Crash Data Modeling with a Generalized Estimator

Zhirui Ye*
Professor, Ph.D.
Jiangsu Key Laboratory of Urban ITS
Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies
Southeast University
2 Sipailou, Nanjing, Jiangsu 210096, China

Yueru Xu
Ph.D. Candidate
Jiangsu Key Laboratory of Urban ITS
Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies
Southeast University
2 Sipailou, Nanjing, Jiangsu 210096, China

Dominique Lord
Professor, Ph.D.
Zachry Department of Civil Engineering
Texas A&M University, 3136 TAMU
College Station, TX 77843-3136

Paper submitted for publication.

April 24th, 2018

ABSTRACT

The investigation of relationships between traffic crashes and relevant factors is important in traffic safety management. Various methods have been developed for modeling crash data. In real world scenarios, crash data often display the characteristics of over-dispersion. However, on occasions, some crash datasets have exhibited under-dispersion, especially in cases where the data are conditioned upon the mean. The commonly used models (such as the Poisson and the NB regression models) have associated limitations to cope with various degrees of dispersion. In light of this, a generalized event count (GEC) model, which can be generally used to handle over-, equi-, and under-dispersed data, is proposed in this study.

This model was first applied to case studies using data from Toronto, characterized by over-dispersion, and then to crash data from railway-highway crossings in Korea, characterized with under-dispersion. The results from the GEC model were compared with those from the Negative binomial and the hyper-Poisson models. The cases studies show that the proposed model provides good performance for crash data characterized with over- and under-dispersion. Moreover, the proposed model simplifies the modeling process and the prediction of crash data.

Key words: crash data analysis; generalized event count model; under-dispersed data.

INTRODUCTION

More than one million people are killed in traffic crashes every year around the world (WHO, 2013). Traffic crashes result in enormous losses to society and the economy. Several researchers have been seeking methods for better understanding contributing factors that influence or are associated with crashes and develop effective strategies to improve road safety.

The relationships between traffic crashes and relative factors have been investigated for more than three decades (Lord and Mannering, 2010). Various kinds of methodologies have been proposed over the years to improve on predicting the likelihood of crashes and determine the variables or factors that significantly influence the number of crashes and their severities.

It has been shown that crash data usually exhibit over-dispersion. Initially, the negative binomial (NB) regression model was proposed to handle such datasets (Miaou, 1994; Poch and Mannering, 1996). The NB model is derived by rewriting the Poisson parameter as $\lambda_i = \mathrm{EXP}(\beta \mathrm{X_i} + \varepsilon_i)$ in which $\mathrm{EXP}(\varepsilon_i)$ is a gamma-distributed error term with mean 1 and variance α. Given important limitations associated with the NB model, highway safety researchers have proposed new and innovative models, such as the random-effects (Hausman et al., 1984; Shankar et al., 1998) and its extension to random parameters models (Anastasopoulos and Mannering, 2009), bi-variate/multivariate models (Maher, 1990; Ma and Kockelman, 2006; Park and Lord, 2007; Barua et al., 2016), multiparameter models (Geedipally et al., 2012; Vangala et al., 2014), generalized additive models (Xie and Zhang, 2008), and semi-parametric models based on the Dirichlet process (Heydari et al., 2016b; Shirazi et al., 2016). These models can handle characteristics commonly found in crash data, such as excess zero responses and datasets with long tails among others. Readers are referred to Lord and Mannering (2010), Mannering and Bhat (2014), and Heydari et al. (2016a), who have provided a comprehensive review of existing methods with their advantages and disadvantages.

In addition to over-dispersion, some researchers have also encountered under-dispersion (Oh et al., 2006; Daniels et al., 2010; Lord et al., 2010). Although the models described above are able to capture or handle over-dispersion or unobserved heterogeneity, they cannot be used efficiently when the data are characterized by under-dispersion, either in the dataset itself or when the observations are conditioned upon the mean (Lord et al., 2010). To handle under-dispersion, Oh et al. (2006) first proposed the gamma model to analyze crash data exhibiting this unique characteristic. Although the gamma model can handle under-dispersion, the model suffers from an important drawback since past observations are assumed to directly influence future observations (e.g., a crash that occurred in one year is directly correlated to a crash that will occur the following or a future year) (Lord et al., 2010). Subsequently, Lord et al. (2007, 2010, 2015) have proposed the Conway-Maxwell-Poisson (COM-Poisson) generalized linear model for analyzing crash data. Recently, Huang (2017) proposed a re-parametrization of the COM-Poisson model, where the mean of the counts is modeled directly rather than

