. The mixed model also did not resolve the issue of omitted variables that could affect the likelihood of crashes. However, joint mixture distributions and random parameters can identify groups of observations with homogeneous variable effects within each group and can allow for the consideration of varying parameters so that the effects of unobserved variables are included (Peng and Lord 2011). Buddhavarapu *et al.* (2016) developed a spatial finite-mixture RPNB model that relaxed the distributional assumptions of RP. The study outlined in this paper pursued the same goal by utilizing the strengths and flexibility of both methods. Although the NB-L does not literally generate multiple homogeneous groups, it offers flexibility to account for skewness in crash observations which occurs when preponderant zero crash sites with a heavy tail are present. The unobserved heterogeneity in explanatory variables is assumed to be addressed when estimated parameters are allowed to vary across observations in NB-L.

3. NB-Lindley GLM

The NB-L distribution re-parameterized in a GLM context can be formulated in Equation 1 (Geedipally *et al.* 2012, Rahman Shaon and Qin 2016):

$$P(Y = y|\ \mu, \phi, \theta) = \int NB(y; \phi, \varepsilon\mu) Lindley(\varepsilon; \theta)\, d\varepsilon \tag{1}$$

In Equation 1, $f\ (u;\ a,\ b)$ is the distribution of the variable u, with parameters a and b. Following this explanation, given ε, the variable Y follows a NB distribution with a mean and inverse-dispersion parameter of $\varepsilon\mu$ and ϕ (ϕ =1/α), respectively. The variable ε follows a Lindley distribution with parameter θ.

The mean response function can be structured as follows, If we assume that the crash count follows the *NB-L(y; μ, ϕ, θ)* distribution (Geedipally *et al.* 2012, Rahman Shaon and Qin 2016):

[2] Refer to 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. for the list of methodological alternatives to account for unobserved heterogeneity.

[3] Finite mixture models, which are a type of latent variable models or latent class models was also explored as another alternative to account for unobserved heterogeneity in literature (Park and Lord 2009, Peng and Lord 2011, Shirazi et al. 2016). This approach express the overall distribution of one or more variables as a mixture of a finite number of component distributions which prescribes the observations from different groups, subpopulations or latent classes, each can be represented by a probability distribution function. Together, a finite mixture model can handle various distributions for different sub-populations in the target dataset.

$$E(Y = y) = \mu \times E(\varepsilon) \tag{2}$$

Where, $\mu = e^{\beta_0+\sum_{i=1}^{q}\beta_i X}$ and $E(\varepsilon) = \frac{\theta+2}{\theta(\theta+1)}$

By replacing the value of μ and E(ε), the mean response function can be written as follows:

$$E(Y) = \left(e^{\beta_0+\sum_{i=1}^{q}\beta_i X}\right) \times \frac{\theta+2}{\theta(\theta+1)} = e^{\left\{\beta_0+log\left[\frac{\theta+2}{\theta(\theta+1)}\right]\right\}+\sum_{i=1}^{q}\beta_i X} = e^{\beta_0'+\sum_{i=1}^{q}\beta_i X}$$

Where, $\beta_0' = \beta_0 + log\left[\frac{\theta+2}{\theta(\theta+1)}\right]$

$$E(Y) = \left(e^{\beta_0+\sum_{i=1}^{q}\beta_i X}\right) \times \frac{\theta+2}{\theta(\theta+1)} = e^{\left\{\beta_0+\log\left[\frac{\theta+2}{\theta(\theta+1)}\right]\right\}+\sum_{i=1}^{q}\beta_i X} = e^{\beta_0'+\sum_{i=1}^{q}\beta_i X} \tag{3}$$

Where, $\beta_0' = \beta_0 + \log\left[\frac{\theta+2}{\theta(\theta+1)}\right]$

The Lindley distribution is a mixture of two gamma distributions. Therefore, the Lindley distribution can be rewritten as (Geedipally *et al.* 2012, Rahman Shaon and Qin 2016):

$$\varepsilon \sim \frac{1}{1+\theta} Gamma(2, \theta) + \left(1 - \frac{1}{1+\theta}\right) Gamma(1, \theta) \tag{4}$$

which can be restructured as:

$$\varepsilon \sim \sum Gamma(1 + z, \theta)\ Bernoulli\left(z; \frac{1}{1+\theta}\right) \tag{5}$$

The NB-L GLM can be written as the following multi-level hierarchical structure using Equations (1) to Equation (5):

$$\begin{gathered} P(Y = y; \phi, \mu|, \varepsilon) = NB(y; \phi, \varepsilon\mu) \\ \mu = e^{\beta_0+\sum_{i=1}^{q}\beta_i X} \\ \varepsilon \sim Gamma(\varepsilon; 1 + Z, \theta) \\ Z \sim Bernoulli\left(z; \frac{1}{1+\theta}\right) \end{gathered} \tag{6}$$

The above formulation is similar to a Generalized Linear Mixed Model (GLMM) (Booth *et al.* 2003) where the mixed effects follow the Lindley distribution. In this modeling structure, the crash count follows a NB distribution which is conditional on a site-specific frailty term. The site-specific frailty term ε was assumed in order to accommodate extra variance in the crash data. The Lindley mixed effect, in hierarchical terms, is formulated by adding a site-specific offset (constant) term in the log-transformed domain of the mean response of the NB distribution.

