Accident Analysis and Prevention 207 (2024) 107711

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

The negative Binomial-Lindley model with Time-Dependent Parameters: Accounting for temporal variations and excess zero observations in crash data

Richard Dzinyela [a,*], Mohammadali Shirazi [b], Subasish Das [c], Dominique Lord [a]

[a] Zachary Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX, 3136 TAMU, College Station, TX 77843-3136, United States
[b] Department of Civil and Environmental Engineering, University of Maine, Orono, ME, 04469, United States
[c] Civil Engineering, Ingram School of Engineering, Texas State University, 601 University Drive, San Marcos, TX, 78666, United States

ARTICLE INFO

Keywords:
Crash count
Temporal variation
Short-duration analysis
Negative binomial Lindley
Excess zeros observation

ABSTRACT

Crash counts are non-negative integer events often analyzed using crash frequency models such as the negative binomial (NB) distribution. However, due to their random and infrequent nature, crash data usually exhibit unique characteristics, such as excess zero observations that the NB distribution cannot adequately model. The negative binomial-Lindley (NBL) and random parameters negative binomial-Lindley (RPNBL) models have been proposed to address this limitation. Despite addressing the issues of excess zero observations, these models may not fully account for unobserved heterogeneity resulting from temporal variations in crash data. In addition, many variables, such as traffic volume, speed, and weather, change with time. Therefore, the analyst often requires disaggregated data to account for their variations. For example, it is recommended to use monthly crash datasets to better account for temporally varying weather variables compared to yearly crash data. Using temporally disaggregated data not only adds the complexity of the temporal variations issue in data but also compounds the issue of excess zero observations. To address these issues, this paper introduces a new variant of the NBL model with coefficients and Lindley parameters that vary by time. The derivations and characteristics of the model are discussed. Then, the model is illustrated using a simulation study. Subsequently, the model is applied to two empirical crash datasets collected on rural principal and minor arterial roads in Texas. These datasets include several time-dependent variables such as monthly traffic volume, standard deviation of speed, and precipitation and exhibit unique characteristics such as excess zero observations. The results of several goodness-of-fit (GOF) measures indicate that using the NBL model with time-dependent parameters enhances the model fit compared to the NB, NBL, and the NB model with time-dependent parameters. Findings derived from crash data collected from both rural minor and principal arterial roads in Texas suggest that the variables denoting the median presence and wider shoulder width are associated with a potential decrease in crash occurrences. Moreover, higher variations in speed and wider road surfaces are linked to a potential increase in crash frequency. Similarly, a higher monthly average daily traffic (Monthly ADT) positively correlates with crash frequency. We also found that it is important to account for temporal variations using time-dependent parameters.

1. Introduction

According to the Centers for Disease Control and Prevention (CDC), an estimated 3,700 people are killed every day globally due to crashes involving cars, buses, motorcycles, bicycles, trucks, or pedestrians. Fatalities due to crash injuries are estimated to be the eighth leading cause of death among people of all age groups, as more people now die in crashes than from HIV/AIDS (Center for Disease Control and Prevention (2023)). In 2022, Texas, U.S. experienced 1.55 deaths per hundred million vehicle miles traveled (VMT) and one reportable crash every 57s (Texas Department of Transportation, 2023). Due to the high number of crashes coupled with the loss of lives and economic consequences associated with crashes, traffic safety is still a significant concern for the state government, transportation planners, and stakeholders.

* Corresponding author.
E-mail addresses: dzinyela_1@tamu.edu (R. Dzinyela), shirazi@maine.edu (M. Shirazi), subasish@txstate.edu (S. Das), dlord@civil.tamu.edu (D. Lord).

https://doi.org/10.1016/j.aap.2024.107711
Received 3 April 2024; Received in revised form 27 June 2024; Accepted 7 July 2024

Improving safety requires the prediction of crashes and identifying factors that affect crashes using robust statistical models (Lord & Mannering, 2010; Mannering & Bhat, 2014; Mannering et al., 2016; Lord et al., 2021; Islam et al., 2023; Dzinyela et al., 2024). Crashes are random, infrequent, and independent events. To model non-negative crash count data, statistical models like Poisson and negative binomial (NB) regression models are preferred over the standard least squares regression models (Washington et al., 2020; Lord et al., 2021). Results from crash frequency models could be used to understand and quantify the association between crash frequency and variables or predictors that could influence safety. In addition, crash frequency models could be used as prediction tools for determining the expected number of crashes on a roadway segment and identifying crash hotspots to implement targeted countermeasures (Gil-Marin et al., 2024). Given the use of crash frequency models, developing a sound and reliable model to analyze crash data is imperative. However, as Lord and Mannering (2010) highlighted, crash frequency models are associated with important data and methodological issues. Therefore, the choice of an appropriate crash frequency model depends on the data and specific objectives of the study.

One prominent issue associated with crash count data is overdispersion. The Poisson model, one of the widely used count models, assumes that the mean and variance of the count data are equal (equidispersion). However, this phenomenon is rare and most crash count datasets have a variance that is greater than the mean, referred to as overdispersion. Using Poisson regression models to model overdispersed data leads to biased and inconsistent parameter estimates (Lord & Mannering, 2010; Lord et al., 2021). To account for overdispersion, the Poisson-gamma (also called negative binomial) and the Poisson-lognormal (PL) models, extensions of the mixed-Poisson models, are generally used. For insights into the choice between the NB and PL models, readers are referred to the work of Shirazi and Lord (2019). Accounting for overdispersions alone is insufficient as crash count data may exhibit other unique features such as excess zero observations, long and heavy tail, low sample mean, time-dependent variables, and temporal and spatial correlations, all of which can profoundly impact the model performance. Among those, addressing the excess zero observations and the unobserved heterogeneity due to time-dependent variables are the two main motivations of this study.

First, let us discuss the first motivation, addressing excess zero observations. Previous studies have examined different models developed to address the problem with data that exhibit excess zero observations. The negative binomial-Lindley (NBL) model was introduced by Zamani and Ismail (2010) to account for the excess zero observations problem. The NBL model combines the negative binomial model and the Lindley distribution. The NBL was first introduced as a two-parameter distribution, which, unlike the zero-inflated models, has a long-term mean that is not equal to zero. Recently, various extensions of the NBL model have been introduced, such as the three and four parameters NBL distributions and the weighted NBL model (Denthet et al., 2016; Tajuddin et al., 2022; Khodadadi et al., 2023). In terms of crash count modeling, Lord and Geedipally (2011) used the NBL distribution to analyze crash datasets that exhibit excess zero observations and compared the performance of this model with results from both the NB and Poisson models. The NBL distribution outperformed the NB and Poisson distributions in terms of the model fit when data exhibits excessive zero observations. Later, Geedipally et al. (2012) formulated the generalized linear model (GLM) extension for the NBL model and compared its performance with the zero-inflated negative binomial (ZINB) model, which the GLM NBL was found to outperform the ZINB in terms of the model fit. Shirazi et al. (2017a) showed that the NBL model outperforms the NB model when the skewness of crash data is large. Several other extensions of the NBL model have also been studied, such as the random parameter NBL model (RPNBL) (Shaon et al., 2018), finite mixture NBL model (FMNBL) (Islam et al., 2022), Grouped RPNBL (G-RPNBL) (Islam et al., 2023) and the empirical Bayes representation of NBL (Khodadadi et al., 2022).

Another concern with crash data modeling involves accounting for the time-dependent variables. Addressing this issue is the second motivation of this study. As crashes are sometimes counted over more extended periods, some explanatory variables such as speed, traffic volume, economic indicators, pavement skid number, and weather patterns may change significantly over time (Ye et al., 2012; Fountas et al., 2020; Islam et al., 2023; Sawtelle et al., 2023). Failing to account for this within-period variation in the explanatory variables may result in a loss of relevant information, also called unobserved heterogeneity (Lord and Mannering, 2010; Mannering et al., 2016; Mannering, 2018). Poch and Mannering (1996) used a panel dataset to investigate rear-end crashes using the NB model without accounting for temporal variability. In response, Lord and Persaud (2000) developed crash prevention models to account for year-to-year variations or trends in crash count. Using the Generalized estimating equations (GEE) model that estimated different models for each year and another that assumed identical coefficients for each year, Lord and Persaud (2000) found that not accounting for temporal correlation considerably underestimates the variance of the parameter estimates. Similarly, Wang and Abdel-Aty (2006) investigated the temporal and spatial effect of rear-end crashes at signalized intersections using the (GEE) with negative binomial models. Wang and Abdel-Aty (2006) found that using GEE with a negative binomial model outperforms the conventional negative binomial model and concluded that the temporal correlation should be considered for the panel data by methods like the GEE. Other studies have also used temporally weighted negative binomial regression model (Fu et al., 2022), multivariate Poisson-lognormal models with temporal random effects model (Cheng et al., 2017), integrated nested Laplace approximation model (Liu and Sharma, 2017), and Poisson-lognormal multivariate conditional autoregressive model (Wang and Kockelman, 2013) among many others to account for temporal variations in panel datasets.

Random parameters models have recently been introduced to address unobserved heterogeneity that may arise due to complex events not documented in the police crash report or other omitted variables (Mannering et al., 2016; Lord et al., 2021). Random parameters allow parameters to vary from one observation to another. The random parameters NB (RPNB) is one of the most used crash frequency models to account for unobserved heterogeneity (see Venkataraman et al., 2011; Rusli et al., 2017; Zhang et al., 2022). Other extensions, such as RPNBL (see Shaon et al., 2018; Rusli et al., 2018), have also gained prominence and have been used to address unobserved heterogeneity and excess zeros problems. Previous studies also show that accounting for correlations in random parameter models provides a better fit. For instance, Coruh et al. (2015) used monthly aggregated data over three years to analyze factors affecting crash frequency using the RPNB panel count data model. These studies revealed that the model that accounts for correlation in the random parameters fits the data better than the model that does not account for the correlation (as long as the correlation exists or is present). Furthermore, to account for spatial variations, the grouped random parameters, a particular case of the random parameters, have been proposed (Washington et al., 2020). For example, Cai et al. (2018) used the grouped random parameters multivariate spatial model to investigate potential observable zonal effects and unobserved heterogeneity in the number of crashes recorded on road segments and intersections. Furthermore, Islam et al. (2023) developed the grouped RPNBL model to address the regional heterogeneity in crash datasets with preponderant zero responses.

Despite the significant advancements in random parameters models, only a few studies have focused on analyzing temporal variations in crash data (Yuan et al., 2021; Hasan and Abdel-Aty, 2024; Rim et al., 2023). To measure temporal variations in crash data, one has to aggregate the data in relatively shorter durations such as hours, days, or months. However, aggregating crash counts every year or monthly (usually called disaggregated data) rather than aggregating the number

of crashes over a more extended period (e.g., five years) may compound the issue of excess zero observations. This problem was highlighted in previous work by Lord and Geedipally (2018), who found that the percentage of zeros increases as the time scale used for aggregation decreases. This could also be applied to the length of roadway segments, where the number of zeros may increase when the analyst collects crash counts over shorter road segments compared to longer stretches of roads (Shirazi et al., 2021; Lord et al., 2021). In this case, the negative binomial may be inappropriate, and developing extensions for other models like the NBL to account for temporal variations may be helpful.

