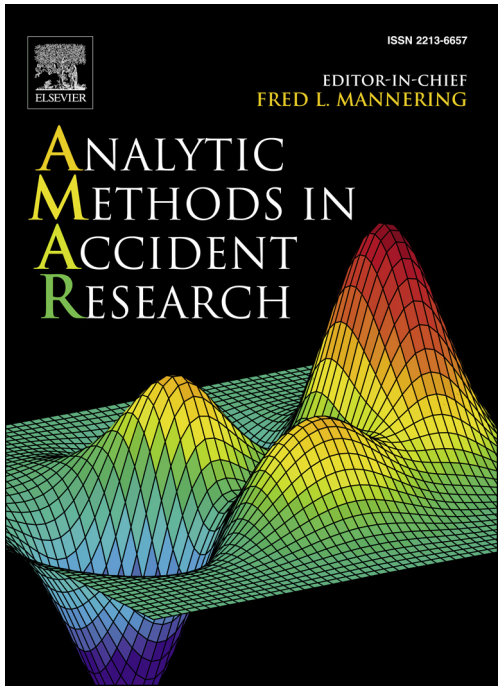

A marginalized random effects hurdle negative binomial model for analyzing refined-scale crash frequency data

Rongjie Yu, Yiyun Wang, Mohammed Quddus, Jian Li

PII: S2213-6657(18)30094-0
DOI: https://doi.org/10.1016/j.amar.2019.100092
Article Number: 100092
Reference: AMAR 100092

To appear in: *Analytic Methods in Accident Research*

Received Date: 16 October 2018
Revised Date: 4 May 2019
Accepted Date: 12 May 2019

Please cite this article as: R. Yu, Y. Wang, M. Quddus, J. Li, A marginalized random effects hurdle negative binomial model for analyzing refined-scale crash frequency data, *Analytic Methods in Accident Research* (2019), doi: https://doi.org/10.1016/j.amar.2019.100092

A marginalized random effects hurdle negative binomial model for analyzing refined-scale crash frequency data

Rongjie Yu[a], **Ph.D., Associate Professor**
E-mail: yurongjie@tongji.edu.cn

Yiyun Wang[a], **Graduate Research Assistant**
E-mail: 1731347@tongji.edu.cn

Mohammed Quddus[b], **Ph.D., Professor**
E-mail: M.A.Quddus@lboro.ac.uk

Jian Li[a], **Ph.D., Associate Professor**
*** Corresponding Author**
E-mail: tongji_lijian@163.com

[a] The Key Laboratory of Road and Traffic Engineering, Ministry of Education, 201804, Shanghai, China
[b] School of Architecture, Building and Civil Engineering, Loughborough University, Loughborough LE11, 3TU, United Kingdom

Revision submitted for potential publication in Analytical Methods in Accident Research

Abstract

Crash frequency prediction models have been an important subject of safety research that unveils a relationship between crash occurrences and their influencing factors. Recently, the hourly-based refined-scale crash frequency analysis becomes attractive since it holds the benefits of introducing time-varying explanatory information (e.g. traffic volume and operating speed). However, crash frequency data with short time intervals possess the analytical issues of excessive zeros and unobserved heterogeneity. In this study, a marginalized random effects hurdle negative binomial (MREHNB) model was developed in which the hurdle modelling structure handles the excessive zeros issue and site-specific random effect terms capture the factors associated with unobserved heterogeneity. Moreover, the marginalized inference approach was first introduced here to obtain the marginal mean inference for the overall population rather than subject-specific estimations. Empirical analyses were conducted based on data from the Shanghai urban expressway system, and the MREHNB model was compared with the HNB (hurdle negative binomial) and the REHNB (random effects hurdle negative binomial) model. In terms of model goodness-of-fits, REHNB and MREHNB model showed substantial improvement compared to the HNB model while there was no distinct difference between the REHNB and MREHNB models. However, as for the estimated parameters, the MREHNB model provided better inference precisions. Furthermore, the MREHNB model provided interesting findings for the crash contributing factors, for example, higher ratios of local vehicles within the volume would enhance the probability of crash occurrence; and a non-linear relationship was concluded between traffic volume and crash frequency with the moderate level of volume held the highest crash occurrence probability. Finally, in-depth analyses about the modeling results and the model technique were discussed.

Keywords: Marginalized model; Site-specific random effects term; Hurdle negative binomial model; Excessive zeros; Unobserved heterogeneity

1 Introduction

Traffic crashes result in significant casualties and economic losses. According to the World Health Organization (WHO), 1.25 million people die each year from road traffic crashes and their associated cost is equal to around 3% Gross Domestic Product (GDP) worldwide (Organization, 2015). In order to improve traffic safety, tremendous efforts have been devoted to identify the crash contributing factors by formulating crash frequency prediction models.

The crash frequency analyses were conducted based upon aggregated crash data and the majority studies aggregate data over long time intervals such as monthly (Wan *et al.*, 2010) and yearly-based (Abdel-Aty and Radwan, 2000). During the crash frequency data preparation procedures, the explanatory variables are also required to be

aggregated. For instance, traffic volume was combined as AADT (Annual Average Daily Traffic) (Aguero-Valverde and Jovanis, 2008) and monthly or even yearly averaged speed was used to represent the operational conditions (Ma *et al.*, 2008). However, the time-varying features, such as weather conditions (e.g. rainfall, snowfall, fog) (Ahmed *et al.*, 2014) and operational statuses (e.g. traffic volume, vehicle occupancy, traffic speed) (Imprialou *et al.*, 2016) have high influences on crash occurrence (Lord and Mannering, 2010). In addition, the highly aggregated crash frequency analysis may lead to the loss of potentially important time-varying explanatory information (e.g. Lord and Mannering 2010; Washington *et al.* 2010).

Recently, as high quality traffic sensing data with smaller time intervals have become much more affluent and conveniently obtainable, the refined-scale crash frequency analyses, mostly based on hourly or daily time intervals, have been emerging (Usman *et al.* 2010). Researchers argued that the analysis based on the small time intervals is capable of revealing the impact of time-varying variables on crashes and providing in-depth understanding of the relationships between the explanatory variables and the crash occurrence (Lord and Mannering, 2010). Taking traffic operational status into account as an example, through obtaining the volume and speed information prior to crash occurrence, studies could analyze the traffic conditions at the time of a crash occurring and reveal the high-risk crash-prone traffic conditions. In addition, the concluded results from the refined-scale crash frequency analyses would benefit the design of Active Traffic Management (AMT) control strategies, such as traffic police force patrol routes scheduling at hourly intervals (Kuo *et al.*, 2013).

