

Developing The Multivariate Negative-Binomial Lindley Generalized Linear Model

Ali Khodadadi^{a,*}, Mohammadali Shirazi^c and Dominique Lord^d

^aTexas A&M University, 3136 TAMU, College Station, TX 77843-3136

^bUniversity of Maine, Orono, Maine, 04469

ARTICLE INFO


Keywords:
Multivariate,
Negative Binomial,
Dependency,
Common shock,
Bayesian,

ABSTRACT

The concept of a multivariate distribution is essential in statistics and econometrics, facilitating the simultaneous analysis of multiple interrelated random variables. In transportation safety research, multivariate models have proven effective in analyzing crash data of different categories, enabling better predictions and assessments of safety measures. In an attempt to extend the application of multivariate count models in transportation safety, this paper aims to redefine an advanced count model within a multivariate framework. In particular, the Negative Binomial Lindley (NB-L) distribution is the subject of interest due to its superior performance over traditional models such as Poisson and Negative Binomial (NB) distributions. A univariate NB-L distribution can explain the crash count variability in highly dispersed or sparse datasets. This motivates the extension of the NB-L model into the multivariate domain where further data breakdowns (e.g. by severity or type in crash data) can be explained and modeled simultaneously. This study defines a multivariate NB-L generalized linear model (GLM) that includes a set of interdependent random variables where each random variable marginally follows the univariate NB-L distribution with a specific mean function. The proposed multivariate NB-L model is hierarchically defined as a mixture of NB and multivariate Lindley distributions and supports a dependence structure (i.e. common shock) to explain the interdependence among random variables in multivariate Lindley distribution. The proposed multivariate model, along with the univariate counterparts, was applied to two crash datasets to demonstrate that (1) a multivariate NB-L can simultaneously define and differentiate between different crash categories that would not be accounted for otherwise (*i.e.*, in the univariate version), and (2) the multivariate model can account for the dependence among random variables which is not possible for univariate models. The results showed that the multivariate model even improved predictive performance. In the end, the authors discuss how the proposed model addresses both the need for advanced count models and the necessity to capture inter-dependencies between random variables. Consequently, it could be applied to real-world crash data to better capture crash data variability and eventually, provide more precise safety assessments.

1. Introduction

The concept of a multivariate distribution is a useful tool in statistics and econometric fields, which enables the simultaneous analysis and understanding of relationships between multiple random variables. A multivariate distribution describes the probability distribution of a random vector—a collection of two or more random variables. This approach is especially useful when the model variables in question are interrelated, allowing analysts to explore the relationship between them, rather than viewing each in isolation. Multivariate models are used in a broad range of fields, including finance, genetics, clinical trials, risk assessment, and transportation, due to their capability to analyze several random variables simultaneously. In particular, multivariate distributions have proven to be very useful in crash data analyses and safety research, leading to more accurate, efficient, and comprehensive insights and predictions, and

 a.khodadadi1994@tamu.edu (A. Khodadadi); shirazi@maine.edu (M. Shirazi); dlord@civil.tamu.edu (D. Lord)
ORCID(s): 0000-0002-3413-8687 (A. Khodadadi)

therefore, thus enhancing the effectiveness of safety assessments (Mannering et al., 2016; Russo et al., 2014; Barua et al., 2016; Anastasopoulos, 2016; Heydari et al., 2017; Fu and Sayed, 2022; Heydari et al., 2016; Ma et al., 2017; Zheng and Sayed, 2020).

Crash datasets can be categorized into multiple subcategories based on the crash type and severity. In the development of crash prediction models, researchers typically encounter one of the following scenarios: (1) a crash prediction model is developed for all crashes regardless of severity/type, or (2) separate distinct crash prediction models are developed for each severity/type. Either approach might introduce biases into the model, or result in the loss of information. On one hand, there might be structured or unstructured differences among categories that, if ignored by combining them, could lead to biased results. For instance, not all the categories may be sufficiently explained by an identical set of covariates (Kim et al., 2006). Similarly, the dispersion level may vary across categories. On the other hand, potential correlations among the categories might be overlooked if treated as separate and independent random variables (Jonathan et al., 2016). One solution to account for both issues is to jointly model different categories.

Joint modeling of crash severities or types has led to the development of various multivariate count models, including multivariate Poisson (Ma and Kockelman, 2006), multivariate Poisson log-normal (Park and Lord, 2007), multivariate Poisson gamma (or Negative Binomial distribution) (Mothafer et al., 2016; Ghitany et al., 2012; Shi and Valdez, 2014), multivariate Sarmanov distributions (Bolancé and Vernic, 2019), and multivariate Poisson Inverse-Gaussian (Ghitany et al., 2012). Gomez-Deniz et al. (2012) extended the Poisson-Lindley distribution from univariate to multivariate cases and discussed how their proposed model is particularly useful when marginal overdispersion is present among several dependent discrete random variables. The numerical results confirmed that their proposed joint model could be an alternative to other bivariate models such as bivariate Poisson-inverse Gaussian, bivariate Negative Binomial, and bivariate Negative Binomial-inverse Gaussian models. Park et al. (2021) discussed the growing interest in modeling correlated multivariate crash counts in road safety research and highlighted the importance of understanding the effects of various countermeasures across different crash types and severity levels, especially given that certain safety interventions may reduce one type of crash while increasing another. The authors introduced a Bayesian copula-based modeling approach and applied it to crash counts categorized by five severity levels from 451 three-leg unsignalized intersections in California. The results showed that compared to independent models, the joint Poisson-Gamma mixture models performed slightly better in terms of goodness-of-fit measures. Mothafer et al. (2016) focused on analyzing crash data from freeway segments in Washington, U.S. They proposed a Multivariate Poisson Gamma Mixture Count Model (MVPGM) to investigate the relationships among various types of crashes (rear-end, sideswipe, fixed object, and others). Using three years of crash frequency data from 274 freeway segments, along with the geometric characteristics and traffic volume information, the authors found that (1) the coefficients of the mean functions (vertical curvature, in particular) may not differ between multivariate and univariate total crash