To summarize, the primary motivations of this study are two fold. First, accounting for temporal variations in crash data, and second, addressing the excess zero observations issue associated with analyzing disaggregated crash data. To address these challenges, we propose developing and exploring the characteristics of an NBL model with time-dependent parameters (NBL-TiDP) using the group random parameters strategy. In this regard, the derivations and characteristics of the model and its power to account for both temporal variations and excess zeros in panel data are documented. Then, a simulation study is conducted to illustrate the model. The model is applied to several datasets collected in Texas and compared with the NB, NBL, and the NB model with time-dependent parameters (NB-TiDP) using various goodness-of-fit (GOF) metrics.

2. Negative binomial Lindley with Time-Dependent parameters

To document the derivation of the proposed model, we first start with explaining the NB GLM. As stated earlier, the NB model is preferred over the Poisson model because it addresses overdispersion in crash data. The NB model can be formulated as a sequence of Bernoulli trials or a mixture of the Poisson and gamma distributions. The NB GLM is formulated as follows (Hilbe, 2011; Geedipally et al., 2012):

$$P(Y=y;\mu,\phi)=NB(y;\mu,\phi)=\frac{\Gamma(\phi+y)}{\Gamma(\phi)\Gamma(y+1)}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\mu}{\mu+\phi}\right)^{y};\phi>0,\mu>0 \tag{1}$$

where, μ = mean response of the observations and ϕ = inverse of the dispersion parameter. The μ is assumed to have a log-linear relationship with a list of covariates (explanatory variables) as shown in Eq. (2):

$$\ln(\mu)=\beta_0+\sum_{j=1}^{m}\beta_j X_j \tag{2}$$

where X_j = j-th explanatory variable considered for the study, β_0 is the intercept and β_j is the regression coefficient for the j-th variable, and m = the total number of variables in the model.

To account for excess zeros, the NBL model was proposed. The NBL is the mixture of the NB and Lindley distributions. The Lindley distribution provides extra flexibility to the NB model to address the excess zeros problems (Shirazi et al., 2016a). The NB-L model can be formulated as follows (Geedipally et al., 2012):

$$P(Y=y;\ \mu,\phi,\theta)=\int NB(y;\ \varepsilon\mu,\ \phi)\text{Lindley}(\varepsilon;\ \theta)d\varepsilon \tag{3}$$

where ε is the frailty term and θ is the Lindley parameter. Eq. (3) however does not have a close form. Therefore, it is often rewritten as a multilevel hierarchical structure as shown in Eq. (4) (Geedipally et al., 2012):

$$y_i|\varepsilon_i\mu_i,\phi\ \sim\ NB(\varepsilon_i\mu_i,\phi) \tag{4.1}$$

$$\varepsilon_i|z_i,\theta\ \sim\ \text{Gamma}(1+z_i,\theta) \tag{4.2}$$

$$z_i|\theta\ \sim\ \text{Bernoulli}\left(\frac{1}{1+\theta}\right) \tag{4.3}$$

$$\ln(\mu_i|\beta_0,\beta_1\cdots\beta_m)=\beta_0+\sum_{j=1}^{m}\beta_j X_{ij} \tag{4.4}$$

As previously discussed, random parameters (RP) models are one of the methods proposed for accounting the unobserved heterogeneity by allowing the parameters to vary from one observation to another (Barua et al., 2016; Hou et al., 2021; Anastasopoulos & Mannering, 2009; Dzinyela et al., 2023; Adanu et al., 2023). The RPNB-L is defined as follows (Shaon et al., 2018):

$$P(y_i\phi,\mu_i|\varepsilon_i)=NB(y_i;\phi,\varepsilon_i\mu_i) \tag{5.1}$$

$$\varepsilon_i\ \sim\ \text{Gamma}(\varepsilon;1+z_i,\theta) \tag{5.2}$$

$$z_i\ \sim\ \text{Bernoulli}\left(z;\ \frac{1}{1+\theta}\right) \tag{5.3}$$

$$\log(\mu_i)=\beta_0+\sum\nolimits_{j=1}^{m}\beta_{ij}X_{ij} \tag{5.4}$$

$$\beta_{ij}=\beta_j+w_{ij} \tag{5.5}$$

$$w_{ij}\ \sim\ \text{Normal}(0,\sigma_j) \tag{5.6}$$

Where the w_{ij} is the random term of the parameter that follows a normal distribution with a mean that is equal to zero and a standard deviation of σ_j.

In this study, to account for temporal variations in data, a special case of the RPNBL is derived and documented using the grouped random parameters strategy (Islam et al., 2023.) In our approach we assume a separate Lindley parameter for each t-th time (θ_t), and time-dependent parameters for the time-dependent covariates. To begin with, let us assume the parameter t as an index to indicate the t-th period, and the parameter T as the total number of temporal instances (e.g., number of months or years). Likewise, let us assume that the model has q time-independent covariates (e.g., shoulder width or lane width, which do not vary by time) and q' time-dependent covariates (e.g., weather, speed, or traffic information, which vary by time). The NBL model with time-dependent parameters (NBL-TiDP) can be formulated as the hierarchical Bayesian model shown in Eq. (6):

$$y_{it}|\varepsilon_{it}\mu_{it},\phi\ \sim\ NB(\varepsilon_{it}\mu_{it},\phi) \tag{6.1}$$

$$\varepsilon_{it}|z_{it},\theta_t\ \sim\ \text{Gamma}(1+z_{it},\theta_t) \tag{6.2}$$

$$z_{it}|\theta_t\ \sim\ \text{Bernoulli}\left(\frac{1}{1+\theta_t}\right) \tag{6.3}$$

$$\ln(\mu_{it}|\beta_1,\cdots\beta_q,\ \psi_{0,t},\ \psi_{1,t},\cdots\cdots\cdots,\ \psi_{q',t}\ )=\psi_{0,t}+\sum_{j=1}^{q}\beta_j X_{ij}+\ \sum_{j=1}^{q'}\psi_{j,t}W_{ijt} \tag{6.4}$$

$$\psi_{0,t}|\mu_0,\ \sigma_0\ \sim\ N(\mu_0,\ \sigma_0) \tag{6.5}$$

$$\psi_{j,t}|\ \mu_j,\ \sigma_j\ \sim\ N\left(\mu_j,\ \sigma_j\right);\ \forall j\in\{1,\cdots\cdots,\ q'\}\text{ and }\forall t\in\{1,\cdots\cdots,\ T\} \tag{6.6}$$

where,

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

X_{ij} = The value of the j-th time-independent covariate for the i-th site.

$\psi_{0,t}$ = The intercept for the time "t".

$\psi_{j,t}$ = The coefficient for the j-th covariate at the t-th time.

W_{ijt} = The value of the j-th time-dependent covariate for the i-th site at the t-th time.

μ_0 = The mean of the random intercepts.

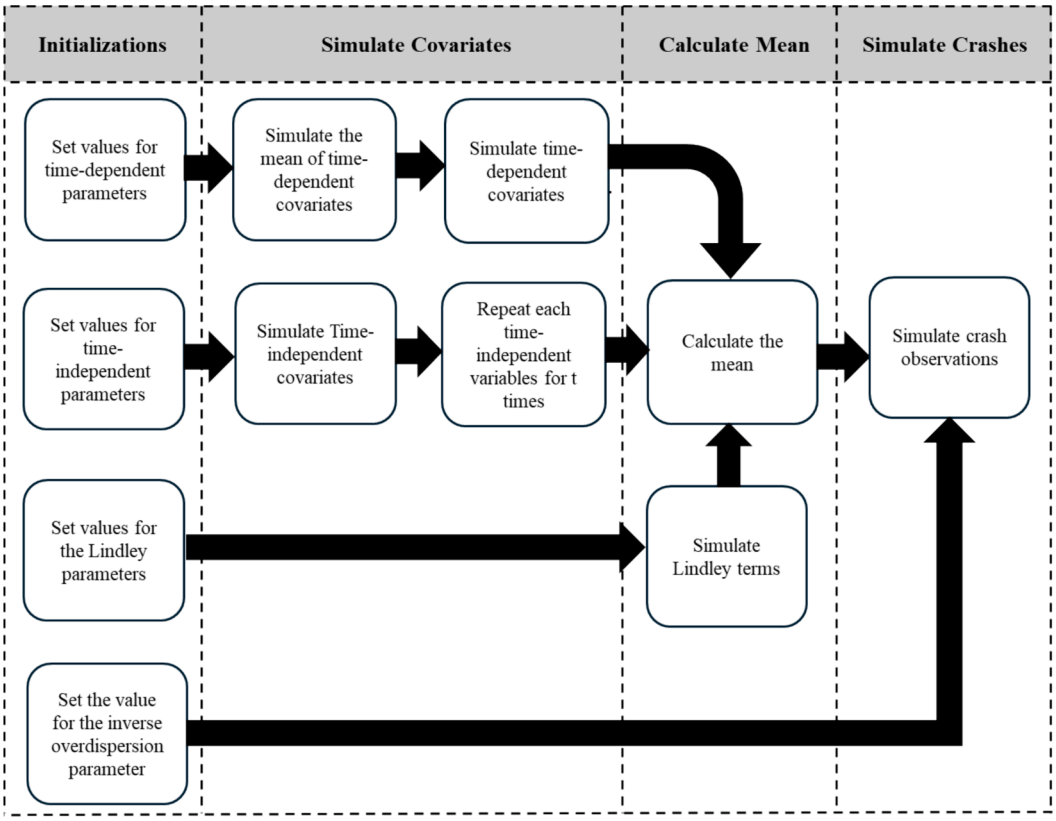

Fig. 1. Graphical representation of the simulation protocol.

σ_0 = The standard deviation of the random intercepts.

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

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

θ_t = The Lindley parameter for the t-th time.