The specification of prior distributions for the parameters is necessary for obtaining the Bayesian estimate. Prior distributions are meant to describe a prior knowledge about the parameters of interest. The site-specific frailty term follows a non-informative prior of the gamma distribution. The shape parameter in the gamma distribution follows a Bernoulli distribution with a probability parameter of 1/(1 + θ). Geedipally *et al.* (2012) used a beta prior on the parameter 1/(1 + θ) in the Bayesian interface. A weakly informative prior may yield a model output in which the parameter estimate for the Lindley distribution may contribute more than the NB distribution. The Markov chain Monte Carlo (MCMC) can suffer from poor mixing due to the correlation between the intercept and the site-specific frailty term. According to the literature, prior knowledge should be used to formulate the informative priors (if known) (Bedrick *et al.* 1996, Schlüter *et al.* 1997). A prior should be used to ensure E(ε) = 1 in order to limit the contribution of the mixed effect from the Lindley distribution. Geedipally *et al.* (2012) suggested using a prior for 1/(1 + θ) that follows a beta distribution. The reasonable choice for prior distribution is Beta (n/3, n/2), where n is the total observations (Geedipally *et al.* 2012).

4. Random Parameters NB-Lindley GLM

Let x_{ij} denote the j-th covariate associated with i-th site. In a RP model, the coefficient β_{ij} is assumed to be random, and is written as:

$$\boldsymbol{\beta_{ij}} = \boldsymbol{b_j} + \boldsymbol{w_{ij}} \tag{10}$$

where b_j denotes the fixed term (the mean parameter estimate), and w_{ij} denotes the random term. The random term is assumed to follow a predefined distribution such as a normal distribution with a mean equal to zero and a variance of σ^2. The random parameter β_{ij} should be used if the standard deviation of the random term w_{ij} is significantly different from 0 (under the frequentist approach; more discussion on that is provided below); otherwise, a fixed parameter or coefficient should be applied over all the individual observations (Anastasopoulos and Mannering 2009, El-Basyouny and Sayed 2009). Considering the above parameterization, the probability mass function (pmf) for the RPNB model can be written as:

$$p(y_i) = \frac{\Gamma(\phi+y_i)}{\Gamma(\phi)\Gamma(y_i+1)}(1-p_i)^{y_i}{p_i}^{\phi}; \quad \phi > 0, 0 < p_i < 1 \tag{11}$$

where, $p_i = \frac{\phi}{\mu_i+\phi}$

Technically, the NB-L GLM itself can also be considered a random parameters model because the intercept (or the mixed effect) that follows the Lindley distribution varies from site-to-site. The coefficients of explanatory variables are considered as random variables when developing a full RPNB-L GLM. In this paper, the NB-L model can be referred to as RPNB-L if any covariate can be considered a random variable. Recalling the hierarchy developed for NB-L GLM, the RPNB-L GLM can be written as the following multi-level structure:

$$\begin{aligned}
&P(y_{i;}\phi, \mu_i|\varepsilon_i) = NB(y_i; \phi, \varepsilon_i\mu_i) \\
&\log(\mu_i) = \beta_0 + \sum_{j=1}^{q} \beta_{ij}x_{ij} \\
&\varepsilon_i \sim Gamma(\varepsilon; 1 + z_i, \theta) \\
&z_i \sim Bernoulli(z; \frac{1}{1+\theta}) \\
&\beta_{ij} = \beta_j + w_{ij} \\
&w_{ij} \sim Normal(0, {\sigma_j}^2)
\end{aligned} \tag{12}$$

The MCMC chains in RPNB-L may suffer from poor mixing due to potential correlations between the intercept and regression coefficients, especially since both coefficients vary across observations. The covariates can be standardized before they are used in the model, which will. The traditional way of standardizing a covariate can be written as follows:

$$x_{ij}^{*} = \frac{x_{ij}-m_j}{s_j} \tag{13}$$

where, i = 1,2…., n denotes the number of observations;
j = 1,2…., q denotes the number of covariates; and
m_j and s_j are the mean and standard deviation of j-th covariate.

The standardized estimated coefficients need to be transformed back to the original scale after convergence, for ease of interpretation and inference. The following formulas describe the transformation (Gelfand *et al.* 1995):

$$\beta_1 = \frac{\beta_1^*}{s_1}$$

$$\bullet\bullet\bullet$$

$$\beta_q = \frac{\beta_q^*}{s_q}$$

$$\beta_0 = \beta_0^* - \sum_{i=1}^{q} \frac{\beta_q m_q}{s_q} \tag{14}$$

Where β_q^* is the standardized coefficient and β_q is the transformed coefficient in the original scale of the covariates.