models, and (2) the joint likelihood (and hence likelihood-based performance measures) appeared to be lower for the joint model compared to the univariate models, which might be due to the restrictions that the joint modeling frameworks apply. Sacchi et al. (2015) introduced a novel technique for identifying and ranking crash hot spots using a multivariate approach. This technique is grounded in the Full Bayesian (FB) context and employs the Mahalanobis distance for multivariate identification and ranking of hot spots, as opposed to traditional univariate methods that focus on a single variable like total collision counts. The proposed method consists of several steps: applying multivariate Poisson–lognormal regression models to data, using Poisson posterior mean estimates for each site to compute the multivariate Mahalanobis distance from the normal Poisson mean for similar sites, and preparing an ordered list of potentially hazardous sites. The authors used a four-year crash dataset for 173 signalized intersections in Vancouver, Canada, to evaluate their proposed approach and found that it outperforms traditional univariate approaches. Fu and Sayed (2022) presented an innovative multivariate extreme value theory (EVT) method for predicting real-time crash risk, overcoming the drawbacks of traditional univariate EVT models that only consider single conflict indicators. By incorporating multiple indicators such as modified time to collision (MTTC), post encroachment time (PET), and deceleration rate to avoid a crash (DRAC), the new multivariate model offers a more precise and comprehensive evaluation of road safety. Four parametric models (tilted Dirichlet, pairwise beta, Husler-Reiss, and extremal) were used to describe the dependency structure among conflict extremes. By effectively capturing the complex dependencies between various conflict indicators, the researchers observed a significant improvement in real-time safety evaluation. Heydari et al. (2016) introduced the use of Bayesian non-parametric Dirichlet process mixture models to better handle unobserved heterogeneity and multimodality in transportation safety data. This study extended the multivariate Poisson-lognormal model to a more adaptable Dirichlet process mixture multivariate model, which accounts for interdependencies between outcomes through a non-parametric random effects density. This innovative approach provides a more robust framework for jointly modeling different outcomes in crash data analysis, enhancing the robustness of parametric distributional assumptions. Ma et al. (2017) focused on improving crash prediction models by addressing the significant challenges posed by spatial and temporal correlations, heterogeneity, and the interdependence of crash frequencies across different injury severity levels. The study introduces a Bayesian multivariate space-time model to effectively capture these complexities. Using daily traffic crash data from the I-70 highway in Colorado, their proposed framework outperformed traditional models by better accommodating unobserved heterogeneity and leveraging correlations between crash types across spatial and temporal dimensions. The model highlighted critical predictors, including geometric features and dynamic variables like daily speed gaps and wet road conditions. Despite its promising results, the study acknowledges limitations due to its specific dataset and calls for further research to validate the model on different road types and incorporate dynamic space-time interactions for a deeper understanding of crash dynamics.

Various techniques have been used to define a correlation structure among multiple random variables and form a multivariate distribution/model. The most common techniques used in crash data analysis include (1) multivariate distributions with specific marginal distributions: this method involves defining a multivariate distribution by specifying the marginal distributions of each random variable, allowing for the modeling of each variable independently while also considering their joint distribution; (2) copula models: copulas are functions that link univariate marginal distribution functions to a multivariate distribution function; and (3) common shocks (or common covariance): this refers to a technique that explicitly defines a dependence structure through shared external influences. This study focuses on the latter technique. The common shock approach introduces dependence among different count variables by assuming that all components share a common random factor. This means that while components might have individual characteristics, they are all simultaneously influenced by a shared random factor. Unlike the normal distribution, where the dependency structure (*i.e.*, covariance structure) is embedded in the mathematical formulation, the original form of many distributions does not have parameters that directly account for the correlation among multivariate count variables. Therefore, the dependence structure should be imposed on the distribution through an external factor or common shock. This approach has been extensively used to form multivariate distributions, including multivariate gamma (Mathai and Moschopoulos, 1991) and multivariate Poisson models (Bermúdez and Karlis, 2011; Karlis and Meligkotsidou, 2005; Tsionas, 2001). The formulation and implementation of the common shock technique are discussed in detail in the next section.

Despite extensive efforts to generalize univariate count distributions to their multivariate counterparts, the emphasis has been solely on a few fundamental count distributions, and little has been done to develop more advanced and flexible multivariate models. In the realm of univariate count distributions, limitations of traditional models (e.g., the Poisson and Negative Binomial) led to the development of more adaptable alternatives. Among them, the Negative Binomial-Lindley (NB-L) distribution is gaining popularity for its effectiveness with long-tailed and sparse datasets (Khodadadi et al., 2021, 2022; Geedipally et al., 2012; Zamani and Ismail, 2010; Islam et al., 2022; Shirazi et al., 2017). Zamani and Ismail (2010) first proposed the NB-L distribution as a mixture of the NB and a one-parameter Lindley distribution. The Lindley distribution, from the lifetime distribution family, is used to simulate the lifetime of a process. Due to its flexibility and broad application, it has been generalized to various parameterizations, including but not limited to one-parameter Lindley (Lindley, 1958, 1965), two-parameter Lindley (Shanker et al., 2013, 2016), three-parameter Lindley (Shanker et al., 2017), and weighted-Lindley (Ghitany et al., 2011). Given that each Lindley parameterization offers a specific degree of flexibility, researchers have attempted to mix more advanced Lindley formulations with the NB distribution. Examples include the mix of NB and two-parameter Lindley distribution (Denthet et al., 2016; Khodadadi et al., 2022), NB and three-parameter Lindley distribution (Tajuddin et al., 2020; Khodadadi et al., 2022), NB and weighted Lindley distribution (Denthet and Promin, 2019; Samutwachirawong, 2017; Khodadadi et al., 2022), and

NB and Quasi Lindley distribution (Khodadadi et al., 2022). All studies collectively concluded that a more advanced Lindley distribution forms a more flexible mixed NB-L distribution. In line with these findings, Khodadadi et al. (2022) conducted a thorough comparative study to assess the efficacy of different NB-L parameterizations in handling datasets with varying means and levels of sparsity. They used both simulated and real-world data to assess the model performance in terms of fitness and prediction accuracy, concluding that the Negative Binomial Weighted-Lindley distribution outperforms the other parameterizations.

Reportedly, no study has developed a multivariate NB-L model (any parameterization of the NB-L). The superior performance of the NB-L in crash prediction models on the one hand, and the necessity of an advanced multivariate model to capture interdependence between random variables, on the other hand, motivate this study to (1) form a multivariate NB-L distribution to extend the application of the multivariate models into the advanced and flexible count models (2) examine the application of the proposed model in real-world scenarios, and (3) explore how the multivariate model can be further adjusted (e.g., extended into a random parameter framework) and thus accommodating even more diverse and heterogeneous datasets. The rest of the paper is organized as follows: The next section presents a detailed explanation of the methodology and the mathematical formulation of the multivariate model. Section Three provides an overview of the data, summarizing key points and presenting descriptive statistics. Finally, the modeling results and related discussions/conclusions are provided.