A normal prior was assumed on the fixed parameters coefficients (β), and a gamma prior on the inverse dispersion parameter (ϕ). Similarly, a normal prior was assumed for μ_0 and μ_j and a gamma prior for the inverse σ_0 and inverse σ_j. In addition, a uniform prior was assumed on the parameters $1/(1+\theta_t)$. Notably, there are correlations between the time-varying intercepts ($\psi_{0,t}$) and the Lindley terms (ε_{it}) which might result in poor mixing. These mixing issues could be addressed by using informative priors in a way that ensures $E(\varepsilon_{it})$ is equal to one (Geedipally et al., 2012; Shaon et al., 2018; Islam et al., 2023). Another way is to drop the intercepts from the model and then calculate the time-dependent intercepts from the Lindley terms after the convergence using Eq. (7) as discussed in Islam et al. (2023).

$$\psi_{0,t} = E(\log(E(\varepsilon_{it}))) = E\left(\log\left(\frac{\theta_t+2}{\theta_t(\theta_t+1)}\right)\right) \quad (7)$$

It is worth noting that the value of $\psi_{0,t}$ can be computed using Markov Chain Monte Carlo (MCMC) samples (see Islam et al., 2023). For this purpose, a sample is drawn from the posterior distribution of $\log\left(\frac{\theta_t+2}{\theta_t(\theta_t+1)}\right)$ in each MCMC iteration. In the end, the parameter estimate for each time duration can be computed using an average over the generated samples.

3. Simulation analysis

Simulation studies are normally used to replicate real-world scenarios. This type of study has been used in past research to analyze the performance of different models under different ranges of sample mean and percentage zeros or other methodological derivations (see Khodadadi et al., 2023; Islam et al., 2022; Islam et al., 2023; Tajuddin et al., 2022; Shirazi et al., 2016b; Shirazi et al., 2017b). This section documents a simulation study used to evaluate the performance of the proposed NBL-TiDP model. We divided the simulation study into two parts. In the first part, the protocol to generate simulated data is discussed. In the second part, the results of the simulation study are discussed.

3.1. Simulation protocol

Without loss of generality, the simulation protocol considers five time periods (t = 5), and 3,000 sites. Two time-dependent covariates (i.e., their values vary during each time), and two time-independent covariates (i.e., their values do not vary during each time) are assumed for the analysis. Fig. 1 shows the general steps, whereas the following instructions provide a step-by-step guide for the simulation protocol.

Step 1. Initialization (set the true parameters).

1.1 Set the values of ψ_{11}, ψ_{12}, ψ_{13}, ψ_{14} and ψ_{15} to represent the value of five time-dependent parameters for the first time-dependent covariate, and ψ_{21}, ψ_{22}, ψ_{23}, ψ_{24} and ψ_{25} to represent the value of five time-dependent parameters for the second time-dependent covariate.

1.2 Set the value of β_1 and β_2 to represent the time-independent parameters for the two time-independent covariates.

Table 1
Characteristics of simulated data.

Total percentage of zeros	Mean	Standard deviation	Percentage of zeros in each period				
			Period 1	Period 2	Period 3	Period 4	Period 5
~50 %	4.26	17.62	40 %	43 %	52 %	59 %	61 %
~60 %	3.90	22.70	52 %	60 %	65 %	67 %	67 %
~70 %	2.31	14.55	58 %	62 %	77 %	79 %	77 %
~80 %	0.83	6.13	75 %	77 %	83 %	85 %	84 %
~90 %	0.25	1.70	88 %	89 %	91 %	92 %	90 %

Table 2
Modeling results for simulated data.

Parameters		Percentage of Zeros									
		~50 %		~60 %		~70 %		~80 %		~90 %	
		True Value	Estimate (S. D.)†	True Value	Estimate (S. D.)†	True Value	Estimate (S. D.)†	True Value	Estimate (S. D.)†	True Value	Estimate (S. D.)†
Fixed Parameters											
β_1		−1.5	−1.493 (0.014)	−1.5	−1.500 (0.016)	−1.5	−1.492 (0.018)	−1.5	−1.482 (0.023)	−1.5	−1.433 (0.032)
β_2		0.5	0.477 (0.012)	0.5	0.501 (0.014)	0.5	0.508 (0.016)	0.5	0.481 (0.021)	0.5	0.475 (0.027)
Time-dependent parameters											
ψ_1	ψ_{11}	0.5	0.477 (0.025)	0.5	0.534 (0.028)	0.5	0.542 (0.029)	−0.5	−0.455 (0.039)	0.5	0.512 (0.055)
	ψ_{12}	−0.2	−0.182 (0.025)	1.5	1.447 (0.034)	1.5	1.510 (0.036)	−1.5	−1.535 (0.046)	1.5	1.405 (0.066)
	ψ_{13}	−0.6	−0.562 (0.028)	1.0	0.969 (0.034)	1.0	1.011 (0.043)	−1.0	−0.979 (0.050)	1.0	1.053 (0.068)
	ψ_{14}	0.2	0.222 (0.030)	−0.5	−0.471 (0.033)	−0.5	−0.447 (0.041)	0.5	0.512 (0.050)	−0.5	−0.438 (0.068)
	ψ_{15}	0.3	0.285 (0.030)	−1.0	−0.982 (0.035)	−1.0	−0.980 (0.044)	1.0	1.013 (0.050)	−1.0	−1.016 (0.068)
ψ_2	ψ_{21}	0.3	0.265 (0.024)	−1.0	−0.948 (0.030)	−1.0	−0.994 (0.032)	1.0	0.999 (0.042)	−1.0	−0.939 (0.058)
	ψ_{22}	−0.5	−0.481 (0.025)	0.5	0.466 (0.032)	0.5	0.449 (0.032)	−0.5	−0.466 (0.041)	0.5	0.497 (0.062)
	ψ_{23}	0.2	0.216 (0.028)	−0.5	−0.515 (0.033)	−0.5	−0.569 (0.041)	0.5	0.473 (0.047)	−0.5	−0.550 (0.067)
	ψ_{24}	−0.8	−0.786 (0.031)	1.0	0.997 (0.035)	1.0	0.977 (0.043)	−1.0	−0.994 (0.054)	1	1.016 (0.074)
	ψ_{25}	−0.2	−0.169 (0.032)	1.5	1.581 (0.038)	1.5	1.472 (0.046)	−1.5	−1.408 (0.055)	1.5	1.527 (0.068)
θ	θ_1	0.8	0.821 (0.018)	1.5	1.471 (0.038)	2.0	2.052 (0.059)	6.0	5.896 (0.254)	21	19.150 (1.317)
	θ_2	1.0	0.990 (0.022)	2.5	2.446 (0.076)	3.0	3.041 (0.106)	8.0	7.944 (0.404)	35	32.450 (2.897)
	θ_3	1.5	1.490 (0.037)	3.0	3.083 (0.101)	7.0	7.153 (0.324)	12	11.560 (0.636)	38	34.930 (3.145)
	θ_4	2.0	2.034 (0.056)	3.5	3.390 (0.113)	8.0	7.776 (0.350)	15	14.720 (0.908)	45	35.690 (3.220)
	θ_5	2.5	2.345 (0.065)	4.0	4.014 (0.154)	10	8.933 (0.467)	20	17.970 (1.201)	50	43.500 (4.288)
Inverse Overdispersion Parameter											
ϕ		5.0	4.890 (0.473)	5.0	5.704 (0.7498)	5.0	6.025 (0.999)	5.0	4.49 (0.8337)	5.0	6.091 (2.328)

† Standard deviation

1.3 Set θ_1, θ_2, θ_3, θ_4, and θ_5 to represent the Lindley parameters for five time periods.
1.4 Set the value of the inverse overdispersion parameter (ϕ)

Step 2. Simulate time-dependent covariates.

2.1 Simulate the mean of time-dependent covariates ($\overline{W}_{i1}$ and $\overline{W}_{i2}$) from the standard normal distributions, as follows
$\overline{W}_{i1} \sim$ Normal(0,1) i = 1, ……, 3000
$\overline{W}_{i2} \sim$ Normal(0,1) i = 1, ……, 3000
2.2 Assuming a standard deviation of 0.1, and the mean values of $\overline{W}_{i1}$ and $\overline{W}_{i2}$, simulate five different values to represent the time-dependent covariates for five periods (t = 5), as follows
$W_{i1t} \sim$ Normal($\overline{W}_{i1}$, 0.1) t = 1, …., 5.
$W_{i2t} \sim$ Normal($\overline{W}_{i2}$, 0.1) t = 1, …., 5.

Step 3. Simulate time-independent covariates.

3.1 Simulate time-independent covariates X_1 and X_2 from the standard normal distributions as follows
$X_{i1} \sim$ Normal(0,1) i = 1, ….., 3000
$X_{i2} \sim$ Normal(0,1) i = 1, ….., 3000
3.2 Repeat each X_{i1} and X_{i2}, five times, accounting for the five time periods
$X_{i1,t} \sim$ repeat (X_{i1}, each = 5) t = 1, …., 5.
$X_{i2,t} \sim$ repeat (X_{i2}, each = 5) t = 1, …., 5.

Step 4. Simulate the Lindley terms and calculate the long-term mean parameter.

4.1 Simulate the Lindley terms for five time periods from Lindley distributions with parameters θ_1, θ_2, θ_3, θ_4, and θ_5 as follows
$\varepsilon_{i1} \sim$ Lindley (θ_1) i = 1, ……., 3000

$\varepsilon_{i2} \sim$ Lindley (θ_2) $\quad i = 1, \ldots\ldots, 3000$
$\varepsilon_{i3} \sim$ Lindley (θ_3) $\quad i = 1, \ldots\ldots, 3000$
$\varepsilon_{i4} \sim$ Lindley (θ_4) $\quad i = 1, \ldots\ldots, 3000$
$\varepsilon_{i5} \sim$ Lindley (θ_5) $\quad i = 1, \ldots\ldots, 3000$

4.2 Calculate the mean from the regression coefficients and simulated covariates as follows

$\mu_{i1} = \exp(\psi_{11}W_{i11} + \psi_{21}W_{i21} + \beta_1X_{i1} + \beta_2X_{i2}) \quad i = 1, \ldots, 3000$
$\mu_{i2} = \exp(\psi_{12}W_{i12} + \psi_{22}W_{i22} + \beta_1X_{i1} + \beta_2X_{i2}) \quad i = 1, \ldots, 3000$
$\mu_{i3} = \exp(\psi_{13}W_{i13} + \psi_{23}W_{i23} + \beta_1X_{i1} + \beta_2X_{i2}) \quad i = 1, \ldots, 3000$
$\mu_{i4} = \exp(\psi_{14}W_{i14} + \psi_{24}W_{i24} + \beta_1X_{i1} + \beta_2X_{i2}) \quad i = 1, \ldots, 3000$
$\mu_{i5} = \exp(\psi_{15}W_{i15} + \psi_{25}W_{i25} + \beta_1X_{i1} + \beta_2X_{i2}) \quad i = 1, \ldots, 3000$

Step 5. Simulate Crash observations.

5.1 Simulate 3,000 observations for each period using the negative binomial distribution as follows

$y_{i1} \sim NB(\varepsilon_{i1}\mu_{i1}, \phi) \quad i = 1, \ldots, 3000$
$y_{i2} \sim NB(\varepsilon_{i2}\mu_{i2}, \phi) \quad i = 1, \ldots, 3000$
$y_{i3} \sim NB(\varepsilon_{i3}\mu_{i3}, \phi) \quad i = 1, \ldots, 3000$
$y_{i4} \sim NB(\varepsilon_{i4}\mu_{i4}, \phi) \quad i = 1, \ldots, 3000$
$y_{i5} \sim NB(\varepsilon_{i5}\mu_{i5}, \phi) \quad i = 1, \ldots, 3000$

5.2 Combine the data, keeping labels for the site id (i) and period (t)

Step 6. Fit the Model.