However, there are several critical issues that need to be addressed when analyzing the crash frequency data with short time intervals. First, due to the rare event and randomness feature of crash occurrence, the refined-scale datasets contain excessive zero counts. For example, Chen *et al.* (2016) revealed that 97% of the crashes occurred in a mountainous highway in Colorado exhibit zero counts with a daily aggregation unit. Traditional crash count analysis models, such as Poisson and negative binomial models, have limited capabilities in dealing with excessive zeros (Dong *et al.*, 2014); where the zero-inflated models have then widely been employed to handle excessive zeros (Carson and Mannering, 2001; Qin *et al.*, 2005). The zero-inflated count model assumes the extra zero counts are arising from two states: a true-zero state where the roadway segments are inherently safe (i.e. no crash would happen on this roadway segment all the time); and a non-zero state, where no crash occurs in the observation periods (Shankar *et al.*, 1997). However, intrinsically safe roadway segments are unlikely to exist in reality, for example, although the geometry design of the roadway segment is nearly perfect, crashes would happen due to the unsafe behavior of a drunken driver. This makes the fundamental assumption of the zero-inflated count model logically fallacious (Lord *et al.*, 2005, 2007). Thus, the hurdle model, also named as two-part model, is employed as an alternative approach to adjust excessive zeros in the dataset (Ma *et al.*, 2016) which assumes the roadway segments with zero crash observation are only safe over the study period, not inherently.

Moreover, the refined-scale crash frequency data exhibits a typical panel data structure, which holds repeated observations at roadway segment level. This suggests that there are potential spatio-temporal correlations between the observations, which is due to the unobserved heterogeneity (i.e. spatial correlation as a result of context specific factors, such as pavement surface conditions and driver population characteristics, are difficult to observe and incorporate into a model) and the observed homogeneity (i.e. serial correlation resulting from the crashes observed over time on the same segment). Such correlations would violate the disturbance independence assumption and may result in erroneous parameter estimates (Washington *et al.*, 2010). Inappropriate treatment of unobserved heterogeneity would result in biased parameter estimates and incorrect inferences (Mannering and Bhat, 2014). To deal with the unobserved heterogeneity issue, a random effects model, which allows for a site-specific disturbance term (in addition to an overall disturbance term) to account for random disturbances specific to each roadway segment were introduced (Mannering *et al.*, 2016).

However, the estimated parameters obtained from the random effects model are subject-specific interpretations. To be more specific, the parameters are conditioned on the random effects related to individuals within the estimated cluster (Diggle, 2002). The obtained influencing factors on crash occurrence could only fit for the specific roadway segments within the sample data, while the more critical marginal mean inference for the overall population cannot be identified (Su *et al.*, 2015). That is to say, they may fail improving the predictive capability for the population due to the poor generalization. In addition, from the model application perspectives, such limitation would impede model transferring to other sites where population-averaged inference approach (e.g. marginalized inference models) is expected rather than the cluster-specific approach.

In this study, a marginalized random effects hurdle negative binomial (MREHNB) modeling approach was proposed to solve the above-mentioned issues. Within the model structure, the hurdle modelling framework was used to deal with the excessive zero issue, while site-specific random effects terms were further introduced to capture the unobserved heterogeneity for each roadway segment. Furthermore, a marginalized inference technique was employed to obtain the population-averaged interpretations for the estimated parameters. The empirical analyses were conducted based on data from the Shanghai urban expressway system. The proposed MREHNB model was then compared with the hurdle negative binomial (HNB) model and the random effects hurdle negative binomial (REHNB) model.

The rest of the paper is organized as follows: the second section discusses relevant studies that handling excessive zeros, the unobserved heterogeneity, and the subject-specific interpretation issue. The next section provides a detailed description of the data preparation, which are followed by the description of the methodologies employed. Then, the fifth section presents the modeling results, and finally, the conclusions and discussions of the work are presented.

2 Literature review

Hurdle negative binomial model

In the literature, zero-inflated Poisson and zero-inflated negative binomial models have been widely adopted to deal with the excessive zeros in crash frequency analyses (Lee and Mannering, 2002; Chin and Quddus, 2003). They have been proved to provide a statistically superior fit to the data in a wide variety of fields (Malyshkina and Mannering, 2009). However, the zero-generating assumptions required by zero-inflated models assumes the existence of inherently safe roadway, were inferred to be logical fallacious (Lord *et al.*, 2005, 2007). Recently, hurdle models have been employed as an alternative analysis approach (Ma *et al.*, 2016). Being a two-part model, the first part of the hurdle model is used to resolve whether the count value is zero or positive; given the value is positive, and then the second part is applied with a truncated count distribution (Cragg, 1971). It is argued that using a 'hurdle' to decide the existence of the zero record is more realistic for the crash frequency analysis (Kassahun *et al.*, 2014). Son *et al.* (2011) utilized a hurdle model to identify the hazardous locations and their contributing factors. Hosseinpour *et al.* (2013) developed a hurdle model to examine the effects of various roadway characteristics for the pedestrian-vehicle crashes. Chen *et al.* (2016) applied a hurdle model to investigate the influence of weather on the traffic safety in real-time. The results showed that the hurdle model outperformed the traditional count models to handle the excessive zeros, and was identified as the best model according to the goodness-of-fit measures.

Marginalized random effects model

In order to address the unobserved heterogeneity issue, random effects models have been introduced. Chin and Quddus (2003) utilized a random effects negative binomial model to consider the site-specific effects when examined traffic crash occurrence at signalized intersections. Chen *et al.* (2016) applied a site-specific random effects model to investigate the short-term impact of driving characteristics on crash occurrence. Anarkooli *et al.* (2017) adopted a random effects model to investigate factors affecting the injury severity of single-vehicle rollover crashes. The concluded results have highlighted the superiority of random effects model for enhancing the model goodness-of-fits (Chin and Quddus, 2003). However, the estimations for the random effects models are conditional-based (i.e. individual-specific) interpretations, and cannot provide a marginal mean (or referred to as a population-averaged interpretation (Iddi, 2013)).

Heagerty (2004) first proposed the marginalized random effects model (MREM) to address the individual-specific issue brought by random effects terms. The marginal means are modeled directly and the random effects are still remained accounting for the unobserved heterogeneity (Lee *et al.*, 2011). It was concluded that the regression coefficients for MREM models do not depend on random effects. To be specific, the offered inferences are not only for the specific individuals within development clusters, but also for the individuals not being considered within the clusters (Pavlou *et al.*,

2015a). Moreover, they are less likely to have biases when the random effects model misspecification occurs (Heagerty, 2004). It was, for example, concluded by Lee *et al.* (2011) and Su *et al.* (2015) through the simulation studies, that the coefficient interpretations obtained from the marginal models were closer to the true estimates. The benefits of MREMs have been validated and utilized in medical (Su *et al.*, 2015) and public health studies (Long *et al.*, 2015). However, to the best of the authors' knowledge, it has not been applied to traffic crash frequency analysis.

To conclude, in this study, the hurdle modeling structure has been adopted to account for excessive zero observations and the site-specific random effects were introduced to handle the unobserved heterogeneity issue. Furthermore, *for the first time*, the marginalized inference technique was introduced to obtain the population-averaged interpretations for the estimated parameters.

3 Data Preparation

In this study, empirical data from the Shanghai urban expressway system were utilized. More specifically, four major expressways, which are Inner ring elevated road, Middle ring elevated road, North-South elevated road, and Yan'an elevated road have been analyzed, given their high quality crash and traffic sensing data. The location and occurrence information for crashes are recorded based on the full coverage of traffic surveillance system. Moreover, the expressway network was equipped with the license plate recognition (LPR) system that has an average spacing of 1 km. In addition, due to the peak hour travel restriction policy for non-local vehicles, and the network's large spatial scale, the traffic flow characteristics (including traffic composition, operating speed, and volume) varies from hour to hour.