2. Methodology

The NB-L distribution, in its original simplest form, is a mixture of the NB and a one-parameter Lindley distribution Zamani and Ismail (2010). More complex parameterizations, however, have shown superior performance. In a thorough comparative study conducted by (Khodadadi et al., 2022), various Lindley parameterizations were examined. Their findings indicated that the mixed Negative-Binomial Weighted-Lindley (NB-WLindley), as it is referred to in the paper, not only outperforms other NB-L parameterizations but also offers an additional parameter to control the bias-variance balance in the predictive model. Additionally, the WLindley distribution possesses a useful mathematical representation, particularly beneficial when generalizing to a multivariate version. The NB-WLindley could be hierarchically defined as follows:

$$\begin{aligned} Y &\sim NB(\phi, \epsilon\mu) \\ \epsilon &\sim WLindley(\theta, c) \end{aligned} \tag{1}$$

The objective is to construct a multivariate NB-WLindley distribution by mixing NB and multivariate WLindley distribution. Let ϵ be a random variable following the WLindley distribution with parameters θ , and c . The pdf and

the first two moments could be written as follows (Ghitany et al., 2011):

$$p(\epsilon) = \frac{\theta^{c+1}}{(\theta + c)\Gamma(c)} \epsilon^{c-1} (1 + \epsilon) e^{-\theta\epsilon}; \quad \epsilon > 0, \theta > 0, c > 0 \quad (2)$$

$$E(\epsilon) = \frac{c(\theta + c + 1)}{\theta(\theta + c)} \quad (3)$$

$$E(\epsilon^2) = \frac{(\theta + c + 2)(c + 1)}{\theta^2(\theta + c)c} \quad (4)$$

where, $\Gamma(c) = \int_0^\infty x^{c-1} e^{-x} dx$.

Now, assume $\epsilon_1, \epsilon_2, \dots, \epsilon_J$ are mutually independent random variables that follow the WLindley distribution with specified parameters as follows:

$$\epsilon_j \sim \text{WLindley}(c_j - \gamma_0, \theta_j); \quad 0 < \gamma_0 < c_j \quad (5)$$

Where c_j and θ_j are specific parameters associated with each random variable ϵ_j , and γ_0 is a positive fixed value that is subtracted from all c_j parameters. The parameter γ_0 is called a common shock (or common effect) and is shared among the random variables ensuring inter-dependency among ϵ_j s. The idea is to form a multivariate WLindely distribution so that we can mix the resulting multivariate WLindely distribution with the NB distribution and form a multivariate NB-WLindley distribution. To do so, let's define a new independent random variable, ϵ_0 such that:

$$\epsilon_0 \sim \text{Gamma}(\gamma_0, 1); \quad \gamma_0 > 0 \quad (6)$$

Next, define a vector of random variables, E_1, E_2, \dots, E_J , such that:

$$E_j = \epsilon_j + \frac{\epsilon_0}{\theta_j}; \quad j = 1, \dots, J \quad (7)$$

According to (Zamani and Ismail, 2010), any Lindley parameterization could be expressed as a summation of two mutually independent gamma-distributed variables. Therefore, each ϵ_j could be re-written as follows:

$$\epsilon_j = b_1 \epsilon_{j1} + b_2 \epsilon_{j2} \quad (8)$$

where: $b_1 = \frac{\theta}{\theta+c}$; $b_2 = \frac{c}{\theta+c}$; $\epsilon_{j1} \sim \text{Gamma}(c_j, \theta_j)$; and $\epsilon_{j2} \sim \text{Gamma}(c_j + 1, \theta_j)$

Replacing ϵ_j with Eq.(8), E_j s could be re-written as follows:

$$E_j = b_1 \left(\epsilon_{j1} + \frac{\epsilon_0}{\theta_j} \right) + b_2 \left(\epsilon_{j2} + \frac{\epsilon_0}{\theta_j} \right) \quad (9)$$

The random variables, ϵ_{j1} , ϵ_{j2} and ϵ_0 are mutually independent and follow gamma distribution. In addition, the summation of two mutually independent gamma distributions is still a gamma distribution. Therefore:

$$\epsilon_1 + \frac{\epsilon_0}{\theta_j} \sim \text{gamma}(c_j, \theta_j) \quad (10)$$

$$\epsilon_2 + \frac{\epsilon_0}{\theta_j} \sim \text{gamma}(c_j - 1, \theta_j) \quad (11)$$

Therefore, E_j s could be re-parameterized as follows:

$$E_j = b_1 \text{gamma}(c_j, \theta_j) + b_2 \text{gamma}(c_j - 1, \theta_j) \quad (12)$$

Eq. (12) defines the WLindley distribution and indicates that each E_j marginally follows the WLindley distribution. Additionally, the distribution of E_j is not affected by the common shock variable (i.e., γ_0); therefore, the random vector E can be interpreted as following a multivariate WLindley distribution, with γ_0 representing the dependence structure that connects all E_j s:

$$E = (E_1, E_2, \dots, E_J) \sim \text{MVWLindley}(C, \Theta, \gamma_0) \quad (13)$$

where: $C = (c_1, c_2, \dots, c_J)$; and $\Theta = (\theta_1, \theta_2, \dots, \theta_J)$; and $\gamma_0 < \text{Min}(C)$

Now that we have developed the multivariate version of the WLindley distribution, we can combine it with the NB distribution to form a multivariate NB-WLindley distribution. Similar to the NB-L parameterization suggested in (Khodadadi et al., 2021; Geedipally et al., 2012), the multivariate NB-WLindley can be hierarchically defined as follows:

$$Y_i \sim \text{NB}(\Phi, E_i M_i) \quad (14)$$

$$E_i \sim \text{MVWLindley}(\Theta, C, \gamma_0)$$

where:

$i \in (1, 2, \dots, N)$, refers to the observation i ;

$j \in (1, 2, \dots, J)$, refers to the individual random variable index j within the random vector Y

$Y_i = (Y_{i1}, Y_{i2}, \dots, Y_{iJ})$, refers to the random vector associated with observation i following a multivariate NB-WLindley distribution (each Y_{ij} follows a univariate NB-WLindely distribution);

$E_i = (E_{i1}, E_{i2}, \dots, E_{iJ})$, refers to the random vector associated with observation i following a multivariate WLindley distribution (each E_{ij} follows a univariate WLindely distribution);

$M_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{iJ})$, refers to the mean vector associated with observation i ;

$\mu_{ij} = \exp(\beta_{0j} + \sum_{k=1}^K \beta_{jk} X_{ik})$, the mean function associated with observation i and random variable j ;

$k \in (1, 2, \dots, K)$, refers to the explanatory variable k

$X_{[N \times k]}$, refers to explanatory variables matrix;

$C = (c_1, c_2, \dots, c_J)$, refers to a random vector storing parameter c_j s from all random variables;

$\Theta = (\theta_1, \theta_2, \dots, \theta_J)$, refers to a random vector storing parameter θ_j s from all random variables;

$\Phi = (\phi_1, \phi_2, \dots, \phi_J)$, refers to a random vector storing parameter ϕ_j s from all random variables;

The next subsection describes the modeling framework and estimation process.

