Applying the Generalized Waring model for investigating sources of variance in motor vehicle crash analysis

Yichuan Peng*
Research Engineer
Department of Civil Engineering
University of Central Florida
4000 Central Florida Blvd, Orlando, FL 32816
Tel: (979) 618-8603
Email: yichuanpeng1982@hotmail.com

Dominique Lord
Associate Professor and Zachry Development Professor I
Texas A&M University, 3136 TAMU
College Station, TX 77843-3136
Phone: 979/458-3949, fax: 979/845-6481
E-mail: d-lord@tamu.edu

Yajie Zou
Ph.D
Zachry Department of Civil Engineering
Texas A&M University
3136 TAMU, College Station, TX 77843-3136
Tel: (979) 595-5985
Email: yajiezou@tamu.edu

July 25th, 2014

*Corresponding author
Abstract

As one of the major analysis methods, statistical models play an important role in traffic safety analysis. They can be used for a wide variety of purposes, including establishing relationships between variables and understanding the characteristics of a system. The purpose of this paper is to document a new type of model that can help with the latter. This model is based on the Generalized Waring (GW) distribution. The GW model yields more information about the sources of the variance observed in datasets than other traditional models, such as the negative binomial (NB) model. In this regards, the GW model can separate the observed variability into three parts: 1) the randomness, which explains the model’s uncertainty; 2) the proneness, which refers to the internal differences between entities or observations; and, 3) the liability, which is defined as the variance caused by other external factors that are difficult to be identified and have not been included as explanatory variables in the model. The study analyses were accomplished using two observed datasets to explore potential sources of variation. The results show that the GW model can provide meaningful information about sources of variance in crash data and also performs better than the NB model.

Key words: Generalized Waring model, negative binomial model, over dispersion, randomness, liability, proneness, crash modeling
1. Introduction

As one of the major safety analysis methods, crash modelling play an important role in traffic safety analysis (Geedipally et al., 2010). They can be used to explain random variations of crashes across sites based on the available information, such as traffic flow and other roadway geometric variables. Despite the introduction of many new models over the last decade or so, very few models have been developed or proposed for investigating the nature of the over-dispersion commonly observed in crash data. So far, models with a varying dispersion parameter (Hauer, 2001; Miaou and Lord, 2003; Park and Lord, 2008; Connors et al., 2013), the latent class/finite mixture/Markov switching models (Park and lord, 2009; Malyskhkina et al., 2009; Eluru et al., 2012; Zou et al., 2013; Yasmin et al., 2013; Mannering and Bhat, 2014) and random parameter models (Anastasopoulos and Mannering, 2009; Shamsunnahar et al., 2013; Xiong and Mannering, 2013; Chen and Tarko, 2014) have been used for investigating over-dispersion. In addition to the count models, Castro et al. (Castro et al., 2012; Castro et al., 2013) proposed a spatial generalized ordered-response model which is one of crash severity models to predict crash frequency at urban intersections to examine highway crash injury severity. Understanding the sources of heterogeneity could help in better strategizing the implementation of treatments for improving safety by focusing efforts on the more important variables that are the major sources of heterogeneity (observations that are far away from the mean).

It has been shown that the variability found in crash data usually result in over-dispersion. The heterogeneity can be explained by various factors, as discussed by Gourvieroux and Visser (1986), Cameron and Trivedi (1998), Poormeta (1999), and Lord et al (2005). More specifically, the cause of the unobserved heterogeneity can be divided further into two components according to Irwin (1968): internal differences across highway entities (i.e., road segments, intersections, etc.) and external factors that have not been incorporated as explanatory variables in the model because they are difficult to collect or identify. In the negative binomial (NB) model, the most popular model used in highway safety, both sources of variance are considered together using a single gamma distribution; the same constraint applies for other mixed-Poisson models, such as the Poisson-lognormal (Miaou et al., 2003; Lord and Miranda-Moreno, 2008) or the Poisson-Weibull (Maher and Mountain, 2009; Cheng et al., 2013; Connors et al., 2013). Given the limitations of the NB model to provide information about sources of over-dispersion, there is a need to evaluate whether alternative count data models could be used for modeling crash data and exploring sources of variance for crash data analysis.