Three datasets were employed to form the crash frequency analysis data: (1) crash data; (2) roadway segment geometric data; and (3) License plate recognition (LPR) data. The study period was set to October 2012, considering the data availability from different sources.

Crash frequency variable

For crash data, Fig. 1 shows the hourly based temporal distribution of crash frequency on the Shanghai urban expressway system. It demonstrates that there are limited crashes reported during 23:00 to 06:00. Non-reporting event over the night period could be included (Amoros *et al.*, 2006), and this may produce biased estimates (Mannering and Bhat, 2014). Thus, the time period of the final dataset was set from 06:00 to 23:00, where a similar time data processing approach utilized in Mo *et al.*'s (2017).

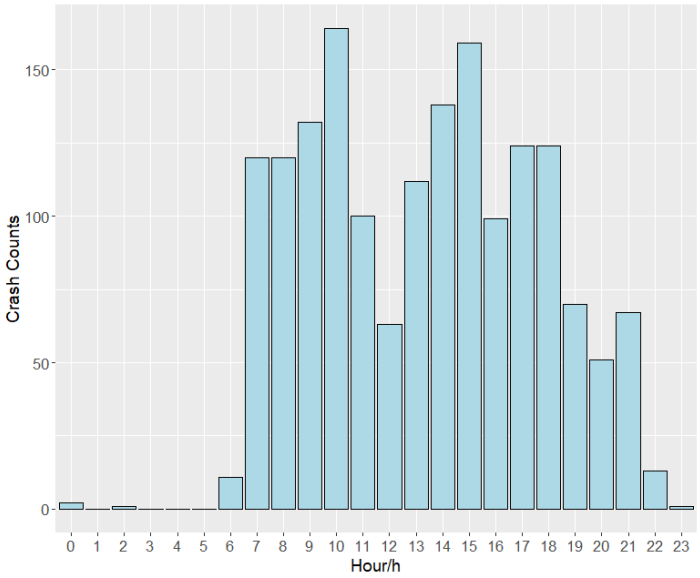

Fig. 1. Hourly distribution of crash counts.

A total of 1,659 crashes occurred on the targeted road segments during October 2012. The crash observations were further aggregated into hourly crash frequency data by 191 roadway segments, which were split using on-ramps and off-ramps as dividing points (Yu *et al.*, 2017). After the data aggregation procedure, crash frequency per hour holds the mean of 0.0176 and the standard deviation of 0.1666, where the 98.63% of the observations possess zero crash count.

The spatio-temporal distribution of crashes for the roadway segments over the analysis time periods is shown in Fig. 2.

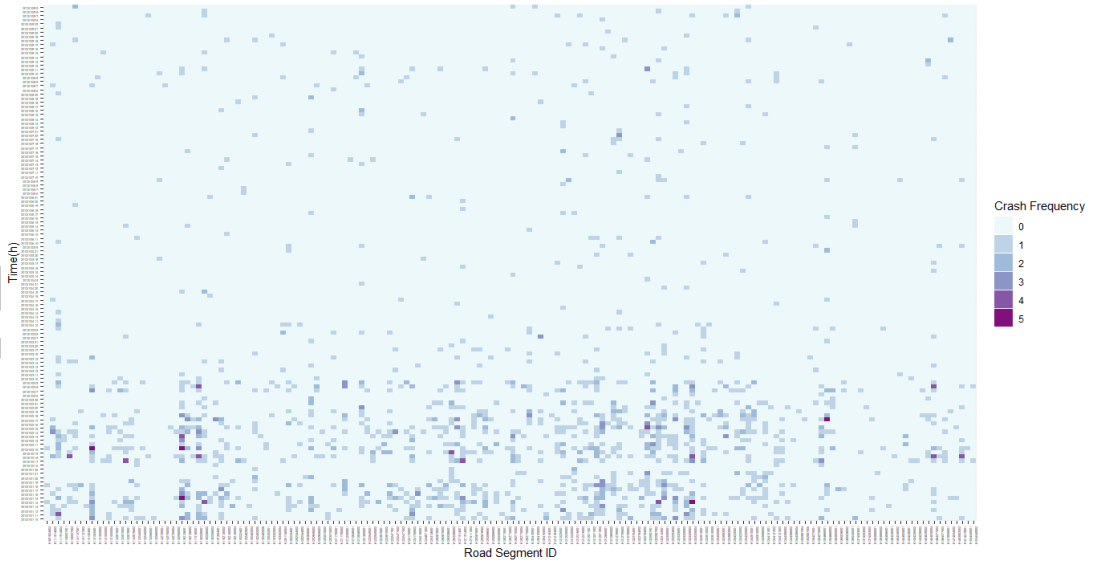

Fig. 2. Crash frequency distribution.

It is noticeable that the distribution has a certain spatial agglomeration effect over some roadway segments and some time periods. Although the crash frequency has excess zeros, the non-zero records are fitted with count data model.

Independent variables

Three groups of variables were selected as independent variables in the models: (1) geometric characteristics; (2) traffic operational statuses and (3) temporal features. Among these, the traffic operational statuses are time varying while the geometric characteristics comprise local specific variables.

For each roadway segment, geometric characteristics data were obtained from on-line street-view map, given there are no detailed design files available. The summary statistics for geometric variables are shown in Table 1, with the ramp combination types indicated in Fig. 3. In final analysis models, geometric variables will be retained according to their statistical significance. Categorical variables (such as Ramp Type) and their potential category combinations between the variables will be explored and the combination with respect to the fit based on the likelihood ratio test (LRT) will be retained.

Table 1 Summary statistics of roadway geometric characteristics.

Variable	Description	Summary Statistics
Length	Roadway section length	Mean: 789 (m)
		Std.Dev.: 534
Ramp Type	Ramp combination type:	
	1. Off-ramp and On-ramp	1: 36
	2. Off-ramp and Off-ramp	2: 65
	3. On-ramp and Off-ramp	3: 22
	4. On-ramp and On-ramp	4: 70
DMS	If there is DMS (Dynamic Message Sign)	
	within the roadway section:	1: 58
	1. Has DMS	0: 135
	0. No DMS	
Curve	If it is a curve section:	
	1. Curve section	1: 93
	0. Straight section	0: 100
Speed Limit Sign	If there is speed limit sign within the roadway	
	section:	
	1. Has speed limit sign	1: 66
	0. No speed limit sign	0: 127
Transition Length	The transition length to the total length of a	Mean: 182 (m)
	roadway section	Std.Dev.: 175
Speed Limit Value	The level of speed limit value:	
	1. 50m/h or 60km/h	1: 80
	2. 80km/h	2: 113
Main Lane	The level of lane numbers for main road:	
	1. Has 2 lanes	1: 62
	2. Has 3 lanes	2: 84
	3. Has 4 lanes and above	3: 47

Access Lane 1	The level of lane numbers for the up-access road: 1. Has 1 lane 2. Has 2 lanes and above	 1: 26 2: 167
Access Lane 2	The level of lane numbers for the down-access road: 1. Has 1 lane 2. Has 2 lanes and above	 1: 27 2: 166

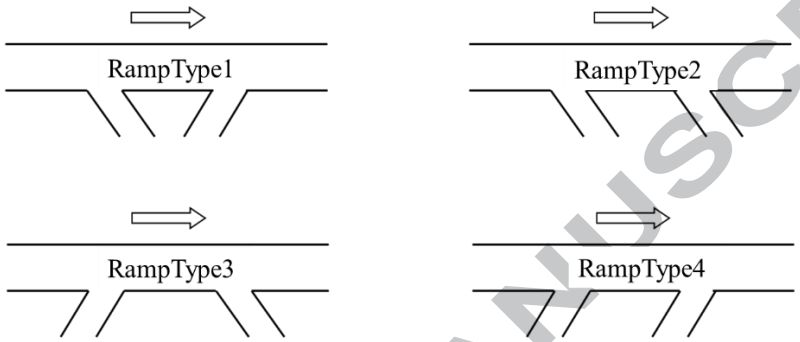

Fig. 3. Ramp type illustration.