using the mode as an approximation of the mean value. The COM-Poisson model can handle both under- and over-dispersion, similar to the gamma model, but without the key limitation of the latter, although it may provide erroneous estimates for very small sample size and low sample mean values (Lord et al., 2010). Along the same line, Zou et al. (2013) examined the applicability of double Poisson (DP) generalized linear model for analyzing crash data and compared its performance with the COM-Poisson model. Khazraee et al. (2014) applied the hyper-Poisson (hP) generalized linear model to analyze under-dispersed crash data. It should be pointed out that the COM-Poisson and hP models both allow the dispersion of the distribution to be observation-specific and dependent on model covariates and both the DP and hP models offered similar statistical performance than those associated with the COM-Poisson. The differences were seen with the complexity for estimating the coefficients of the models.

The research documented in this paper therefore continues the work performed on the development of tools that would allow the analysis of both over- and under-dispersion. More specifically, the main goal is to apply the generalized event count (GEC) model developed by King (1989) for crash analysis and prediction. Similar to the COM-Poisson, DP and hP models introduced above, this model also handles over-, under- and equi-dispersion and has been shown to provide good statistical performances in other fields, such as the evaluation of congressional challenges of presidential votes and superpower conflicts (King, 1989). So far, this model has not been applied for analyzing crash data. Overall, the GEC model is easy to implement since the coefficients can be estimated using maximum likelihood estimation (MLE) and can handle over-, under- and equi-dispersion with good performance. The next section presents the GEC model for crash data analysis. Subsequently, case studies are presented; they were used to evaluate the performance of the proposed method by comparing the model with existing models, such as NB regression model or HP model. Finally, the findings and conclusions are summarized.

METHODOLOGY

This section first briefly introduces the Poisson model as background. It is followed by a more detailed description of the GEC model.

The Poisson Regression Model

The Poisson regression model is the basic model for analyzing count data. It aims at modeling a count (or crash) variable Y, which is assumed to follow a Poisson distribution with a parameter (or mean) λ (Lord and Mannering, 2010; Myers et al., 2012). The Poisson distribution usually implies that the probability of an event occurring at any instant is constant and independent of all previous events during the observation period (King, 1988). In highway safety, the probability that the number of crashes takes the value y_i on the *i*th entity can be expressed as Equation 1.

$$P(Y_i = y_i) = f(y_i|\lambda_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!},\ i = 1,2,\dots\dots n. \quad (1)$$

In the Poisson regression model, the mean can be written as $\lambda_i = \mathrm{EXP}(\beta X_i)$, where X_i is a vector of k explanatory variables and β is a $1\times k$ parameter vector that indicates the effect of the explanatory variables on the dependent variable. To estimate β, the method of maximum likelihood can be used. The likelihood function is presented in Equation 2.

$$L(\boldsymbol{\beta}|y) = \prod f(y_i|\lambda_i) = \prod \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \quad (2)$$

For a Poisson regression model, the variance of Y_i is equal to its expected value: $V(Y_i) = E(Y_i) = \lambda_i$. In practice, this model is not used frequently in safety research since the main assumption between the mean and variance is violated. This model is presented here since its characteristics are expanded in the next section.

The Generalized Event Count Model

In most cases, road crash data display the characteristic of over-dispersion and, on rare occasions, could exhibit under-dispersion. Considering all possible situations, the relationship between mean and variance is defined by $V(Y_i) = \lambda_i\sigma^2$ for $\lambda_i > 0$ and $\sigma^2 > 0$. σ^2 is called the dispersion parameter. If the crash variable follows a Poisson distribution, then $\sigma^2 = 1$ and $V(Y_i) = E(Y_i) = \lambda_i$; if $\sigma^2 > 1$, the data are over-dispersed; and if $0 < \sigma^2 < 1$, the data are regarded as under-dispersed. With the introduction of the parameter σ^2, the GEC model is developed and is able to model event counts with unknown degrees of dispersion. To construct this model, a GEC probability distribution with parameters λ_i and σ^2 is established. In this model, σ^2 can take on any value greater than zero. Special cases occur when the dispersion parameter falls into different ranges. When $0 < \sigma^2 < 1$, the GEC distribution can handle under-dispersed data. When $\sigma^2 = 1$, the model has the same probability function as the Poisson regression model; and when $\sigma^2 > 1$, it's probability function is similar to the NB regression model. This GEC's probability distribution offers smooth transitions between these scenarios. To derive the GEC's probability distribution, a concept taken from theoretical statistics called "bilinear recurrence relationship" was introduced (Katz, 1965). The relationship is shown in Equation 3.

$$\frac{f_k(y_i+1|\theta_i,y_i)}{f_k(y_i|\theta_i,y_i)} = \frac{\theta_i+\gamma_i y_i}{y_i+1} \quad \text{for } y_i=0,1,2,\dots \text{and } \theta_i + \gamma_i y_i \geq 0 \quad (3)$$

where θ_i and γ_i are ancillary parameters. In this case, Equation 3 should be re-parameterized in order to make the relationship suitable for the previous definitions.