This research aims to introduce the Generalized Waring (GW) model and evaluate its performance for analyzing over-dispersed crash data, especially related to the characteristics of the variance. The GW model was first proposed by Irwin to theoretically analyze accidents (Irwin, 1968; Xekalaki, 1983) and falls under other recently introduced three-parameter models, such as the Negative Binomial-Lindley (NB-L), Negative Binomial-Generalized Exponential (NB-GE) and the Sichel (SI) models (see Lord and Geedipally, 2014). The main advantage of this model over the NB is that the former provides more specific information about the sources of variance. It can be used to further distinguish the observed over-dispersion that is caused by internal factors inherent to each road segment or intersection
from those caused by external factors that are difficult to be observed or measured, and have not been included in the model. To accomplish the objective of the study, the GW model is examined using two empirical crash datasets to explore potential sources of variation.

2. Methodology

According to the classic accident theory (Greenwood and Yule, 1920; Newbold, 1925; Newbold, 1927), it was assumed that the overall population of highway entities is all subjected to the same external factors, but has different levels of proneness to crashes. Accordingly, the traditional NB model was developed based on the assumption that the mean number of crashes $\lambda$ incurred under given exposure conditions is assumed to have a continuous distribution, and the distribution of crashes among locations with the same levels of proneness is assumed to be Poisson distributed (Irwin, 1968).

However, it is very difficult to truly assume that all the entities will be exposed to exactly the same external risk of accident. Differences caused by external factors from one site to the next are known as differences in accident liability, which is distinguished from internal differences known as differences in proneness. The variance caused by these external factors which includes driver behaviors and distractions in addition to the non-geometric and environmental factors, such as weather and lighting conditions while proneness are associated directly with the internal characteristics of each site which were also difficult to be observed, that is, with their internal probability to cause accident. There are several internal characteristics of road segments related to the proneness. For example, the road surface friction, the pavement condition of road segments (rutting, cracks, etc.), the presence and maintenance condition of rumble strips or pavement markings, location, types and preservation of roadside devices, and so on.

In practice, the effects of proneness and liability are considered together when the NB is used. This combination was originally referred to as "susceptibility" by Newbold (1925; 1927). After that, Irwin (1968) proposed the univariate Generalized Waring distribution (UGWD) in a proper way that the gamma distribution accounts for liability and assumed a beta distribution to accommodate proneness. In this research, the GW GLM model which was created based on UGWD will be applied for traffic crash analysis.

The GW generalized linear model (GLM) model was first developed and introduced by Rodriguez-Avi et al (2009). The dependent variable for the GW model applied in this paper is the mean number of crashes for each highway entity. The model specification assumes that the proneness is independent and its beta distribution is the same for all levels of covariates. Based on this assumption, the GW GLM model was developed using the following three-level hierarchical framework (Rodríguez-Avi et al., 2009):

1. $(Y | X) \sim \text{Poisson} \left( \lambda_x \right)$ \hspace{1cm} (1)
2. $\lambda_x \sim \text{Gamma} \left( \alpha_x, \varphi \right)$ \hspace{1cm} (2)
3. $\varphi \sim \text{Beta} \left( \rho, k \right)$

It is noted that the beta distribution is not a standard beta distribution but a beta prime distribution.

Therefore, the probability density function (PDF) of $Y | X$ c:

$$f(Y | x) = \frac{\Gamma(\alpha_s + \rho)\Gamma(k + \rho)}{\Gamma(\alpha_s + k + \rho)\Gamma(\rho)} \frac{(\alpha_s + k + \rho)_y (k + \rho)_y}{\alpha_s + k + \rho + \rho} \frac{1}{y!}, y = 0, 1, \ldots, n$$

(4)

where $n$ is the sample size.

In addition, the equation of log-linearity for the mean number of crashes is shown in the following equation which is created using the traditional functional forms used for modeling crash data:

$$E(Y|x) = \mu_x = f(X;\beta) = e^{\beta_0 + \mathbf{x}^T\beta} \quad \text{with} \quad \beta' = (\beta_1, \ldots, \beta_p)$$

(5)

$f(.)$ is a function of the site specific covariates ($X$) and $p$ is the number of parameters in the model.