The current formulation of the random part w_{ij} is defined with a prior that follows a normal distribution with a zero mean value. However, even though the prior is considered to have a mean value of zero for w_i, the posterior mean of the parameter will not necessarily be zero. Hence, this causes a conflict with the fixed effect parameter estimate of β which results in poor mixing in MCMC chains as well as issues identifying parameter estimates. A simple but effective method of centering the fixed effect parameter in the mean of the defined random coefficient can help to overcome this issue. The random coefficient definition in the model can be structured as:

$$\beta_{ij} \sim Normal(\beta_j, \sigma_j^2)$$

$$1/\sigma_j^2 \sim Gamma(0.01, 0.01) \tag{15}$$

Previous literature has explored several distributions such as normal, lognormal, uniform, triangular, gamma etc. The normal distribution was found to provide the best statistical fit (Li *et al.* 2008, Anastasopoulos and Mannering 2009). Thus, normal distribution was adopted for this study. The above formulation helped to achieve a good mixing in the MCMC chains.

5. Model Estimation

The RPNB-L model was formulated and estimated in a Bayesian framework using WinBUGS (Lunn *et al.* 2000). The traditional fixed-coefficients NB, the random parameters NB, and the fixed-coefficients NB-L models were also implemented in a Bayesian framework for comparison purposes. A total of three (3) Markov chains were used in the model estimation process with 80,000 iterations per chain for each model. In order to reduce autocorrelation, a thinning factor of three (3) was used in WinBUGS. The first 25,000 iterations were discarded as burn-in samples. The remaining iterations were used for estimating the model coefficients. The Gelman-Rubin (G-R) convergence statistic and Monte Carlo (MC) error were used to verify that the simulation runs converged properly. In the analysis, the research team ensured that the G-R statistic was less than 1.1. Mitra and Washington (2007) suggested that convergence was achieved when the G-R statistic was less than 1.2. The MC error of each parameter estimate was tested to ensure it was less than 3 percent of the estimated posterior standard deviation.

It is important to note that the estimation of RP models in a Bayesian framework is somewhat different compared to the frequentist or Maximum Likelihood Estimate (MLE) approach. In a Bayesian framework, the RP approach provides additional modeling flexibility by adding another level of hierarchy in the model parameterization. The variance in model parameters is assumed to come from unobserved data heterogeneity and is estimated by adding another level of hierarchy for the variance. Thus, unlike the MLE estimates, any parameter defined as random in a Bayesian framework will have a positive variance. In short, although the parameters may have the same mean estimates, the identification of which variables are random will be completely different. The parameters will always be random in Bayesian models if the Bayesian hierarchical model is defined as such, but the variables in MLE are considered random only if they meet a specific statistical criterion (i.e., $\sigma^2 > 0$ at a 5% significance level for example). The goodness-of-fit (GOF) of the models under investigation is also influenced by this difference in parameters.

Marginal effects are used to determine the impact of each covariate on the expected mean value of the dependent variable [4]. The marginal effect represents the effect of a unit change in the independent variable on the expected mean of the dependent variable. The marginal effect can be estimated as $\frac{\delta\mu_i}{\delta x_{ik}} \times \frac{x_{ik}}{\mu_i} = \beta_{ik}x_{ik}$, where μ_i is the expected mean outcome in each modeling approach (Washington *et al.* 2010). In the case of RP models, it is important to note that the marginal effects were estimated considering variation in estimated model parameters. The parameter means for each site were estimated after the MCMC chains converged in WinBUGS, and were then used to estimate the marginal effect of each observation.

6. Data Description

The characteristics of the two datasets used in this study are described in this section, which is divided into two subsections. The first subsection summarizes the characteristics of the data collected at 338 rural interstate roadway segments in Indiana. The second subsection describes the characteristics and summary statistics of the data collected at rural two-lane two-way highways in South Dakota. Both datasets are highly dispersed and characterized by a heavy tail, and both contain several variables which were used in model development to minimize the omitted-variable bias problem that can plague the development of crash prediction models (Lord and Mannering 2010).

6.1 Indiana Data

The Indiana dataset contains crash, roadway geometry, and traffic data collected over a five-year period (from 1995 to 1999) on 338 rural interstate roadway segments in the state of Indiana. The Indiana dataset has been used in several previous research studies, such as Washington *et al.* (2010), Geedipally *et al.* (2012). and Shirazi *et al.* (2016). In this dataset, 120 out of the 338 highway segments did not have any reported crashes over the five-year period (~36% are 0s). Table 1 presents the summary statistics of the variables used for developing the models in this study.

[4] The marginal effects were estimated for each observation and the mean value of all marginal effects are represented in Table 4 and Table 6. It is important to note that, the marginal effect for each covariate significantly varies from site-to-site.

Table 1 Summary Statistics for the Indiana Dataset.