Statistical analysis reveals that the expected value E (Y_i) and variance V (Y_i) of a random variable Y_i that adheres to the relationship in Equation 3 are as follows (Lee, 1986):

$$E(Y_i) = \lambda_i = \frac{\theta_i}{1-\gamma_i} \quad (4)$$

$$V(Y_i) = \lambda_i\sigma^2 = \frac{\theta_i}{(1-\gamma_i)^2} \quad (5)$$

Solving the above two equations, we then get:

$$\gamma_i = 1 - \frac{1}{\sigma^2}, \quad \theta_i = \frac{\lambda_i}{\sigma^2} \quad (6)$$

Then Equation 3 becomes:

$$f_{gec}(y_i|\lambda_i,\sigma^2) = \left(\frac{\lambda_i+(\sigma^2-1)(y_i-1)}{\sigma^2 y_i}\right) f_{gec}(y_i-1|\lambda_i,\sigma^2) \quad (7)$$

In Equation 7, f_{gec} represents the GEC distribution. The expected value and variance of the distribution is consistent with the previous definitions: E (Y_i) =λ_i and V (Y_i) =$\lambda_i\sigma^2$.

At this point, Equation 7 is still in the form of a recurrence relationship, so it should be transformed into more traditional probability distributions. To achieve this, the term $f_{gec}(y_i-1|\lambda_i,\sigma^2)$ must be recursively substituted by λ_i and σ^2. To better explain this process, Equation 8 shows a simple example where given $y_i = 2$ (King, 1989)

$$Pr(Y_i = 2|\lambda_i,\sigma^2) = f_{gec}(2|\lambda_i,\sigma^2) \quad (8)$$

$$= \left(\frac{\lambda_i + (\sigma^2-1)\times 1}{\sigma^2\times 2}\right) f_{gec}(1|\lambda_i,\sigma^2)$$

$$= \left(\frac{\lambda_i+(\sigma^2-1)\times 1}{\sigma^2\times 2}\right)\left(\frac{\lambda_i+(\sigma^2-1)\times 0}{\sigma^2\times 1}\right) f_{gec}(0|\lambda_i,\sigma^2)$$

In the example, the above equation still contains $f_{gec}(0|\lambda_i,\sigma^2)$, so further analysis is required before the distribution density function can be assigned numerical values. By adapting Equation 8 to a general situation, the GEC probability density function becomes (King, 1989):

$$f_{gec}(y_i|\lambda_i,\sigma^2) = f_{gec}(0|\lambda_i,\sigma^2)\prod_{j=1}^{y_i}\left(\frac{\lambda_i+(\sigma^2-1)(j-1)}{\sigma^2\times j}\right) \quad (9)$$

In theoretical statistics, $\sum_{m=0}^{\infty} f_{gec}(\mathrm{m}|\lambda_i,\sigma^2)=1$ is one of the basic axioms of probability, which means that the probability of all situations in period i is 1. Thus, we can solve the problem by using this axiom:

$$f_{gec}(0|\lambda_i,\sigma^2)=[1+\sum_{m=1}^{\infty}\prod_{j=1}^{m}\left(\frac{\lambda_i+(\sigma^2-1)\times j-1}{\sigma^2\times j}\right)]^{-1} \quad (10)$$

Then, using standard results on the convergence of infinite series leads to the final expression for the probability function $\mathrm{f_{gec}}(0|\lambda_i,\sigma^2)$ (King, 1989)

$$f_{gec}(0|\lambda_i,\sigma^2)=\begin{cases} e^{-\lambda_i} & for\ \sigma^2=1 \\ (\sigma^2)^{-\lambda_i/(\sigma^2-1)} & for\ \sigma^2>1 \\ (\sigma^2)^{-\lambda_i/(\sigma^2-1)}D_i^{-1} & for\ \sigma^2<1 \end{cases} \quad (11)$$

in which $D_i^{-1}=\sum_{m=0}^{\left[\frac{\lambda_i}{\sigma^2-1}\right)+1}\left(\frac{\Gamma\left(\frac{\lambda_i}{\sigma^2-1}+1\right)}{m!\Gamma\left(\frac{\lambda_i}{\sigma^2-1}-m+1\right)}(1-\sigma^2)^m(\sigma^2)^{\frac{\lambda_i}{\sigma^2-1}-m}\right)$.