According to the above hierarchical framework, the variance in this GW model can be calculated into three parts (Rodríguez-Avi et al., 2009; Irwin 1968):

$$\sigma_a^2 = \frac{\alpha_s k}{(\rho - 1)} = E(Y | x) = \mu_x$$

(6a)

$$\sigma_b^2 = \frac{\alpha_s k(k + 1)}{(\rho - 1)(\rho - 2)} = \frac{k + 1}{\rho - 2} \mu_x$$

(6b)

$$\sigma_c^2 = \frac{\alpha_s^2 k(\rho + k - 1)}{(\rho - 1)^2(\rho - 2)} = \frac{k + \rho - 1}{(\rho - 2)k} \mu_x^2$$

(6c)

The first part of the variance (Eq. 6a) represents that the variability that comes from the randomness associated with the assumed Poisson distribution or model. The other two parts are the liability (Eq. 6b) and proneness (Eq. 6c), respectively. The relationships between Eqs. 1 to 3 and Eq. 6 are explained in great details on pages 207 and 208 of Irwin (1968).

The estimation of the coefficients for the GW model can be accomplished using the maximum likelihood (ML) method. In order to use the ML for the GW model, the first step is to specify the joint density function. The probability density function of the GW model was defined in Equation (4) above. Therefore, the joint density function of GW model or the likelihood function can be defined as follows (Rodríguez-Avi et al., 2007):
\[ \Pi_{i=1}^{n} f(Y_i | k, \rho, \alpha_x) = \frac{\Gamma(\alpha_x + \rho) \Gamma(k + \rho) \Gamma(\alpha_x + k + \rho) \Gamma(\alpha_x + k + \rho)}{\Gamma(\alpha_x + k + \rho) \Gamma(\rho) (\alpha_x + k + \rho)^y y!} \]  

(7)

where \( n \) is the sample size.

The logarithm of this equation is shown in the following equation:

\[ \ln \Pi_{i=1}^{n} f(X_i | k, \rho, \alpha_x) = \sum_{i=1}^{n} \left[ \ln \Gamma(\alpha_x + \rho) + \ln \Gamma(k + \rho) + \ln(\alpha_x) + \ln(k) - \ln(\alpha_x + k + \rho) - \ln(\rho) - \ln(\alpha_x + k + \rho) - \ln x_i! \right] \]  

(8)

Once the log-likelihood was defined, several functions in R (Rodríguez-Avi et al., 2009) were used for the parameter estimation of the GW GLM. These functions were created using three different kinds of algorithms: the Nelder-Mead, Quasi-Newton, and conjugate-gradient algorithms (R Development Core Team, 2007).

The Nelder-Mead method is a commonly used nonlinear optimization technique, and the quasi-Newton methods are algorithms based on the Newton's method for finding the local maxima and minima of functions (John and Kurtis, 2004). The quasi-Newton and conjugate gradient methods are the other two algorithms that can be used for finding the numerical solution of particular linear equations by applying an iterative method and can be applied to equations too complex to be handled by direct methods. Several initial values were used for the GW model in order to get a global optimum solution for the parameter estimation (Rodríguez-Avi et al., 2009).

Finally, it is noted that the GW model converges to the NB model, as illustrated in the following equation (Rodriguez-Avi et al., 2009):

when \( k, \rho \to \infty \), which means the parameter \( \varphi \) is considered as constant value so that the effects of proneness and liability are considered together then

\[ f(y \mid x) \propto \frac{(k)^y \theta_x^{y \rho^x + o(\rho^{(r-1)})}}{y!} \frac{(\theta_x (\rho - 1))^y}{(k + (1 + \theta_x)\rho - \theta_x)} \]  

\[ = \frac{(k)^y \theta_x^{y \rho^x + o(\rho^{(r-1)})}}{y! (1 + \theta_x)^y \rho^x + o(\rho^{(r-1)})} \rightarrow \frac{(k)^y \theta_x^{(\rho^x)}}{y! (1 + \theta_x)^y} \]  

(9)

The equation above is in fact the NB II model (as defined in Cameron and Trivedi, 1998), which is the most commonly used model in traffic safety (Lord and Mannering, 2010; Mannering and Bhat, 2014). In sum, it can be concluded that the NB model is nested in the
GW GLM. Curious readers are referred to Rodríguez-Avi et al (2007; 2009) for additional information.