2.1. Modeling framework

A Full Bayesian (FB) technique was employed to estimate the unknown parameters and produce credible posterior inferences. The FB approach is capable of integrating all available information and prior knowledge into one hierarchical model, providing relatively accurate estimates even with low sample sizes. It should be noted that an alternative parameterization of the Lindley distribution was used, which is expressed as a combination of two gamma distributions, with the mixture components following a Bernoulli distribution (see Khodadadi et al. (2022); Geedipally et al. (2012)). Additionally, the FB method requires specifying prior distributions for all unknown hyper-parameters to combine the data likelihood with previous evidence. This study adopted a non-informative normal prior for the coefficients (β s) in the mean function (M), gamma distributions for the dispersion vector Φ , uniform distributions for transformed parameterizations of parameters C , θ , and γ_0 . The rationale behind the choice of prior distributions and their effects on the model's overall convergence is explored in the discussion section.

The vector-level hierarchical representation of the multivariate model is shown in Eq. (14). To parameterize the proposed model into the modeling language supported by Markov Chain Monte Carlo (MCMC) software (e.g., BUGS and JAGS), the vector representation needs to be decomposed as follows:

$$Y_{ij} \sim NB(\phi_j, E_{ij} \mu_{ij})$$

$$E_{ij} = \epsilon_{ij} + \frac{\epsilon_0}{\theta_j}$$

$$\mu_{ij} = \exp(\beta_{0j} + \sum_{k=1}^K \beta_{jk} X_{ik})$$

$$\begin{aligned}
\epsilon_{ij} &\sim \text{Gamma}(c_j + z_{ij} - \gamma_0, \theta_j) \\
\epsilon_0 &\sim \text{Gamma}(\gamma_0, 1) \\
z_{ij} &\sim \text{Bernoulli}\left(\frac{c_j}{c_j + \theta_j}\right) \\
\phi_j &\sim \text{Gamma}(0.01, 0.01) \\
\beta_{jk} &\sim \text{Normal}(0.01, 0.01) \\
\frac{\gamma_0}{c_j} &\sim \text{Uniform}(0, 1) \\
\frac{c_j}{\theta_j} &\sim \text{Uniform}(0, 1)
\end{aligned} \tag{15}$$

After establishing the hierarchical joint model, random samples from the posterior distribution of the unknown parameters were generated using the MCMC method. Depending on whether the full conditional distribution of the unknown parameter was available, either the Gibbs sampling method or the Metropolis-Hastings algorithm was employed to obtain these random samples. For the MCMC analysis in this study, an open-source R package named 'Rjags' was utilized. To ensure sufficient convergence, three separate Markov chains were initiated, each with 100,000 iterations. The initial 20,000 iterations from each chain were discarded as burn-in, and the subsequent samples were used for estimating the unknown coefficients. Additionally, to reduce sample auto-correlation, out of 10 samples, only one was retained for estimation purposes.

3. Data

Two crash datasets were selected to examine the capabilities of the proposed multivariate model. Both datasets categorize crash frequencies into multiple sub-categories (by crash type or crash severity), necessitating the use of a multivariate distribution that can handle all these random variables at once and accounts for the interdependence structure among various categories.

The first dataset is extracted from the Texas Department of Transportation (TxDOT) research project 0-6991 (Avelar et al., 2020). The dataset includes 220 uniform segments from rural two-lane highways in Texas. Variables such as average daily traffic, truck percentage, shoulder widths, lane widths, and speed limit were obtained from the TxDOT's Road-Highway Inventory Network (RHINo) database. In addition, information on the posted speed limit and number of horizontal curves was gathered using Google's Imagery and Street View services. Data on crashes that occurred from 2014 to 2018 were sourced from the TxDOT Crash Records Information System (CRIS). This dataset, specifically, analyzes run-off-road incidents, with crash data categorized and reported separately for (1) guardrail hit crashes, (2) rollovers, and (3) other fixed object crashes. Summary statistics for significant variables in the modeling results

Table 1

Summary Statistics of Variables - Texas run of the road crash data

Variables	Min	Max	Average	Std. Dev.	Skewness
AADT (veh/day)	9	9651	1892	2164	-
Segment length (mi)	0.1	6.34	1.01	1.18	-
Paved shoulder width (ft)	0	32	5.75	7.68	-
Number of curves	0	9	1.24	1.93	-
Speed limit (mph)	30	75	58	11	-
Guardrail crashes	0	4	0.11	0.47	5.38
Rollover crashes	0	5	0.27	0.76	3.69
Fixed object crashes	0	21	0.72	1.90	6.21

Table 2

Summary Statistics of Variables - California intersection crash data

Variables	Min	Max	Average	Std. Dev.	Skewness
Major AADT (veh/day)	7.80	11.27	9.42	0.75	-
Minor AADT (veh/day)	2.30	10.05	4.92	1.51	-
Number of Lanes	2	4	3.69	- 0.73	-
Painted Left Turn (Yes: 1, No: 0)	-	-	0.39	0.49	-
Rolling (Yes: 1, No: 0)	-	-	0.36	0.48	-
Lighting (Yes: 1, No: 0)	-	-	0.35	0.48	-
Mountain (Yes: 1, No: 0)	-	-	0.14	0.35	-
Fatal (K) crashes	0	5	0.17	0.52	4.17
Incapacitating injury (A) crashes	0	6	0.45	0.96	3.02
Non-Incapacitating injury (B) crashes	0	20	1.64	2.51	2.98
Minor injury (C) crashes	0	28	1.92	3.55	3.82
Property-damage-only (O) crashes	0	88	6.33	9.94	3.86

are provided in Table 1.

The second dataset utilized in this study was derived from crash count data from five different severity levels: fatal injuries, (denoted as K), incapacitating injuries (denoted as A), non-incapacitating injuries (denoted as B), minor injuries (denoted as C), property damage only (denoted as O or PDO). These data were collected from 451 three-legged unsignalized intersections in California. The dataset was previously analyzed using a multivariate Poisson-Lognormal (MVPLN) modeling approach as reported in (Park and Lord, 2007). Over a span of 10 years, the dataset includes 77 fatal injuries, 202 accidents of incapacitating injuries, 738 accidents of non-incapacitating injuries, 865 accidents of minor injuries, and 2,857 PDO accidents at the 451 intersections. Table 2 provides summary statistics of multivariate crash frequency by severity, with the unit of crash frequency represented as the number of crashes per intersection over the 10-year period.