Since the function is continuous in both σ^2 and λ_i, the piecewise function can be regarded as a whole and substituted into Equation 9.

Finally, the full probability distribution can be written by combining all the above equations, as shown in Equation 12. The probability distribution is continuous in σ^2 .

$$Pr(Y_i=y_i)=f_{gec}(y_i|\lambda_i,\sigma^2) \quad (12)$$

$$=\begin{cases} f_{gec}(0|\lambda_i,\sigma^2)\prod_{j=1}^{y_i}\left(\frac{\lambda_i+(\sigma^2-1)\times(j-1)}{\sigma^2\times j}\right) & for\ y_i=1,2,\ldots\ldots \\ e^{-\lambda_i} & for\ y_i=0\ and\ \sigma^2=1 \\ (\sigma^2)^{-\frac{\lambda_i}{(\sigma^2-1)}} & for\ y_i=0\ and\ \sigma^2>1 \\ (\sigma^2)^{-\frac{\lambda_i}{(\sigma^2-1)}}D_i^{-1} & for\ y_i=0\ and\ 0<\sigma^2<1 \\ & and\ y_i\le\left[\frac{\lambda_i}{\sigma^2-1}\right)+1 \\ 0 & otherwise \end{cases}$$

For the estimation of the GEC's probability distribution, a more general maximum likelihood estimator should be derived. The log-likelihood can be summarized as the following equation (King, 1989):

$$ln\,L(\beta,\sigma^2|y)=\sum_{i=1}^{n}\{C_i-y_i\,ln(\sigma^2)+\sum_{j=1}^{y_i}ln[exp(\beta X_i)+(\sigma^2-1)(j-1)]\} \quad (13)$$

where

$$C_i = \begin{cases} -exp(\beta X_i) & for\ \sigma^2 = 1 \\ -exp(\beta X_i)\, ln(\sigma^2)\, (\sigma^2 - 1)^{-1} & for\ \sigma^2 > 1 \\ -exp(\beta X_i)\, ln(\sigma^2)\, (\sigma^2 - 1)^{-1} - ln(D_i) & for\ 0 < \sigma^2 < 1 \end{cases}$$

The likelihood function is simultaneously maximized with respect to σ^2 and β. The virtue of this approach is that one do not need to choose among the three sets of assumptions about the process generating the counts ahead of time. Contagion, heterogeneity and other processes will produce an event count with a particular type of dispersion.

This model produces estimates of β and σ^2 in a single step simultaneously without specifying if the crash data are over-, under- or equi-dispersed. Since no additional parameters are introduced in the GEC distribution, this model actually reduced the chance for inconsistency with no need for additional assumptions. With the GEC model, the safety analysist does not need to take into account the choice of statistical models, since the value of σ^2 is estimated rather than assumed.

As we all know, highway crashes are complicated events that involve complex interactions between human, vehicle, roadway condition, and traffic-related factors. It usually not possible to have access to all of the data that could potentially determine the likelihood of a crash. In this context, these unavailable factors can lead to unobserved heterogeneity, which may cause biased estimates and inconsistent predictions. This problem can be minimized by various statistical approaches, such as using random parameters, latent-class models or latent-class models with random parameters within each class (Mannering et al., 2016) and multiparameter models with and without random parameters (Lord and Geedipally, 2018; Shaon et al., 2018). Some of these statistical approaches could eventually be incorporated with the GEC model to help minimize the unobserved heterogeneity in future research.

CASE STUDIES

To verify and evaluate the performance of the GEC model, three datasets with varying degrees of dispersion were used for crash modeling. The most commonly used NB regression model and an hP model were also applied to the datasets for comparison purposes.

Data Collection

Data collected in Toronto and Korea were used in this study. The Toronto crash data was collected from 868 four-legged signalized intersections in Toronto and the Korea dataset contains crash data collected at 162 rail-highway crossings (RHXs) in Korea.

The statistics of the datasets are summarized in Table 1. As will be shown in the following data modeling, the Toronto data set was found to be of good quality by several research studies (Lord, 2000; Miaou and Lord, 2003) and are characterized by

over-dispersion. Although the raw Korea dataset show signs of slight over-dispersion or even equi-dispersion (mean=0.33, variance=0.36), several previous research studies--Oh et al., 2006; Lord et al., 2010; Khazraee et al., 2015--have noted that under-dispersion was observed when conditioned upon the mean). These datasets provide the opportunity to comprehensively evaluate the performance of the GEC method.

