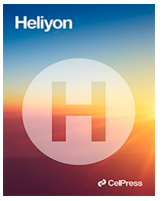
Heliyon 10 (2024) e28900

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

A nested grouped random parameter negative binomial model for modeling segment-level crash counts

Omar Almutairi

Civil Engineering Department, College of Engineering, Al Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

ARTICLE INFO

Keywords:
Random parameters
Negative binomial model
Longitudinal data
Correlations along county routes
Temporal correlation
Total crashes

ABSTRACT

In this study, a nested grouped random parameter negative binomial framework is proposed to model crash counts at the segment level, a three-level longitudinal framework. The proposed model accounts for correlations along county routes and over time and thus includes a time variable, the year index, to analyze crash counts. The model is applied to crashes on undivided two-lane arterial roads in Ohio from 2012 to 2017. The results present two variants of the model: one with varying intercepts and fixed slopes and the other with varying intercepts and slopes. Both variants have comparable interpretations concerning the fixed parameters, but the latter variant exhibits a significantly improved fit and provides additional information on the interpretations. The results show a significant quadratic relationship between the time variable and the crash count, indicating that, on average, the crash count of segments increases with a decreasing rate as time variable increases. Regarding random parameters, the findings show that 17% of segments within routes and 2% of routes exhibit crash counts that decrease at accelerating downward trend as time variable increases. The effect of the natural logarithm of the segment length varies significantly across different routes, with an increase in this value primarily leading to an increase in crashes. On the other hand, the effect of the total shoulder width also varies across routes, but unlike the former, an increase in this value generally results in a decrease in crashes. The proposed model shows high forecast accuracy for crash count prediction, making it a valuable tool for informed decision-making in safety improvement.

1. Introduction

One challenge in analyzing road crash data is to account for the heterogeneity of observations, which means that the number or severity of crashes may vary depending on various factors that are not observed or measured. This heterogeneity can be derived from multiple sources, such as endogeneity, risk compensation, and spatial or temporal correlation [1]. Endogeneity refers to the correlation of an observed variable to other important observed or unobserved variables that have a large impact on the dependent variable and can affect parameter estimates and causal inferences. For example, if the relationship between the increase in traffic tickets and the risk of crashes in a particular place is not controlled by the duration of deployment, it may lead to erroneous conclusions about the effectiveness of speed enforcement [2,3]. Risk compensation refers to the driver's behavior due to changes in perceived risk and safety, which can reduce or offset the expected benefits of safety interventions or policies. For example, if the driver wears the seatbelt but drives faster or aggressively, the net effect of the use of the seatbelt on the severity of the crash may be different from expected [4]. Spatial or temporal correlation refers to the dependence or similarity of the results of collisions over space and time, which is contrary

E-mail address: OAlhjlih@imamu.edu.sa.

https://doi.org/10.1016/j.heliyon.2024.e28900
Received 27 November 2023; Received in revised form 22 March 2024; Accepted 26 March 2024
Available online 28 March 2024

Table 1
Summary of recent related studies that applied random parameter models in different applications.

Model Type	Summary	Year Article
Random parameters vary across observations		
Multinomial logit model	Studied factors that influenced the severity of injuries from crashes at highway-rail crossings, using a random parameter model with heterogeneity in mean and variance.	2023 [10]
Poisson/negative binomial regression Models	Developed and evaluated a data-driven crash feature-based approach for boundary crash allocation with random parameter models.	2023 [11]
Ordered logit model	Studied factors that affect non-motorists' safety perception of autonomous vehicles (AV) after hearing about a fatal AV crash using a random parameters model.	2023 [12]
Logistic regression models	Developed and evaluated spatiotemporal logistic regression models with random parameters to analyze the severity of motor vehicle–pedestrian crashes at urban intersections.	2023 [13]
Negative binomial Lindley model	Analyzed the impact of geometric design consistency on run-off-road crashes using a random parameter model.	2023 [14]
Multinomial logit model	Studied various factors that influenced the rear-end and non-rear-end crash severity using random parameter models with heterogeneity in mean and variance.	2022 [15]
Bivariate tobit model	Studied the effect of speed limit increase on both non-injury and injury rates using a correlated random parameter model.	2022 [16]
Binary logit model	Applied a random parameter model to account for the heterogeneity and unobserved factors that affected the severity of vulnerable road users in crashes with motor vehicles across different seasons.	2022 [17]
Multinomial logit model	Compared and analyzed contributing factors to the severity of large truck crashes in-state and out-of-state using random parameters with heterogeneity in means and variances.	2022 [18]
Random parameters vary across groups		
Linear regression model	Studied the factors that influenced a driver's intention to avoid traffic violations, such as speeding, illegal overtaking, and running red lights, using a grouped random parameter model, with parameters that vary across individuals.	2022 [19]
Poisson regression model	Investigated the contributing factors to right-turn crashes at intersections using a grouped random parameter model, with parameters that varied across intersections.	2022 [20]
Binary logit model and exponential regression model	Investigated the effect of sleepiness in truck drivers on their time headway using a grouped random parameter model, with the parameters varying across individuals.	2022 [21]
Multinomial logit model	Analyzed the effect of traffic, geometric, and context variables on urban crash types using a grouped random parameter model, with the parameters varied across segments or intersections.	2020 [22]
Linear regression model	Analyzed the free-flow speed data before and after the reduction of the speed limit using a grouped random parameter model (3-level model), where parameters varied across sites within communities and varied across communities.	2015 [23]

to the hypothesis of independent observations in standard models. For example, if crashes are spatially distributed at specific locations or temporally related at a specific time, disregarding this structure can lead to inaccurate and inconsistent parameter estimates and predictions. Therefore, the random parameter model can effectively deal with these heterogeneous sources in road crash data [5].

1.1. Random parameter models

Several studies have used random parameter models to account for heterogeneity across observations by introducing random terms in the estimates of explanatory variables. The main difference between random parameter and fixed parameter models is that the effect of an explanatory variable on the outcome variable varies across observations in the former but is fixed in the latter. Considering a parameter that varies across observations could catch the heterogeneous effect of an explanatory variable, leading to an improvement of the model performance. For instance, one study compared the performance of random parameter and fixed parameter negative binomial models in modeling road crash frequencies on four rural interstate highways in Indiana [6]. The data covered five years and included pavement, geometric, and traffic characteristics as explanatory variables. The results showed that the random parameter model fitted the data better and had more accurate predictions than the fixed parameter model. One notable difference between the models was that the average annual daily traffic (AADT) variable had a negative coefficient in the fixed parameter model but a positive coefficient in the random parameter model. The authors explained that this could reflect the complex interaction between traffic volume, driver behavior, and crash frequency, which the random parameter model could better capture. The study also concluded that ignoring the heterogeneity of the effects of explanatory variables could lead to significant biases in both marginal effects and the magnitude of the effects of contributing variables on crash counts. Another study evaluated the performance of three types of negative binomial models in modeling road injury and non-injury crash counts on 826 multilane highway segments over a period of three years [7]. The models were uncorrelated random parameter, correlated random parameter, and fixed-parameter. The study found that the uncorrelated and correlated random parameter models were comparable, and both outperformed the fixed-parameter model. The significant variables for both injury and non-injury models included exposure-related measures (AADT, segment length), geometric features (lane width, left shoulder width, median width, tangent segments), road facility type (undivided and divided with different numbers of lanes), and pavement surface conditions. Only two variables, namely left shoulder width and tangent segment indicator, were found to vary significantly and randomly across observations in both models. A study investigated how ramp type, alignment, traffic volume, and interchange geometry affect the frequency of ramp crashes [8]. The study used a random parameter negative binomial model that captured two sources of heterogeneity: one in the mean and the other in the variance of crash counts. The study