4. Modeling Results

The authors used the FB approach to estimate the parameters in the hierarchical multivariate NB-WLindley model defined in Eq. (15). Unlike univariate models, the multivariate model generates a distinct set of estimates for each random variable. Therefore, to better understand how the proposed multivariate model works, it would be beneficial to monitor each category individually. To this end, along with the multivariate model, a separate univariate model for each category was also developed in order to see how and to what extent transitioning from the univariate to the multivariate domain enhances the estimates.

In addition, as mentioned in the introduction, this study also aims to determine how the proposed model would behave if equipped with other techniques to overcome challenges such as unobserved heterogeneity in crash data. To this end, the authors developed a Random Parameter (RP) version of the proposed multivariate NB-WLindley model. According to (Shaon et al., 2018), random parameter models are capable of identifying clusters of observations where variable effects are homogeneous within each cluster, thus allowing the parameters to vary from one observation to the next (Shaon et al., 2018). To construct a random parameter model, the regression coefficients in the mean function, β_{jk} , are assumed to be random variables specific to each observation (therefore β_{jk} would be changed to β_{ijk} , where the i index refers to the observations), indicating that each consists of a fixed part, b_{jk} , and a random part, w_{ijk} :

$$\begin{aligned}\beta_{ijk} &= b_{jk} + W_{ijk} \\ b_{jk} &\sim Nnormal(0.01, 0.01) \\ W_{ijk} &\sim Nnormal(0, \sigma_k^2) \\ \frac{1}{\sigma_k^2} &\sim Gamma(0.01, 0.01)\end{aligned}\tag{16}$$

where; $i \in [1, N]$ denotes the observation, $j \in [1, J]$ denotes the category of the dependent variable, and $k \in [1, K]$ denotes the covariate corresponding to the coefficient. Similar to other studies (Shaon et al., 2018), the parameter W_{ijk} is assumed to be normally distributed with mean zero and standard deviation following an inverse gamma distribution.

Table 3 summarizes the modeling results for run-off-the-road crashes in Texas. The Texas dataset categorizes 10-year crash counts by five crash severity levels namely: fatal (k), incapacitating injury (A), non-incapacitating injury (B), minor injury (C), and property damage only (PDO). The modeling outputs include the results from one multivariate NB-WLindley model, one random parameter multivariate NB-WLindley model, and a set of univariate NB-WLindley models separately developed for each crash category. As seen in the table, each category has its own set of covariates that have a statistically significant effect on that specific crash category. Collectively, six covariates appeared to significantly affect various crash categories. Larger AADT values in both major and minor roads, mountainous and rolling

terrains, and a higher number of main lanes showed an increasing effect on crash counts. In addition, the existence of a painted left-turn lane at the intersection showed contradicting effects on different crash severity. It showed a decreasing effect on property damage crashes but an increasing effect on major injury crashes, which is consistent with the observation made in other studies (Park and Lord, 2007; Park et al., 2021). For each crash category, a different set of covariates was found significant highlighting the fact that multivariate models are capable of seeing all categories separately and simultaneously. As seen in Table 3, the difference between the univariate and multivariate estimates are mostly coming from (1) the NB and Lindley distribution parameters C , θ , and ϕ , as (2) the dependence parameter, γ_0 . It should be noted that these parameters are highly correlated and if one changes, the others will change accordingly. This matter is more discussed in the next section.

Collectively, four performance measures were selected to determine the model's fitness as well as its predictive accuracy: two Bayesian metrics, Widely Applicable Information Criterion (WAIC) and Leave-One-Out Cross-Validation (LOO), and two commonly used measures to assess the predictive capabilities of the model, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). Table 3 shows the performance measures for (1) univariate NB-WLindley models, (2) the multivariate NB-WLindley model, and (3) the random parameter multivariate NB-WLindley model. The univariate NB-WLindley model was considered as the benchmark model for comparison purposes. It treats each category as a separate (and independent) random variable that follows a univariate NB-WLindley distribution. Even though a multivariate and a univariate model are not directly comparable, including univariate results can help to better understand how developing a multivariate model and forming a dependency structure can affect the estimates and enhance the model performance for each category individually, and overall. In the majority of cases, the multivariate model reports better (or similar) performance measures. As expected, the random parameter multivariate NB-WLindley model showed slightly better results in all categories, compared to the fixed-parameter multivariate model. Table 4 summarizes the modeling results for run-off-the-road crashes in Texas. The Texas dataset categorizes crash counts by three crash types namely: (1) overturn crashes, (2) guardrail crashes, and (3) other fixed-object crashes. All modeling results confirmed that (1) as the traffic volume increases at intersecting roads, the number of crashes also goes higher, which is supported by previous studies, and (2) wider roadways result in more crashes. This could be due to the fact that wider roads with more visible sight distances tend to encourage higher speeds and more aggressive driving behaviors. Wider roads can also give drivers a false sense of security, leading to less caution and greater risk-taking. Additionally, wider roads often have multiple lanes, increasing the complexity of navigating intersections due to the potential for more conflict points, such as lane merging and changing. Similar to the observations for the Texas dataset, the multivariate model reported either better or very similar performance measures compared to the univariate version. In all categories, both fitness and accuracy performance measures showed the superiority of the multivariate model over the univariate version suggesting that there is a dependency structure among the crash types