Table 1 Statistics of Variables for Selected Datasets

Route.	Variables	Max.	Min.	Mean	SD	Frequency
Toronto Data	Major approach AADT (veh./day)	72178	5469	28044.81	10660.39	868
	Minor approach AADT (veh./day)	42644	53	11010.18	8599.40	868
	Crashes (crash/year)	54	0	11.56	10.02	868
Korean Data	Highway ADT (veh./day)	61200	10	4617	10391.57	162
	Average daily railway traffic	203	32	70.29	37.34	162
	Train detector distance	1329	0	824.5	328.38	162
	Time duration btw activation of warning signals and gates	232	0	25.46	25.71	162
	Presence of commercial area	1(yes)	-	-	-	149(91.98%)
		0(no)	-	-	-	13(8.02%)
	Presence of a speed hump	1(yes)	-	-	-	134(82.72%)
		0(no)	-	-	-	28(17.28%)
	Presence of a track circuit controller	1(yes)	-	-	-	113(69.75%)
		0(no)	-	-	-	49(30.25%)
	Presence of a guide	1(yes)	-	-	-	126(77.78%)
		0(no)	-	-	-	36(22.22%)
	Crashes (crash/year)	3	0	0.33	0.60	162

In order to evaluate the models, standard errors and the P-value of t-test were chosen as measures of effectiveness (MOEs). The standard error represents the standard deviation of the sampling distribution of a statistic. The t-test is used to evaluate if the associated explanatory variables have a significant influence on the dependent variables. In addition, the AIC and MPB are also used to assess goodness of fit of the developed models.

Toronto Data Analysis

The GEC model was first applied to the Toronto dataset. Since the Toronto crash data were shown to be over-dispersed, the hP and Negative Binomial models were also estimated for the purpose of the comparison analysis. The results of these models were estimated in R using maximum likelihood method (R Core Team, 2013). The values in parentheses represent the standard error of the estimate values. Table 2 shows the modeling results for these three different models.

Table 2 Modeling Results for the NB, HP and GEC model

Estimate.	Negative Binomial Model	Hyper-Poisson Model	Generalized Event Count Model
Ln($\boldsymbol{\beta_0}$)	-10.25(0.4626)	-10.22(0.4464)	-10.24(0.4601)
$\boldsymbol{\beta_1}$	0.6207(0.04652)	0.6076(0.0462)	0.6172(0.0464)
$\boldsymbol{\beta_2}$	0.6853(0.02152)	0.6981(0.02205)	0.6897(0.02173)
$\boldsymbol{\lambda}$	-	25.7492	-
$\boldsymbol{\alpha}$	0.1398(0.0122)	-	-
$\boldsymbol{\delta^2}$	-	-	1.60
AIC[a]	5077.3	5157.3	5089.4
MPB[b]	-0.045	0.033	-0.041
MAD[c]	4.142	4.142	4.142
MSPE[d]	32.70	32.62	32.66

a Akaike information criterion.
b Mean prediction bias.
c Mean absolute deviance.
d Mean squared predictive error.
– = not applicable.

As Table 2 shows, the parameter δ^2 in the GEC model is 1.60, which is larger than 1. It indicates that the Toronto data and modeling results are over-dispersed. The results are consistent with the conclusion of previous research studies that used the same dataset.

The MPB, MAD and MSPE of the models considered for the Toronto data are relatively similar. The MAD and MSPE measures of fit vary only slightly between these three models. This is mainly attributed to the similar estimates of relevant parameters. Since the MPB, MAD and MSPE are only dependent on the mean function and not on the dispersion parameter. The similar coefficients can result in similar values for these measures of fit.

On the other hand, the AIC measure depends on the model likelihood function. Therefore, it is influenced by the dispersion parameter, which is presents for these three kinds of models. The models with similar mean function parameters may have significantly different AICs.

From Table 2, we can find that the AIC of the hP model is a little higher than other two models. The AIC of GEC model is 5091.4, which is very close to the value of NB model. Therefore, we can conclude that the GEC model and NB model could fit the over-dispersed data from Toronto a little better than the hP model. As a rule of thumb, when the change in AIC is less than 5, the difference is usually deemed to be insignificant (Spiegelhalter et al., 2002). In addition, the AIC of hP model is 5157.3 in this case. This value is larger than that of the GEC model. Therefore, the performance of GEC model can be regarded as well as the NB model and better than the hP model.