identified some variables that had significant and varying effects on ramp crash frequency, such as ramp length, curvature, and truck volume. Another recent study developed an improved mixture model framework called joint negative binomial-multinomial logit fractional split to model crash counts by crash type for traffic analysis zones. The model accommodated zero crashes and incorporated random parameters to account for heterogeneity across traffic analysis zones [9]. Clearly, the random-parameter model is widely used in different applications. Table 1 shows some recent studies that applied random parameter models to different types of data. In the next section, the grouped random parameter model is introduced, which allows the parameters to vary across groups of observations.

1.2. Grouped random parameter models

A grouped random parameter model is similar to a random parameter model, except that it allows parameters to vary across groups of observations instead of individual observations. The grouping of observations can be based on criteria such as counties, cities, traffic analysis zones, and corridors [24]. This type of model is also known as a multilevel or nested model because it implies that observations are nested within groups. For example, segments can be nested within counties, cities, traffic analysis zones, corridors, etc. Such a grouping of observations assumes that there is a correlation within groups or unobserved heterogeneity across groups [24,25]. A study compared five negative binomial models to analyze intersection crashes using five years of data [26]. The models were standard, grouped random parameters, and three variations of grouped random parameters with different weight factors. The models accounted for the correlation of intersections within traffic analysis zones and the effects of boundary intersections between adjacent zones. The results showed that the grouped random parameter models performed better than the standard model. The study concluded that group-level variables can reduce unobserved heterogeneity across traffic analysis zones. Another study proposed two approaches that accounted for the effect of independent variables at the segment level or the intersection level in the estimation of crash frequency at the traffic analysis zone level [27]. The study used a grouped random parameter negative binomial model to model crash frequency at the segment level and intersection level. The study accounted for correlations within the traffic analysis zone by assuming that the segments or intersections located in the same traffic analysis zone were correlated. Among the models tested, the study found that the model improved performance by accommodating correlations within the traffic analysis zone, which was measured by the correlation between segments or intersections in the same traffic analysis zone. The study concluded that there were unobserved factors between segments or intersections in the same traffic analysis zone. A more recent study proposed a grouped random parameter negative binomial-Lindley model (G-RPNB-L) to account for unobserved heterogeneity in crash counts with a high percentage of zero occurrences [28]. The study applied the proposed model to lane departure crashes from rural interstate segments and compared it with standard negative binomial (NB), negative binomial-Lindley (NB-L), and grouped random parameter negative binomial (G-RPNB). The grouping variable was counties, and the group-level variables differed considerably across counties. The study found that the proposed model outperformed all its counterparts. In addition, the results showed that G-RPNB performed better than NB, and G-RPNB-L performed better than NB-L. The results also showed that NB-L performed better than G-RPNB. The authors explained that this was because 90% of the observations had zero values. The authors concluded that G-RPNB could better account for unobserved heterogeneity. Table 1 shows recent studies that have applied grouped random parameter models to different types of data. However, most of the studies mentioned above accounted for correlations within geographical units. The following section presents illustrative studies that have employed grouped random parameter models as a basis for longitudinal analysis, thereby accounting for a range of correlations.

1.3. Longitudinal data analysis

Longitudinal data refer to multiple observations taken on the same identity, which vary in time or location. For example, an annual crash count on the same segment or a crash count on multiple segments located along the same corridor. These multiple observations are usually assumed to be correlated. A grouped random parameter model can handle these correlations by grouping the observations into one group. Within this group, the observations are assumed to be correlated. A study innovatively utilized a grouped random parameters model to analyze multiple crash frequency variables by crash type for traffic analysis zones [29]. This study treated six crash types as repeated observations for the same traffic analysis zone. In this context, the ‘groups’ refer to traffic analysis zones. Consequently, unobserved heterogeneity is accounted for across these zones. In a subsequent study, the authors developed latent class segmentation for their proposed approach, aiming to account for heterogeneity across estimates of exogenous variables [30]. Another study used a grouped random parameter negative binomial model to analyze monthly crash counts at the city level [31]. Multiple observations were collected for each city, and these observations were grouped together. In other words, the monthly crash counts for one city were treated as one group. Several variables were included in the model to describe the groups or cities, group-level variables, and dummy variables indicating the year, season, and region in which the city was located, and other variables that change annually or monthly. The study compared three models: a random intercept model, an uncorrelated random intercept and slope model, and a correlated intercept and slope model. The study concluded that the correlated intercept and slope model outperformed all its counterparts. In another study, a grouped random parameter negative binomial model was used to analyze 9 years of crash counts on interstate highways in Washington state [32]. The study allowed the parameters to vary across directional segments and assumed a correlation between two consecutive years by employing a one-lag autocorrelated structure. The study concluded that the grouped random parameter negative binomial model performed better than the standard negative binomial model. These findings show that grouped random parameter models can be used to account for a range of correlations among observations.

Random parameter models are versatile tools for dealing with heterogeneity across groups of observations or homogeneity among groups of observations. These models have proven to be highly effective in capturing and analyzing complex data, as shown in the

studies discussed above and in Table 1. Section 1.2 presents compelling studies that used grouped random parameter models to capture spatial heterogeneity while investigating the effects of heterogeneity across traffic analysis zones, intersections, and counties. Section 1.3 presents innovative studies that used grouped random parameter models for longitudinal data to account for correlations. This study proposes a three-level longitudinal model, known as the nested grouped random parameter model, to analyze crash count rates over time while accounting for correlations among segments along county routes. Crash counts over the years are considered repeated measurements at the lowest level, nested within segments at the second level, and these segments are nested within county routes at the third level. Three types of variables exist in this model: time variables, time-varying variables, and time-invariant variables. The time variable is simply an index of the year for that observation. Time-varying variables, such as AADT, are predictors that change yearly. Time-invariant variables, such as segment length, are predictors that remain fixed over the years. This advanced approach enables the concurrent modeling of both intraindividual change (how crash counts rate for a segment change over time) and inter-individual change (the heterogeneity of this temporal change across segments) accounting for correlation along county routes. This article contributes to the field by providing a novel perspective on the analysis of crash count rates at the segment level. The proposed model could serve as a valuable tool for decision makers, aiding them in their efforts to improve road safety based on rigorous statistical analysis.

