Extension of the Application of Conway-Maxwell-Poisson Models: Analyzing Traffic Crash Data Exhibiting Under-Dispersion

Dominique Lord
Assistant Professor
Zachry Department of Civil Engineering
Texas A&M University
3136 TAMU
College Station, TX 77843-3136
Tel. (979) 458-3949
Fax. (979) 845-6481
Email: d-lord@tamu.edu

Srinivas Reddy Geedipally
Engineering Research Associate
Zachry Department of Civil Engineering
Texas A&M University
3136 TAMU
College Station, TX 77843-3136
Tel. (979) 862-3446
Fax. (979) 845-6481
Email: sринivas8@tamu.edu

Seth D. Guikema
Assistant Professor
Department of Geography and Environmental Engineering
Johns Hopkins University
Baltimore, MD 21218
Tel. (410) 516-6042
Fax. (410) 516-8996
E-mail: sguikema@jhu.edu

March 10, 2010

Published in Risk Analysis

1Corresponding author
ABSTRACT

The objective of this paper is to evaluate the performance of the COM-Poisson GLM for analyzing crash data exhibiting under-dispersion (when conditional on the mean). The COM-Poisson distribution, originally developed in 1962, has recently been re-introduced by statisticians for analyzing count data subjected to either over- or under-dispersion. Over the last year, the COM-Poisson GLM has been evaluated in the context of crash data analysis and it has been shown that the model performs as well as the Poisson-gamma model for crash data exhibiting over-dispersion. To accomplish the objective of this study, several COM-Poisson models were estimated using crash data collected at 162 railway-highway crossings in South Korea between 1998 and 2002. This dataset has been shown to exhibit under-dispersion when models linking crash data to various explanatory variables are estimated. The modeling results were compared to those produced from the Poisson and gamma probability models documented in a previous published study. The results of this research show that the COM-Poisson GLM can handle crash data when the modeling output shows signs of under-dispersion. Finally, they also show that the model proposed in this study provides better statistical performance than the gamma probability and the traditional Poisson models, at least for this dataset.

Key Words: Conway-Maxwell-Poisson, Under-dispersion, Regression models, Gamma models, Negative binomial models
1. INTRODUCTION

Over the last two decades, there has been a considerable amount of research performed on statistical methods for analyzing motor vehicle crashes. Some of the most recent methods include the application of neural and Bayesian neural networks, latent class or mixture model, gamma probability and support vector machine models among others. In highway safety, the traditional Poisson and mixed-Poisson models remain the most common probabilistic models utilized for analyzing crash data. This type of data has been found to often exhibit over-dispersion (i.e., the variance is larger than the mean) and thus mixed-Poisson models (such as the Poisson-gamma or negative binomial) are generally preferred over the traditional Poisson model. Although very rare, crash data have sometimes shown characteristics of under-dispersion (i.e. the variance is smaller than the mean value under an assumed probability model) especially in cases where the sample mean is very low. It has been shown that Poisson and Poisson-gamma models experience important limitations when the sample mean value is low and the sample size is small. In addition, other studies have demonstrated that both types of models have significant difficulties handling (or cannot handle) data characterized by under-dispersion. In the light of these limitations, several researchers have started examining the application of new and innovative methods for analyzing crash data.

As part of these new methods, the Conway-Maxwell-Poisson (COM-Poisson) distribution has very recently been re-introduced by statisticians for modeling count data that are characterized by either over- or under-dispersion. This distribution was first introduced in 1962, but has only been evaluated in the context of a generalized linear model (GLM) by Guikema and Coffelt,
Lord et al. and Sellers and Shmueli. The COM-Poisson distribution has been used in many studies, such as analyzing word length, births, the prediction of purchase timing and quantity decisions, quarterly sales of clothing, internet search engine visits, the timing of bid placement and extent of multiple bidding, modeling electric power system reliability and motor vehicle crashes. Only a handful of studies have applied the COM-Poisson distribution to observed or simulated data characterized by under-dispersion.