Korean Data Analysis

In this section, the Korea RHXs data was used to evaluate the performance of GEC model on data with under-dispersion. The NB model was previously applied to

this dataset and it was shown to be inappropriate. Therefore, the GEC model in this section was compared to the gamma and hP models, which were found to be successful in handling under-dispersion (note: the gamma model was only included since it was used in the original study by Oh et al., 2006). Table 3 shows the results for the three models. All of the three models were estimated using the MLE method.

Table 3 Modeling Results for the Gamma, HP and GEC model

Variables	Gamma Model	Hyper-Poisson Model	Generalized Event Count Model
Constant	-3.438(1.008)	-5.513(0.756)	-5.782(0.942)
Ln(ADT)	0.230(0.076)	0.472(0.057)	0.451(0.071)
Average daily railway traffic	0.004(0.0024)	-	-
Presence of commercial area	0.651(0.287)	0.965(0.370)	1.084(0.393)
Train detector distance	0.001(0.0004)	0.0017(0.0006)	0.0019(0.0007)
Time duration between the activation of warning signals and gates	0.004(0.002)	-	-
Presence of a track circuit controller	-	-0.924(0.303)	-1.036(0.403)
Presence of a guide	-	-0.665(0.294)	-0.796(0.477)
Presence of a speed hump	-1.58(0.859)	-1.080(0.441)	-1.147(0.519)
Shape parameter	2.062(0.758)	-	-
Dispersion parameter	-	0.298(0.189)	-
δ^2	-	-	0.351(0.301)
AIC	211.38	209.54	210.05
MPB	0.179	0.004	-0.09
MAD	0.459	0.357	0.351
MSPE	0.308	0.246	0.239

– = not applicable.

The Poisson, gamma probability, and hP models for the Korea data were originally developed using 31 candidate explanatory variables by previous researchers (Oh et al., 2006; Lord et al., 2010). According to the results, eight of the variables were found significant in at least one model. Therefore, the eight variables were applied as the explanatory variables used in the GEC model. The explanation of these eight variables were listed in Table 3.

From Table 3, we can find that the dispersion parameter of hP model is 0.298 and δ^2 of the GEC model is 0.351, which both indicates that the Korea data are under-dispersed (when conditioned upon the mean). In addition, the significant variables of hP model are the same as the GEC model, but not the same as for the gamma model. The difference between some of the significant variables in one model and not in the other is probably due to the fact that the main assumption of the gamma model affected the estimation of the coefficients.

Similar to the Toronto data application, the MPB, MAD and MSPE were used to compare the performance of the three models. The MPB is very close to 0, which means the model neither over-predicts nor under-predicts crashes. The MPB and

MSPE are as low as hP model, and a little better than those of the gamma model. Therefore, the GEC model performs as well as hP model. By comparing the AIC values, we can also conclude that the GEC model performs as well as the hP model (we have already ruled out the gamma model).

CONCLUSIONS

The investigation of relationships between traffic crashes and relative factors is important in traffic safety management. Various methods have been developed for the modeling of crash data. In real world cases, crash data usually display the characteristics of over-dispersion, but on rare occasions, can display under- dispersion. The commonly used models (such as the Poisson and the NB regression models) have associated limitations to account for various degrees of dispersion. In light of this, a GEC model developed by Gary King (1989) was applied in this study. This model can be generally used without considering the degrees of dispersion to simplify the process of crash data analysis.

Two case studies were carried out to evaluate the performance of the proposed model. Data collected from Toronto and Korea were used in this study. The Toronto intersection data are characterized by over-dispersion while the Korea dataset was shown to exhibit under-dispersion (when conditioned upon the mean). The NB and hP models were applied to Toronto data while the gamma and Hyper-Poisson models were used for Korea data.

The modeling results showed that when handling over-dispersed data, the GEC model performs as well as the NB model and better than the hP model (at least for this dataset). The authors admit that the variance-mean relationship structure of the GEC model is better than the hP model when the variance increases very rapidly with the increase in the mean.

When applying the Korean dataset, the GEC model has an equally good performance compared to the hP model and better than the gamma model (putting aside the direct correlation between observations through time). Therefore, the GEC model can handle under-dispersed data well while the NB model can result in unstable and unreliable parameter estimates in this situation. The case studies collectively showed that the GEC model can be used for modeling crash data without considering the degrees of dispersion.

In addition, the GEC model is found to be easier to code for the case studies described in this paper. The analyses are easier to implement based on the MLE. Almost all statistical software can be used for estimating this model.