It can be seen that the main advantage of this model over the NB model is that it provides more specific information about the source of variance. It can be used to further distinguish the observed over-dispersion that is caused by internal factors inherent to each road segment or intersection from those caused by external factors that are difficult to be observed includes driver behaviors and distractions in addition to the non-geometric and environmental factors.

3. Empirical Crash Data Analysis

This section described the application of the GW GLM to two observed crash datasets for the purpose of comparing the performance of GW GLM with the most commonly used NB model.

3.1 Description

As discussed above, two empirical datasets were analyzed to compare the performance of the GW and NB models and investigate the sources of over-dispersion for crashes occurring on highway segments.

The first dataset was assembled as a part of a National Cooperative Highway Research Program (NCHRP) research project (Lord and Park, 2008). This dataset contained 1,499 rural undivided 4-lane rural highway segments located in Texas. The data spanned 5 years. The length of the segments ranged from 0.10 to 6.28 miles, with an overall mean equal to 0.55 mile. The mean and the variance of the crashes are 2.84 and 32.4, respectively. The summary statistics for the variables contained in the dataset are presented in Table 1.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean(SD)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of crashes</td>
<td>0</td>
<td>97</td>
<td>2.84(5.69)</td>
<td>4253</td>
</tr>
<tr>
<td>Annual Average daily traffic</td>
<td>42</td>
<td>24,800</td>
<td>6,613(4,010)</td>
<td>-</td>
</tr>
<tr>
<td>(AADT†)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lane width (LW) (feet)</td>
<td>9.75</td>
<td>16.5</td>
<td>12.57(1.59)</td>
<td>-</td>
</tr>
<tr>
<td>Total Shoulder Width (SW) (feet)</td>
<td>0</td>
<td>40</td>
<td>9.96(8.02)</td>
<td>-</td>
</tr>
<tr>
<td>Segment Length (miles)</td>
<td>0.1</td>
<td>6.28</td>
<td>0.55(0.67)</td>
<td>830.49</td>
</tr>
</tbody>
</table>

† Annual Average Daily Traffic
The second dataset was collected from 1995 to 1999 and includes traffic and other covariates for interstate highways in Indiana (see Table 2). The dataset were collected from the Indiana Department of Transportation and the data were divided into homogeneous roadway segments (defined by roadway geometrics and pavement type). The segment-defining information included shoulder characteristics (inside and outside shoulder presence and width, and rumble strips), pavement characteristics (pavement type), median characteristics (median width, type, condition, barrier presence and location), number of lanes and speed limit. A total of 338 roadway segments were defined and the number of police-reported vehicle accidents occurring on each segment over the 5-year period was obtained from the Indiana State Patrol accident-data files. The Indiana dataset contains more explanatory variables. This dataset has been analyzed by other researchers and confirmed to be of good quality (Washington et al., 2011). It should be noted that 120 of the 338 segments did not have any reported crashes over the five-year study period.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean(SD)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of crashes (5 years)</td>
<td>0</td>
<td>329</td>
<td>16.97 (36.3)</td>
<td>5737</td>
</tr>
<tr>
<td>Average daily traffic over the 5 years (ADT†)</td>
<td>9,442</td>
<td>143,422</td>
<td>30,237 (28776.4)</td>
<td>-</td>
</tr>
<tr>
<td>Minimum friction reading in the road segment over the 5-year period (Friction) (FR)</td>
<td>15.9</td>
<td>48.2</td>
<td>30.51 (6.67)</td>
<td>-</td>
</tr>
<tr>
<td>Pavement surface type (1 if asphalt, 0 if concrete) (PS)</td>
<td>0</td>
<td>1</td>
<td>0.77 (0.42)</td>
<td>-</td>
</tr>
<tr>
<td>Median width (MW) (in feet)</td>
<td>16</td>
<td>194.7</td>
<td>66.98 (34.17)</td>
<td>-</td>
</tr>
<tr>
<td>Presence of median barrier (1 if present, 0 if absent) (MB) (BARRIER)</td>
<td>0</td>
<td>1</td>
<td>0.16 (0.37)</td>
<td>-</td>
</tr>
<tr>
<td>Interior rumble strips (IRS) (RUMBLE)</td>
<td>0</td>
<td>1</td>
<td>0.72 (0.45)</td>
<td>-</td>
</tr>
<tr>
<td>Segment length (miles)</td>
<td>0.009</td>
<td>11.53</td>
<td>0.89 (1.48)</td>
<td>300.09</td>
</tr>
</tbody>
</table>