2. Methodology

This study proposes a three-level longitudinal model, known as the nested grouped random parameter model, to account for correlations along county routes and over years when modeling crash counts at the segment level. The proposed model assumes that the annual crash counts are nested within the segments and that these segments are nested within the routes. This nested structure implies that the multiple annual crash counts for the same segment are correlated and that the crash counts for segments located along the same route are also correlated. The following sections present the model framework and the statistical testing of the competing models.

2.1. Three-level longitudinal framework model

A three-level longitudinal modeling was applied to model the crash count per year at the segment level. Explanatory variables can be classified into three types: time variables, time-varying variables, and time-invariant variables. A time variable is a time index indicating the year. A time-varying variable changes over time. A time-invariant variable remains constant over time [33]. Thus, the regression models for each level are expressed as follows:

Level 1

$$y_{ijk} = e^{\left(\beta_{0jk}+\beta_{1jk}x_{1ijk}+\beta_{2jk}x_{2ijk}+\ldots+\beta_{pjk}x_{pijk}+\varepsilon_{ijk}\right)} \tag{1}$$

Level 2

$$\begin{aligned}
\beta_{0jk} &= \gamma_{00k} + \gamma_{01k}Z_{1jk} + \gamma_{02k}Z_{2jk} + \ldots + \gamma_{0qk}Z_{qjk} + U_{0jk}\\
\beta_{1jk} &= \gamma_{10k} + \gamma_{11k}Z_{1jk} + \gamma_{12k}Z_{2jk} + \ldots + \gamma_{1qk}Z_{qjk} + U_{1jk}\\
\beta_{2jk} &= \gamma_{20k} + \gamma_{21k}Z_{1jk} + \gamma_{22k}Z_{2jk} + \ldots + \gamma_{2qk}Z_{qjk} + U_{2jk}\\
&\vdots\\
\beta_{pjk} &= \gamma_{p0k} + \gamma_{p1k}Z_{1jk} + \gamma_{p2k}Z_{2jk} + \ldots + \gamma_{pqk}Z_{qjk} + U_{pjk}
\end{aligned} \tag{2}$$

Level 3

$$\begin{aligned}
\gamma_{00k} &= \delta_{000} + \delta_{001}R_{1k} + \delta_{002}R_{2k} + \ldots + \delta_{00l}R_{lk} + V_{00k}\\
&\vdots\\
\gamma_{0qk} &= \delta_{0q0} + \delta_{0q1}R_{1k} + \delta_{0q2}R_{2k} + \ldots + \delta_{0ql}R_{lk} + V_{0qk}\\
&\vdots\\
\gamma_{p0k} &= \delta_{p00} + \delta_{p01}R_{1k} + \delta_{p02}R_{2k} + \ldots + \delta_{p0l}R_{lk} + V_{p0k}\\
&\vdots\\
\gamma_{pqk} &= \delta_{pq0} + \delta_{pq1}R_{1k} + \delta_{pq2}R_{2k} + \ldots + \delta_{pql}R_{lk} + V_{pqk}
\end{aligned} \tag{3}$$

The dependent variable, y_{ijk}, is the crash count for year i ($i = 1, 2, \ldots, I$) at segment j ($j = 1, 2, \ldots, J$) and route k ($k = 1, 2, \ldots, K$). The explanatory variables that are indicated by X and indicated by the subscript p ($p = 1, 2, \ldots, P$) are the variables at the first level. These variables contained information that described specific observations. The explanatory variables denoted by Z and indicated by the subscript q ($q = 1, 2, \ldots, Q$) are the variables at the second level. These variables contained information that described specific segments. These variables can be called time-invariant because they are constant over the years. The explanatory variables denoted by R and indicated by subscript l ($l = 1, 2, \ldots, L$) are the variables at the third level. These variables contain information that describes the specific route or are fixed-parameter variables. However, the dataset used in this study did not have a route describing variables. The

last term in each equation and in each level indicates the random parameter. The others, β, γ, and δ, are the regression parameters.

Equations from all levels can be consolidated into a single equation. This is achieved by first substituting the equations from Level 3, shown in (3), into those from Level 2, shown in (2). Subsequently, the resulting equations from Level 2 are substituted into the equation from Level 1, shown in (1). This process yields the following consolidated equation:

$$y_{ijk} = e^{\left(\delta_{p00}X_{pijk}+\delta_{pql}R_{lk}Z_{qij}X_{pijk}+U_{pjk}X_{pijk}+V_{pqk}Z_{qjk}X_{pijk}+\varepsilon_{ijk}\right)} \tag{4}$$

where $X_{0ijk} = 1$, ($p = 0, 1, 2, \ldots, P$), $Z_{0ij} = 1$, and ($q = 0, 1, 2, \ldots, Q$), and others are as defined previously.

Since the crash count is a non-negative value, and most often its variance is greater than its mean, a negative binomial model is typically applied [7]. Thus, the last term, $e^{(\varepsilon_{ijk})}$, in equation (1) is the random part in the first level and gamma-distributed with a mean of one and variance of $1/k$, the dispersion parameter. The last terms in Level 2 are random parameters and have a multivariate normal distribution with a mean of zero. Similarly, the last terms in Level 3 are random parameters and have a multivariate normal distribution with a mean of zero. As shown in Equations (1) and (4), there is a link function between the mean of the crash count and the predicted crash count to ensure that the fitted values are always non-negative. This is achieved using a log link.

The negative binomial density function is as follows:

$$f\left(y_{ijk};k,\mu_{ijk}\right) = \frac{\Gamma\left(y_{ijk}+k\right)}{\Gamma(\mathrm{k})\times y_{ijk}!}\times\left(\frac{k}{\mu_{ijk}+k}\right)^{k}\times\left(\frac{\mu_{ijk}}{\mu_{ijk}+k}\right)^{y_{ijk}} \tag{5}$$

where $\Gamma(.)$ is a gamma function. The relationship between the variance (σ^2), and the mean (μ) is described by the equation $\sigma^2 = \mu + \frac{\mu^2}{k}$, indicating that the variance increases in a quadratic manner with the mean. If the dispersion parameter k tends toward infinity, the mean and variance become equal. In this case, the Poisson model becomes a more appropriate choice than negative binomial model, as shown in equation (5), [25]. All model estimations were performed using the glmmTMB R package, which uses maximum likelihood estimation and Laplace approximation to integrate over random parameters [34].