The objective of this paper is to evaluate the performance of the COM-Poisson GLM for analyzing crash data exhibiting under-dispersion, in cases where Poisson and Poisson-gamma models cannot be used. This paper is continuation of the work done by Guikema and Coffelt and Geedipally et al. and Lord et al. on the COM-Poisson distribution and GLM. To accomplish the objective of this study, several COM-Poisson models were estimated using crash data collected at 162 railway-highway crossings (RHX) in South Korea between 1998 and 2002. This dataset has been identified as being characterized by under-dispersion when the observations were modeled using regression methods. To model such dataset, Oh et al. have proposed the gamma probability model. The results of this study will show that the COM-Poisson GLM can handle crash data when the modeling output shows signs of under-dispersion. The results also show that the model provides better statistical performance than the gamma probability and the traditional Poisson model.

The paper is organized as follows. The first section provides a brief background on the COM-Poisson and gamma probability models. The second section describes the methodology used for estimating and comparing the various models. The third section presents the characteristics of the
data used in this study. The fourth section summarizes the results of the parameter estimates and comparison analysis. The last section offers important concluding remarks as well as ideas for further research.

2. BACKGROUND

This section provides a brief description of the characteristics of the COM-Poisson and the gamma probability models, respectively.

2.1. COM-Poisson Model

The COM-Poisson distribution is a generalization of the Poisson distribution and was first introduced by Conway and Maxwell for modeling queues and service rates.\(^{(26)}\) Shmueli et al. further elucidated the statistical properties of the COM-Poisson distribution using the formulation given by Conway and Maxwell, and Kadane et al. developed the conjugate distributions for the parameters of the COM-Poisson distribution.\(^{(22, 23, 26)}\) Its probability mass function (PMF) can be given by Equations (1) and (2).

\[
P(Y = y) = \frac{1}{Z(\lambda, \nu)} \frac{\lambda^y}{(y!)^\nu} \tag{1}
\]

\[
Z(\lambda, \nu) = \sum_{n=0}^{\infty} \frac{\lambda^n}{(n!)^\nu} \tag{2}
\]
where, \( Y \) is a discrete count; \( \lambda \) is a centering parameter that is approximately the mean of the observations in many cases; and, \( \nu \) is defined as the shape parameter of the COM-Poisson distribution. The centering parameter \( \lambda \) is approximately the mean when \( \nu \) is close to one, it differs substantially from the mean for small \( \nu \). Given that \( \nu \) would be expected to be small for over-dispersed data, this would make a COM model based on the original COM formulation difficult to interpret and use for over-dispersed data.

To circumvent this problem, Guikema and Coffelt proposed a re-parameterization of the COM-Poisson distribution by substituting \( \frac{1}{\nu} \) to provide a clear centering parameter.\(^{(24)}\) This new formulation of the COM-Poisson is summarized in Equations (3) and (4) below.

\[
P(Y = y) = \frac{1}{S(\mu, \nu)} \left( \frac{\mu^y}{y!} \right)^\nu
\]

\[
S(\mu, \nu) = \sum_{n=0}^{\infty} \left( \frac{\mu^n}{n!} \right)^\nu
\]

The mean and variance of \( Y \) are given in terms of the new formulation as \( E[Y] = \frac{1}{\nu} \frac{\partial \log S}{\partial \log \mu} \) and \( Var[Y] = \frac{1}{\nu^2} \frac{\partial^2 \log S}{\partial \log^2 \mu} \) with asymptotic approximations \( E[Y] \approx \mu + 1/2\nu - 1/2 \) and \( Var[Y] \approx \mu/\nu \) especially accurate once \( \mu > 10 \). With this new parameterization, the integral part of \( \mu \) is now the mode leaving \( \mu \) as a reasonable centering parameter. The substitution \( \mu = \lambda^{1/\nu} \) also allows \( \nu \) to keep its role as a shape parameter. That is, if \( \nu < 1 \), the variance is greater than the mean while \( \nu > 1 \) leads to under-dispersion.
Guikema and Coffelt developed a COM-Poisson GLM framework using Bayesian framework in WinBUGS for modeling discrete count data. Equations (5) – (6) describe this modeling framework. The framework is a dual-link GLM in which both the mean and the variance depend on the covariates. In Equations (5) and (6), $x_i$ and $z_j$ are covariates, and there are assumed to be $p$ covariates used in the centering link function and $q$ covariates used in the shape link function (similar to the varying dispersion parameter of the Poisson-gamma model proposed by Miaou and Lord, Hauer, and Heydecker and Wu). The sets of parameters used in the two link functions do not necessarily have to be identical.