† Average Daily Traffic

### 3.2 Model Estimation

The following commonly used functional form was used for both models:

$$
\mu_i = \beta_0 \times L_i \times \pi_i^\lambda \times y \times e^{\sum_{j=1}^d x_i \beta_j}
$$

(8)
Where,  
\(\mu_i\) = the estimated number of crashes per year for site \(i\);  
\(F_i\) = vehicles per day (Average Daily Traffic/ADT or Annual Average Daily Traffic/AADT) for segment \(i\);  
\(L_i\) = length of segment \(i\) in miles;  
\(y\) = number of years of crash data;  
\(x_i\) = a series of covariates (e.g., shoulder width, lateral clearance, etc.) for site \(i\);  
\(n\) = number of covariates; and,  
\(\beta_1, \beta_2, ..., \beta_n\) = estimated coefficients.

3.3 Modeling Results

This section summarizes the modeling results for both datasets.

3.3.1 Modeling Results- Texas Rural Multilane Highways

This section documents the modeling results for the Texas dataset. Three covariates were considered in this analysis. They are average daily traffic (ADT), lane width (LW) and shoulder width (SW). The length of segment was considered as an offset term in order to be consistent with previous research (Geedipally et al., 2012). Table 3 shows that the coefficient values for all three covariates for both models have the same sign and are very close to each other. The GOF statistics show that the GW performs better than the NB model, as expected.

<table>
<thead>
<tr>
<th>variable</th>
<th>NB</th>
<th>GW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-5.89(0.56)</td>
<td>-6.81(0.59)</td>
</tr>
<tr>
<td>Ln(AADT)</td>
<td>1.0063(0.06)</td>
<td>0.95(0.055)</td>
</tr>
<tr>
<td>Lane Width(LW)</td>
<td>-0.1316(0.02)</td>
<td>-0.097(0.02)</td>
</tr>
<tr>
<td>Shoulder Width(SW)</td>
<td>-0.0316(0.005)</td>
<td>-0.019(0.005)</td>
</tr>
<tr>
<td>(\rho)</td>
<td>-</td>
<td>3.01</td>
</tr>
<tr>
<td>k</td>
<td>-</td>
<td>1.71</td>
</tr>
<tr>
<td>(\phi^1)</td>
<td>0.66(0.03)</td>
<td>-</td>
</tr>
<tr>
<td>-2LL (the smaller the better)</td>
<td>6015.4</td>
<td>5936.2</td>
</tr>
<tr>
<td>AIC (the smaller the better)</td>
<td>6025.4</td>
<td>5948.2</td>
</tr>
<tr>
<td>BIC (the smaller the better)</td>
<td>6051.9</td>
<td>5980.0</td>
</tr>
<tr>
<td>MAD</td>
<td>2.75</td>
<td>2.44</td>
</tr>
<tr>
<td>MSPE</td>
<td>32.8</td>
<td>28.4</td>
</tr>
<tr>
<td>Pearson (\chi^2)</td>
<td>12148.3</td>
<td>11156.4</td>
</tr>
</tbody>
</table>
NOTE: ( ) indicate the standard deviation. Bold characters indicate a better fit.

\[ \text{Inverse dispersion parameter: } \text{VAR}(Y) = \mu + \frac{\mu^2}{\phi} \]

As discussed in methodology section, the other more important advantage of the GW model is that it can provide more information about the sources of variance in crash data.

In this context, it can be seen that the proposed GW model was more useful to identify sources of variability and provide more valuable information to quantify the variance of each component compared to a model based on the NB distribution. Specific to this Texas crash dataset, it is assumed in the GW model that the proneness represents the over-dispersion due to between-segments variation in their internal probability to cause accident with the same AADT, lane and shoulder width, while liability is related to the over-dispersion caused by missing external covariates which would affect segments with the same AADT, lane width and shoulder width identically.

Table 4 shows the variance of each part or component for this crash dataset. It is quantified using the equations described in Rodríguez-Avi et al. (2009).