Fit the simulated data using the proposed model (Eq. 6) and compare the estimated coefficients with the true parameters set in Step 1.

3.2. Simulation results

Table 1 shows the characteristics of the simulated datasets. To assess the proposed model considering different percentages of zeros, the simulated datasets were adjusted to contain a range of 50 % to 90 % of zero observations. This was achieved by controlling the regression coefficients and the Lindley Parameters. As shown in Table 1, the mean and standard deviation of the simulated crash data varied from 0.25 to 4.26 and 1.70 to 17.62, respectively. Specifically, the mean (and standard deviation) for simulated datasets containing 50% zeros were 4.26 (17.62), for those with 60% zeros were 3.90 (22.70%), and for datasets with 70% zeros were 2.31 (14.55). Moreover, simulated datasets with 80% zeros exhibited a mean (and standard deviation) of 0.83 (6.13), while those with 90% zeros showed a mean (and standard deviation) of 0.25 (1.70). These values are compatible with the mean and standard deviation information gathered from empirical data used in previous studies (Geedipally et al, 2012; Shaon et al., 2018; Islam et al., 2023). We also ensured that different periods constituted different but close percentages of zeros. From Table 1, it was observed that the dataset with approximately 50% zero observations (across all periods) has 40%, 43%, 52%, 59% and 61% of zeros in periods 1–5, the dataset with approximately 60% zero observations has 52%, 60%, 65%, 67% and 67 % of zeros in period 1–5, the dataset with 70% zero observations has 58%, 62%, 77%, 79% and 77% of zeros in periods 1–5, the dataset with 80% zero observations has 75%, 77%, 83%, 85% and 84% of zeros in periods 1–5, and the dataset with 90% zero observations has 88%, 89%, 91%, 92% and 90% of zeros in periods 1–5.

The proposed model was implemented in WinBUGS software (Spiegelhalter et al., 2003), and the parameters of the model were estimated using the MCMC method. Three MCMC chains, each with 30,000 iterations, were used for analysis. The first 5,000 posterior samples were considered as burn-in and discarded from the analysis. A thinning of 10 was also considered to ensure adequate mixing. The mean and standard deviation of the remaining 25,000 posterior samples are reported in Table 2. As shown in this table, all estimated parameters are statistically significant at the 95% confidence interval and close to the true parameters. The difference between the true and estimated values of the inverse dispersion parameter may stem from the low mean and sample size problem. This problem was highlighted by Lord (2006), noting that the smaller the sample mean could result in a greater difference between the values of the estimated and true inverse overdispersion parameter. The standard deviations associated with the estimated values indicate that the true value can be encompassed within the 95% credible interval of the estimated value.

4. Empirical applications

This section documents the applications of the proposed NBL-TiDP model to two empirical datasets collected in Texas. This section is divided into two subsections. The first subsection describes the datasets used for the analysis. Then, the second subsection covers the application of the NBL-TiDP model to these datasets and how the performance of the model is compared with other models such as the NB, NBL, and NB-TiDP.

4.1. Data description

Crash data collected in 2021 on rural minor and principal arterial roads in Texas were used for the analysis. The 2021 crash data was gathered in 12 months of data. The 10-minute interval speed data were collected from INRIX (INRIX, n.d), and hourly precipitation data were collected from the North American Land Data Assimilation System (NLDAS). The Monthly Average Daily Traffic (Monthly ADT) values for each road segment were estimated based on monthly adjustment factors from the Texas Department of Transportation. The covariates used are divided into two main categories:

Table 3
Characteristics of rural principal and minor arterial in Texas used in the crash frequency models.

Facility Type	Variables	Min	Max	Mean	Standard Deviation
Rural Principal Arterials	Number of crashes	0	19	0.602	1.141
	Median presence (1 if median is present, 0 otherwise)	0	1	0.340	0.473
	Surface width (without shoulders) (in feet)	20	128	53.450	13.580
	Shoulder width (both sides) (in feet)	0	64	14.214	10.494
	Concrete pavement (1 if pavement is concrete, 0 otherwise)	0	1	0.117	0.321
	Segment length (in miles)	0.1	4.59	0.415	0.370
Rural Minor Arterials	Number of crashes	0	10	0.442	0.871
	Median presence (1 if median is present, 0 otherwise)	0	1	0.085	0.279
	Surface width (without shoulders) (in feet)	20	114	40.34	16.918
	Shoulder width (in feet)	0	50	10.65	8.586
	Segment length (in miles)	0.1	4.10	0.461	0.397

Table 4
Mean and standard deviation of time-dependent variables.

Facility type	Months	Monthly ADT[†]	Standard Deviation of speed (mph)[†]	Precipitation Sum (inch)[†]
Rural Principal Arterials	January	22,314 (13035)	4.7 (1.149)	0.780 (0.389)
	February	24,362 (14231)	6.3 (1.964)	0.716 (0.406)
	March	18,755 (10956)	4.8 (1.441)	0.982 (0.580)
	April	19,042 (11124)	4.4 (1.444)	1.223 (0.748)
	May	19,056 (11132)	4.5 (1.491)	2.341 (1.027)
	June	18,426 (10764)	4.3 (1.275)	1.204 (0.629)
	July	18,491 (10802)	4.2 (1.329)	1.021 (0.663)
	August	19,357 (11308)	4.3 (1.229)	0.674 (0.238)
	September	19,458 (11367)	4.6 (1.429)	0.427 (0.645)
	October	18,999 (11098)	4.4 (1.487)	1.037 (0.519)
	November	18,445 (10775)	4.5 (1.442)	0.508 (0.265)
	December	18,402 (10749)	4.4 (1.336)	0.368 (0.272)
Rural Minor Arterials	January	13,278 (9365)	4.7 (1.122)	0.731 (0.369)
	February	14,249 (10050)	5.9 (1.656)	0.697 (0.382)
	March	11,246 (7932)	4.7 (1.269)	0.865 (0.516)
	April	11,235 (7924)	4.2 (1.199)	1.267 (0.754)
	May	11,306 (7974)	4.3 (1.328)	2.354 (0.967)
	June	10,897 (7686)	4.3 (1.201)	1.063 (0.530)
	July	11,127 (7848)	4.2 (1.333)	1.097 (0.706)
	August	11,409 (8047)	4.4 (1.349)	0.703 (0.222)
	September	11,362 (8014)	4.2 (1.124)	0.399 (0.561)
	October	11,167 (7877)	4.3 (1.357)	1.006 (0.505)
	November	10,971 (7738)	4.4 (1.505)	0.537 (0.259)
	December	11,383 (8029)	4.1 (1.092)	0.374 (0.252)

[†] The number in the parathesis shows the standard deviation.

1. Time-dependent covariates, like the Monthly ADT, the standard deviation of speed, and precipitation.
2. Time-independent covariates include the median presence, surface width, shoulder width, etc.

The rural minor and principal arterial roads in Texas consist of 20,184 and 23,508 road segments, respectively. For the analysis, crash frequencies were aggregated per month. The average monthly crashes are 0.442 for minor arterials and 0.612 for the principal arterial roads. Furthermore, the minor and principal arterial datasets include 70.55% and 64.56% of zero observations, respectively. In addition, the number of zero observations varies within the months of the year for both datasets. The number of zeros ranged from 66.9% to 75.3% for the minor arterials and from 61.6% to 69.0% for the principal arterials. Table 3 shows the information on crashes for both facility types and the time-independent covariates used in the study, including the median presence, surface width (the width of the road without shoulders), shoulder width, concrete pavement (for rural principal arterial roads), and the segment length. Only segments with lengths equal to or greater than 0.1 miles were used for this study since data on crash locations with shorter roadway segments may not be accurate enough (Geedipally et al., 2009).

As mentioned earlier, variables such as speed, precipitation, and traffic volume vary with time, and aggregating them could lead to a significant loss of information (Lord and Mannering, 2010; Shirazi et al., 2021). Therefore, in this study, we used the disaggregated monthly information for these time-varying covariates. The information about traffic volume, the standard deviation of speed, and precipitation was sourced externally from the crash database, before being merged with the crash frequency data. The mean and standard deviation of speed during the month of t was calculated as follows:

$$V^{t}_{mean} = \frac{\sum_{m} V^{t}_{m}}{M} \tag{8-1}$$

$$V^{t}_{s.d.} = \sqrt{\frac{(V^{t}_{m} - V^{t}_{mean})^{2}}{M-1}} \tag{8-2}$$

where,

V^{t}_{mean} = the mean of 10 min operational speed for the month of t.

$V^{t}_{s.d.}$ = Standard deviation (s.d.) of 10 min operational speed for the month of t.

V^{t}_{m} = operational speed during the m-th 10 min interval for the month t.

M = number of 10-minute intervals in a month.

The average and standard deviation of speed are highly correlated. Therefore, only one of them should be considered in the final models. The standard deviation of speed was preferred over the average speed because it exhibited greater variations across the months of the year and resulted in a better model fit. Table 4 shows the monthly variations of traffic, standard deviation of speed, and total monthly precipitation. Higher traffic volumes were observed during January, February, and September (See Fig. 2). Larger standard deviations of speed were observed in January, February, and March (see Fig. 3). Regarding weather, significant rain was recorded in May, as reflected in both datasets (See Fig. 4). For a better model fit, a dummy variable was created from the total precipitation variable with total precipitation equal to or greater than 1, having a mean (standard deviation) of 0.325 (0.468) for rural minor arterials and 0.338 (0.473) for rural principal arterials. Finally, the distribution of crashes over each month is shown in Fig. 5, with October, November, and December recording the highest number of crashes for rural principal and minor arterial roads.

4.2. Modeling results

This section discusses the results of applying the proposed NBL-TiDP model. The first subsection discusses the modeling results for the rural principal arterial road crash data while the second subsection discusses the modeling results for the rural minor arterial road crash data. The models were implemented in WinBUGS software, and the Bayesian Inference and MCMC were used to estimate model parameters. As described in the methodological section of the paper, the model intercepts were estimated separately to remove correlations between the intercepts and the Lindley terms. 30,000 MCMC iterations were run for each model; the first 5,000 samples were considered as burn-in and discarded. Estimated parameters were based on the remaining 25,000 posterior samples. The autocorrelation was also removed by using a thinning value of 10. The performance of the proposed NBL-TiDP model is compared with that of the NB, NBL, and NB-TiDP models.