\[
\ln(\mu) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i
\]  
\[
\ln(\nu) = \gamma_0 + \sum_{j=1}^{q} \gamma_j z_j
\]

The GLM framework can model under-dispersed data sets, over-dispersed data sets, and data sets that contain intermingled under-dispersed and over-dispersed counts (for dual-link models only, since the dispersion characteristic is captured using the covariate-dependent shape parameter). The variance is allowed to depend on the covariate values, which can be important if high (or low) values of some covariates tend to be variance-decreasing while high (or low) values of other covariates tend to be variance-increasing. The parameters have a direct link to either the mean or the variance, providing insight into the behavior and driving factors in the problem, and the mean and variance of the predicted counts are readily approximated based on the covariate values and regression parameter estimates.
Recently, Sellers and Shmueli derived the likelihood function for the COM-Poisson GLM. This derivation greatly simplifies the estimation of the parameters of a COM GLM when full posteriors are not needed for the parameters, as opposed to the Bayesian estimating method.\(^{(24, 28)}\)

However, by the time this paper was prepared, the MLE formulation did not allow for a varying shape parameter, as described in Eq. (6). The interested reader can find the MLE codes in R for the COM-Poisson GLM here: [http://cran.r-project.org/web/packages/compoisson/index.html](http://cran.r-project.org/web/packages/compoisson/index.html).\(^{(36)}\)

### 2.2. Gamma Probability Model

The gamma probability distribution can be used for analyzing under-dispersed and over-dispersed datasets. Oh et al. used this distribution to analyze crashes collected at RHX in South Korea.\(^{(15)}\) They found the gamma probability model to provide a good statistical fit for the railway-highway crossing crash data under study. The gamma probability model for the count data is given by:\(^{(15, 37)}\)

\[
P(y_i = k) = \text{Gamma}(\alpha k, \lambda_i) - \text{Gamma}(\alpha k + \alpha, \lambda_i) 
\]

Where \(\lambda_i\) is the mean of the crashes at \(i^{th}\) site and is given by \(\lambda_i = \exp(\mathbf{x}_i \mathbf{\beta})\)

\[
\text{Gamma}(\alpha k, \lambda_i) = 1 \text{ if } k=0; 
\]

\[
\text{Gamma}(\alpha k, \lambda_i) = \frac{1}{\Gamma(\alpha k)} \int_0^{\lambda_i} t^{\alpha k-1} e^{-t} dt \text{ if } k>0; 
\]

Where \(\alpha\) is the dispersion parameter; for \(\alpha >1\), the model shows under-dispersion; for \(\alpha <1\), the model exhibits over-dispersion; for \(\alpha =1\), it is equi-dispersion which means that the gamma model reduces to Poisson model. It should be noted that the gamma probability model assumes a
dual-state process, one for $y_i = 0$ and one for $y_i > 0$; this is the formulation of a Hurdle function (note: $\lambda_i = 0$ since $y_i$ is always equal to zero). As discussed in previous studies, this kind of process may not be appropriate for analyzing crash data.

3. METHODOLOGY

This section briefly describes the methodology used for comparing the different models. The same functional form used by Oh et al. was utilized for fitting all the models:

$$
\mu_i = \exp(\beta_0 + \beta_1 \ln(F_i) + \sum_{j=1}^{n} \beta_j x_{ij})
$$

Where,

$\mu_i =$ the mean number of crashes for site $i$;

$F_i =$ average daily vehicle traffic on site $i$ (vehicles/day);

$x_{ij} =$ estimated covariates such as average daily railway traffic, detector distance etc; and,

$\beta_i =$ estimated regression coefficients.