Variables	Description	Mean	Standard Deviation	Minimum	Maximum
Crash	Number of Crashes in 5 years	16.973	36.297	0	329
Log(ADT)	Logarithm of Average daily traffic over the 5 years	10.036	0.681	9.153	11.874
Friction	Minimum friction reading in the road segment over the 5-year period	30.514	6.674	15.9	48.2
Pavement	Pavement surface type (1 if asphalt, 0 if concrete)	0.769	0.422	0	1
Median Width	Median width in feet	66.984	34.169	16	194.7
Barrier	Presence of median barrier (1 if present, 0 if absent)	0.160	0.367	0	1
Rumble	Interior rumble strips	0.725	0.447	0	1
Length	Segment length in miles	0.710	1.225	0.009	4.054

6.2 South Dakota Data

The South Dakota dataset is characterized by a preponderant number of zero responses and a heavy tail. In this dataset, the roadway geometric characteristics and traffic data elements were collected from the South Dakota Department of Transportation (SDDOT). Multiple event tables from the SDDOT Roadway Inventory System (RIS) were combined to generate homogeneous segments. Crash data between 2008 and 2012 were spatially joined with the roadway data according to their distance. The original dataset for rural two-lane two-way highway segments in South Dakota contains 16,827 segments. A sample of 10,000 observations from the total segments was used to evaluate the performance of the RPNB-L model in this study. The rural two-lane two-way segment database was previously used by Shaon and Qin to evaluate the performance of the NB-L model (Rahman Shaon and Qin 2016). The summary statistics of the sample data from South Dakota data are provided in Table 2.

Table 2 Summary Statistics for the South Dakota Dataset.

Variable	Definition	Mean	Standard Deviation	Minimum	Maximum
Crash	Count of Crashes	0.614	2.493	0	88
AADT	Annual Average Daily Traffic	917.933	913.790	45.00	21396.00
Segment Length	Segment Length in Miles	0.383	1.035	0.010	16.494
Speed Limit	Posted Speed Limit	57.273	10.712	20.00	65.00
Radius	Radius of curvature in miles	0.081	0.184	0.00	1.084
Lane Width	Lane width in feet	12.955	2.098	9.00	24.00
Shoulder Width	Shoulder width in feet	3.046	2.553	0.00	15.00
Vertical Grade	Yes	21.58%			
	No	78.42%			

In the South Dakota dataset, 78 percent of the 10,000 sample segments did not experience any crashes during the study period. The mean and standard deviation of the crash count for the 10,000 sample observations are equal to 0.614 and 2.493, respectively. Due to preponderant zero crash sites, the estimated skewness of the crash count was equal to 11.624, which shows that the crash count is highly skewed to the right. Annual average daily traffic (AADT), segment length, lane width, shoulder width, speed limit and radius of curvature of the horizontal curve were used as continuous explanatory variables to model crash data. Vertical grade is the only binary variable (1 if Yes, 0 if No) included in the model.

7. Results and Discussions

Detailed modeling results from the application of the RPNB-L GLM to both Indiana and South Dakota datasets are presented in this section. The first subsection that follows documents the modeling results for the Indiana dataset. The second subsection provides the modeling results for the South Dakota dataset. The performance of the RPNB-L model was compared to the NB, RPNB, and the NB-L GLMs for both datasets.

7.1 Indiana Data Results

Table 3 and Table 4, respectively, summarize the modeling results and the estimated marginal effects for the Indiana dataset. The segment length variable was considered as an offset variable in all modeling approaches, as developed in previous studies that utilized this dataset (listed above). Therefore, it is assumed that the number of crashes will increase linearly as the segment length increases. In Table 3, the results of the RPNB-L model were compared to the fixed and random parameters NB and the fixed parameters NB-L model. In all models, the estimated 95 percent marginal posterior credible intervals for all coefficients did not include zero. Hence, it can be concluded that all coefficients are statistically significant at a 5 percent significance level. In this section, only the modeling results for the application of the RPNB-L GLM are discussed. Anastasopoulos and Mannering (2009) and Geedipally et al. (2012) provide further discussions on the parameter estimates for random parameters NB and fixed parameters NB-L, respectively.

Table 3 Modeling Results for the Indiana Dataset.

Parameters	NB		RPNB		NB-L		RPNB-L	
	Value	Std. Dev.	Value	Std. Dev.	Value	Std. Dev.	Value	Std. Dev.
Parameter Mean								
Intercept	-4.449	0.067	-5.486	0.035	-3.947	0.162	-4.443	0.206
Log(ADT)	0.689	0.133	0.816	31.750	0.651	0.145	0.717	0.231
Friction	-0.027	0.011	-0.029	0.133	-0.027	0.012	-0.032	0.015
Pavement	0.422	0.189	0.588	0.012	0.445	0.210	0.605	0.281
Median Width	-0.005	0.002	-0.012	0.240	-0.006	0.002	-0.012	0.004
Barrier	-3.031	0.308	-6.614	0.003	-3.282	0.338	-6.152	0.898
Rumble	-0.405	0.186	-0.288	0.437	-0.404	0.207	-0.329	0.260
$\alpha = 1/\phi$	0.950	0.122	0.137	0.035	0.239	0.083	0.128	0.028
θ					1.464	0.180	1.414	0.173
Std. Deviation of Random Parameters								
Log(ADT)			0.302	0.172			0.232	0.137
Friction			0.057	0.011			0.056	0.011
Pavement			0.326	0.216			0.291	0.200
Median Width			0.028	0.003			0.028	0.003
Barrier			2.390	0.399			1.925	0.709
Rumble			0.379	0.242			0.310	0.183