2.2. Fitting performance of competing models

This study used the likelihood ratio test in conjunction with Akaike's information criterion (AIC) and Schwarz's Bayesian information criterion (BIC) to assess the significance of the contribution of a variable to the dependent variable. The test statistic for the likelihood ratio was calculated from the difference in deviances between the two competing models, equation (6). Here, deviance is defined as negative two times the log of the likelihood, with the likelihood being the value of the likelihood function at convergence [7]. The difference in deviances follows a chi-square distribution, with degrees of freedom equal to the difference in the number of estimated parameters between the two models [33]. The test statistic is expressed as follows:

$$\chi^2 = deviance_{Model\ 1} - deviance_{Model\ 2} \tag{6}$$

The AIC test utilizes deviance but imposes a penalty for additional estimated parameters. This is expressed as follows:

$$AIC = deviance + 2\times q \tag{7}$$

Here, equation (7), q represents the number of estimated parameters. The BIC imposes a stricter penalty on complex models and is expressed as follows:

$$BIC = deviance + q\times ln\ (N) \tag{8}$$

In equation (8), N represents the total number of observations, with other variables defined previously. Lower AIC and BIC values indicate a better fit for the model.

This study used a bottom-up approach to test and construct the model using sequential likelihood ratio tests. Whenever a variable was introduced into the model, its significance was assessed. If found to be significant, it was retained in the model; otherwise, it was excluded. The initial step involved comparing a model with only a fixed intercept to a model with varying intercepts. Following this, each variable was tested as a fixed parameter until all variables were tested. Subsequently, the variables were tested in a similar manner to determine whether they varied significantly. The results section presents the model with varying intercepts and variables as fixed parameters, as well as models with varying intercepts and significantly varying slopes. These models were further evaluated using root mean squared errors (RMSE), equation (9), and plotting predicted values against observed values [35]. The RMSE is calculated as follows [7]:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(Observed\ Values_i - Predicted\ Values_i\right)^2}{N}} \tag{9}$$

These tools serve as effective measures for evaluating the accuracy of model forecasting.

Table 2
Summary statistics of variables used to describe road segments.

Variables	Mean	Range	SD
Total crashes	2.4388	0–73	4.3030
Time variable	2.1971	0–5	1.6245
Time variable^2	7.4660	0–25	8.1247
Natural log of annual average daily traffic (AADT)	8.5757	4.6051–10.2234	0.6093
Natural log of segment length (miles)	−1.4924	−4.6052–2.6838	1.3949
Total shoulder width (feet)	8.5440	0–60	5.7278
Roadway width (feet) (without shoulder width)	23.8962	18–28	2.4215
Area indicator (1 if the segment located in rural area, zero urban area)	0.4831	0–1	0.4997
Arterial type indicator (1 if principal, 0 minor)	0.3275	0–1	0.4693
International roughness index indicator, IRI (1 if IRI reading greater than 95 and less than or equal to 170, 0 otherwise)	0.4206	0–1	0.4937
International roughness index indicator, IRI (1 if IRI reading greater than 170, 0 otherwise)	0.1247	0–1	0.3304

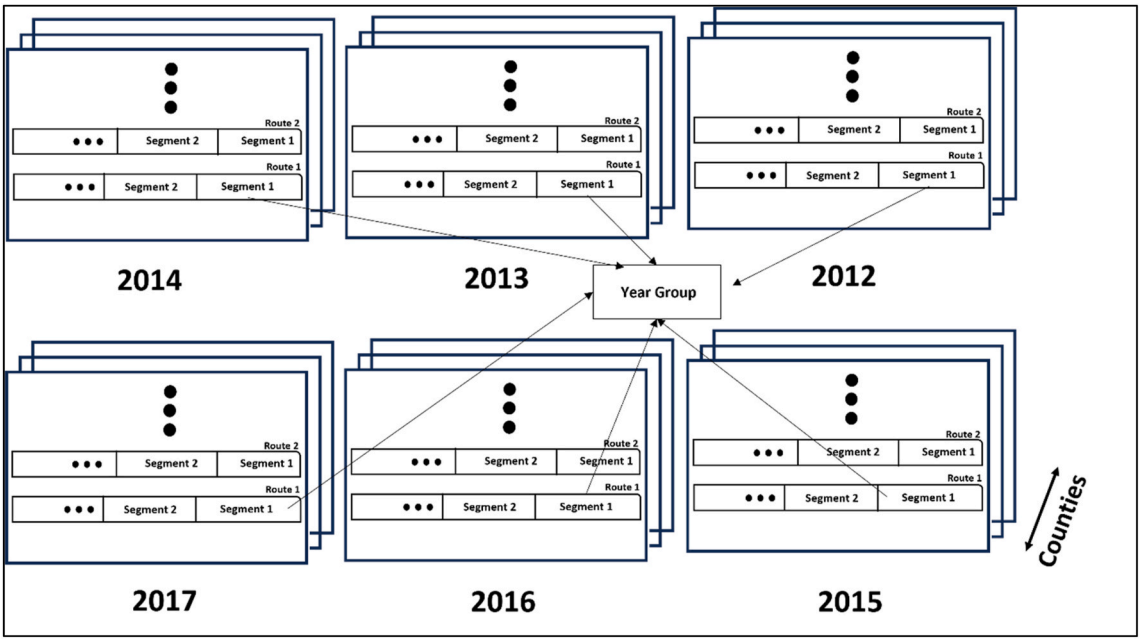


Fig. 1. The nested structure of the dataset.

3. Data preparation and description

The data used in this study are for the state of Ohio and were provided by the Highway Safety Information System (HSIS). Data were received in two separate Excel files for each year from 2012 to 2017, one for the segment information file and one for the crash data. The segment file mainly contained information on state routes. There were multiple routes within each county, and each route was divided into homogeneous segments. Within each route, the mileposts denote the starting and end points of each segment. Thus, both the route name and milepost in the crash data were utilized to effectively associate each crash record with the corresponding segment based on the starting and ending milepost values, as well as the route name provided in the segment file. This process facilitated the determination of crash counts within each individual segment. The target population for this study was undivided two-lane arterial roads. After filtering, there were 38,184 records for segment information and crash counts for six years combined. This information describes undivided two-lane arterial segments in terms of annual average daily traffic, segment length, roadway width excluding shoulder width, total shoulder width, and several indicators. Table 2 shows the summary statistics for the variables. Several variables were created based on existing variables, or transformed. The time variable was indexed from zero to five, representing the years 2012 (index zero) to 2017 (index five). The total shoulder width was calculated by adding both outside shoulder widths. The International Roughness Index (IRI) is classified into three groups: good pavement condition (IRI reading less than or equal to 95), fair pavement condition (IRI reading greater than 95 but less than or equal to 170), and poor pavement condition (IRI reading greater than 170). Two grouping variables were created: one that contained each segment over the years, and another that contained each route. The year group had 9,104 distinct groups. Among these, 3,569 groups had 6 observations, 125 groups had 5 observations, 2,677 groups had 4 observations, 557 groups had 3 observations, 1,590 groups had 2 observations, and 586 groups had 1 observation. The route groups, which marked segments from the same routes, comprised a total of 621 routes (see Fig. 1). These grouping variables define the nested

Table 3
Model estimation results for varying intercept models and varying intercept and slope models.