Different methods were used for evaluating the goodness-of-fit (GOF) and predictive performance of the models. The methods used in this research include the following:

*Akaike Information Criterion (AIC)*
The AIC is a measure of the goodness of fit of an estimated statistical model and is defined as\(^{(40)}\)

\[
AIC = -2 \log L + 2p \tag{10}
\]

Where \(L\) is the maximized value of the likelihood function for the estimated model, and \(p\) is the number of parameters in the statistical model. The AIC methodology attempts to find the model that best explains the data with a minimum of free parameters and thus it penalizes models with a large number of parameters. The model with the lowest AIC is considered to be the best model among all available models.

\textit{Mean Prediction Bias (MPB)}

MPB provides a measure of the magnitude and direction of the average model bias.\(^{(41)}\) If the MPB is positive then the model over-predicts crashes and if the MPB is negative then the model under-predicts crashes. It is computed using the following equation:

\[
\text{Mean Prediction Bias (MPB)} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i) \tag{11}
\]

Where \(n\) is the sample size, \(\hat{y}_i\) and \(y_i\) are the predicted and observed crashes at site \(i\) respectively.
Mean Absolute Deviance (MAD)

MAD provides a measure of the average mis-prediction of the model. The model closer to zero is considered to be the best among all the available models. It is computed using the following equation:

\[
\text{Mean Absolute Deviance (MAD)} = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|
\]  

Mean Squared Predictive Error (MSPE)

MSPE is typically used to assess the error associated with a validation or external data set. The model closer to zero is considered to be the best among all the available models. It can be computed using Equation (12):

\[
\text{Mean Squared Predictive Error (MSPE)} = \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2
\]  

Since the models estimated by Oh et al. were done using the likelihood method, the coefficients of the COM-Poisson GLMs were also estimated using the MLE code developed by Sellers and Shmueli. This way, all the model comparisons are performed using the same estimation method.

4. DATA DESCRIPTION

This section provides an overview of the characteristics of the dataset used in this study. This dataset was previously used to develop Poisson and gamma probability models by Oh et al.
The characteristics of the dataset used in this study are described in Table I. Only the variables that are found to be significant in this study are presented in Table I. For the characteristics of all other variables, the reader is referred to the Table 2 of Oh et al.\textsuperscript{(15)} It should be noted that looking at the raw observations, the crash data exhibit over-dispersion (mean=0.33, variance=0.36). The under-dispersion is in fact noticed when the observed values are modeled conditional on the mean, as described in the next section.
Table I: Summary Statistics of the Dataset

<table>
<thead>
<tr>
<th>Variables</th>
<th>Min.</th>
<th>Max.</th>
<th>Average (std. dev)</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crashes</td>
<td>0</td>
<td>3</td>
<td>0.33 (0.60)</td>
<td>162</td>
</tr>
<tr>
<td>AADT</td>
<td>10</td>
<td>61199</td>
<td>4617 (10391.57)</td>
<td>162</td>
</tr>
<tr>
<td>Average daily railway traffic</td>
<td>32</td>
<td>203</td>
<td>70.29 (37.34)</td>
<td>162</td>
</tr>
<tr>
<td>Presence of a commercial area</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (yes)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>149 (91.98%)</td>
</tr>
<tr>
<td>0 (no)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>13 (8.02%)</td>
</tr>
<tr>
<td>Train detector distance</td>
<td>0</td>
<td>1329</td>
<td>824.5 (328.38)</td>
<td>162</td>
</tr>
<tr>
<td>Time duration between the activation of warning signals and gates</td>
<td>0</td>
<td>232</td>
<td>25.46 (25.71)</td>
<td>162</td>
</tr>
<tr>
<td>Presence of a speed hump</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (yes)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>134 (82.72%)</td>
</tr>
<tr>
<td>0 (no)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>28 (17.28%)</td>
</tr>
<tr>
<td>Presence of a track circuit controller</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (yes)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>113 (69.75%)</td>
</tr>
<tr>
<td>0 (no)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>49 (30.25%)</td>
</tr>
<tr>
<td>Presence of a guide</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 (yes)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>126 (77.78%)</td>
</tr>
<tr>
<td>0 (no)</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>36 (22.22%)</td>
</tr>
</tbody>
</table>