Model Performance				
Dbar	1891.93	1481.09	1585.93	1422.70
Dhat	1883.01	1296.86	1469.51	1276.00
pD	8.92	184.22	116.41	146.30
DIC	1900.84	1665.31[†]	1702.34	1569.00
MAD[5]	6.92	6.90	6.88	6.71

Note: [†] With the MLE RPNB, only three variables (logarithm of ADT, presence of median barrier and interior rumble strips) were found to be random. This increased the Deviance Information Criterion or DIC to 1736.

The parameter mean for the traffic flow variable was estimated using the RPNB-L model to be less than one, indicating that the crash risk increases at a decreasing rate as the value of the traffic flow variable increases. A similar or consistent trend was observed for all other modeling approaches. The estimated marginal effect of the traffic flow variable also indicates that this variable has a positive influence on crash occurrence. Although the magnitude of coefficient can vary from site-to-site using the RPNB-L GLM, all estimated coefficients for the traffic flow variable have a value that is greater than zero.

The sign of the parameter mean estimates for both the roadway geometry and pavement-related variables are consistent with those found in Geedipally *et al.* (2012) using the same dataset. In this study, the RPNB-L helps to provide more details about the parameter estimates by combining the RP structure with the NB-L framework. The friction variable, which represents the minimum friction reading on the road segment over the five-year period, shows that the majority of sites (71.6 percent of normal density function) have estimated model coefficients with a value of less than zero while the rest of the sites have a coefficient that is greater than zero; this indicates that the friction variable has a mixed (both positive and negative) effect on crash risk. The marginal effect illustrates that the overall impact of the friction variable has a decreasing effect on crash risk. A similar pattern can also be observed with the median width variable. In this case, 66.6 percent of the estimated coefficients have a negative value while the rest are positive. More than 98 percent of the normal density function for pavement type has a value greater than zero with an estimated parameter mean of 0.422, meaning a change in pavement type from concrete to asphalt almost always increases the probability of a crash. A similar observation can also be obtained for the median barrier variable, which supports the effect of the median barrier variable as observed by Anastasopoulos and Mannering (2009).

Table 4 Average marginal effects for the Indiana Dataset.

Variables	**Model**			
	NB	**RPNB**	**NB-L**	**RPNB-L**
Log(ADT)	6.915	8.189	6.533	7.537
Friction	-0.812	-0.897	-0.824	-0.896
Pavement	0.325	0.452	0.343	0.578
Median Width	-0.351	-0.771	-0.412	-0.785
Barrier	-0.484	-1.057	-0.524	-1.181
Rumble	-0.293	-0.209	-0.293	-0.378

[5] Mean Absolute Deviance (MAD) provides a measure of the average miss-prediction of the model which can be estimated as $\frac{1}{n}\sum_{i=1}^{n}|Predicted\ value - Observed\ value|$. A value close to 0 suggests that, on average, the model predicts the observed data well.

7.2 South Dakota Data Results

The model parameter estimates and marginal effects of the covariates for the South Dakota data are provided in Table 5 and Table 6, respectively. The first part of Table 5 provides the estimates of the parameter means, and the second part of the table provides the estimated standard deviation of the random parameters. Unlike the model development for the Indiana data, the segment length variable was defined as a random parameter rather than as an offset. All covariates were defined as random parameters in both RPNB and RPNB-L models. The estimated standard deviation of all random parameters was found to be statistically significant at a 5 percent significance level.

Table 5 Modeling Results for the South Dakota Dataset.