In addition, for the traffic operational status, LPR data were collected from 268 detectors installed along the expressway network. The LPR system recorded each specific vehicle's information when it passes through the detector. Table 2 lists the recorded parameters and their descriptions. Before analyzing, the raw LPR data were processed to remove the missing observation, failure recognitions, error identifications, and other abnormal values.

Table 2 Descriptions of LPR data.

Parameters	**Descriptions**
TIME	Time when a vehicle passes through the detector section
TRANSECT ID	ID number of detector sections
LANE NO	Number of lanes of the segment on which a vehicle is traveling
VEHICLE ID	License plate number of vehicles
COLOR	License plate color of vehicles: 01-yellow: large vehicle 02-blue: light vehicle 06-black: foreign enterprise or embassy vehicle 25-white: special vehicle
VEHICLE REGION	Region of vehicles
SPEED	Vehicle speed: km/h

After the data cleaning process, traffic volumes for each time interval at each roadway segments were obtained and the average traveling speeds were calculated. Based on the VEHICLE ID variable in LPR data, the ratio of vehicles with Shanghai license plate

were calculated. It is noteworthy to state that since no prior assumption is considered about the functional relationships between traffic operational parameters and crash frequency for the Shanghai urban expressway system, different functional forms were tested for the traffic volume and speed variables. This includes both continuous and categorical forms. In the final models, only the significant variables and the best functional forms were kept.

Besides, temporal feature variables were also obtained. Several temporal variables were modelled and only the significant variables were retained. They are unique time-of-day periods: morning, afternoon and evening.

Table 3 shows the summary statistics for the traffic operational and temporal feature variables.

Table 3 Description of traffic operational and temporal feature characteristics.

Variable	Description	Summary Statistics
Volume	The level of volume during each time interval on specific roadway section:	
	1. If section volume is less than 2500	1: 19602
	2. If section volume is between 2500 and 5000	2: 60069
	3. If section volume is more than 5000	3: 15685
ShanghaiVehicleRatio	Ratio of vehicles with Shanghai license plate	Mean: 0.182 s.d.: 0.280
Workday	If it is on workday or not:	
	1. Workday	1: 54118
	0. Holiday or weekend	0: 41238
Period	The time period of the day:	
	1. Morning: 6:00 to 12:00	1: 33665
	2. Afternoon: 12:00 to 18:00	2: 33672
	3. Evening: 18:00 to 23:00	3: 28019

4 Methodology

In this study, three models were employed for the empirical crash frequency analyses. HNB model was firstly utilized as the baseline model, and then it was extended to include site-specific random effects terms for the binary and count part of hurdle structure respectively. Moreover, the MREHNB model was introduced to obtain the population-averaged interpretation for the model.

Random effects hurdle negative binomial model

For REHNB model, let Y_{it} represents the crash frequency with its observed value y_{it} for roadway segment *i* at time interval *t*. Within the Hurdle model structure, the first part models only the zero state using a Bernoulli model with probability π_{it} through a logit link function; the second part handles non-zero counts, which are assumed to

follow a truncated-at-zero probability mass function, in this case, a truncated Random Effects Negative Binomial model (Molenberghs *et al.*, 2010). Thus, the specification is as follows:

$$p(Y_{it} = y_{it}|\boldsymbol{b_i}, \boldsymbol{\xi}, \theta_{it}, \phi, \pi_{it}^c) = \begin{cases} \pi_{it}^c & \text{if } y_{it} = 0 \\ (1-\pi_{it}^c)\frac{f_i(y_{it}|\lambda_{it}^c, \theta_{it})}{1-f_i(0|\lambda_{it}^c, \theta_{it})} & \text{if } y_{it} > 0 \end{cases} \quad (1)$$

where $\pi_{it}^c = \Phi(\Delta_{it1} + b_{i1})$, $\lambda_{it}^c = \theta_{it} exp(\Delta_{it2} + b_{i2})$, $\lambda_{it} = \theta_{it}\kappa_{it}$. Δ_{it1} and Δ_{it2} are connector functions of the zero part and the positive count part. $\text{logit}(\pi_{it}) = \boldsymbol{x}_{1it}^T\boldsymbol{\gamma} + b_{i1}$, $\ln(\lambda_{it}) = \ln(\theta_{it}) + \boldsymbol{x}_{2it}^T\boldsymbol{\xi} + b_{i2}$.

$\boldsymbol{x}_{1it}^T$ is the regressor vector for the fixed parameter $\boldsymbol{\gamma}$ in the binary part in the form of:

$$\boldsymbol{x}_{it1}^T = \gamma_1 RampType2 + \gamma_2 RampType3 + \gamma_3 RampType4 + \gamma_4 DMS + \gamma_5 SpeedLimitSign + \gamma_6 Workday + \gamma_7 Morning + \gamma_8 Afternoon + \gamma_9 Volume2 + \gamma_{10} Volume3 + \gamma_{11} ShanghaiVehicleRatio \quad (2)$$

and $\boldsymbol{x}_{2it}^T$ is for the fixed parameter $\boldsymbol{\xi}$ in the count part:

$$\boldsymbol{x}_{it2}^T = \xi_1 RampType2 + \xi_2 RampType3 + \xi_3 RampType4 + \xi_4 Curve + \xi_5 logLength + \xi_6 Workday \quad (3)$$

b_{i1} and b_{i2} are the site-specific random intercepts for the binary and count part respectively. They are unobserved random effects of roadway segment *i*, which is constant within each roadway segment and different across roadway segments. In additional, the site-specific random effects are assumed to be independent to each other, and follow normal distribution with mean zero and variance $\boldsymbol{\sigma_b^2}$. Thus, they obey the distribution:

$$\boldsymbol{b_i} = (b_{i1}, b_{i2})^T \sim N(\mathbf{0}, \boldsymbol{\sigma_b^2}) \quad (4)$$

$$\boldsymbol{\sigma_b^2} = \begin{pmatrix} \sigma_{b1}^2 & 0 \\ 0 & \sigma_{b2}^2 \end{pmatrix} \quad (5)$$

θ_{it} is the over-dispersion parameter for the negative binomial model, and $\theta_{it} \sim Gamma(\alpha, 1/\alpha)$, with the constraint of $\beta = 1/\alpha$. The over-dispersion effect in the analysis data was tested in advance through the LRT (likelihood ratio test). The results indicated the over-dispersion feature and the negative binomial counterpart other than Poisson counterpart was adopted.