<table>
<thead>
<tr>
<th>Source of variability</th>
<th>Variance</th>
<th>Variance fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomness</td>
<td>( \mu_s )</td>
<td>( \frac{0.4642}{1.71 + \mu_s} )</td>
</tr>
<tr>
<td>Liability</td>
<td>2.683 ( \mu_s )</td>
<td>( \frac{1.2457}{1.71 + \mu_s} )</td>
</tr>
<tr>
<td>Proneness</td>
<td>2.1539 ( \mu_s^2 )</td>
<td>( \frac{\mu_s}{1.71 + \mu_s} )</td>
</tr>
<tr>
<td>Total</td>
<td>3.683 ( \frac{\mu_s + \mu_s^2}{1.71} )</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 1 shows the relationship between fraction of each component of the variance and different mean of crashes \( \mu_s \) which was determined by all the values of the covariates included in the model. The relationship between the covariate AADT and the fractions of each component is also illustrated in Figure 2 when shoulder width equals to 16 feet (8 feet on each side) and the lane width equals to 12 feet. These two values represent the largest percentage of segments with these characteristics.
It can be seen from the above figures that randomness and liability decrease from 0.27 to 0.03 and from 0.72 to 0.1 separately as the crash mean on each segment $\mu_x$ increases from 0 to 10, which means that the randomness and liability decrease when AADT increases and the shoulder width or lane width decreases, whereas proneness increases. As it is well-known that $\mu_x$ increases with the increase in segment length (usually in a linear fashion), the randomness and liability decrease with the segment length whereas proneness increases.
The fraction of proneness is more important on segments with higher AADT, which means the source of over-dispersion comes more from proneness for those segments. The fraction of proneness increases from 0 to 0.79 with the increase in AADT from 0 to 24,000. As mentioned earlier, several internal characteristics of road segments such as those listed above (i.e., road surface conditions, etc.) are related to the proneness. Therefore, the effect of these internal characteristics of road segments on the over-dispersion of traffic crashes seems to be more significant on the segments with higher AADT. Accordingly, for those segments with lower crash mean and lower AADT, the source of over-dispersion comes more from the randomness and liability which is related to external factors. Therefore, it is meaningful for traffic engineers, transportation safety specialists or managers to take effective measures to minimize variance of traffic crashes based on better understanding the source of variance using the GW model. Transportation managers should pay more attention to the factors such as the pavement-related problems or the quality of rumble strips (note: the Texas Department of Transportation or TxDOT has installed several miles of rumble strips, but their location was not known at the time this study was conducted) and pavement markings on segments with higher AADT, which is the main source of variance of crashes when a dataset exhibits highly over-dispersed characteristic. In short, the GW model can narrow down sites where traffic engineers could focus his or her efforts to improve safety (even when the sites have not been identified as dangerous or experiencing more crashes than expected).

3.3.2 Modeling Results- Indiana Interstate Highways

Table 5 summarizes the model estimation results for the Indiana dataset. The segment length variable is still considered an offset. This table is also shown in this table that the GW model performs better in terms of fit than the NB model. The estimated coefficients of all the covariates between the two models have the same sign which indicates similar results with regards to the effects of these variables, though there are some obvious differences between the values of these variables.
Table 5. Modeling results for the Indiana data

<table>
<thead>
<tr>
<th>Variable</th>
<th>NB</th>
<th>GW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-4.456(1.292)</td>
<td>-3.81(1.326)</td>
</tr>
<tr>
<td>Ln (ADT)</td>
<td>0.688(0.120)</td>
<td>0.711(0.124)</td>
</tr>
<tr>
<td>Minimum Friction on segment</td>
<td>-0.027(0.010)</td>
<td>-0.034(0.012)</td>
</tr>
<tr>
<td>Pavement Surface</td>
<td>0.43(0.185)</td>
<td>0.441(0.214)</td>
</tr>
<tr>
<td>Median Width</td>
<td>-0.005(0.002)</td>
<td>-0.008(0.003)</td>
</tr>
<tr>
<td>Median Barrier</td>
<td>-3.026(0.283)</td>
<td>-3.161(0.295)</td>
</tr>
<tr>
<td>Interior Rumble Strips</td>
<td>-0.398(0.180)</td>
<td>-0.413(0.191)</td>
</tr>
<tr>
<td>$\rho$</td>
<td>-</td>
<td>2.64</td>
</tr>
<tr>
<td>$k$</td>
<td>-</td>
<td>0.36</td>
</tr>
<tr>
<td>$\phi^\dagger$</td>
<td>1.071(0.075)</td>
<td>-</td>
</tr>
<tr>
<td>-2LL (the smaller the better)</td>
<td>1884.5</td>
<td>1846.4</td>
</tr>
<tr>
<td>AIC</td>
<td>1900.5</td>
<td>1864.4</td>
</tr>
<tr>
<td>BIC</td>
<td>1931.0</td>
<td>1898.8</td>
</tr>
<tr>
<td>MAD</td>
<td>6.78</td>
<td>6.48</td>
</tr>
<tr>
<td>MSPE</td>
<td>204.68</td>
<td>198.61</td>
</tr>
<tr>
<td>Pearson $\chi^2$</td>
<td>1170.2</td>
<td>1150.8</td>
</tr>
</tbody>
</table>