Parameters	**NB**		**RPNB**		**NB-L**		**RPNB-L**	
	Value	Std. Dev.	Value	Std. Dev.	Value	Std. Dev.	Value	Std. Dev.
Parameter Mean								
Intercept	-7.609	0.027	-7.879	0.052	-7.546	0.038	-7.676	0.043
log(AADT)	0.751	0.031	0.744	0.032	0.754	0.031	0.738	0.032
Segment Length	0.674	0.020	0.745	0.026	0.658	0.018	0.740	0.026
Speed Limit	0.025	0.003	0.033	0.003	0.026	0.003	0.030	0.003
Lane Width	***-0.006***	***0.011***	-0.037	0.013	-0.010	0.009	-0.026	0.013
Shoulder Width	***-0.001***	***0.010***	***-0.009***	***0.011***	***-0.002***	***0.010***	***-0.002***	***0.012***
Radius	-0.501	0.129	-0.564	0.124	-0.516	0.131	-0.506	0.121
Vertical Grade	-0.992	0.073	-1.389	0.133	-1.013	0.073	-1.066	0.089
α = 1/ϕ	1.228	0.063	0.406	0.083	0.260	0.049	0.114	0.014
θ					1.501	0.033	1.495	0.034
Std. Deviation of Random Parameters								
log(AADT)			0.317	0.057			0.121	0.049
Segment Length			0.235	0.022			0.195	0.021
Speed Limit			0.038	0.004			0.033	0.003
Lane Width			0.117	0.021			0.101	0.020
Shoulder Width			0.092	0.019			0.069	0.013
Radius			0.437	0.301			0.384	0.136
Vertical Grade			1.274	0.180			0.550	0.176
Model Performance								
Dbar	14321		13450		12238.6		11550	
Dhat	14310.3		12780		11166.8		10150	
pD	8.981		669.9		1071.81		1393	
DIC	14330		14120†		13310.4		12940	
MAD	6.92		6.88		6.72		6.64	

Note: Parameter estimates not significant under 5 percent significance level are shown in italic and bold fonts.
† With the MLE RPNB, all the variables except speed limit were found to be random. This increased the DIC to 14132.

The parameters mean for the lane width variable is not statistically significant at a 5 percent significance level when the NB distribution is used, as indicated in Table 5. Yet, for the purpose of comparison between different models used in the analysis, this variable was kept in the model. The parameters mean for lane width did become significant when more advanced modeling alternatives (i.e.: RPNB, NB-L, and RPNB-L) were applied to this dataset. In addition, the results in Table 5 indicate that the parameters mean for the shoulder mean variable is not significant for all modeling approaches; however, since the standard deviation of the parameters is significant, this variable was kept in the model. The location of the mean of the coefficient distribution is not necessarily critical as long as the likelihood function improves with the significant standard deviation of the parameters (Anastasopoulos and Mannering 2009). While the parameters mean for all explanatory variables have a similar sign in all applied models, the magnitude of the estimates is different. Interestingly, the standard deviations of parameters for lane width and shoulder width are both statistically significant at a 5 percent significance level. The parameters mean is significant at a 5 percent confidence level for all other variables.

The RPNB-L model has smaller standard deviation estimates for all model coefficients (random parameters). A smaller standard deviation for the random parameter estimates means that the normal distribution of a covariate parameter is more centered around the mean value when the RPNB-L model is used; this may be a result of the site-specific frailty term used in the NB-L formulation that accounts for a portion of data variation.

Table 6 Average marginal effects for the South Dakota Dataset.

Variables	**Model**			
	NB	**RPNB**	**NB-L**	**RPNB-L**
log(AADT)	4.83	4.769	4.85	4.69
Segment Length	0.258	0.269	0.252	0.262
Speed Limit	1.419	2.027	1.473	1.672
Lane Width	-0.078	-0.566	-0.131	-0.361
Shoulder Width	-0.003	-0.025	-0.006	-0.005
Radius	-0.041	-0.05	-0.042	-0.04
Vertical Grade	-1.698	-3.189	-1.754	-2.127

The segment length and the AADT variables in the RPNB-L model showed a positive relationship with crash count for almost all segments, but with varying magnitude. The estimated marginal effect for AADT also emphasizes the positive effect AADT has on crash occurrence. A similar trend is also observed for the segment length variable. More than 81.8 percent of the sites have parameter estimates that are greater than zero for the speed limit variable. The estimated marginal effect for the speed limit variable indicates that there is an overall increase in crash occurrence with a unit increase in the speed limit variable.

The distribution of Radius of curvature has a crash count that decreases when the radius of curvature increases, but the magnitude varies among sites, as expected. The standard deviation of parameter estimate for the radius of curvature indicates that more than 90 percent of sites have negative coefficients. Similar observations can be made for the lane width variable, where 60.2 percent of the random parameter estimates have a value of less than 0. This trend also applies to the shoulder width variable, where more than 51 percent of the parameter estimates have a value of less than zero. Shaon and Qin used the same dataset and made similar observations (Rahman Shaon and Qin 2016). The authors noted that lane width may have mixed safety effects, and an increasing lane width or shoulder width or combination of both may not always bring additional safety benefits. Further research should look into whether or not an increase in lane width leads to an increase in safety. One interesting finding is that the estimated marginal effect of the shoulder width variable is quite similar between NB, NB-L, and RPNB-L (between -0.003 to -0.006), whereas it is quite different for the RPNB model. One possible explanation for this variation could be that the mean estimate of the shoulder width itself is not statistically significant in all models. The model parameters estimate and marginal effect of the grade variable indicates that the presence of vertical grade reduces crash occurrence for almost all sites (97.4 percent of the distribution has value less than zero). The estimate of the dispersion parameter is also the smallest for the RPNB-L model. The Poisson regression is a limiting case of the NB regression because the dispersion parameter approaches zero. The mean estimates in the RPNB-L model are less affected by the data dispersion, which means this model captures more variation in the data than the other three models.