Figure 1 shows the comparison between the actual crash data distribution with the values predicted by the Poisson distribution. This figure illustrates that the Poisson distribution predicts almost the same number of crashes as the actual data.
5. RESULTS

This section describes the results of the analysis. Several models were estimated using the variables documented in Oh et al.\(^{(15)}\) To evaluate the characteristics of the variance function, Poisson-gamma models were first estimated using the six variables that were reported to be significant by the gamma probability model in the original study.\(^{(15)}\) Figures 2 and 3 show the output of the Poisson-gamma models for the MLE and Bayesian estimating methods, respectively. The Bayesian estimating method was used to confirm the results of the MLE method for this particular dataset. For the MLE, Figure 2 illustrates that the Poisson-gamma model cannot handle the data very well, as determined by the negative value of the dispersion parameter and its confidence interval (e.g., \(Var(Y) = \mu + \alpha \mu^2\), where \(\alpha\) = the dispersion parameter of the Poisson-gamma model). In addition, the model provides unreliable parameter estimates, since most of the variables are not significant at the 10% level. For the Bayesian model, Figure 3 shows that the inverse dispersion parameter of the Poisson-gamma model

![Figure 1: Observed Crash Data versus Values Estimated using the Poisson Distribution](image-url)
becomes unstable and tends towards infinity (i.e., reverts back to a Poisson model), both when
vague ($\phi \sim \text{gamma}(0.01,0.01)$) and less-vague hyper-priors ($\phi \sim \text{gamma}(0.2,0.1)$) are used.

Similar to the MLE, most of the parameter estimates were not significant at the 10% level. In
sum, these plots confirm that the modeling results are characterized by under-dispersion when
the model is estimated using the six original explanatory variables or covariates.

**Figure 2: SAS Output of the Poisson-gamma Model**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald 95% Confidence Limits</th>
<th>Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-2.7048</td>
<td>0.6447</td>
<td>-3.9683, -1.4413</td>
<td>17.60</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>log_MNT</td>
<td>1</td>
<td>0.1518</td>
<td>0.0636</td>
<td>0.0271, 0.2765</td>
<td>5.63</td>
<td>0.0171</td>
</tr>
<tr>
<td>Railway_traffic</td>
<td>1</td>
<td>0.0027</td>
<td>0.0027</td>
<td>-0.0027, 0.0000</td>
<td>0.35</td>
<td>0.5542</td>
</tr>
<tr>
<td>Presence_of_comm_arcs</td>
<td>1</td>
<td>0.2332</td>
<td>0.3188</td>
<td>0.0000, 1.3501</td>
<td>5.17</td>
<td>0.0229</td>
</tr>
<tr>
<td>Distance_of_train_de</td>
<td>1</td>
<td>0.0005</td>
<td>0.0004</td>
<td>-0.0002, 0.0002</td>
<td>2.00</td>
<td>0.1568</td>
</tr>
<tr>
<td>Warning_time_difference</td>
<td>1</td>
<td>0.0052</td>
<td>0.0021</td>
<td>0.0012, 0.0092</td>
<td>6.36</td>
<td>0.0117</td>
</tr>
<tr>
<td>Presence_of_speed_hu</td>
<td>1</td>
<td>-0.4906</td>
<td>0.3180</td>
<td>-1.958, 0.8245</td>
<td>1.58</td>
<td>0.2091</td>
</tr>
<tr>
<td>Dispersion</td>
<td>0</td>
<td>-0.3333</td>
<td>0.0000</td>
<td>-0.3333, -0.3333</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**NOTE:** The negative binomial dispersion parameter was estimated by maximum likelihood.
In the initial step, a COM-Poisson model was estimated with all thirty one explanatory variables documented in Table 2 of Oh et al.\textsuperscript{(15)} Only six variables were found to be significant at the 10% confidence level. In the subsequent step, a Poisson model was estimated, also in R, by considering all thirty one variables to compare this model with the COM-Poisson model in identifying significant variables and assess its statistical fit. Then, the final Poisson and COM-Poisson models were compared with the gamma probability model in Oh et al.\textsuperscript{(15)}

The comparison analysis related to the significant variables is summarized in Table II. As seen in this table, each model gives different significant variables. The variables AADT, \textit{presence of a commercial area}, \textit{train detector distance} and \textit{presence of a speed hump} are significant in all the
three models. The presence of a guide is significant only in the COM-Poisson model, while the presence of a track circuit controller is significant in both the Poisson and COM-Poisson models. The variables average daily railway traffic and warning time duration are only significant in gamma probability model.