Several performance metrics have been used to compare models in roadway safety analysis. Among them, the Deviance Information Criterion (DIC) value is a popular goodness of fit metric used in the Bayesian framework to compare models. However, model parameterizations can influence the estimation of the DIC value (Geedipally et al., 2014). Hence, the DIC value should not be the only criterion for model

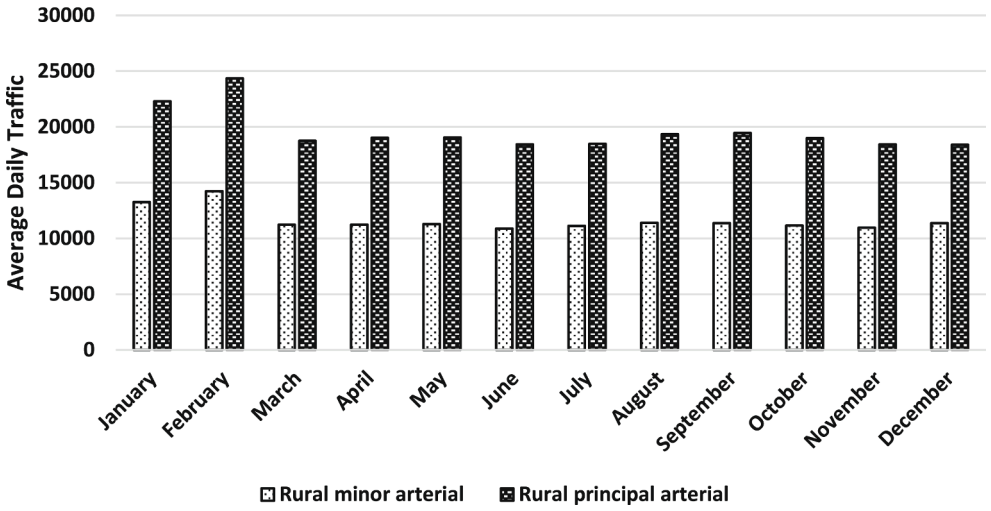

Fig. 2. Distribution of monthly ADT.

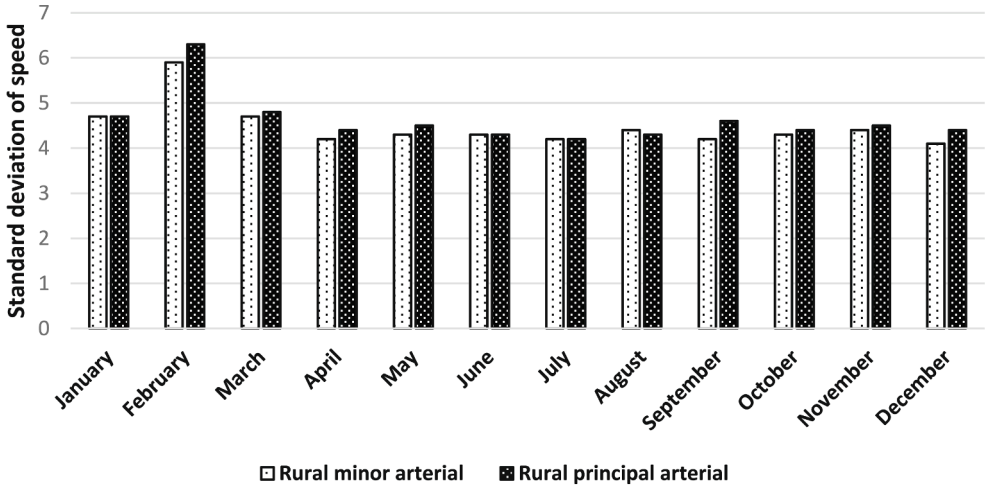

Fig. 3. Distribution of the standard deviation of observed speeds.

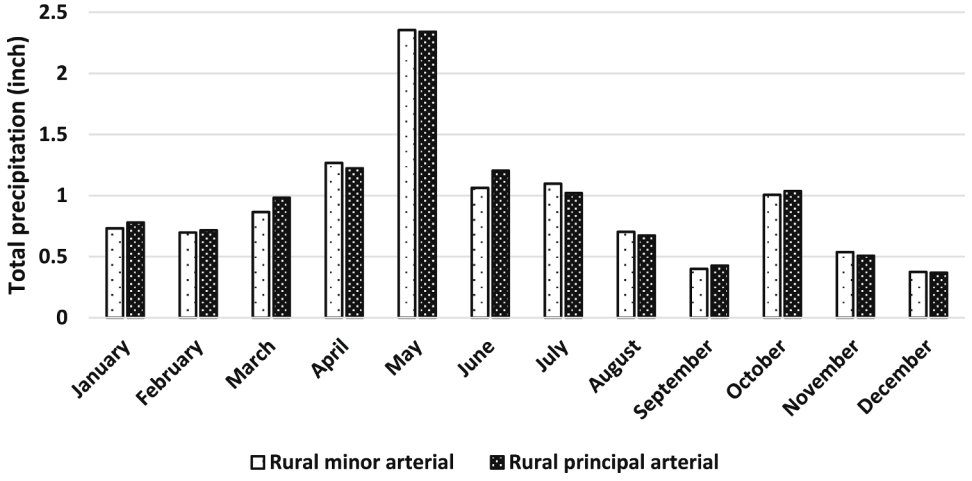

Fig. 4. Distribution of the observed total precipitation.

selection, but it can be appropriately used to compare only models with similar parameterizations (Geedipally et al., 2012; Geedipally et al.,2014). Consequently, we also considered other performance metrics, such as the Leave-one-out Cross-validation Information Criterion (LOOIC) and Widely Applicable Information Criterion (WAIC), which are the more robust alternatives to DIC (Vehtari et al., 2017) and have been used in several recent roadway safety studies (Islam et al., 2023; Khodadadi et al., 2023).

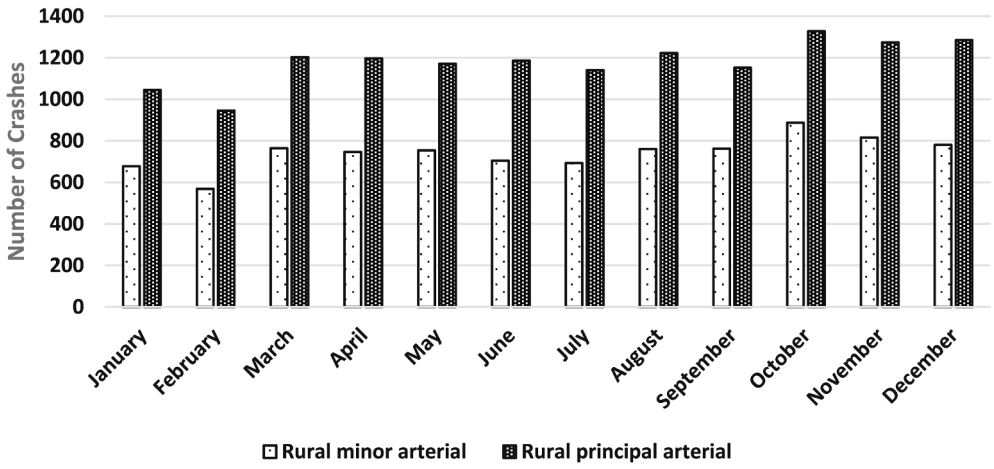

Fig. 5. Distribution of observed monthly crashes.

Table 5
Modeling results of rural principal arterial roadway data.

	NB		NB-TiDP		NBL		NBL-TiDP	
	Mean	S.D. †	Mean	S.D. †	Mean	S.D. †	Mean	S.D. †
Mean of Parameters								
Intercept	−0.598	0.017	−0.582	0.016	−6.048	0.197	−6.381	0.205
Ln (Monthly ADT)	1.137	0.05	1.147	0.087	1.135	0.051	1.195	0.086
Standard deviation of speed	0.072	0.007	*0.088*‡	*0.048*	0.072	0.007	0.097	0.045
Precipitation (1 if monthly sum > 1 in., 0 otherwise)	*−0.036*‡	*0.025*	*−0.046*‡	*0.068*	*−0.034*‡	*0.024*	*−0.039*‡	*0.072*
Median presence (1 if median is present, 0 otherwise)	−0.15	0.028	−0.148	0.026	−0.153	0.028	−0.158	0.028
Surface width (in feet)	0.004	0.001	0.004	0.001	0.004	0.001	0.003	0.001
Shoulder width (in feet)	−0.015	0.001	−0.014	0.001	−0.015	0.001	−0.015	0.001
Concrete pavement (1 if pavement is concrete, 0 otherwise)	*0.065*‡	*0.035*	*0.057*‡	*0.035*	*0.067*‡	*0.035*	*0.042*‡	*0.036*
Inverse dispersion parameter	1.001	0.03	1.023	0.032	9.301	1.695	11.29	2.414
Standard Deviation of Random Parameters								
Ln (Monthly ADT)			0.238	0.066			0.226	0.062
Standard deviation of speed			0.154	0.037			0.149	0.034
Precipitation (1 if monthly sum > 1 in., 0 otherwise)			0.2	0.052			0.21	0.055
Model Performance								
DIC	47,300		47,220		43,890		**43,730** ˆ	
WAIC	47,310		47,230		45,050		**44,860** ˆ	
LOOIC	47,311		47,231		4593		**45,780** ˆ	
MAD	1.282		1.291		**0.819** ˆ		0.839	
MSPE	2.783		2.800		**1.643** ˆ		1.677	

† Standard deviation
‡ Italic font shows non-significant at the 95% confidence level.
ˆ Bold values indicate a better fit.

4.2.1. *Rural principal arterial roadways*

Table 5 shows the modeling results for the principal arterial road datasets. The variable for segment length was used as an offset, indicating that the number of crashes has a linear relationship with the segment length. In other words, the increase in the segment length is expected to have the same proportional increase in the number of crashes. As shown in this table, the NBL-TiDP outperformed all the other models (i.e., NB, NBL, and NB-TiDP) in terms of DIC, WAIC, and LOOIC metrics. Although the objective of this study is not related to predicted values, the authors briefly examined the predictive capabilities of the models. The MAD and MPSE values for the NB and NB-L models, with and without time-dependent parameters, are shown in Table 5. Given the uncertainty associated with the prediction (see Ash et al., 2021; Lord et al., 2021), the predicted values for the NB-L and NB-L-TiDP models are considered similar based on the MAD and MPSE metrics, and they outperform those of the NB and NB-TiDP models. Note that it is possible that the 95% confidence interval of the predicted values could be smaller when the temporal variations are included. However, this is beyond the scope of this study.

The results from the NBL-TiDP model were used for the discussion.

Table 6
Estimate and standard deviations of the time-dependent coefficients for the rural principal arterial roadways.