7.3 Model Performance

The last section of Table 3 and Table 5 provides the model performance estimates based on the Deviance Information Criterion (DIC) for the Indiana and South Dakota datasets, respectively[6]. The DIC is a widely used GOF statistic for comparing models in a Bayesian framework (Spiegelhalter *et al.* 2002). It is worth pointing out that the model parameterization can influence the estimation of the DIC value, and the comparisons with DIC should be made only between models that have similar parameterizations (Geedipally *et al.* 2014). All developed models can be adequately compared using the DIC measure because both the NB-L and RPNB-L models are developed based on the NB model parameterization. The DIC consists of two components: (a) measures of how well the model fits the data, Dbar ($\overline{D(\theta)}$) and (b) a measure of model complexity (pD). Thus, DIC can provide a better comparison between models that are characterized by different complexities.

A comparison of the DIC values between models illustrated that the RPNB-L model performed better than the NB-L and RPNB. Table 3 and Table 5 show that the DIC value is highest in the traditional NB model. The small pD value illustrates that the NB model is less complex than other model alternatives used in this study. According to the estimated pD value, the RPNB-L model is the most complex of all the models due to its mixed distribution and random components in the explanatory variables. The point estimate of deviance illustrated by Dhat shows that the RPNB-L model has the smallest deviance in both datasets. Dbar represents almost the same information as Dhat except that it represents the posterior mean of deviance rather than a point estimate. The RPNB-L model, despite having the highest penalty value of pD, has a 5.8 percent and 7.8 percent improvement in DIC values for the Indiana dataset when compared with the RPNB and fixed parameters NB-L model, respectively. The MAD estimates indicate that the fixed-coefficient NB-L model has better predictive ability than RPNB even though the estimated DIC value is smaller with RPNB compared to the fixed parameters NB-L model. The improvement in DIC with the RPNB-L model compared to the RPNB and NB-L models for the South Dakota dataset are 8.4 percent and 2.8 percent, respectively. The MAD estimates illustrate that RPNB-L has the lower mean absolute error compared to other models in both datasets. Due to the frailty terms that explain additional data heterogeneity along with random parameter, RPNB-L compensates for increased model complexity by improving the predictive modeling ability, which is reflected in the MAD that considers both bias and variance.

8. Summary and Conclusions

Researchers can experience challenges when it comes to understanding the underlying crash generating process, producing reliable model coefficients, and making statistical inferences from crash data. This study proposed the application of a RPNB-L GLM for analyzing crash data by implementing an NB-L model

[6] DIC is a hierarchical modeling generalization of the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), defined as $DIC = \overline{D(\theta)} + pD \; and \; pD = \overline{D(\theta)} - D(\bar{\theta})$, where θ represents the collection of parameters.

with coefficients that varied from site to site. The model was applied to two observed datasets, one collected in Indiana and the other in South Dakota. The model results were compared to the traditional NB, RPNB, and fixed parameters NB-L models. Results showed that both the fixed coefficient NB-L (especially compared to the MLE RPNB) and newly developed RPNB-L GLMs performed better than a fixed and random parameters NB GLM. The estimated effects of covariates using RPNB-L were less dispersed when compared to the RPNB model, according to the standard deviation of random parameters. The RPNB-L model's proficiency in accounting for highly dispersed data led to its ability to achieve around 6 percent and more than 8 percent improvement in DIC, respectively, for the Indiana and South Dakota datasets when compared to the RPNB model. The estimated skewness of the crash count was 11.624 for the South Dakota data. Shirazi et al. (Shirazi *et al.* 2017) recommended that the NB-L (and RPNB-L) should be used over the NB when the skewness value exceeds 1.92. In conclusion, both the fixed and random parameters of NB-L GLMs offer a viable alternative to the traditionally fixed and random parameters NB GLMs when analyzing over-dispersed crash datasets.

The random parameters defined in this study were independent and characterized by a single normal distribution to account for unobserved heterogeneity in crash occurrences. The independence assumption restricts the interaction between random parameters. It is possible that the sources of heterogeneity are correlated due to the interactions between explanatory variables (Mannering *et al.* 2016). Mannering *et al.* (2016) suggested developing a random parameters model with correlated parameters to account for correlation among random parameters; however, using a simple distribution to characterize the random parameter mean and variance may not fully capture the underlying nature of unobserved heterogeneity in the dataset which could result in erroneous model inferences. Unobserved heterogeneity can be tracked in a more sophisticated manner when heterogeneity is included in the mean and variance, as additional flexibility is included in the capturing process (Behnood and Mannering 2017b, a, Seraneeprakarn *et al.* 2017). The proposed model should be developed further, and more reliable parameter estimates should be obtained by applying an RPNB-L with correlated random parameters and an RPNB-L with heterogeneity in the mean and variance. Additionally, more work should be performed to examine the "identification" of random parameters under the Bayesian framework in order to match those identified under the frequentist approach.