To evaluate the performance of the GEC model in a more comprehensive way, simulations of crash datasets can be used in the future to generate more samples representing various degrees of dispersion as a function of the sample size and sample mean values. Furthermore, the basic characteristics of the GEC model should be expanded to analyze multivariate data; incorporate latent class and random parameters framework; and, to account for spatial heterogeneity.

ACKNOWLEDGEMENTS

This study was sponsored by the National Science Foundation of China (No. 51308114) and the Project Sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry.

REFERENCES

Abdelwahab, H., & Abdel-Aty, M. (2002). Artificial neural networks and logit models for traffic safety analysis of toll plazas. *Transportation Research Record,* (1784), 115-125.

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

Barua, S., El-Basyouny, K., Islam, M. D. (2016) Multivariate random parameters collision count data models with spatial heterogeneity. Analytic Methods in Accident Research, 9, 1–15

Bijleveld, F. D. (2005). The covariance between the number of accidents and the number of victims in multivariate analysis of accident related outcomes. *Accident Analysis & Prevention, 37*(4), 591-600.

Chang, L. Y. (2005). Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. *Safety Science, 43*(8), 541-557.

Francis, R.A., S.R. Geedipally, S.D. Guikema, S.S. Dhavala, D. Lord, and S. Larocca (2011). Characterizing the performance of the Conway-Maxwell-Poisson generalized linear model. *Risk Analysis*, 32(1), 167-183.

Daniels, S., Brijs, T., Nuyts, E., & Wets, G. (2010). Explaining variation in safety performance of roundabouts. *Accident Analysis & Prevention*, 42(2), 393-402.

Gouriéroux, C., Monfort, A., & Trognon, A. (1984). Pseudomaximum likelihood methods: theory. *Econometrica, 52*(3), 681-700.

Guo, G. (1996). Negative multinomial regression models for clustered event counts. *Sociological Methodology, 26*(26), 113-132.

Hausman, J., Hall, B. H., & Griliches, Z. (1984). Econometric models for count data with an application to the patents-r & d relationship, *Econometrica, 52*(4), 909-938.

Heydari, S., Fu, L., Jopseph, L., Miranda-Moreno, L.F. (2016a) Bayesian nonparametric modeling in transportation safety studies: Applications in univariate and multivariate settings. *Analytic Methods in Accident Research*, 12, 18-34.

Heydari, S., L. Fu, D. Lord, and B.K. Mallick (2016b) Multilevel Dirichlet process mixture analysis of railway grade crossing crash data. *Analytic Methods in Accident Research,* 9, 27-43.

Huang, A., 2017. Mean-parametrized Conway–Maxwell–Poisson regression models for dispersed counts. *Statistical Modelling*, 17(6), 1–22.

Johansson, P. (1996). Speed limitation and motorway casualties: a time series count data regression approach. *Accident Analysis & Prevention,28*(1), 73-87.

Katz, L. (1965). Unified treatment of a broad class of discrete probability distributions. *Classical and Contagious Discrete Distributions*, 1, 175-182.

Khazraee, S. H., Sáez‐Castillo, A. J., Geedipally, S. R., & Lord, D. (2015). Application of the hyper‐Poisson generalized linear model for analyzing motor vehicle crashes. *Risk Analysis*, 35(5), 919-930.

King, G. (1988). Statistical models for political science event counts: bias in conventional procedures and evidence for the exponential Poisson regression model. *American Journal of Political Science, 32*(3), 838-863.

King, G. (1989). Variance specification in event count models: From restrictive assumptions to a generalized estimator. *American Journal of Political Science*, 762-784.

Knuiman, M. W., Council, F. M., & Reinfurt, D. W. (1993). *Association of median width and highway accident rates*.

Lee, L. F. (1986). Specification test for Poisson regression models. *International Economic Review, 27*(3), 689-706.

Liang, F. (2005). Bayesian neural networks for nonlinear time series forecasting. *Statistics and Computing, 15*(1), 13-29.

Lord, D. (2000). The prediction of accidents on digital networks: characteristics and issues related to the application of accident prediction models. Ph.D. Dissertation, *Department of Civil Engineering, University of Toronto, Toronto, Ont.*

Lord, D., & Persaud, B. (2000). Accident prediction models with and without trend: application of the generalized estimating equations procedure. *Biochemistry International, 1717*(1), 102-108.

Lord, D., & Geedipally, S.R. (2018) Safety Prediction with Datasets Characterised with Excess Zero Responses and Long Tails, in Dominique Lord, Simon Washington (ed.) Safe Mobility: Challenges, Methodology and Solutions (Transport and Sustainability, Volume 11) Emerald Publishing Limited, 297 – 323.