	Monthly ADT		The standard deviation of speed	
	NB − TiDP	NBL − TiDP	NB − TiDP	NBL − TiDP
January	0.970 (0.13)	1.211 (0.139)	0.112 (0.028)	0.139 (0.029)
February	0.999 (0.136)	1.194 (0.140)	−0.044 (0.02)	*0.028 (0.024)* ‡
March	1.195 (0.136)	1.226 (0.136)	0.105 (0.022)	0.111 (0.023)
April	1.198 (0.136)	1.202 (0.134)	0.103 (0.025)	0.106 (0.026)
May	1.068 (0.138)	1.096 (0.136)	0.081 (0.026)	0.082 (0.026)
June	1.440 (0.145)	1.434 (0.146)	0.118 (0.026)	0.121(0.027)
July	1.162 (0.132)	1.168 (0.131)	0.116 (0.027)	0.119 (0.027)
August	1.273 (0.136)	1.276 (0.134)	0.117 (0.026)	0.117 (0.026)
September	1.159 (0.132)	1.211 (0.136)	0.094 (0.025)	0.093 (0.026)
October	1.107 (0.133)	1.109 (0.133)	0.070 (0.023)	0.070 (0.024)
November	1.178 (0.136)	1.171 (0.132)	0.087 (0.024)	0.088 (0.024)
December	1.045 (0.133)	1.036 (0.134)	0.093 (0.025)	0.094 (0.025)

‡ Italic font shows non-significant at the 95% confidence level.

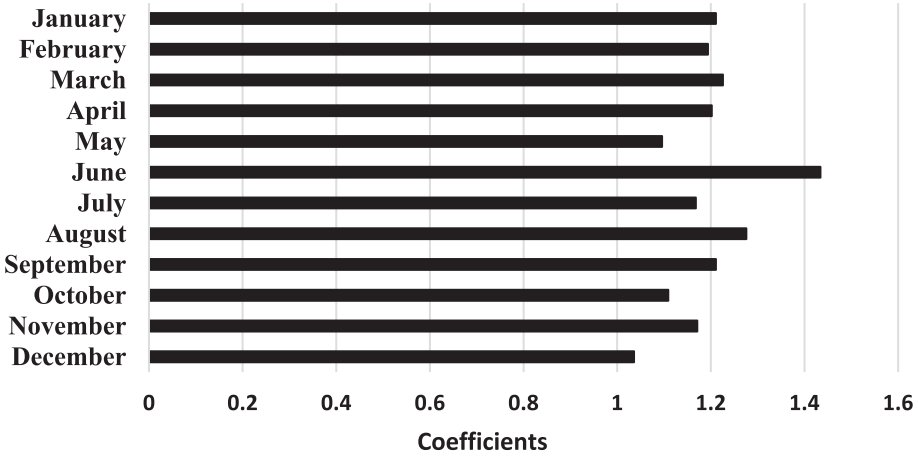

Fig. 6. Coefficients of the "monthly ADT" variable for different months of the year (Rural principal arterial model).

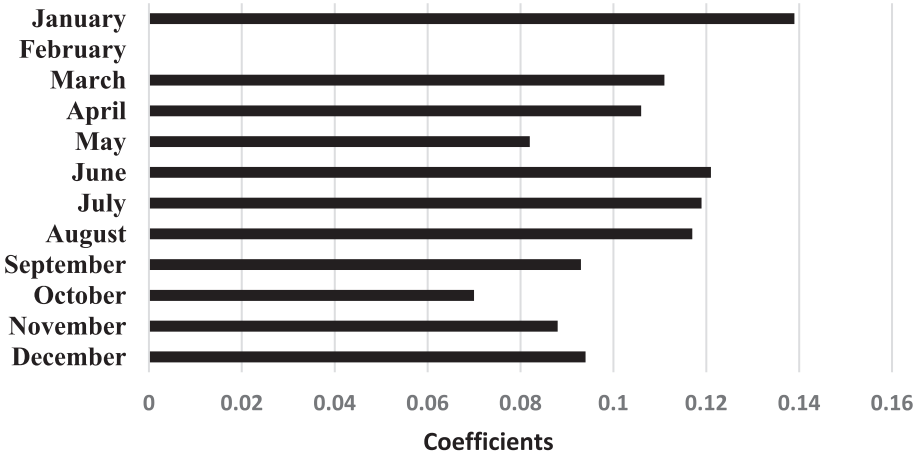

Fig. 7. Coefficients of the "standard deviation of speed" variable for different months of the year (Rural principal arterial model).

Table 7
Modeling results of rural minor arterial roadway data.

	NB		NB-TiDP		NB-L		NBL-TiDP	
	Mean	**S.D.**[†]	**Mean**	**S.D.**[†]	**Mean**	**S.D.**[†]	**Mean**	**S.D.**[†]
Mean of Parameters								
Intercept	−1.015	0.016	−1.001	0.017	−7.793	0.202	−8.107	0.206
Ln (Monthly ADT)	1.534	0.049	1.550	0.086	1.524	0.052	1.576	0.086
Standard deviation of speed	0.026	0.010	*0.041*[‡]	*0.048*	0.026	0.011	*0.055*[‡]	*0.046*
Precipitation (1 if monthly sum ≥ 1 in., 0 otherwise)	*0.005*[‡]	*0.027*	*−0.016*[‡]	*0.070*	*0.003*[‡]	*0.030*	*−0.006*[‡]	*0.074*
Median presence (1 if median is present, 0 otherwise)	−0.234	0.045	−0.234	0.046	−0.235	0.048	−0.243	0.048
Surface width (in feet)	0.008	0.001	0.008	0.001	0.008	0.001	0.007	0.001
Shoulder width (in feet)	−0.012	0.002	−0.012	0.002	−0.012	0.002	−0.012	0.002
Inverse dispersion parameter	1.357	0.063	1.391	0.066	45.54	14.99	48.83	15.53
Standard Deviation of Random Parameter								
Ln (Monthly ADT)			0.235	0.065			0.228	0.063
Standard deviation of speed			0.160	0.039			0.152	0.036
Precipitation (1 if monthly sum > 1 in., 0 otherwise)			0.198	0.052			0.211	0.058
Model Performance								
DIC	33,490		33,450		31,370		**31,290**[^]	
WAIC	33,490		33,450		32,100		**31,990**[^]	
LOOIC	33,490		33,450		32,858		**32,748**[^]	
MAD	1.487		1.499		**0.718**[^]		0.732	
MSPE	2.963		3.002		**1.007**[^]		1.033	

[†] Standard deviation
[‡] Italic font shows non-significant at 95% confidence level.
[^] Bold values indicate a better fit.

Table 8
Monthly estimates and standard deviations of time-dependent parameters for rural minor arterial roadways.

	Monthly ADT		Standard deviation of speed	
	NB- TiDP	NBL TiDP	NB-TiDP	NBL TiDP
January	1.454 (0.123)	1.682 (0.140)	*0.052 (0.035)* ‡	0.076 (0.038)
February	1.348 (0.133)	1.611 (0.145)	−0.111 (0.029)	*−0.002 (0.036)* ‡
March	1.379 (0.131)	1.385 (0.137)	*0.064 (0.033)* ‡	*0.062 (0.034)* ‡
April	1.527 (0.130)	1.541 (0.136)	0.079 (0.035)	0.078 (0.037)
May	1.526 (0.132)	1.535 (0.133)	0.080 (0.035)	0.082 (0.037)
June	1.702 (0.136)	1.726 (0.143)	*0.044 (0.037)* ‡	*0.042 (0.038)* ‡
July	1.688 (0.132)	1.714 (0.143)	*0.061 (0.037)* ‡	*0.054 (0.038)* ‡
August	1.535 (0.129)	1.520 (0.137)	*0.028 (0.036)* ‡	*0.029 (0.037)* ‡
September	1.462 (0.128)	1.442 (0.138)	0.102 (0.032)	0.102 (0.034)
October	1.711 (0.133)	1.611 (0.130)	*0.047 (0.035)* ‡	*0.053 (0.034)* ‡
November	1.622 (0.130)	1.535 (0.134)	*0.058 (0.033)* ‡	*0.067 (0.035)* ‡
December	1.662 (0.129)	1.617 (0.138)	*−0.003 (0.036)* ‡	*0.003 (0.037)* ‡

‡ Italic font shows non-significant at the 95% confidence level.

All coefficients for the time-independent variables were statistically significant at 95% credible intervals except the variable indicating the concrete pavement. However, this variable was kept in the final model since it improves the overall model fit. As shown in Table 5, the variables indicating median presence and shoulder width each had a negative relationship with the number of crashes. This suggests that an increase in shoulder width or the presence of an unprotected median decreases the number of crashes. These relationships align with previous findings by Islam et al. (2023) and Shaon et al. (2018). Furthermore, the surface width variable (excluding the shoulder width) is positively associated with the number of crashes, indicating that an increase in road surface width increases the frequency of the crash. This presumably could be due to increased instances of speeding on roads with wider lanes (see Marshall et al., 2023; Vergara et al., 2024). The modeling results show that the standard deviation of time-dependent parameters (i.e., the standard deviation of speed, precipitation indicator, and monthly ADT) is significant. This outcome highlights the importance of accounting for time (i.e., the month of the year) in estimating parameters for these variables.

Table 6 indicates the estimated parameters for each month of the year for each time-dependent variable in the NB-TiDP and NBL-TiDP models developed for the rural principal arterials. For the parameters that varied temporally, the mean of random parameters was significant for only the monthly ADT and the standard deviation of speed. However, the standard deviation of random parameters was significant for all the time-dependent variables. This indicates considerable variations across the months of the year which highlights the importance of accounting for time-dependent coefficients. Given the significance of the standard deviations of the parameters, these variables should be kept in the model. The modeling results show that the monthly ADT variable has a positive relationship with crash frequency, indicating an increase in the traffic volume will result in an increased number of crashes Examining the estimated parameters for each month for the standard deviation of speed, all the parameters except the one for February were significant and indicated a positive relationship with the crash count. Therefore, for months with a significant estimated mean, we can infer that the higher the variations in the vehicle speed, the higher the expected number of crashes. This finding aligns with previous studies by Quddus (2013), who found that the presence of slower and faster vehicles given a set speed limit might lead to significant speed variations and higher crash risk. However, Davis (2002) reported in his study that this relationship can only be expected where individual risk is either an increasing or decreasing or a U-shaped speed function. The results for monthly ADT and standard deviation of speed are also shown in Figs. 6-7. These figures only show the estimated parameters for months that were statistically significant. As shown in the figures, the estimated parameters could vary significantly for each month of the year for the time-dependent variables considered in the analysis. This could be due to unobserved heterogeneity, such as variations in other factors not considered in the model at different times of the year. Finally, the estimated mean for the precipitation variable was not significant at the 95% confidence level for all months of the year; hence, no inference could be made from them, and they were not shown in Table 6.