Table 3
Modeling results - California intersection crash data

Crash severities	Variables	Univariate NB-WLindley		Multivariate NB-WLindley		RP multivariate NB-WLindley	
		Mean	Std.dev	Mean	Std.dev	Mean	Std.dev
K	Intercept	-15.80	1.62	-15.61	1.59	-15.89	1.64
	Log (Major AADT)	1.15	0.22	1.14	0.22	1.16	0.24
	Log (Minor AADT)	0.24	0.10	0.23	0.10	0.25	0.11
	ϕ	5.73	2.59	5.40	2.60	5.82	2.54
	c/θ	0.72	0.10	0.62	0.10	0.75	0.08
	γ_0	-	-	0.40	0.09	0.58	0.18
	WAIC	380.94	-	385.86	-	383.44	-
	LOO	388.73	-	390.74	-	389.59	-
	MAE	0.21	-	0.21	-	0.19	-
	RMSE	0.10	-	0.10	-	0.09	-
A	Intercept	-13.68	1.41	-13.80	1.45	-13.74	1.45
	Log (Major AADT)	1.08	0.15	1.07	0.15	1.08	0.15
	Log (Minor AADT)	0.21	0.07	0.21	0.07	0.22	0.07
	Painted Left Turn	0.57	0.21	0.56	0.21	0.50	0.23
	Rolling	0.49	0.20	0.50	0.20	0.47	0.20
	ϕ	6.08	2.34	5.80	2.38	5.82	2.37
	c/θ	0.79	0.06	0.76	0.08	0.81	0.07
	γ_0	-	-	0.40	0.09	0.58	0.18
	WAIC	683.33	-	686.68	-	686.22	-
	LOO	691.81	-	694.10	-	694.18	-
B	MAE	0.39	-	0.40	-	0.39	-
	RMSE	0.25	-	0.26	-	0.25	-
	Intercept	-10.27	1.12	-10.24	1.13	-10.32	1.16
	Log (Major AADT)	0.98	0.09	0.98	0.09	0.98	0.09
	Log (Minor AADT)	0.17	0.04	0.17	0.04	0.17	0.05
	Mountain	0.56	0.18	0.55	0.17	0.47	0.20
	ϕ	6.16	2.20	5.79	2.20	6.65	2.04
	c/θ	0.82	0.04	0.78	0.06	0.82	0.07
	γ_0/c	-	-0.76	0.09	0.67 0.15	-	-
	WAIC	1377.15	-	1368.79	-	1367.05	-
C	LOO	1408.10	-	1398.84	-	1396.20	-
	MAE	0.84	-	0.83	-	0.75	-
	RMSE	1.11	-	1.07	-	0.81	-
	Intercept	-11.62	1.24	-11.95	1.23	-11.81	1.29
	Log (Major AADT)	1.11	0.10	1.15	0.11	1.13	0.11
	Log (Minor AADT)	0.21	0.04	0.20	0.04	0.21	0.05
	Lighting	0.54	0.14	0.52	0.15	0.46	0.16
	Mountain	0.47	0.19	0.44	0.19	0.43	0.20
	ϕ	4.25	2.11	6.06	2.20	6.17	2.23
	c/θ	0.80	0.05	0.62	0.06	0.74	0.07
PDO	γ_0	-	-	0.40	0.09	0.58	0.18
	WAIC	1410.45	-	1386.22	-	1383.98	-
	LOO	1442.81	-	1432.00	-	1430.58	-
	MAE	0.98	-	0.77	-	0.75	-
	RMSE	1.57	-	0.90	-	0.83	-
	Intercept	-9.80	1.07	-9.97	1.09	-10.08	1.10
	Log (Major AADT)	1.00	0.07	1.02	0.07	1.02	0.08
	Log (Minor AADT)	0.23	0.03	0.23	0.03	0.24	0.04
	Lighting	0.48	0.11	0.47	0.11	0.44	0.12
	Painted Left Turn	-0.24	0.11	-0.23	0.11	-0.22	0.11
PDO	Number of Lanes	0.14	0.07	0.14	0.07	0.15	0.07
	Mountain	0.57	0.14	0.57	0.14	0.50	0.16
	ϕ	5.09	1.95	5.63	2.05	5.80	2.03
	c/θ	0.83	0.03	0.76	0.06	0.76	0.07
	γ_0	-	-	0.40	0.09	0.58	0.18
	WAIC	2228.24	-	2212.41	-	2210.60	-
	LOO	2283.59	-	2278.33	-	2282.28	-
	MAE	2.13	-	1.83	-	1.66	-
	RMSE	7.41	-	5.36	-	4.34	-

which could enhance the overall model performance (in terms of both accuracy and fitness) if taken into account. The random parameter model also showed superior predictive abilities; however, similar to what was observed in the other dataset, improvements are little.

5. Discussion

As observed in the previous section, modeling results for both datasets showed that the proposed multivariate NB-WLindley model can account for the hidden dependency among different crash categories and yield superior performance overall relative to its univariate counterpart. In the following, some interesting findings and conclusions, as well as the methodology limitations, are discussed in more detail.

Table 4
Modeling results - Texas run of the road crash data

Crash types	Variables	Univariate NB-WLindley		Multivariate NB-WLindley		RP multivariate NB-WLindley	
		Mean	Std.dev	Mean	Std.dev	Mean	Std.dev
Overturn Crashes	Intercept	-9.15	0.29	-9.07	0.23	-9.14	0.31
	Log (AADT)	0.94	0.22	0.91	0.20	0.92	0.20
	Length (mi)	1.03	0.24	1.00	0.22	1.00	0.22
	ϕ	5.44	2.61	5.28	2.68	5.41	2.61
	θ	1.78	0.92	1.39	0.80	1.57	0.83
	C	1.37	0.88	1.00	0.76	1.16	0.80
	γ_0	-	-	0.16	0.07	0.17	0.10
	WAIC	200.48	-	195.99	-	196.73	-
	LOO	205.99	-	204.82	-	203.59	-
	MAE	0.43	-	0.42	-	0.38	-
	RMSE	0.11	-	0.11	-	0.10	-
Guardrail Crashes	Intercept	-10.31	0.47	-10.99	0.52	-10.72	0.47
	Log (AADT)	0.92	0.34	0.97	0.34	0.94	0.33
	Length (mi)	1.08	0.37	1.16	0.37	1.08	0.34
	ϕ	5.27	2.74	5.14	2.80	5.20	2.79
	θ	1.30	1.00	0.93	0.80	1.41	0.94
	C	0.93	0.94	0.60	0.74	1.02	0.89
	γ_0	-	-	0.16	0.07	0.17	0.10
	WAIC	95.04	-	97.45	-	94.16	-
	LOO	96.46	-	97.71	-	95.84	-
	MAE	0.15	-	0.15	-	0.14	-
	RMSE	0.01	-	0.01	-	0.01	-
Other Fixed-object Crashes	Intercept	-7.41	0.22	-7.03	0.19	-7.21	0.23
	Log (AADT)	0.83	0.16	0.75	0.16	0.77	0.15
	Length (mi)	1.11	0.19	1.21	0.20	1.17	0.19
	ϕ	3.89	2.60	5.28	2.52	4.01	2.63
	θ	1.64	0.73	0.56	0.19	0.75	0.38
	C	1.22	0.70	0.26	0.16	0.42	0.35
	γ_0	-	-	0.16	0.07	0.17	0.10
	WAIC	334.23	-	330.62	-	324.77	-
	LOO	345.61	-	342.28	-	339.34	-
	MAE	1.01	-	0.99	-	0.95	-
	RMSE	0.51	-	0.46	-	0.43	-

In both the California and Texas datasets, the multivariate model exhibited superior performance compared to the univariate model in most cases. The authors provided a comparative analysis of the multivariate model's performance across different crash categories against a corresponding univariate model tailored to each specific category. While the data in both tables demonstrate the superiority of the multivariate model, it should be noted that it is neither guaranteed nor expected to always see better metrics for the multivariate model. Assuming a dependency structure among different random variables is intended to improve the overall model flexibility and performance for all categories (*i.e.* random variable) collectively and may result in improvements in some categories while showing no significant enhancement (or even declines) in others. The random parameter multivariate models showed similar results as regular multivariate models. The likelihood performance measures showed a small improvement over fixed parameter multivariate models indicating that the model flexibility is not further improved by changing the fixed parameter to random. However, the predictive performance measures, MAE and RMSE, showed significantly better results for random parameter models.