Table II: Significant Variables with Three Different Distributions

<table>
<thead>
<tr>
<th>Variables</th>
<th>COM-Poisson</th>
<th>Poisson</th>
<th>Gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td>AADT</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Average daily railway traffic</td>
<td>--</td>
<td>--</td>
<td>✓</td>
</tr>
<tr>
<td>Presence of a commercial area</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Train detector distance</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Time duration between the activation of warning signals and gates</td>
<td>--</td>
<td>--</td>
<td>✓</td>
</tr>
<tr>
<td>Presence of a track circuit controller</td>
<td>✓</td>
<td>✓</td>
<td>--</td>
</tr>
<tr>
<td>Presence of a guide</td>
<td>✓</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Presence of a speed hump</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

✓ - Significant at 10% confidence level

Table III compares the parameter estimates and the associated standard errors for the three models. Different measures of GOF are also presented in Table 3 to compare the models’ performance. As a rule of thumb, if the difference in the AIC value is less than 10, then the model fit is assumed to be not significantly different from one another. Thus, using the AIC values, the COM-Poisson is not significantly different from Poisson and gamma probability model. The MPB values, on the other hand, show that the COM-Poisson slightly under-predicts crashes whereas the Poisson model slightly over-predicts crashes. The value of MPB for the gamma probability model shows that it highly over-predicts crashes.
Table III also shows that the shape parameter of COM-Poisson and gamma probability model clearly indicates the modeling output exhibits under-dispersion. Thus, the Poisson model cannot be used for this dataset, even if it fits the data relatively well. Although, fitting a Poisson model to such data will not significantly influence the mean of the regression coefficients, it will have a significant effect on the standard errors. This can be seen in the comparison of the standard errors for each model. The Poisson model underestimates the standard errors and in turn produces inflated t-values and confidence intervals of the coefficients. It is important to note that the coefficients estimated for COM-Poisson model in Table 3 are for centering parameter “$\lambda$” and not for mean “$E[Y]$”, as in the case of Poisson and gamma model. Finally, the MAD and MSPE values in Table 3 show that the COM-Poisson model fits the data much better than the gamma model.

**Table III: Parameter Estimates with Three Different Distributions (MLE)**

<table>
<thead>
<tr>
<th>Variables</th>
<th>COM-Poisson</th>
<th>Poisson</th>
<th>Gamma</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-6.657 (1.206)$^c$</td>
<td>-5.326 (0.906)$^c$</td>
<td>-3.438 (1.008)$^c$</td>
</tr>
<tr>
<td>Ln(AADT)</td>
<td>0.648 (0.139)</td>
<td>0.388 (0.076)</td>
<td>0.230 (0.076)</td>
</tr>
<tr>
<td>Average daily railway traffic</td>
<td>--</td>
<td>--</td>
<td>0.004 (0.0024)</td>
</tr>
<tr>
<td>Presence of commercial area</td>
<td>1.474 (0.513)</td>
<td>1.109 (0.367)</td>
<td>0.651 (0.287)</td>
</tr>
<tr>
<td>Train detector distance</td>
<td>0.0021 (0.0007)</td>
<td>0.0019 (0.0006)</td>
<td>0.001 (0.0004)</td>
</tr>
<tr>
<td>Time duration between the activation of warning signals and gates</td>
<td>--</td>
<td>--</td>
<td>0.004 (0.002)</td>
</tr>
<tr>
<td>Presence of track circuit controller</td>
<td>-1.305 (0.431)</td>
<td>-0.826 (0.335)</td>
<td>--</td>
</tr>
<tr>
<td>Presence of guide</td>
<td>-0.998 (0.512)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Presence of speed hump</td>
<td>-1.495 (0.531)</td>
<td>-1.033 (0.421)</td>
<td>-1.58 (0.859)</td>
</tr>
<tr>
<td>Shape Parameter ($v_0$)</td>
<td>2.349 (0.634)</td>
<td>--</td>
<td>2.062 (0.758)</td>
</tr>
<tr>
<td>AIC</td>
<td>210.70</td>
<td>196.55</td>
<td>211.38</td>
</tr>
<tr>
<td>MPB</td>
<td>-0.007</td>
<td>0.004</td>
<td>0.179</td>
</tr>
<tr>
<td>MAD</td>
<td>0.348</td>
<td>0.359</td>
<td>0.459</td>
</tr>
<tr>
<td>MSPE</td>
<td>0.236</td>
<td>0.252</td>
<td>0.308</td>
</tr>
</tbody>
</table>