4.2.2. *Rural minor arterial roadways*

Table 7 shows the modeling results for the rural minor arterials. Again, the segment length was considered as an offset in these models. As discussed earlier, the DIC, WAIC, and LOOIC metrics were used to compare the models. Similar to the previous empirical dataset, the NBL with time-dependent parameters showed superior model fit compared to other estimated models. Based on results from the NBL-TiDP, variables representing median presence and shoulder width were found to be significant and exhibited a negative relationship with the number of crashes. This indicates that increasing shoulder width reduces crash frequency. Likewise, the median presence is associated with a reduction in crashes. The modeling results indicate that the surface width variable

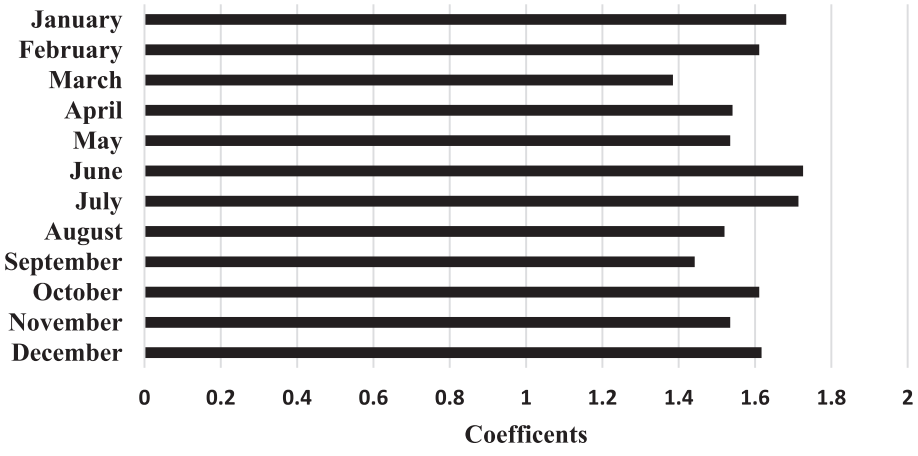

Fig. 8. Coefficients of the "monthly ADT" variable for different months of the year (rural minor arterial model).

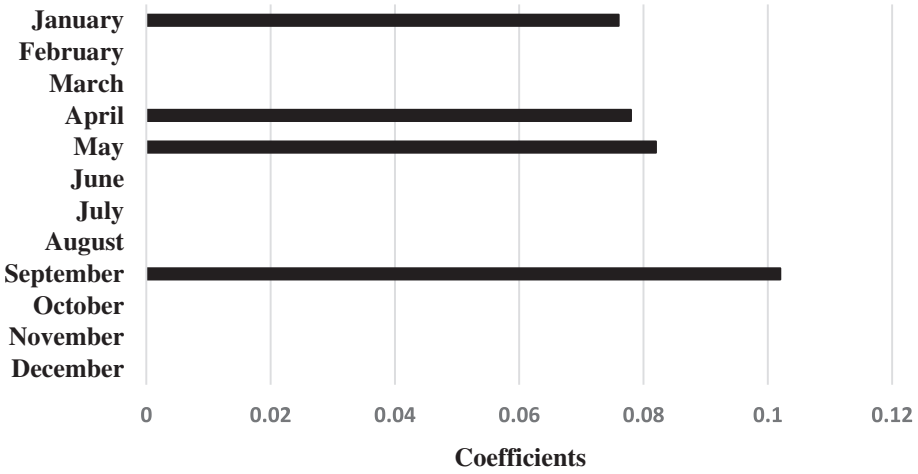

Fig. 9. Coefficients of the "standard deviation of speed" variable for different months of the year (Rural minor arterial model).

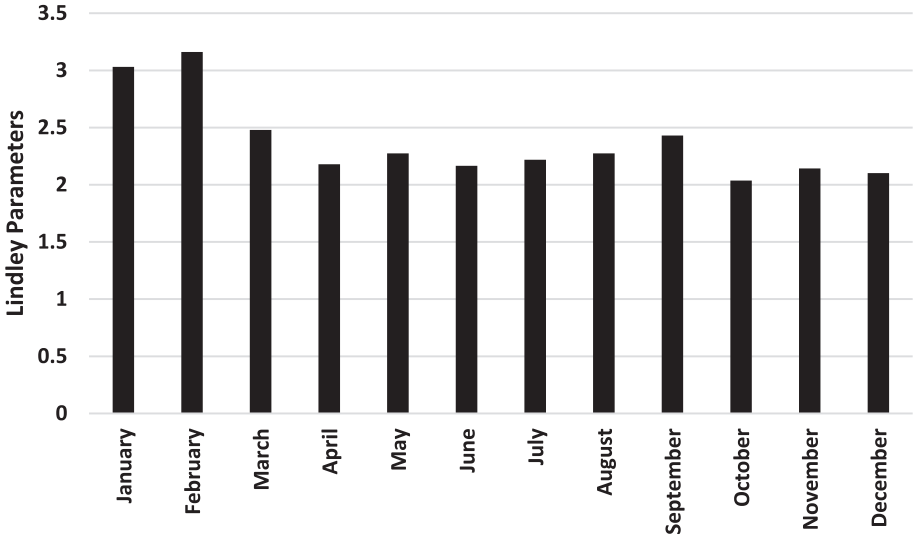

Fig. 10. Lindley parameters for different months of the year (rural principle arterial model).

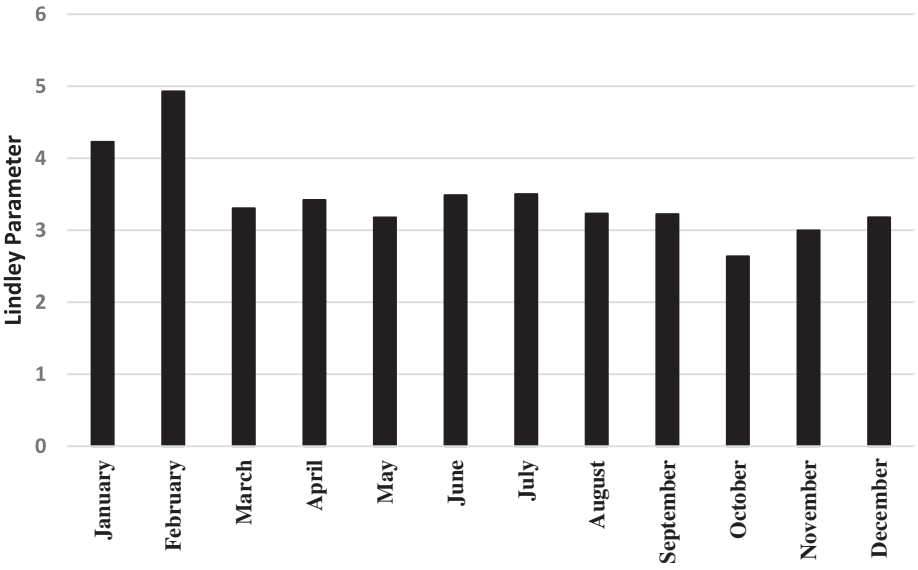

Fig. 11. Lindley parameters for different months of the year (rural minor arterial model).

has a positive effect on the number of crashes. An increase in this variable leads to a higher frequency of crashes, similar to what is observed on principal arterial roads.

Regarding the time-dependent variables, only the mean of the monthly ADT variable was significant. However, since the standard deviations of all the time-dependent variables were significant, all these variables should be retained in the final model. Table 8 shows the coefficients of the monthly ADT and the standard deviation of speed for all months of the year. The coefficients of the estimated parameters for the monthly ADT were significant for all months of the year, as shown in Table 8. The estimated monthly parameters for the standard deviation of the speed were significant for only January, April, May, and September. The estimated monthly parameters for Monthly ADT and standard deviation of speed are shown in Fig. 8 and Fig. 9, respectively. These figures only show the estimated parameters for months that were statistically significant. As shown in Figs. 8-9, the estimated parameters for monthly ADT and the standard deviation of speed vary in each month of the year; therefore, it is crucial to account for time-dependent parameters in the model. As explained earlier, this variation could be due to unobserved heterogeneity in data. For the variable indicating the level of precipitation, the mean of the random parameters was not significant at the 95% level; hence, the monthly parameters were not reported in Table 8.

Finally, it is worth highlighting that, unlike the NBL model which assumes a single Lindley parameter for all months of the year, the NBL-TiDP model considers different Lindley parameters for each month of the year. As discussed earlier, this is a better way to account for the issue of excess zero responses, as crash data in different months of the year may exhibit different percentages of zero observations. The variations of the Lindley parameters are shown in Figs. 10 and 11, for rural principal and rural minor arterials, respectively.

5. Summary and conclusions

Crash data are usually characterized by overdispersion, excess zero observations, and long tails. These characteristics are exacerbated when crash data are collected over shorter durations like days or months. One popularly used model to address excess zero observations is the NBL model. However, the NBL model does not account for the temporal variations associated with crash data collected over time. This study developed an extension of the NBL model to account for temporal variations related to variables such as traffic volume, standard deviation of speed, and weather information, which change from time to time. Using simulated crash data, the proposed NBL model with time-dependent variables estimated coefficients close to the actual values. Crash data with variables that varied over time from Texas were used to evaluate the model's performance. Using several GOF performance metrics, the study showed that the proposed model provides a superior model fit compared to all the other analyzed models. This observation revealed that using temporally varying coefficients in the model is important and could enhance the overall model fit. Results from crash data obtained from principal and minor arterial roads in Texas indicate that the median presence and an increase in shoulder width will likely reduce the number of crashes. The results also indicate that as the variation of speed increases, more crashes are observed. Finally, the monthly ADT has a positive relationship with crash frequency. The results from the time-dependent variables reveal that aggregating these variables over more extended periods rather than shorter durations may lead to loss of information and biased parameter estimates. Overall, this study sought to address the gap in addressing the issue of excess zero observations and accounting for temporal variations using time-dependent parameters in NBL models. However, this study is not without limitations. Future studies are still needed to more comprehensively assess the impact of temporal variations in crash frequency models, particularly those characterized by an excess of zero observations. This research was methodological, and mainly focused on the derivations and characteristics of the NB-TiDP model to account for excess zero observations and temporal variations. We demonstrated the method through two empirical examples using datasets collected in Texas by incorporating a few time-dependent variables that vary by month. However, without loss of generality, this model can be applied to data collected in seasonal, yearly, or other intervals. Further research is recommended to explore the application of the model using data from other states, to account for seasonal, yearly, or other temporal variations, and to include additional time-dependent variables, such as visibility.

CRediT authorship contribution statement

Richard Dzinyela: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Mohammadali Shirazi:** Writing – review & editing, Writing – original draft, Formal analysis, Conceptualization, Methodology, Supervision. **Subasish Das:** Writing – review & editing, Supervision. **Dominique Lord:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Adanu, E.K., Dzinyela, R., Agyemang, W., 2023. A comprehensive study of child pedestrian crash outcomes in Ghana. Accid. Anal. Prev. 189, 107146.

Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accid. Anal. Prev. 41 (1), 153–159.

Ash, J.E., Zou, Y., Lord, D., Wang, Y., 2021. Comparison of confidence and prediction intervals for different mixed-Poisson regression models. Journal of Transportation Safety & Security 13 (3), 357–379.

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.

Cai, Q., Abdel-Aty, M., Lee, J., Wang, L., Wang, X., 2018. Developing a grouped random parameters multivariate spatial model to explore zonal effects for segment and intersection crash modeling. Analytic Methods in Accident Research 19, 1–15.

Center for Disease Control and Prevention, (2023). Available at : https://www.cdc.gov/injury/features/all-features-by-category.html.

Cheng, W., Gill, G.S., Dasu, R., Xie, M., Jia, X., Zhou, J., 2017. Comparison of Multivariate Poisson lognormal spatial and temporal crash models to identify hot spots of intersections based on crash types. Accident Anal. Prev. 99, 330–341.

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

Davis, G.A., 2002. Is the claim that 'variance kills' an ecological fallacy? Accid. Anal. Prev. 34 (3), 343–346.

Denthet, S., Thongteeraparp, A., Bodhisuwan, W., 2016, October. 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.

Dzinyela, R., Adanu, E.K., Lord, D., Islam, S., 2023. Analysis of factors that influence injury severity of single and multi-vehicle crashes involving at-fault older drivers: A random parameters logit with heterogeneity in means and variances approach. Transp. Res. Interdisciplinary Perspectives 22, 100974.

Dzinyela, R., Alnawmasi, N., Adanu, E.K., Dadashova, B., Lord, D., Mannering, F., 2024. A multi-year statistical analysis of driver injury severities in single-vehicle freeway crashes with and without airbags deployed. Analytic Methods in Accident Research 41, 100317.

Fountas, G., Fonzone, A., Gharavi, N., Rye, T., 2020. The joint effect of weather and lighting conditions on injury severities of single-vehicle accidents. Analytic Methods in Accident Research 27, 100124.

Fu, X., Liu, J., Jones, S., Barnett, T., Khattak, A.J., 2022. From the past to the future: Modeling the temporal instability of safety performance functions. Accid. Anal. Prev. 167, 106592.

Geedipally, S.R., Lord, D., Park, B.J., 2009. Analyzing different parameterizations of the varying dispersion parameter as a function of segment length. Transp. Res. Rec. 2103 (1), 108–118.

Geedipally, S.R., Lord, D., Dhavala, S.S., 2012. The negative binomial-Lindley generalized linear model: Characteristics and application using crash data. Accid. Anal. Prev. 45, 258–265.
Geedipally, S.R., Lord, D., Dhavala, S.S., 2014. A caution about using deviance information criterion while modeling traffic crashes. Saf. Sci. 62, 495–498.
Gil-Marin, J.K., Shirazi, M., Ivan, J.N., 2024. Assessing the Negative Binomial-Lindley model for crash hotspot identification: Insights from Monte Carlo simulation analysis. Accid. Anal. Prev. 199, 107478.
Hasan, T., Abdel-Aty, M., 2024. Short-term safety performance functions by random parameters negative binomial-Lindley model for part-time shoulder use. Accid. Anal. Prev. 199, 107498.
Hilbe, J.M., 2011. Negative binomial regression. Cambridge University Press.
Hou, Q., Huo, X., Tarko, A.P., Leng, J., 2021. Comparative analysis of alternative random parameters count data models in highway safety. Analytic Methods in Accident Research 30, 100158.
INRIX. n.d. INRIX XD. Available at: https://inrix.com/products/speed/ [Accessed 01/25/2024].
Islam, A.M., Shirazi, M., Lord, D., 2022. Finite mixture Negative Binomial-Lindley for modeling heterogeneous crash data with many zero observations. Accid. Anal. Prev. 175, 106765.
Islam, A.M., Shirazi, M., Lord, D., 2023. Grouped Random Parameters Negative Binomial-Lindley for accounting unobserved heterogeneity in crash data with preponderant zero observations. Analytic Methods in Accident Research 37, 100255.
Khodadadi, A., Tsapakis, I., Shirazi, M., Das, S., Lord, D., 2022. Derivation of the Empirical Bayesian method for the Negative Binomial-Lindley generalized linear model with application in traffic safety. Accid. Anal. Prev. 170, 106638.
Khodadadi, A., Shirazi, M., Geedipally, S., Lord, D., 2023. Evaluating alternative variations of Negative Binomial-Lindley distribution for modelling crash data. Transportmetrica a: Transport Science 19 (3), 2062480.
Liu, C., Sharma, A., 2017. Exploring spatio-temporal effects in traffic crash trend analysis. Analytic Methods in Accident Research 16, 104–116.
Lord, D., 2006. Modeling motor vehicle crashes using Poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Accident Anal. Prev. 38 (4), 751–766.
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. Accid. Anal. Prev. 43 (5), 1738–1742.
Lord, D., Geedipally, S.R., 2018. Safety prediction with datasets characterised with excess zero responses and long tails. In: Safe Mobility: Challenges, Methodology and Solutions, Vol. 11. Emerald Publishing Limited, pp. 297–323.
Lord, D., Mannering, F., 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. Transp. Res. A Policy Pract. 44 (5), 291–305.
Lord, D., Persaud, B.N., 2000. Accident prediction models with and without trend: application of the generalized estimating equations procedure. Transport. Res. Rec. 1717 (1), 102–108.
Lord, D., Qin, X., Geedipally, S.R., 2021. Highway safety analytics and modeling. Elsevier.
Mannering, F., 2018. Temporal instability and the analysis of highway accident data. Analytic Methods in Accident Research 17, 1–13.
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.
Marshall, E., Shirazi, M., Shahlaee, A., Ivan, J.N., 2023. Leveraging probe data to model speeding on urban limited access highway segments: Examining the impact of operational performance, roadway characteristics, and COVID-19 pandemic. Accid. Anal. Prev. 187, 107038.
Poch, M., Mannering, F., 1996. Negative binomial analysis of intersection-accident frequencies. J. Transp. Eng. 122 (2), 105–113.
Quddus, M., 2013. Exploring the relationship between average speed, speed variation, and accident rates using spatial statistical models and GIS. Journal of Transportation Safety & Security 5 (1), 27–45.
Rim, H., Abdel-Aty, M., Mahmoud, N., 2023. Multi-vehicle safety functions for freeway weaving segments using lane-level traffic data. Accid. Anal. Prev. 188, 107113.
Rusli, R., Haque, M.M., King, M., Voon, W.S., 2017. Single-vehicle crashes along rural mountainous highways in Malaysia: An application of random parameters negative binomial model. Accid. Anal. Prev. 102, 153–164.
Rusli, R., Haque, M.M., Afghari, A.P., King, M., 2018. Applying a random parameters Negative Binomial Lindley model to examine multi-vehicle crashes along rural mountainous highways in Malaysia. Accid. Anal. Prev. 119, 80–90.
Sawtelle, A., Shirazi, M., Garder, P.E., Rubin, J., 2023. Exploring the impact of seasonal weather factors on frequency of lane-departure crashes in Maine. Journal of Transportation Safety & Security 15 (5), 445–466.
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.
Shirazi, M., Lord, D., 2019. Characteristics-based heuristics to select a logical distribution between the Poisson-gamma and the Poisson-lognormal for crash data modelling. Transportmetrica a: Transport Science 15 (2), 1791–1803.
Shirazi, M., Lord, D., Geedipally, S.R., 2016a. Sample-size guidelines for recalibrating crash prediction models: Recommendations for the highway safety manual. Accid. Anal. Prev. 93, 160–168.
Shirazi, M., Lord, D., Dhavala, S.S., Geedipally, S.R., 2016b. A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: Characteristics and applications to crash data. Accid. Anal. Prev. 91, 10–18.
Shirazi, M., Dhavala, S.S., Lord, D., Geedipally, S.R., 2017a. A methodology to design heuristics for model selection based on the characteristics of data: Application to investigate when the Negative Binomial Lindley (NB-L) is preferred over the Negative Binomial (NB). Accid. Anal. Prev. 107, 186–194.
Shirazi, M., Geedipally, S.R., Lord, D., 2017b. A Monte-Carlo simulation analysis for evaluating the severity distribution functions (SDFs) calibration methodology and determining the minimum sample-size requirements. Accid. Anal. Prev. 98, 303–311.
Shirazi, M., Geedipally, S.R., Lord, D., 2021. A simulation analysis to study the temporal and spatial aggregations of safety datasets with excess zero observations. Transportmetrica a: Transport Science 17 (4), 1305–1317.
Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). WinBUGS user manual.
Tajuddin, R.R.M., Ismail, N., Ibrahim, K., 2022. Several two-component mixture distributions for count data. Communications in Statistics-Simulation and Computation 51 (7), 3760–3771.
Texas Department of Transportation, (2023). Available at: https://ftp.txdot.gov/pub/txdot-info/trf/crash_statistics/2022/01.pdf.
Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. 27, 1413–1432.
Venkataraman, N.S., Ulfarsson, G.F., Shankar, V., Oh, J., Park, M., 2011. Model of relationship between interstate crash occurrence and geometrics: exploratory insights from random parameter negative binomial approach. Transp. Res. Rec. 2236 (1), 41–48.
Vergara, E., Aviles-Ordonez, J., Xie, Y., Shirazi, M., 2024. Understanding speeding behavior on interstate horizontal curves and ramps using networkwide probe data. J. Saf. Res. in press.
Wang, X., Abdel-Aty, M., 2006. Temporal and spatial analyses of rear-end crashes at signalized intersections. Accid. Anal. Prev. 38 (6), 1137–1150.
Wang, Y., Kockelman, K.M., 2013. A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighborhoods. Accident Anal. Prev. 60, 71–84.
Washington, S., Karlaftis, M.G., Mannering, F., Anastasopoulos, P., 2020. Statistical and econometric methods for transportation data analysis. CRC Press.
Ye, F., Garcia, T.P., Pourahmadi, M., Lord, D., 2012. Extension of negative binomial GARCH model: analyzing effects of gasoline price and miles traveled on fatal crashes involving intoxicated drivers in Texas. Transp. Res. Rec. 2279 (1), 31–39.
Yuan, J., Abdel-Aty, M., Fu, J., Wu, Y., Yue, L., Eluru, N., 2021. Developing safety performance functions for freeways at different aggregation levels using multi-state microscopic traffic detector data. Accident Analysis & Prevention 151, 105984.
Zamani, H., Ismail, N., 2010. Negative binomial-Lindley distribution and its application. J. Math. Stat. 6 (1), 4–9.
Zhang, P., Wang, C., Chen, F., Cui, S., Cheng, J., Bo, W., 2022. A random-parameter negative binomial model for assessing freeway crash frequency by injury severity: Daytime versus nighttime. Sustainability 14 (15), 9061.