Model	Model A			Model B		
Fixed parameters	*Estimate*	*Std. error*	*Z-stat*	*Estimate*	*Std. error*	*Z-stat*
Intercept	−4.6370	0.1718	−26.99	−4.6941	0.1728	−27.16
Time variable	0.0584	0.0080	7.33	0.0633	0.0084	7.56
Time variable^2	−0.0089	0.0016	−5.47	−0.0111	0.0016	−6.76
Ln(AADT)	0.7083	0.0197	36.03	0.7159	0.0197	36.32
Ln(Segment length)	0.9002	0.0078	114.82	0.9087	0.0104	87.56
Total outside shoulder width (in feet)	−0.0163	0.0019	−8.54	−0.0172	0.0021	−8.03
International roughness index (IRI), indicators						
Fair pavement conditions $95 < IRI \leq 170$	0.0659	0.0126	5.23	0.0704	0.0129	5.48
Poor pavement conditions $IRI > 170$	0.1659	0.0195	8.52	0.1693	0.0198	8.54
Area	−0.2255	0.0218	−10.34	−0.2215	0.0222	−9.98
Random parameters						
Standard deviation of intercept (at level 2)	0.6617	0.0087	75.79	0.6194	0.0116	53.23
Negative sign percentages	≈100%			≈100%		
Standard deviation of intercept (at level 3)	0.2767	0.0162	17.04	0.2509	0.0195	12.87
Negative sign percentages	≈100%			≈100%		
Standard deviation of the time variable (at level 2)	–	–	–	0.0663	0.0055	12.06
Negative sign percentages				17%		
Standard deviation of the time variable (at level 3)	–	–	–	0.0316	0.0045	6.95
Negative sign percentages				2%		
Standard deviation of Ln(Segment length) (varying across routes only)	–	–	–	0.1328	0.0105	12.70
Negative sign percentages				≈0%		
Standard deviation of total outside shoulder width (varying across routes only)	–	–	–	0.0141	0.0020	7.21
Negative sign percentages				89%		
Goodness of fit measures						
Deviance	115644.7			115445.6		
AIC	115668.7			115481.6		
BIC	115771.3			115635.5		
Forecasting accuracy						
RMSE	1.582006			1.49796		
Likelihood ratio test	Model A vs. Model B					
Degree of freedom	6					
Chi-square Statistics	199.06					
P-value	<000.1					

structure discussed in the methodology.

4. Results and discussions

Crash counts typically exhibit overdispersion, a condition in which the variance exceeds the mean. As illustrated in Table 2, the mean crash count was 2.44, while the variance was 18.52 (SD: 4.3030). This case leads to the favoring of the negative binomial model over the Poisson model. Another factor to consider is the presence of excess zeros, which contribute to overdispersion and heterogeneity [1]. To address this, the intercept-only zero inflated negative binomial model was compared with the intercept-only negative binomial model. Both models yielded a deviance of 152265, suggesting that excess zeros did not significantly influence the results. The nested structure described in the methodology was tested by comparing a varying intercept-only model with a fixed intercept-only model. The reduction in deviance was 27155, indicating a significant improvement in the model's fit. This suggests a significant correlation among observations on the same segment and segments along the same route. While the correlation over the years is expected because crash counts are in the same segment, the correlation of segments on the same routes is less intuitive. To investigate this, a varying intercept-only model that varied only across segments was compared with one that varied across segments within routes and across routes. The latter model demonstrated a significant reduction in deviance (641), indicating a significant correlation between segments on the same route. Subsequently, a comparison was made between a varying intercept-only model and another with all significant variables added as fixed parameters. The reduction in deviance was 9,423.3, suggesting that adding explanatory variables as fixed parameters significantly improved the model's fit. The last step was to test the improvement of a model that added significant variables as random parameters compared with a varying intercept-only model with explanatory variables added as fixed parameters. This model showed a reduction of 199 in deviance, indicating that adding significant variables as random parameters significantly improved the model's fit.

Table 3 presents the results of the final modeling step. Model A represents the varying intercept-only model, with explanatory variables added as fixed parameters, while Model B represents the varying intercept and significant varying slope model. The likelihood ratio test indicated that Model B significantly outperformed Model A in terms of fit. Furthermore, the RMSE for Model B was