$^\dagger$ Inverse dispersion parameter: $\text{VAR}(Y) = \mu + \frac{\mu^2}{\phi}$

Table 6 shows the variance of each component for this crash dataset, using the same equations as used in Table 4.
Table 6. Variance of each component for the Indiana data

<table>
<thead>
<tr>
<th>Source of variability</th>
<th>Variance</th>
<th>Variance fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomness</td>
<td>$\mu_x$</td>
<td>$\frac{1}{3.125+8.680\mu_x}$</td>
</tr>
<tr>
<td>Liability</td>
<td>$2.125\mu_x$</td>
<td>$\frac{2.125}{3.125+8.680\mu_x}$</td>
</tr>
<tr>
<td>Proneness</td>
<td>$8.680\mu_x^2$</td>
<td>$\frac{8.680\mu_x}{3.125+8.680\mu_x}$</td>
</tr>
<tr>
<td>Total</td>
<td>$(3.125\mu_x+8.680\mu_x^2)$</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 3 shows the proportions of variance attributed to randomness, liability, and proneness for different crash means $\mu_x$ which was determined by all of the values of the covariates included in the model. The relationship between the covariate ADT and the fractions of each component is indicated in Figure 4, and the relationship between the covariate friction (FR) and fractions of each component when the other variables included in the model are equal to their means (as determined in the data) is illustrated in Figure 5. The relationships for the other covariates are not shown here due to space constraints.

![Figure 3](image_url)

**Figure 3. Relationship between the fraction of each component and the mean of crashes for the Indiana data**
Figure 4. Relationship between the fraction of each component and the average daily traffic for Indiana data

Figure 5. Relationship between the fraction of each component and friction reading for Indiana data

It can be seen from above figures that both randomness and liability decrease from 0.32 to 0.01 and from 0.68 to 0.03, separately, as the mean of crashes on each segment $\mu_i$ increases from 0 to 10, which signifies that both randomness and liability decrease when the ADT and pavement type increases. Both randomness and liability slightly decrease from 0.29 to 0.23 and from 0.62 to 0.49, separately, as the ADT on each segment increases from 5,000 to 15,000. On the other hand, proneness decreases from 0.83 to 0.67 as the friction increases from 15 to 40. Accordingly, both randomness and liability increase from 0.05 to 0.10 and from 0.11 to 0.22, separately, as the friction increases from 15 to 40. The proneness also
increases with the presence of interior rumble strips and a median barrier. It can be seen that the effect of the friction reading on the fraction of each component is not as significant or important as is the effect of the ADT.

It should be noted that there are more variables included in this model compared to the model used for the Texas data. Therefore, the over-dispersion is smaller than those models that include fewer independent geometric or environmental variables, as expected (Miaou and Song, 2005). Accordingly, the fraction of variance for the randomness is higher when fewer variables are included in the model (e.g., Figures 1 and 2 vs Figure 3). The fraction of variance for the liability could increase and the fraction of variance for the proneness could decrease when more covariates are included in the model (note: variable specific). This also means that the GW model could be used to analyze the role of different combinations of covariates that are included in the model on the variance and could potentially be utilized to identify models that suffer from the omitted variable bias (Lord and Mannering, 2010; Wu et al., 2014).