Vuong test (1989) is often used to justify the appropriateness of zero-inflated count data model (a zero-inflated model or a hurdle model) over standard count model (a Poisson model or a negative binomial model). Here it is used to compare HNB and NB model. In this test, a statistic m_{it} is firstly computed:

$$m_{it} = \ln\left(\frac{f_1(y_{it}|X_{it})}{f_2(y_{it}|X_{it})}\right) \quad (6)$$

where $f_1(y_{it}|X_{it})$ is the probability density function of the HNB model and $f_2(y_{it}|X_{it})$ is the probability density function of the NB model.

The Vuong's statistic is tested as:

$$V = \frac{\bar{m}\sqrt{N}}{S_m} \quad (7)$$

where $\bar{m}$ and S_m are the mean and the standard deviation of m_{it}, N is the sample size. If V > 1.96 it favors the HNB model while V < −1.96 it favors the NB model but otherwise neither model is preferred. In this study, a pscl package in R was utilized to conduct Vuong's test. The test results (as shown in Table 4) indicated that hurdle model is preferred with Vuong z-statistic of 7.

Table 4 Vuong test results for HNB and NB models.

	Vuong z-statistic	**H_A**	**p-value**
Raw	7.0084	Model1>Model2	<0.0001
AIC-corrected	6.3968	Model1>Model2	<0.0001
BIC-corrected	3.5024	Model1>Model2	0.0002

Marginal random effects hurdle negative binomial model

The marginal specification is

$$p(Y_{it} = y_{it}) = \begin{cases} \pi_{it}^m & \text{if } y_{it} = 0 \\ (1-\pi_{it}^m)\frac{f_i(y_{it}|\lambda_{it}^m)}{1-f_i(0|\lambda_{it}^m)} & \text{if } y_{it} > 0 \end{cases} \quad (8)$$

where $\text{logit}(\pi_{it}^m) = \boldsymbol{x}_{it1}^T\boldsymbol{\gamma^m}$ and $\ln(\lambda_{it}^m) = \boldsymbol{x}_{it2}^T\boldsymbol{\xi^m}$, with known regressors $\boldsymbol{x}_{it1}$ and $\boldsymbol{x}_{it2}$ and a vector of zero-inflation coefficients $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$. Specifying a logit link for the marginal model and a probit link for the conditional model leads to computational advantages from the probit-normal relationship, with the marginal parameters still having the odds-ratio interpretation. Hence, the connector functions are as follows (Kassahun *et al.*, 2014). For the logit,

$$\Delta_{it1l} = \sqrt{1+\frac{1}{2}\sqrt{1+\sigma_{b1}^2}}\,\Phi^{-1}[\text{expit}(\boldsymbol{x}_{it1}^T\boldsymbol{\gamma^m})] \quad (9)$$

$$\text{With } \text{expit}(\boldsymbol{x}_{it1}^T\boldsymbol{\gamma^m}) = \int \Phi(\Delta_{it1} + b_{i1})f(b_{i1})db_{i1} \quad (10)$$

The connector function for the positive counts part is as follows:

$$\Delta_{it2} = lnE(\theta_{it}) + \boldsymbol{x}_{2it}^T\boldsymbol{\xi^m} - \frac{1}{2}\sqrt{1+\sigma_{b2}^2} \quad (11)$$

Estimation

The likelihood function for REHNB and MREHNB model is given by the following:

$$L(\xi,\gamma,D,\phi) = \prod_{i=1}^{N}\int\prod_{t=1}^{n_i}\pi_{it}^c(\boldsymbol{b_i})^{I(y_{it}=0)}\{1-\pi_{it}^c(\boldsymbol{b_i})f_i(\boldsymbol{b_i})\}^{1-I(y_{it}=0)}\phi(\boldsymbol{b_i}|\boldsymbol{D})d\boldsymbol{b_i} \quad (12)$$

$$f_i(\boldsymbol{b_i}) = \binom{\alpha+y_{it}-1}{\alpha-1}\left(\frac{\beta}{1+\kappa_{it}^c\beta}\right)^{y_{it}}\left(\frac{1}{1+\kappa_{it}^c\beta}\right)^{\alpha}\kappa_{it}^{cy_{it}} \quad (13)$$

Where $\phi(\boldsymbol{b_i}|\boldsymbol{D})$ is the zero-mean normal density with variance-covariance matrix $\boldsymbol{D}$.

The application of numerical techniques to obtain the maximum likelihood estimates

was proposed using the adaptive Gauss-Hermite quadrature in the SAS procedure NLMIXED (Wolfinger, 2007).

Model Goodness-of-fit

For the model goodness-of-fit comparisons, Akaike Information Criteria (AIC) was adopted. The AIC is defined as:

$$AIC = -2l + 2p \quad (14)$$

where l is log-likelihood and p is the number of parameters. The smaller AIC value indicates a better fitted model.

5 Modeling Results

The final results for HNB, REHNB and MREHNB models are shown in Table 5. Statistically significant variables were grouped into three classes: roadway geometry characteristics, temporal features and traffic operational statuses. In general, the significant variables are consistent in terms of signs for their estimated parameters among the three models. For brevity, the following section introduces the modeling results of MREHNB model and discusses the similarities and differences between the three models.

Marginalized Random Effects Hurdle Negative Binomial Model

Binary part

The binary part of the MREHNB model has identified the influencing factors for the possibility of having crashes within the observation period, where a positive regression coefficient implies an increase in the probability of crash occurring.

Roadway geometry characteristics

Three roadway geometry characteristics variables are found to be significant: Ramp Type, DMS and Speed Limit Sign. As stated earlier, the Ramp Type variable was tested using two mechanisms: (1) a categorical variable with four ramp type categories (i.e. Ramp Type 1, Ramp Type 2, Ramp Type 3 and Ramp Type 4 where Ramp Type 1 is the reference category), and (2) a dummy variable (e.g. whether the segment is Ramp Type 1 or others).

Since there are two modeling parts (i.e. the binary part and the count part) in a single model and two mechanisms for the Ramp Type variable (i.e. dummy and categorical), there are a total of four pairwise combinations between the parts and the variables. The combinations are (1) Combination 1: using the dummy variable in the binary part and the count part; (2) Combination 2: using the dummy variable in the binary part and the categorical variable in the count part; (3) Combination 3: using the categorical variable in the binary part and the dummy variable in the count part; and (4) Combination 4: using the categorical variables in the binary part and the count part.

Firstly, the HNB model was estimated for each of the four combinations and the Likelihood Ratio Test (LRT) was utilized to obtain the combination that provided the best fit. It was found that Combination 3 (i.e. using the categorical variable in the binary part and the dummy variable in the count part) provided the best goodness-of-fit for the data. The same combination was also provided the best fit for the other two models (i.e. REHNB and MREHNB).

Therefore, Ramp Type variable used Ramp Type 1 as reference while Ramp Type 2, 3 and 4 are all statistically significant with positive coefficients. And Ramp Type 3 holds the largest estimated coefficient, which indicates that the off-on ramp type combination (Ramp Type 1) is relatively safer than the other ramp types and sections with the on-off ramp type combination (Ramp Type 3) are the most hazardous one.