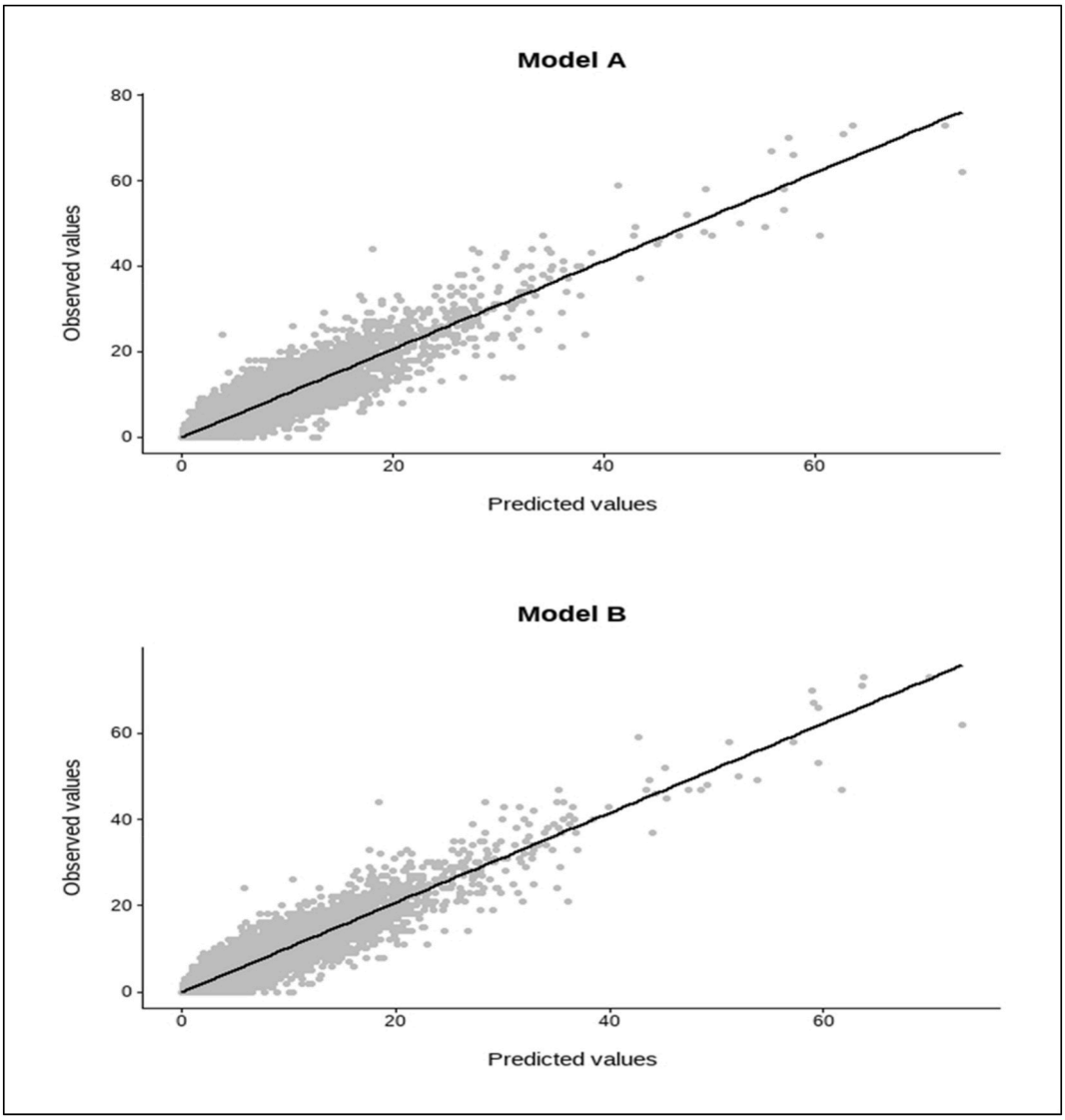


Fig. 2. Predicted values versus observed values for Models A and B.

marginally lower than that for Model A, suggesting a slight improvement in forecasting accuracy. Fig. 2 provides a visual comparison of the predicted values versus the observed values for both models, further demonstrating that Model B offers a marginally better overall fit compared to Model A; the points were closer to the straight line in Model B compared to Model A. Thus, Model B exhibits a better fit than Model A

In Model A, the intercept, which almost always took negative values, as shown in Table 3, varied across segments within and across different routes. This model retained only significant variables and identified five significant fixed-parameter variables: time variable, the natural logarithm of AADT, the natural logarithm of the segment length, the international roughness index (categorized into three groups), and area. Model B incorporated the same set of variables as Model A but also included significant random slopes for the time variable, the natural logarithm of segment length, and total outside shoulder width. The key difference between the two models was in these three variables, which significantly enhanced the fit of the model. The time variable was used to analyze individual crash count rates over the years. In Model A, both the linear and quadratic terms of the time variable were significant. This suggests that, on average, crash counts increased over time, but the increase decelerated. Model B exhibited a similar pattern, but the linear term of the time variable varied across segments within routes and across routes. At Level 2, the time variable coefficient followed a normal distribution, with a mean of 0.0633 and a standard deviation of 0.0663. At Level 3, it followed a normal distribution, with a standard deviation of 0.0316. Fig. 3 illustrates this quadratic relationship with the time variable on the x-axis and the incidence rate ratio on the y-axis. It depicts the incidence rate ratio of the time variable (including both the linear and quadratic terms), with all other variables held at their means, for Model B. If the sign of the linear term in the time variable were to change, Fig. 3 would then depict a rate that decreases, with an accelerating downward trend as the time variable increases. Consequently, 17% of the segments within the routes

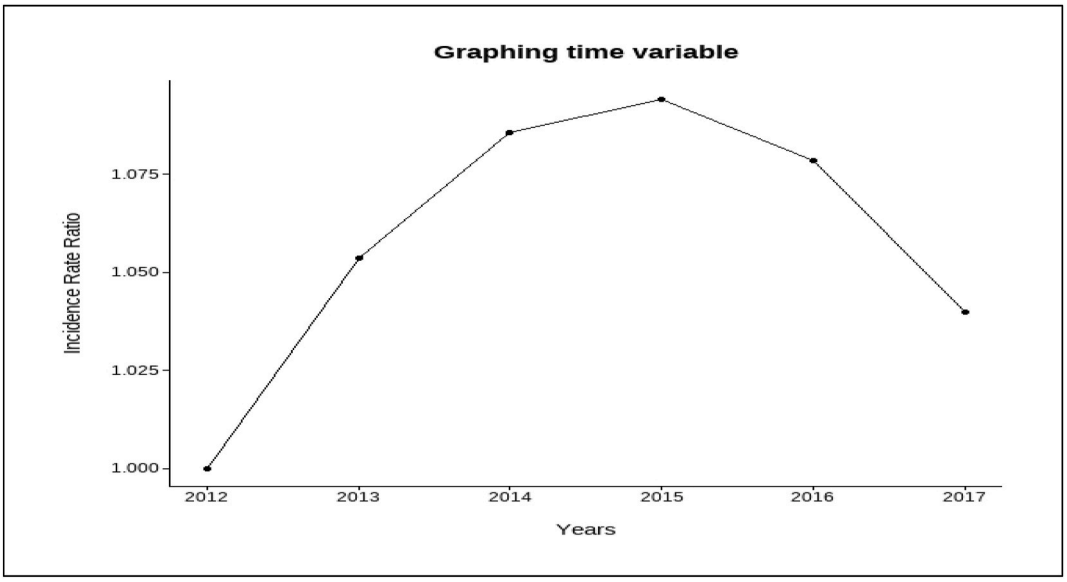


Fig. 3. Time variable versus incidence rate ratio for Model B.

Table 4
Average marginal effects for Models A and B.

Variables	Model A	Model B
Time variable	0.1413	0.1630
Time variable^2	−0.0214	−0.0269
Natural log *of annual average daily traffic*	1.7131	1.7321
Natural log *of segment length (miles)*	2.1774	2.1249
Total shoulder width (feet)	−0.0395	−0.0366
Area indicator (1 if the segment located in rural area, zero urban area)	−0.5487	−0.5389
International roughness index indicator, IRI (1 if IRI reading greater than 95 and less than 170, 0 otherwise)	0.1603	0.1716
International roughness index indicator, IRI (1 if IRI reading greater than 170, 0 otherwise	0.4279	0.4373

exhibited this latter pattern, while only 2% of the routes did so. In other words, for 17% of segments within routes and 2% of routes, crash counts decreased at accelerating downward trend as the time variable increases. In Model B, the coefficient of the natural logarithm of segment length was found to vary significantly across routes, and an increase in the natural logarithm of the segment almost always led to an increase in the crash count. The coefficient of the total outside shoulder also varied significantly across routes, with 89% of the coefficients taking negative values. This implies that as the shoulder width increased, the number of crashes decreased. These findings are consistent with past studies [6,7].