$^c$ Standard error.
Figure 4 shows the estimated number of crashes for the Poisson, COM-Poisson, and gamma models as a function vehicular traffic flow. These values were estimated at the average value of all other continuous variables that are significant in the three models. Thus, a direct comparison between the curves should not be conducted, since each predicted value is estimated using different input variables. Figure 4 illustrates that the trends shown by the Poisson and COM-Poisson models (i.e., the shape of the curve) are very similar, but are much different than the trend shown by gamma model. The number of crashes predicted by the gamma model sharply increases with the increase in traffic flow until 2,000 and the rate per unit of exposure decreases significantly for larger flows.

Figure 4: Estimated Values for Gamma, Poisson and COM-Poisson GLMs
6. SUMMARY AND CONCLUSIONS

The objective of this paper was to evaluate the performance of the COM-Poisson GLM for analyzing crash data exhibiting under-dispersion (when conditional on the mean). The COM-Poisson distribution, originally developed in 1962, has recently been re-introduced by statisticians for analyzing count data subjected to either over- or under-dispersion. Over the last year, the COM-Poisson GLM has been evaluated in the context of crash data analysis and it has been shown that the model performs as well as the Poisson-gamma model for crash data exhibiting over-dispersion. To accomplish the objective of this study, several COM-Poisson models were estimated using crash data collected at 162 railway-highway crossings in South Korea between 1998 and 2002. This dataset has been shown to exhibit under-dispersion when models linking crash data to various explanatory variables are estimated. The modeling results were compared to those produced from the Poisson and gamma probability models documented in Oh et al.\(^{(15)}\)

The results of this study show that the COM-Poisson GLM can handle crash data when the modeling output shows signs of under-dispersion. They also show that the model analyzed in this study provides better statistical performance than the gamma probability and the traditional Poisson models, at least for this dataset. Similarly, the COM-Poisson GLM offers a more defensible approach than the gamma probability model, since the former does not assume that the observed data follow a dual-state generating process, one of which has a long-term equal to zero. Although interesting results were found in this study, further research should be conducted using different datasets to confirm those found in this research. The proposed analysis should also include looking at the robustness of the COM-Poisson GLM for data characterized by very
low sample values and small sample size, since under-dispersion is usually observed for data characterized by these two conditions (see some interesting findings in Geedipally et al. \(^{(25)}\)). Finally, COM-Poisson models with a dual-link should be estimated for datasets exhibiting similar characteristics as the one used in this research. Given the changes in the nature of the variance function when variables are included or excluded, it is possible that such datasets could contain intermingled over- and under-dispersed counts.

ACKNOWLEDGEMENTS

The authors wish to thank Dr. Simon Washington from Arizona State University and Dr. Jutaek Oh from the Korean Transport Institute for providing the data used in their study. Their answers to our numerous questions are also greatly appreciated.

REFERENCES


3 Miaou, S-P. The Relationship between Truck Accidents and Geometric Design of Road Sections: Poisson versus Negative Binomial Regressions. Accident Analysis & Prevention, 1994, 26 (4); 471-482.


14 Park, B.-J., Lord, D. Application of finite mixture models for vehicle crash data analysis. Accident Analysis & Prevention, 2009, 41 (4); 683-691.


34 Hauer, E. Overdispersion in modelling accidents on road sections and in Empirical Bayes estimation. Accident Analysis & Prevention, 2001, 33 (6); 799- 808.