9. Acknowledgements

The authors are very grateful for the comments received from Dr. Soma Sekhar Dhavala regarding the characteristics of random parameters models under a Bayesian framework.

10. References

Aguero-Valverde, J., Jovanis, P., 2008. Analysis of road crash frequency with spatial models. Transportation Research Record: Journal of the Transportation Research Board (2061), 55-63.

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

Bedrick, E.J., Christensen, R., Johnson, W., 1996. A new perspective on priors for generalized linear models. Journal of the American Statistical Association 91 (436), 1450-1460.

Behnood, A., Mannering, F., 2017a. 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., 2017b. The effect of passengers on driver-injury severities in single-vehicle crashes: A random parameters heterogeneity-in-means approach. Analytic Methods in Accident Research 14, 41-53.

Booth, J.G., Casella, G., Friedl, H., Hobert, J.P., 2003. Negative binomial loglinear mixed models. Statistical Modelling 3 (3), 179-191.

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

Chen, E., Tarko, A.P., 2014. Modeling safety of highway work zones with random parameters and random effects models. Analytic methods in accident research 1, 86-95.

Cheng, L., Geedipally, S.R., Lord, D., 2013. The poisson–weibull generalized linear model for analyzing motor vehicle crash data. Safety science 54, 38-42.

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

Garnowski, M., Manner, H., 2011. On factors related to car accidents on german autobahn connectors. Accident Analysis and Prevention 43 (5), 1864-1871.

Geedipally, S.R., Lord, D., Dhavala, S.S., 2012. The negative binomial-lindley generalized linear model: Characteristics and application using crash data. Accident Analysis and Prevention 45, 258-265.

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.

Gelfand, A.E., Sahu, S.K., Carlin, B.P., 1995. Efficient parametrisations for normal linear mixed models. Biometrika 82 (3), 479-488.

Guo, J.Q., Trivedi, P.K., 2002. Flexible parametric models for long-tailed patent count distributions.

Hallmark, S.L., Qiu, Y., Pawlovitch, M., Mcdonald, T.J., 2013. Assessing the safety impacts of paved shoulders. Journal of Transportation Safety & Security 5 (2), 131-147.

Li, W., Carriquiry, A., Pawlovich, M., Welch, T., 2008. The choice of statistical models in road safety countermeasure effectiveness studies in iowa. Accident Analysis and Prevention 40 (4), 1531-1542.

Lindley, D.V., 1958. Fiducial distributions and bayes' theorem. Journal of the Royal Statistical Society. Series B (Methodological), 102-107.

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 and Prevention 43 (5), 1738-1742.

Lord, D., Guikema, S.D., Geedipally, S.R., 2008. Application of the conway–maxwell–poisson generalized linear model for analyzing motor vehicle crashes. Accident Analysis and Prevention 40 (3), 1123-1134.

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

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. Accident Analysis and Prevention 37 (1), 35-46.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. Winbugs-a bayesian modelling framework: Concepts, structure, and extensibility. Statistics and computing 10 (4), 325-337.

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

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

Mitra, S., Washington, S., 2007. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis and Prevention 39 (3), 459-468.

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

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 and Prevention 42 (2), 741-749.

Peng, Y., Lord, D., 2011. Application of latent class growth model to longitudinal analysis of traffic crashes. Transportation Research Record: Journal of the Transportation Research Board (2236), 102-109.

Rahman Shaon, M.R., Qin, X., 2016. Use of mixed distribution generalized linear models to quantify safety effects of rural roadway features. Transportation Research Record: Journal of the Transportation Research Board (2583), 134-141.

Schlüter, P., Deely, J., Nicholson, A., 1997. Ranking and selecting motor vehicle accident sites by using a hierarchical bayesian model. Journal of the Royal Statistical Society: Series D (The Statistician) 46 (3), 293-316.

Seraneeprakarn, P., Huang, S., Shankar, V., Mannering, F., Venkataraman, N., Milton, J., 2017. Occupant injury severities in hybrid-vehicle involved crashes: A random parameters approach with heterogeneity in means and variances. Analytic Methods in Accident Research 15, 41-55.

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

Shirazi, M., Dhavala, S.S., Lord, D., Geedipally, S.R., 2017. A methodology to design heuristics for model selection based on characteristics of data: Application to investigate when the negative binomial lindley (nb-l) is preferred over the negative binomial (nb). Accident Analysis and Prevention Forthcoming.

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

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64 (4), 583-639.

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.

Venkataraman, N., Ulfarsson, G., Shankar, V., Oh, J., Park, M., 2011. Model of relationship between interstate crash occurrence and geometrics: Exploratory insights from random parameter negative binomial approach. Transportation research record: journal of the transportation research board (2236), 41-48.

Washington, S.P., Karlaftis, M.G., Mannering, F., 2010. Statistical and econometric methods for transportation data analysis CRC press.

Xu, J., Sun, L., 2015. Modeling of excess zeros issue in crash count andysis. Journal of Jilin University (Engineering and Technology Edition) 45 (3), 769-775.

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

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.