In the modeling process, the authors faced challenges in ensuring the convergence of the parameters across models. This issue primarily stemmed from the high correlation among parameters (especially those introduced by the Lindley distribution) in the NB-WLindley model. On the one hand, according to (Khodadadi et al., 2022), the parameters

C , θ , and β_0 are highly correlated, which makes it challenging to achieve convergence in their MCMC chains. On the other hand, the multivariate model introduces an additional hyper-parameter, γ_0 , which adjusts the Lindley mean function (and hence affects all the parameters inside the function), making the mixed model even more complex. The high correlation among these parameters suggests that they are likely interdependent, meaning that a change in one is expected to be mirrored by a change in the others. This relationship is indicative of a complex system where each parameter does not act in isolation but rather in interaction with others, influencing and being influenced dynamically. The convergence plots also showed poor mixing for these parameters even after running for many iterations and high thinning. To avoid the mentioned convergence issues, the authors made a series of adjustments to mitigate them and achieve convergence faster.

Since both the conditional mean and variance of the multivariate NB-WLindley distribution are adjusted by ϵ , previous studies have assumed an informative prior distribution on parameters θ and C to ensure that ϵ s have their mean equal to one, $E(\epsilon_{ij}) = 1$. This ensures the uniqueness of the parameter estimates and, therefore, the identifiability of the model. According to the definition of the WLindley distribution, the mean function could be defined as follows in terms of two parameters, C and θ :

$$E(\epsilon) = \frac{c(\theta + c + 1)}{\theta(\theta + c)} = 1 \quad (17)$$

Having the mean function equal to one, the following non-linear relationship between the two parameters could be expressed as follows:

$$\theta^2 = c^2 + c \quad (18)$$

Therefore, c should be essentially smaller than θ . This led the researchers to create a new joint parameter, $(\frac{c}{\theta})_j$, for each c_j and θ_j , and then assumed a prior distribution on the joint variable instead.

$$(\frac{c}{\theta})_j \sim \text{uniform}(0, 1) \quad (19)$$

From Eq. (17) to Eq. (19), simulations for parameters c_j and θ_j are then calculated as follows:

$$c_j = \frac{1}{(\frac{c}{\theta})_j^2 - 1} \quad (20)$$

$$\theta_j = (\frac{c}{\theta})_j * c_j \quad (21)$$

As mentioned before, the common shock variable, γ_0 appears in the mean function and affects the convergence

of other parameters in the mean function including c , θ , and ϕ . This led the authors to take a similar procedure and instead of parameterizing γ_0 itself, define a joint variable first, $\frac{\gamma_0}{c}$ and then use the joint variable simulations to estimate γ_0 . Note that from the definition of the multivariate WLindley distribution in Eq. (5), γ_0 should be essentially smaller than all c_j s. Therefore, we can set a uniform prior distribution for the new joint parameter, $\frac{\gamma_0}{c}$:

$$\left(\frac{\gamma_0}{c}\right) \sim \text{uniform}(0, 1) \quad (22)$$

Then, the γ_0 simulations can be calculated by multiplying c and $\frac{\gamma_0}{c}$ simulations.

6. Summary and Conclusion

A multivariate distribution describes the probability distribution of a random vector—a collection of two or more random variables. This approach is especially useful when the model variables in question are interrelated, allowing analysts to explore the relationship between them, rather than viewing each in isolation. In transportation safety research, multivariate models have proven effective in analyzing crash data, enabling better predictions and assessments of safety measures. In an attempt to extend the application of multivariate count models in transportation safety, this paper aims to redefine an advanced count model within a multivariate framework. In particular, the NB-L distribution is the subject of interest due to its superior performance over traditional models such as Poisson and NB distributions. A univariate NB-L distribution can explain the crash count variability in highly dispersed or sparse datasets. The superior performance of the NB-L in crash prediction models on the one hand, and the necessity of an advanced multivariate model to capture interdependence between random variables, on the other hand, motivates this study to (1) form a multivariate NB-L distribution to extend the application of the multivariate models into the advanced and flexible count models (2) examine the application of the proposed model in real-world scenarios, and (3) explore how the multivariate model can be further adjusted (e.g., extended into a random parameter framework).

The proposed multivariate NB-WL model is hierarchically defined as a mixture of NB and multivariate WLindley distributions and supports a dependence structure (i.e. common shock) to explain the interdependence among random variables in multivariate WLindley. Two crash datasets were selected to examine the capabilities of the proposed multivariate model. Both datasets categorize crash frequencies into multiple sub-categories (by crash type or crash severity), necessitating the use of a multivariate distribution that can handle all these random variables at once and accounts for the interdependence structure among various categories. Collectively, four performance measures, WAIC, LOO, MAE, and RMSE were selected to determine the model's fitness as well as its predictive accuracy. Even though multivariate and univariate models are not directly comparable, the authors developed a univariate model for each random variable to better display how developing a multivariate model and forming a dependency structure can affect

the estimates and enhance the model performance for each category individually, and overall. In the majority of cases, the multivariate model reports better (or similar) performance measures. As expected, the random parameter multivariate NB-WLindley model showed slightly better results in all categories, compared to the fixed-parameter multivariate model.

In conclusion, this study demonstrated how to effectively extend the application of an advanced count model into a multivariate context. The results indicated that the multivariate models, especially with random parameters, offer superior performance in analyzing crash data, emphasizing the importance of considering inter-dependencies among variables. This approach provides a valuable tool for transportation safety research, with potential for further refinement and application in other fields where multivariate count data is relevant.