In addition, DMS variable is significant with a positive coefficient, which indicates that segments with DMS are more likely to incur crashes. The DMS (Dynamic Message Sign) on Shanghai urban expressway system provides traffic flow operational information for the following roadway segments. The positive estimated coefficients may indicate that the DMS have brought too much information for the drivers passing by. The potential distractions or the additional work load to the drivers could increase the probability of crash occurrence. Similarly, estimated parameter for Speed Limit Sign variable is positive. It implies that the segments with speed limit signs have larger probability of crash occurrence.

Temporal features

For temporal features, Workday and Period variables are found to be significant. The estimated coefficient for Workday variable is negative, indicating that workdays are less likely to have crashes as opposed to weekdays and holidays. In terms of temporal periods, Evening Period is chosen as the reference category, Morning Period and Afternoon Period variables both have positive signs, which indicate that crashes are more likely to occur during morning and afternoon periods.

Traffic operational statuses

With respect to traffic operational statuses, Volume variable is discretized to three levels: Volume 1 (less than 2500), Volume 2 (between 2500 and 5000) and Volume 3 (more than 5000). Using Volume 1 as the reference category, the estimated parameter for Volume 2 are found to be significantly positive while Volume 3 holds a negative coefficient. It implies that crashes are much more possible to be observed for a moderate-level volume (Volume 2), rather than a low-level volume (Volume 1) or a high-level volume (Volume 3). Moreover, the proportion of vehicles with a Shanghai license plate is found to be significant with a negative sign. This can be concluded as that roadway segments with a higher composition of local vehicles are expected to have a lower possibility of crash occurrence.

Count part

The count part of the MREHNB model reveals the influencing factors for the crash frequency, where positive regression coefficients indicate the increasing number of crash occurrence.

Roadway geometry characteristics

For roadway geometry characteristics, Ramp Type, Log Length and Curve variables are found to be statistically significant. The binary Ramp Type variable exhibits the statistically significant result with a positive sign, indicating that the off-on ramp type (Ramp Type 1) holds lower crash frequency compared to the other types of ramps. The results are consistent with previous studies (e.g. Yu *et al.*, 2017). In addition, Log Length variable has a positive coefficient, which indicates that the number of crashes would increase along with the length of roadway segment. Similar modeling results have found in Chen *et al.*'s (2016). Furthermore, the estimated parameter for Curve variable holds a positive sign, indicating that the existence of curve would increase the number of crashes.

Temporal features

In regard to temporal features, only Workday variable was found to be significant with a negative sign. It indicates that fewer crashes would incur on workdays compared to weekends and holidays. Period variables are significant in the binary part while not in the count part, it indicates that the period variables have contribution to the likelihood of crash occurrence other than the number of crash frequency.

Model Comparison

From the aspect of model goodness-of-fits, the AIC values have reduced substantially for REHNB and MREHNB model compared to the HNB model. It indicates that REHNB and MREHNB model provide better fit relative to the HNB model. In addition, the standard deviation of random effects for binary part (σ_{b_1}) and count part (σ_{b_2}) are highly significant in both the REHNB model and the MREHNB model. It indicates that the proposed site-specific random effects model worked well on capturing unobserved heterogeneity among roadway segments.

Comparing REHNB model and its counterpart MREHNB, some estimates for binary part seem to differ as a result of marginalization, while the estimates corresponding to the count part appear similar. This follows from the nature of the connector function (Kassahun *et al.*, 2014).

With respect to the binary part of the REHNB and MREHNB models, non-negligible differences in the estimated parameters are found. For example, estimated parameters for Speed Limit Sign, Time Period of the Day, Shanghai Vehicle Proportion (highlighted in bold in the Table 5) reduced in the MREHNB model compared to the REHNB model. It can be explained that the marginalization diminished differences in the binary part of the models (Kassahun *et al.*, 2014). In regard to the count part between

the REHNB and MREHNB models, the estimated parameters appear similar. The phenomenon reflects in the variation of random effects parameter σ_{b_1} (for the count part) and σ_{b_2} (for the binary part). Between REHNB model and its marginal counterpart MREHNB model, the random effects variation changed substantially for the binary part while it remained similar for the count part. Generally, the marginal model leads to estimates that are relatively superior in precision, and suggests similar inferences for all covariates.

Table 5 Modeling results for HNB model, REHNB model and MREHNB model.

	HNB			REHNB			MREHNB		
Variable	**Estimate**	**S.D.**	**p value**	**Estimate**	**S.D.**	**p value**	**Estimate**	**S.D.**	**p value**
Binary part									
Intercept	-7.9459	0.1477	<0.0001	-8.1836	0.1974	<0.0001	-7.9724	0.1998	<0.0001
Roadway geometry characteristics									
Ramp Type 2	0.8462	0.0953	<0.0001	0.7765	0.2146	0.0004	0.7784	0.2092	0.0003
Ramp Type 3	1.0126	0.1034	<0.0001	0.9695	0.2489	0.0001	0.9650	0.2379	<0.0001
Ramp Type 4	0.7729	0.1000	<0.0001	0.7494	0.2281	0.0012	0.7307	0.2200	0.0011
DMS	0.3510	0.0717	<0.0001	0.4009	0.1936	0.0389	0.4123	0.1807	0.0236
Speed Limit Sign	**0.2157**	0.0600	0.0003	**0.3234**	0.1575	0.0415	**0.2736**	0.1486	0.0672
Temporal features									
Workday	-2.7907	0.1053	<0.0001	-2.7730	0.1059	<0.0001	-2.7703	0.1079	<0.0001
Morning Period	0.5159	0.0780	<0.0001	0.5391	0.0787	<0.0001	0.5409	0.0786	<0.0001
Afternoon Period	**0.6391**	0.0786	<0.0001	**0.7310**	0.0802	<0.0001	**0.7240**	0.0802	<0.0001
Traffic operational statuses									
Volume 2	0.3003	0.0738	<0.0001	0.1386	0.0803	0.0860	0.1369	0.0800	0.0885
Volume 3	-0.6148	0.1289	<0.0001	-1.0536	0.1446	<0.0001	-1.0798	0.1488	<0.0001
Shanghai Vehicle Proportion	**0.9743**	0.0933	<0.0001	**1.1067**	0.2535	<0.0001	**1.0624**	0.2415	<0.0001
Count part									
Intercept	-5.4641	1.1476	<0.0001	-5.2010	1.1564	<0.0001	-5.2099	1.1624	<0.0001
Roadway geometry characteristics									
Ramp Type	0.6476	0.2108	0.0021	0.5825	0.2266	0.0109	0.5902	0.2276	0.0103
Curve	0.2748	0.1336	0.0397	0.2537	0.1594	0.1131	0.2539	0.1598	0.1139
Log Length	0.3029	0.1292	0.0190	0.3370	0.1526	0.0284	0.3369	0.1531	0.0289
Temporal features									
Workday	-1.0433	0.3472	0.0027	-0.9854	0.3391	0.0041	-0.9903	0.3404	0.0041
α	0.4792	0.3802	0.2076	1.4169	0.9010	0.1175	1.3319	0.8403	0.1146
σ_{b_1}				**0.4363**	0.0889	<0.0001	**0.2179**	0.0446	<0.0001
σ_{b_2}				**0.8396**	0.0660	<0.0001	**0.2452**	0.0184	<0.0001

-2log-likelihood	13482	13105	13101
AIC	13519	13145	13141

1

6 Discussions and Conclusions

The refined-scale crash frequency analysis models have been widely applied to explore the relationships between time-varying contributing factors and crash occurrence. However, due to the small time intervals of the crash counts, issues such as preponderant zero counts, unobserved heterogeneities raised. Ignoring these issues would lead to biased parameter estimations and incorrect inferences. In this study, a MREHNB model was applied to deal with the abovementioned issues. At the same time, the proposed approach is targeted at the poor generalization issue for the random effects models. Empirical analyses were conducted based on data from the Shanghai urban expressway system. The proposed MREHNB model were further compared with traditional used HNB model and REHNB model.