The two models offered comparable interpretations with respect to the fixed parameters. In both models, the coefficient of the indicator of fair pavement conditions suggested an increase in crash count compared to good pavement conditions. Furthermore, the indicator of poor pavement conditions had a higher coefficient, implying a higher increase in crashes. Segments located in urban areas also exhibited more crashes than those in rural areas. Table 4 presents the average marginal effects for all variables for both models. Both models yielded comparable results and interpretations, all of which are logical and consistent with previous studies [6,7]. For example, the average marginal effect for the natural log of segment length indicated that a unit increase in this variable resulted in an average increase in crash count of 2.12 for Model B (and 2.18 for Model A). Thus, while the two models were comparable with respect to the fixed parameters in terms of interpretation, the random parameters provided additional information and significantly improved the model's fit.

5. Conclusion

Recent methodological advances in crash count modeling have made random parameter models an effective tool for capturing dependencies between observations. These random parameters are typically used to capture heterogeneity across observations, while grouped random parameters are used to capture heterogeneity across groups of observations. From another perspective, the effects of these parameters are fixed on observations within the same group under the assumption that these observations are correlated. Several studies have utilized grouped random parameters to account for a range of correlations. This study proposed a nested grouped random parameter negative binomial three-level longitudinal framework for modeling crash counts at the segment level for undivided two-

lane arterial roads. The proposed model accounts for correlations among segments located on the same route and the temporal correlation of annual crash counts on the same segments. The time variable, the year index, is included in the proposed model to analyze the rate of change over the years.

The proposed model was applied to Ohio data from 2012 to 2017. Two variants of the model are presented in Table 3: one with varying intercepts and a fixed slope and the other with varying intercepts and slopes. The findings reveal that the latter significantly improved the model's fit and provided additional valuable information in terms of interpretation. Furthermore, the varying time variable provided more insights into the rate of change of crashes on segments over the years. The findings reveal a significant quadratic relationship of the time variable, indicating that, on average, the crash counts of segments increase over time, but the increase rates decelerated as the time variable increases. Further, findings regarding the random aspect of the model show that 17% of segments within routes and 2% of routes had crash counts decreasing at accelerating downward trend as the time variable increases. The effect of the natural logarithm of the segment length varied significantly across routes and almost always increased crashes as it increased. Similarly, the effect of the total outside shoulder varied significantly across routes but with 89% of coefficients taking negative values, indicating that most of the time, its increase decreased crashes. The two variants of the proposed model had comparable interpretations with respect to fixed parameters, but the random parameters improved the model's fit and provided additional information about its interpretation. Both variants exhibited decent forecasting accuracy, as indicated by the RMSE and Fig. 2.

However, there is always room for improvement. The dataset used did not include variables that describe specific routes, which could improve the model's accuracy. These variables can be easily added to a third level that, for example, explains variations between routes. Another assumption is that the crash counts for the segments within each route are equally correlated. Future studies could consider assuming that adjacent segments are more correlated than far segments. Despite this potential for further refinement, the proposed model offers substantial insights into interpretation, with decent forecasting accuracy for crash count predictions, making it a powerful tool for reaching informed decisions to improve safety.

Data availability statement

The authors do not have permission to share data.

CRediT authorship contribution statement

Omar Almutairi: Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, 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.

Acknowledgments

The author would like to express sincere gratitude to the Highway Safety Information System (HSIS) for providing the dataset that significantly contributed to this study. Special acknowledgment is extended to Dr. Kristin Kersavage, the HSIS Lab Manager, whose assistance in delivering the data was invaluable.

References

[1] F.L. Mannering, C.R. Bhat, Analytic methods in accident research: methodological frontier and future directions, Anal Methods Accid Res 1 (2014) 1–22, https://doi.org/10.1016/J.AMAR.2013.09.001.
[2] S. Yasmin, N. Eluru, M.M. Haque, Addressing endogeneity in modeling speed enforcement, crash risk and crash severity simultaneously, Anal Methods Accid Res 36 (2022), https://doi.org/10.1016/j.amar.2022.100242.
[3] X. Wang, X. Zhang, Y. Pei, A systematic approach to macro-level safety assessment and contributing factors analysis considering traffic crashes and violations, Accid. Anal. Prev. 194 (2024), https://doi.org/10.1016/j.aap.2023.107323.
[4] A. Cohen, L. Einav, The effects of mandatory seat belt laws on driving behavior and traffic fatalities, Rev. Econ. Stat. 85 (2003) 828–843, https://doi.org/10.1162/003465303772815754.
[5] D. Lord, F. Mannering, The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives, Transp Res Part A Policy Pract 44 (2010) 291–305, https://doi.org/10.1016/J.TRA.2010.02.001.
[6] P.C. Anastasopoulos, F.L. Mannering, A note on modeling vehicle accident frequencies with random-parameters count models, Accid. Anal. Prev. 41 (2009) 153–159, https://doi.org/10.1016/j.aap.2008.10.005.
[7] T.U. Saeed, T. Hall, H. Baroud, M.J. Volovski, Analyzing road crash frequencies with uncorrelated and correlated random-parameters count models: an empirical assessment of multilane highways, Anal Methods Accid Res 23 (2019), https://doi.org/10.1016/J.AMAR.2019.100101.
[8] N. Feknssa, N. Venkataraman, V. Shankar, T. Ghebrab, Unobserved heterogeneity in ramp crashes due to alignment, interchange geometry and truck volume: insights from a random parameter model, Anal Methods Accid Res 37 (2023) 100254, https://doi.org/10.1016/J.AMAR.2022.100254.
[9] M.I. Jahan, T. Bhowmik, N. Eluru, Enhanced aggregate framework to model crash frequency by accommodating zero crashes by crash type, Transp Res Rec (2023), https://doi.org/10.1177/03611981231175887.
[10] S.S. Ahmed, F. Corman, P.C. Anastasopoulos, Accounting for unobserved heterogeneity and spatial instability in the analysis of crash injury-severity at highway-rail grade crossings: a random parameters with heterogeneity in the means and variances approach, Anal Methods Accid Res 37 (2023), https://doi.org/10.1016/j.amar.2022.100250.