References

- Anastasopoulos, P.C., 2016. Random parameters multivariate tobit and zero-inflated count data models: addressing unobserved and zero-state heterogeneity in accident injury-severity rate and frequency analysis. *Analytic methods in accident research* 11, 17–32.
- Avelar, R., Geedipally, S., Das, S., Wu, L., Kutela, B., Lord, D., Tsapakis, I., et al., 2020. Evaluation of Roadside Treatments to Mitigate Roadway Departure Crashes: Technical Report. Technical Report. Texas A&M Transportation Institute.
- Barua, S., El-Basyouny, K., Islam, M.T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. *Analytic methods in accident research* 9, 1–15.
- Bermúdez, L., Karlis, D., 2011. Bayesian multivariate poisson models for insurance ratemaking. *Insurance: Mathematics and Economics* 48, 226–236.
- Bolancé, C., Vernic, R., 2019. Multivariate count data generalized linear models: Three approaches based on the sarmanov distribution. *Insurance: Mathematics and Economics* 85, 89–103.
- Dentheth, S., Promin, P., 2019. The negative binomial-weighted lindley distribution. *Decision Science Letters* 8, 317–322.
- Dentheth, S., Thongteeraparp, A., Bodhisuwan, W., 2016. Mixed distribution of negative binomial and two-parameter lindley distributions, in: 2016 12th International Conference on Mathematics, Statistics, and Their Applications (ICMSA), IEEE. pp. 104–107.
- Fu, C., Sayed, T., 2022. A multivariate method for evaluating safety from conflict extremes in real time. *Analytic methods in accident research* 36, 100244.
- 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.
- Ghitany, M., Alqallaf, F., Al-Mutairi, D.K., Husain, H., 2011. A two-parameter weighted lindley distribution and its applications to survival data. *Mathematics and Computers in simulation* 81, 1190–1201.
- Ghitany, M., Karlis, D., Al-Mutairi, D., Al-Awadhi, F., 2012. An em algorithm for multivariate mixed poisson regression models and its application. *Applied Mathematical Sciences* 6, 6843–6856.
- Gomez-Deniz, E., Sarabia, J.M., Balakrishnan, N., 2012. A multivariate discrete poisson-lindley distribution: Extensions and actuarial applications. *ASTIN Bulletin: The Journal of the IAA* 42, 655–678.
- Heydari, S., Fu, L., Joseph, L., Miranda-Moreno, L.F., 2016. Bayesian nonparametric modeling in transportation safety studies: applications in univariate and multivariate settings. *Analytic methods in accident research* 12, 18–34.

- Heydari, S., Fu, L., Miranda-Moreno, L.F., Jopseph, L., 2017. Using a flexible multivariate latent class approach to model correlated outcomes: A joint analysis of pedestrian and cyclist injuries. *Analytic methods in accident research* 13, 16–27.
- Islam, A.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.
- Jonathan, A.V., Wu, K.F.K., Donnell, E.T., 2016. A multivariate spatial crash frequency model for identifying sites with promise based on crash types. *Accident Analysis & Prevention* 87, 8–16.
- Karlis, D., Meligkotsidou, L., 2005. Multivariate poisson regression with covariance structure. *Statistics and Computing* 15, 255–265.
- Khodadadi, A., Shirazi, M., Geedipaly, S., Lord, D., 2022. Evaluating alternative variations of negative binomial-lindley distribution for modeling crash data. *Transportmetrica A: Transport Science*.
- 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.
- Kim, D.G., Washington, S., Oh, J., 2006. Modeling crash types: New insights into the effects of covariates on crashes at rural intersections. *Journal of Transportation Engineering* 132, 282–292.
- Lindley, D.V., 1958. Fiducial distributions and bayes' theorem. *Journal of the Royal Statistical Society. Series B (Methodological)*, 102–107.
- Lindley, D.V., 1965. Introduction to probability and statistics: from a Bayesian viewpoint. 2. Inference. CUP Archive.
- Ma, J., Kockelman, K.M., 2006. Bayesian multivariate poisson regression for models of injury count, by severity. *Transportation Research Record* 1950, 24–34.
- Ma, X., Chen, S., Chen, F., 2017. Multivariate space-time modeling of crash frequencies by injury severity levels. *Analytic Methods in Accident Research* 15, 29–40.
- 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.
- Mathai, A.M., Moschopoulos, P.G., 1991. On a multivariate gamma. *Journal of Multivariate Analysis* 39, 135–153.
- Mothafer, G.I., Yamamoto, T., Shankar, V.N., 2016. Evaluating crash type covariances and roadway geometric marginal effects using the multivariate poisson gamma mixture model. *Analytic methods in accident research* 9, 16–26.
- Park, E.S., Lord, D., 2007. Multivariate poisson-lognormal models for jointly modeling crash frequency by severity. *Transportation Research Record* 2019, 1–6.
- Park, E.S., Oh, R., Ahn, J.Y., Oh, M.S., 2021. Bayesian analysis of multivariate crash counts using copulas. *Accident Analysis & Prevention* 149, 105431.
- Russo, B.J., Savolainen, P.T., Schneider IV, W.H., Anastasopoulos, P.C., 2014. Comparison of factors affecting injury severity in angle collisions by fault status using a random parameters bivariate ordered probit model. *Analytic methods in accident research* 2, 21–29.
- Sacchi, E., Sayed, T., El-Basyouny, K., 2015. Multivariate full bayesian hot spot identification and ranking: New technique. *Transportation research record* 2515, 1–9.
- Samutwachirawong, S., 2017. A negative binomial-new weighted lindley distribution for count data and its application to hospitalized patients with diabetes at ratchaburi hospital, thailand, in: *The 2nd International Conference of Multidisciplinary Approaches on UN Sustainable Development Goals (UNSDGs)*.
- Shanker, R., Fesshaye, H., Sharma, S., 2016. On two parameter lindley distribution and its applications to model lifetime data. *Biom. Biostat. Int. J* 3, 00056.
- Shanker, R., Sharma, S., Shanker, R., 2013. A two-parameter lindley distribution for modeling waiting and survival times data.

- Shanker, R., Shukla, K.K., Shanker, R., Tekie, A., 2017. A three-parameter lindley distribution. *American Journal of Mathematics and Statistics* 7, 15–26.
- 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.
- Shi, P., Valdez, E.A., 2014. Multivariate negative binomial models for insurance claim counts. *Insurance: Mathematics and Economics* 55, 18–29. URL: <https://www.sciencedirect.com/science/article/pii/S0167668713001947>, doi:<https://doi.org/10.1016/j.insmatheco.2013.11.011>.
- Shirazi, M., Dhavala, S.S., Lord, D., Geedipally, S.R., 2017. 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.
- Tajuddin, R.R.M., Ismail, N., Ibrahim, K., Bakar, S.A.A., 2020. A four-parameter negative binomial-lindley distribution for modeling over and underdispersed count data with excess zeros. *Communications in Statistics-Theory and Methods* , 1–13.
- Tsionas, E.G., 2001. Bayesian multivariate poisson regression. *Communications in Statistics-Theory and Methods* 30, 243–255.
- Zamani, H., Ismail, N., 2010. Negative binomial-lindley distribution and its application. *Journal of Mathematics and Statistics* 6, 4–9.
- Zheng, L., Sayed, T., 2020. A bivariate bayesian hierarchical extreme value model for traffic conflict-based crash estimation. *Analytic methods in accident research* 25, 100111.