From the modeling results, it can be seen that the significant variables are consistent among the three developed models. With respect to the model goodness-of-fits, the REHNB model provides much better fit compared to the HNB model; which further confirms the site-specific random effects terms effectively incorporate the unobserved heterogeneity. However, slightly better fit was observed between the REHNB and MREHNB models; this finding is consistent with several previous studies (Griswold and Zeger, 2004; Su *et al.*, 2015). As for the estimated parameters, the values varied substantially when the HNB model was extended to the REHNB model. As the MREHNB model was employed, some estimates for binary part seem to differ as a result of marginalization, while the estimates corresponding to the count part appear similar. Moreover, the marginal model leads to estimates that are relatively superior in precision, and suggests similar inferences for all covariates.

As for the crash occurrence influencing factors, in this study, the characteristics of roadway users were extracted and transferred as the proportion of local and non-local drivers as explanatory variables. The modeling results indicated that higher ratio of vehicles with Shanghai local license plate would increase the probability of crash occurrence. It is illustrated that the local drivers may drive much more recklessly relative to the non-local driver.

In addition, different from some previous studies, this study identified a non-linear relationship between traffic volume and crash frequency. Table 6 shows the studies that analyzed relationship between volume and crash frequency. In the previous studies, high traffic flow (e.g. AADT) was found to increase the risk of crashes (Abdel-Aty and Radwan, 2000). While some studies had identified low traffic volume to be a significant crash precursor (e.g. Garber and Ehrhart, 2000). The inconclusive impact was partially resulted from the highly aggregated data (Imprialou *et al.*, 2016). In this study, similarly, a continuous form of volume variable was adopted first, but it was not statistically significant. Therefore, further explorations of the impact of traffic volumes on crash frequency had been conducted. In Fig. 4, the high columns of crashes were clustered at a moderate-level of traffic volume. Then, discretized traffic volume variables were used

and included in the final model. The modeling results verifies the conjecture, that a moderate level of volume will increase the crash occurrence likelihood.

Table 6 Literature that analyzed relationship between volume and crash frequency.

Authors & year	Volume information assembled in the analysis	Key finding on the relationship between volume and crash frequency
Abdel-Aty and Radwan, 2000	Annual average daily traffic (AADT)	An increase in AADT per lane has a positive impact on the likelihood of accidents
Garber and Ehrhart, 2000	-	low traffic volume to be a significant crash precursor
Garber and Subramanyan, 2001	Vehicle occupancy ratio (real-time)	non-linear relationship
Elvik et al., 2004	Odds ratio of change in volume	lower flows increase the crash occurrence
Chang, 2005	Average daily traffic (ADT)	higher flows increase the crash likelihood

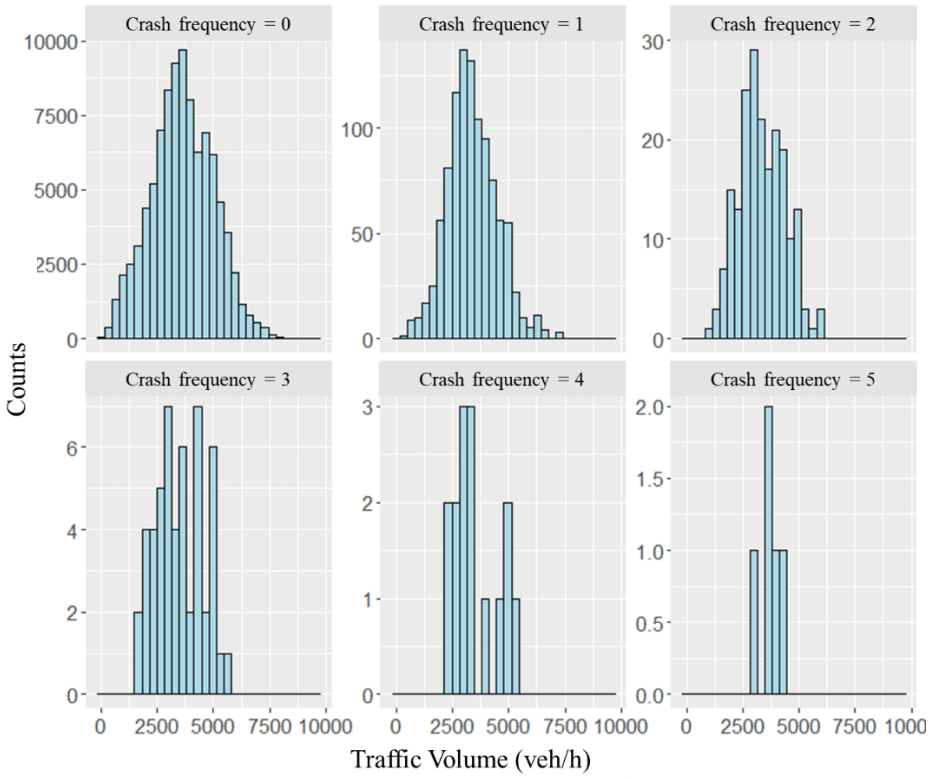

Fig. 4. Stratification histogram of crashes and volume.

Apart from the in-depth analysis of crash contributing factors, to the best of our knowledge, this study is the *first* to introduce MREHNB model to traffic safety research area. Retaining the advantage of REHNB for capturing the unobserved heterogeneity, MREHNB model could transform the conditional effects of estimated coefficients into

the marginal ones. Through revealing the population-averaged interpretation, the more accurate estimation could be provided for road segments not only within but also outside one of the clusters in developed model (Pavlou *et al.*, 2015a). Moreover, this study applied a marginalized random effects model (MREM) which integrates over the estimated distribution of random effects (Kassahun *et al.*, 2014). However, there are several other marginalized approaches that would obtain the marginal means, such as 'RE-approx' (scaling the estimated conditional coefficients using approximation of Zeger) and GEE (fitted by GEE with exchangeable correlation) (Pavlou *et al.*, 2015a). Therefore, future studies could try to compare the different marginalized approaches and further unveil the benefits of marginal inference techniques.

Finally, the analyzed results could shed lights on the traffic safety management. Firstly, as crashes are clustered at a moderate-level of traffic volume (2500-5000 veh/h), active traffic safety management such as variable message signs should be implemented based on the degree of traffic volume by giving drivers a warning of potentially hazardous traffic conditions. In addition, police force could be assigned more efficiently according to proportion of local/non-local vehicles on different roadway segment. Roadway segments with larger proportion of Shanghai vehicles have higher crash risk, where more police force could supply the effort on supervision and control.

Acknowledgements

This study was jointly sponsored by the Chinese National Natural Science Foundation (NSFC 71771174 and 71531011) and the 111 Project (B17032). In addition, the authors would like to sincerely thank Mrs. Xiaohan Yang, who provided us with valuable statistical advice.