[11] H. Ding, Y. Lu, N.N. Sze, C. Antoniou, Y. Guo, A crash feature-based allocation method for boundary crash problem in spatial analysis of bicycle crashes, Anal Methods Accid Res 37 (2023), https://doi.org/10.1016/j.amar.2022.100251.
[12] A. Ogungbire, S.S. Pulugurtha, Does non-motorists' safety perception of autonomous vehicles vary across opinion change stemming from crash occurrence? Investigating perceptions using fixed and random parameter ordered logit models, Heliyon 9 (2023) E19913, https://doi.org/10.1016/j.heliyon.2023.e19913.
[13] Q. Zeng, Q. Wang, K. Zhang, S.C. Wong, P. Xu, Analysis of the injury severity of motor vehicle–pedestrian crashes at urban intersections using spatiotemporal logistic regression models, Accid. Anal. Prev. 189 (2023), https://doi.org/10.1016/j.aap.2023.107119.
[14] S.A. Khan, A.P. Afghari, S. Yasmin, M.M. Haque, Effects of design consistency on run-off-road crashes: an application of a Random Parameters Negative Binomial Lindley model, Accid. Anal. Prev. 186 (2023), https://doi.org/10.1016/j.aap.2023.107042.
[15] C. Wang, F. Chen, Y. Zhang, S. Wang, B. Yu, J. Cheng, Temporal stability of factors affecting injury severity in rear-end and non-rear-end crashes: a random parameter approach with heterogeneity in means and variances, Anal Methods Accid Res 35 (2022), https://doi.org/10.1016/j.amar.2022.100219.
[16] S.S. Ahmed, N. Alnawmasi, P.C. Anastasopoulos, F. Mannering, The effect of higher speed limits on crash-injury severity rates: a correlated random parameters bivariate tobit approach, Anal Methods Accid Res 34 (2022), https://doi.org/10.1016/j.amar.2022.100213.
[17] Z. Sun, Y. Xing, J. Wang, X. Gu, H. Lu, Y. Chen, Exploring injury severity of vulnerable road user involved crashes across seasons: a hybrid method integrating random parameter logit model and Bayesian network, Saf. Sci. 150 (2022), https://doi.org/10.1016/j.ssci.2022.105682.
[18] S. Okafor, E.K. Adanu, S. Jones, Severity analysis of crashes involving in-state and out-of-state large truck drivers in Alabama: a random parameter multinomial logit model with heterogeneity in means and variances, Heliyon 8 (2022) E11989, https://doi.org/10.1016/j.heliyon.2022.e11989.
[19] M. Shukri, F. Jones, M. Conner, Theory of planned behaviour, psychological stressors and intention to avoid violating traffic rules: a Multi-Level modelling analysis, Accid. Anal. Prev. 169 (2022), https://doi.org/10.1016/j.aap.2022.106624.
[20] S. Manirul Islam, S. Washington, J. Kim, M. Haque, A comprehensive analysis on the effects of signal strategies, intersection geometry, and traffic operation factors on right-turn crashes at signalised intersections: an application of hierarchical crash frequency model, Accid. Anal. Prev. 171 (2022), https://doi.org/10.1016/j.aap.2022.106663.
[21] A.P. Afghari, E. Papadimitriou, F. Pilkington-Cheney, A. Filtness, T. Brijs, K. Brijs, A. Cuenen, B. De Vos, H. Dirix, V. Ross, G. Wets, A. Lourenço, L. Rodrigues, Investigating the effects of sleepiness in truck drivers on their headway: an instrumental variable model with grouped random parameters and heterogeneity in their means, Anal Methods Accid Res 36 (2022), https://doi.org/10.1016/j.amar.2022.100241.
[22] P. Intini, N. Berloco, A. Fonzone, G. Fountas, V. Ranieri, The influence of traffic, geometric and context variables on urban crash types: a grouped random parameter multinomial logit approach, Anal Methods Accid Res 28 (2020), https://doi.org/10.1016/j.amar.2020.100141.
[23] M.T. Islam, K. El-Basyouny, Multilevel models to analyze before and after speed data, Anal Methods Accid Res 8 (2015) 33–44, https://doi.org/10.1016/J.AMAR.2015.10.001.
[24] H. Huang, M. Abdel-Aty, Multilevel data and Bayesian analysis in traffic safety, Accid. Anal. Prev. 42 (2010) 1556–1565, https://doi.org/10.1016/j.aap.2010.03.013.
[25] F.L. Mannering, V. Shankar, C.R. Bhat, Unobserved heterogeneity and the statistical analysis of highway accident data, Anal Methods Accid Res 11 (2016) 1–16, https://doi.org/10.1016/j.amar.2016.04.001.
[26] H.C. Park, B.J. Park, P.Y. Park, A multiple membership multilevel negative binomial model for intersection crash analysis, Anal Methods Accid Res 35 (2022) 100228, https://doi.org/10.1016/J.AMAR.2022.100228.
[27] S. Pervaz, T. Bhowmik, N. Eluru, Integrating macro and micro level crash frequency models considering spatial heterogeneity and random effects, Anal Methods Accid Res 36 (2022) 100238, https://doi.org/10.1016/J.AMAR.2022.100238.
[28] A.S.M.M. Islam, M. Shirazi, D. Lord, Grouped Random Parameters Negative Binomial-Lindley for accounting unobserved heterogeneity in crash data with preponderant zero observations, Anal Methods Accid Res 37 (2023) 100255, https://doi.org/10.1016/J.AMAR.2022.100255.
[29] T. Bhowmik, S. Yasmin, N. Eluru, Do we need multivariate modeling approaches to model crash frequency by crash types? A panel mixed approach to modeling crash frequency by crash types, Anal Methods Accid Res 24 (2019), https://doi.org/10.1016/j.amar.2019.100107.
[30] T. Bhowmik, S. Yasmin, N. Eluru, Accommodating for systematic and unobserved heterogeneity in panel data: application to macro-level crash modeling, Anal Methods Accid Res 33 (2022), https://doi.org/10.1016/j.amar.2021.100202.
[31] E. Coruh, A. Bilgic, A. Tortum, Accident analysis with the random parameters negative binomial panel count data model, Anal Methods Accid Res 7 (2015) 37–49, https://doi.org/10.1016/j.amar.2015.07.001.
[32] N.S. Venkataraman, G.F. Ulfarsson, V. Shankar, J. Oh, M. Park, Model of Relationship between Interstate Crash Occurrence and Geometrics (2011) 41–48, https://doi.org/10.3141/2236-05.
[33] E. Dupont, H. Martensen, Multilevel modelling and time series analysis in traffic research – methodology, Deliverable D7.4 of the EU FP6 project SafetyNet (2007). http://www.dacota-project.eu/Links/erso/safetynet/fixed/WP7/D7_4/D7.4.pdf.
[34] M.E. Brooks, K. Kristensen, K.J. van Benthem, A. Magnusson, C.W. Berg, A. Nielsen, H.J. Skaug, M. Mächler, B.M. Bolker, glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling, R Journal 9 (2017) 378–400, https://doi.org/10.32614/RJ-2017-066.
[35] S. Washington, M. Karlaftis, F. Mannering, Statistical and Econometric Methods for Transportation Data Analysis, 2011.