Lord, D., Geedipally, S. R., & Guikema, S. D., (2010). Extension of the application of Conway-Maxwell-Poisson models: analyzing traffic crash data exhibiting underdispersion. *Risk Analysis, 30*(8), 1268–1276.

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

Lord, D., Guikema, S. D., & Geedipally, S. R. (20082). Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes. *Accident Analysis & Prevention, 40*(3), 1123-1134.

Ma, J., Kockelman, K. M., Boothe, C., & Luce Associate. (2006). Bayesian multivariate Poisson regression for models of injury count, by severity. *Transportation Research Record: Journal of the Transportation Research Board, 1950*(1).

Maher, M. J. (1990). A bivariate negative binomial model to explain traffic accident migration. *Accident Analysis & Prevention, 22*(5), 487-98.

Maher, M. J. (1990). A bivariate negative binomial model to explain traffic accident migration. *Accident Analysis & Prevention, 22*(5), 487-98.

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

Mannering, F. L., Shankar, V., & Bhat, C. R. (2016). Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic methods in accident research*, 11, 1-16.

Miaou, S. P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. *Accident Analysis & Prevention, 26*(4), 471-482.

Miaou, S. P., & Lord, D. (2003). Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. *Transportation Research Record: Journal of the Transportation Research Board*, (1840), 31-40.

Milton, J. C., Shankar, V. N., & Mannering, F. L. (2008). Highway accident severities and the mixed logit model: an exploratory empirical analysis. *Accident Analysis & Prevention, 40*(1), 260.

Myers, R. H., Montgomery, D. C., Vining, G. G., & Robinson, T. J. (2012). Generalized Linear Models: with Applications in Engineering and the Sciences, 2nd Edition. *Physics of Electronic Materials*.

Oh, J., Washington, S. P., & Nam, D. (2006). Accident prediction model for railway-highway interfaces. *Accident Analysis & Prevention, 38*(2), 346.

Organization, W. H. (2013). Global status report on road safety - 2013: supporting a decade of action. *Injury Prevention, 15*(4), 286-286.

Park, E.S., and D. Lord (2007) Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity. *Transportation Research Record 2019*, 1-6.

Poch, M., & Mannering, F. L. (1996). Negative binomial analysis of intersection-accident frequencies. *Journal of Transportation Engineering, 122*(2), 105-113.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (*http://www.R-project.org/*)

Riviere, C., Lauret, P., Ramsamy, J. F. M., & Page, Y. (2006). A Bayesian neural network approach to estimating the energy equivalent speed. *Accident Analysis & Prevention, 38*(2), 248-59.

Shaon, M.R.R., Qin, X., Shirazi, A., Lord, D., & Geedipally, S.R. (2018) Development of a Random Parameters Negative Binomial-Lindley Generalized Linear Model to analyze Over-Dispersed Data. *Analytic Methods in Accident Research*, in press. (https://doi.org/10.1016/j.amar.2018.04.002)

Shankar, V., Albin, R., Milton, J., & Mannering, F. (1998). Evaluating median crossover likelihoods with clustered accident counts: an empirical inquiry using the random effects negative binomial model. *Transportation Research Record, 1635*(1), 44-48.

Shirazi, M., Lord, D., Dhavala, S. S., Geedipally, S. R. (2016). A Semiparametric Negative Binomial Generalized Linear Model for Modeling Over Dispersed Count Data with a Heavy Tail: Characteristics and Applications to Crash Data. *Accident Analysis & Prevention,* 91, 10-18.

Song, J. J., Ghosh, M., Miaou, S., & Mallick, B. (2006). Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis, 97*(1), 246-273.

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

Tsai, C. P., & Lee, T. L. (1999). Back-propagation neural network in tidal-level forecasting. *Journal of Waterway Port Coastal & Ocean Engineering, 125*(4), 689.

Xie, Y., Lord, D., & Zhang, Y. (2007). Predicting motor vehicle collisions using Bayesian neural network models: an empirical analysis. *Accident Analysis & Prevention, 39*(5), 922-933.

Xie, Y., & Zhang, Y. (2008). Crash frequency analysis with generalized additive models. *Transportation Research Record Journal of the Transportation Research Board,* (2061), 39-45.

Zou, Y., Geedipally, S. R., & Lord, D. (2013). Evaluating the double Poisson generalized linear model. *Accident; analysis and prevention, 59*(5), 497.

Zou, Y., Wu, L., & Lord, D. (2015). Modeling over-dispersed crash data with a long tail: examining the accuracy of the dispersion parameter in negative binomial models. *Analytic Methods in Accident Research*, 5, 1-16.