References

Abdel-Aty, M.A., Radwan, A.E., 2000. Modeling traffic accident occurrence and involvement. Accident Analysis & Prevention 32 (5), 633-642.

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

Ahmed, M.M., Abdel-Aty, M., Lee, J., Yu, R., 2014. Real-time assessment of fog-related crashes using airport weather data: A feasibility analysis. Accident Analysis & Prevention 72, 309-317.

Amoros, E., Martin, J.-L., Laumon, B., 2006. Under-reporting of road crash casualties in france. Accident Analysis & Prevention 38 (4), 627-635.

Anarkooli, A.J., Hosseinpour, M., Kardar, A., 2017. Investigation of factors affecting the injury severity of single-vehicle rollover crashes: A random-effects generalized ordered probit model. Accident Analysis & Prevention 106, 399-410.

Carson, J., Mannering, F., 2001. The effect of ice warning signs on ice-accident frequencies and severities. Accident Analysis & Prevention 33 (1), 99-109.

Chen, F., Ma, X., Chen, S., Yang, L., 2016. Crash frequency analysis using hurdle models with random effects considering short-term panel data. International Journal of Environmental Research and Public Health 13 (11).

Chin, H.C., Quddus, M.A., 2003. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. Accident Analysis and Prevention 35 (2), 253-259.

Cragg, J.G., 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica 39 (5), 829-844.

Diggle, P.J., 2002. The analysis of longitudinal data. Journal of the American Statistical Association 90 (431), 1231–1232.

Dong, C., Richards, S.H., Clarke, D.B., Zhou, X., Ma, Z., 2014. Examining signalized intersection crash frequency using multivariate zero-inflated poisson regression. Safety Science 70, 63-69.

Garber, N., Ehrhart, A., 2000. Effect of speed, flow, and geometric characteristics on crash frequency for two-lane highways.

Griswold, M.E., Zeger, S.L., 2004. On marginalized multilevel models and their computation. Bepress.

Heagerty, P.J., 2004. Marginally specified logistic-normal models for longitudinal binary data. Biometrics 55 (3), 688-698.

Hosseinpour, M., Prasetijo, J., Yahaya, A.S., Ghadiri, S.M.R., 2013. A comparative study of count models: Application to pedestrian-vehicle crashes along malaysia federal roads. Traffic Injury Prevention 14 (6), 630-638.

Iddi, S., 2013. A marginalized model for zero-inflated, overdispersed, and correlated count data. Electronic Journal of Applied Statistical Analysis 6 (2), 149-165.

Imprialou, M.-I.M., Quddus, M., Pitfield, D.E., Lord, D., 2016. Re-visiting crash–speed relationships: A new perspective in crash modelling. Accident Analysis & Prevention 86, 173-185.

Kassahun, W., Neyens, T., Molenberghs, G., Faes, C., Verbeke, G., 2014. Marginalized multilevel hurdle and zero-inflated models for overdispersed and correlated count data with excess zeros. Statistics in Medicine 33 (25), 4402-4419.

Kuo, P.-F., Lord, D., Walden, T.D., 2013. Using geographical information systems to organize police patrol routes effectively by grouping hotspots of crash and crime data. Journal of Transport Geography 30, 138-148.

Lee, J., Mannering, F., 2002. Impact of roadside features on the frequency and severity of run-off-roadway accidents: An empirical analysis. Accident Analysis & Prevention 34 (2), 149-161.

Lee, K., Kang, S., Liu, X., Seo, D., 2011. Likelihood-based approach for analysis of longitudinal nominal data using marginalized random effects models. Journal of Applied Statistics 38 (8), 1577-1590.

Long, D.L., Preisser, J.S., Herring, A.H., Golin, C.E., 2015. A marginalized zero-inflated poisson regression model with overall exposure effects. Statistics in Medicine 33 (29), 5151-5165.

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

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

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

Ma, L., Yan, X., Wei, C., Wang, J., 2016. Modeling the equivalent property damage only crash rate for road segments using the hurdle regression framework. Analytic Methods in Accident Research 11, 48-61.

Ma, J., Kockelman, K.M., Damien, P., 2008. A multivariate poisson-lognormal regression model for prediction of crash counts by severity, using bayesian methods. Accident Analysis & Prevention 40 (3), 964-975.

Malyshkina, N., Mannering, F., 2010. Zero-state Markov switching count-data models: An empirical assessment. Accident Analysis and Prevention 42(1), 122-130.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Analytic Methods in Accident Research 1 (Complete), 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.

Mo, B., Li, R., Zhan, X., 2017. Speed profile estimation using license plate recognition data. Transportation Research Part C Emerging Technologies, 82:358-378.

Organization, W.H., 2015. Global status report on road safety 2015. Injury Prevention 15 (4), 286-286.

Pavlou, M., Ambler, G., Seaman, S., Omar, R.Z., 2015a. A note on obtaining correct marginal predictions from a random intercepts model for binary outcomes. BMC Medical Research Methodology,15,1(2015-08-05) 15 (1), 59.

Qin, X., Ivan, J.N., Ravishanker, N., Liu, J., 2005. Hierarchical bayesian estimation of safety performance functions for two-lane highways using markov chain monte carlo modeling. Journal of Transportation Engineering 131 (5), 345-351.

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

Son, H.D., Kweon, Y.-J., Park, B.B., 2011. Development of crash prediction models with individual vehicular data. Transportation Research Part C: Emerging Technologies 19 (6), 1353-1363.

Su, L., Tom, B.D., Farewell, V.T., 2015. A likelihood-based two-part marginal model for longitudinal semicontinuous data. Statistical Methods in Medical Research 24 (2), 194.

Usman, T., Fu, L., Miranda-Moreno, L.F., 2010. Quantifying safety benefit of winter road maintenance: Accident frequency modeling. Accident Analysis & Prevention 42 (6), 1878-1887.

Vuong, Q. H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 57(2), 307-333.

Wan, F.W.Y., Lazim, M.A., Wah, Y.B., 2010. Evaluating spatial and temporal effects of accidents likelihood using random effects panel count model. International Conference on Science and Social Research. IEEE, 960-964.

Washington, S., Karlaftis, M., Mannering, F., 2010. Statistical and econometric methods for transportation data analysis, 2nd edition. Maritime Economics & Logistics 6 (2), 187-189.

Wolfinger, R.D., 2007. Fitting nonlinear mixed models with the new nlmixed procedure.

Yu, R., Wang, X., Abdel-Aty, M., 2017. A hybrid latent class analysis modeling approach to analyze urban expressway crash risk. Accident Analysis & Prevention 101, 37-43.

Highlights

- Utilized marginalized random effects hurdle negative binomial (MREHNB) model to address the issue of excessive zeros, unobserved heterogeneity, and subject-specific of interpretation for hourly-based crash frequency analyses.

- Compared MREHNB model with hurdle negative binomial (HNB) and random effects hurdle negative binomial (REHNB) model.
- Conducted an empirical analysis with license plate recognition (LPR) data.
- REHNB and MREHNB model showed substantial model goodness-of-fit improvement compared to the HNB model.
- Identified nonlinear relationship between volume and crash frequency and moderate level of volume held the highest crash occurrence likelihood.