4. Summary and Further Research

In this study, a regression model developed based on the GW distribution was introduced and applied for analyzing crash data. There are several conclusions drawn from the above analysis:

1. Two empirical over-dispersed datasets were analyzed in this paper to compare the performance of GW model and NB model as well as to investigate the source of over-dispersion of crashes occurred on road segments. It was shown that the GW model provides a better performance than NB model based on the modeling results.

2. Besides the advantage that the GW model provided a better fit for the both empirical datasets, the other more important advantage of GW model is that it provided more information about the sources of variance for crashes that occurred on each segment. This model distinguished the observed variability for the crash dataset into three parts: randomness, proneness and liability. This information is valuable because it can help transportation safety professionals to better control the variance found in traffic crashes by implementing more cost-effective safety countermeasures without having to conduct a full identification of hazardous sites (such as the method proposed in the Highway Safety Manual—AASHTO, 2010).

3. As for the Texas dataset, the randomness and liability decrease from 0.27 to 0.03 and from 0.72 to 0.10 separately as mean of crashes on each segment increases from 0 to 10, whereas proneness increases. The randomness and liability decrease with the length of segment whereas proneness increases. As for the Indiana dataset which contains more variables, both randomness and liability decrease from 0.29 to 0.23 and from 0.62 to 0.49, separately, as the AADT on each segment increases from 5,000 to 15,000, separately, as the AADT on each segment increases from 5,000 to 15,000. On the other hand, proneness decreases from 0.83 to 0.67 as the friction increases from 15 to 40.
Accordingly, both randomness and liability increase from 0.05 to 0.10 and from 0.11 to 0.22, separately, as the friction increases from 15 to 40.

4. The fraction of proneness is more important on segments with higher ADT. Therefore, for those segments with higher number of crashes and ADT, more attention about source of over-dispersion should be paid to factors related to the proneness of segments, such as the roadway friction and the pavement condition of road segments, as discussed above, in order to reduce the variance of crashes effectively.

Another important issue discussed in previous work is the one related to ranking or selecting the best model based on GOF statistics. As discussed by others (Miaou and Lord, 2010; Lord and Bonneson, 2005; Lord et al., 2007), GOF measures should not be the sole criterion in selecting statistical models. It is also important to determine whether or not the proposed distribution or model is logically or theoretically sound (Geedipally et al., 2012). Miaou and Lord (2003) referred the latter attribute as “goodness-of-logic.” Compared to some simpler models, such as the NB or the Poisson model, the proposed GW model better describes the crash data generating process, because it separates the liability from the proneness. It should be noted that although the usefulness of the GW model appears in cases where the proneness and liability seem to be distinguishable, in other circumstances, the GW model can still be used effectively to capture the variance when these two components of the variance are not distinguishable (Rodríguez-Avi et al., 2007). When this happens, the GW model converges to the NB model.

For future work, there are several avenues that can be explored. First, the GW model can be seen as a three-layer hierarchical model, as discussed above, with a standard distribution at each stage. Therefore, the GW model can be developed based on a Bayesian modeling framework. Second, further work needs to be done to relate the liability and proneness to other mixed-distributions and models with a varying dispersion parameter. For the former, examining other continuous (and fixed) distributions for modeling the liability and proneness should be attempted, while, for the latter, GW models where the liability and/or proneness are analyzed as a function of the covariates, similar to the recent work done for two-parameter models (see Geedipally et al., 2009, and Connors et al., 2013), should be examined. Third, similar to the Sichel model (Zou et al., 2013), this research can be extended by developing an empirical Bayes (EB) modeling framework for the GW model to examine whether the EB framework is feasible for this model. In addition, there are several other models such as random parameter models that have also been used to describe sources of dispersion. The difference between these models with the GW is worthwhile to be investigated and compared in order to better understand the source of variance of traffic vehicle crashes. Finally, there have been some discussions about the assumption related to the covariate-dependent dispersion for NB models. Similar to the NB model with a varying dispersion parameter, further research should be conducted to examine the effect of a covariate-dependent parameter with the GW model. It is hoped that the GW model could be added to the panoply of tools that are currently used for managing highway safety.
Acknowledgements

The authors would like to thank Dr. José Rodríguez-Avi, from Universidad de Jaén, for answering various questions related to the Waring distribution and its model.
References


