An Application of the Negative Binomial-Generalized Exponential Model for Analyzing Traffic Crash Data with Excess Zeros

Prathyusha Vangala
Graduate Student
Zachry Department of Civil Engineering
Texas A&M University
Email: prathyusha89@neo.tamu.edu

Dominique Lord[1], Ph.D., P.Eng.
Associate Professor & Zachry Development Professor I
Zachry Department of Civil Engineering
Texas A&M University
Tel. (979) 458-3949
Email: d-lord@tamu.edu

Srinivas Reddy Geedipally, Ph.D., P.E.
Assistant Research Engineer
Texas A&M Transportation Institute
Texas A&M University System
Tel. (817) 462-0519
Email: srinivas-g@ttimail.tamu.edu

August 23, 2014

[1] Corresponding author

ABSTRACT

In order to analyze crash data, many new analysis tools are being developed by transportation safety analysts. The Negative Binomial-Generalized Exponential distribution (NB-GE) is such a tool that was recently introduced to handle datasets characterized by a large number of zero counts and are over-dispersed. As the name suggests, this three-parameter distribution is a combination of both Negative binomial and Generalized Exponential distributions. So far, nobody has used this distribution in the context of a regression model for analyzing datasets with excess zeros. This paper therefore describes the application of the NB-GE generalized linear model (GLM). The distribution and GLM were applied to four datasets known to have large dispersion and/or a large number of zeros. The NB-GE was compared to the Poisson, NB as well as the Negative Binomial-Lindley (NB-L) model, another three-parameter recently introduced in the safety literature. The study results show that for datasets characterized by a sizable over-dispersion and contain a large number of zeros, the NB-GE performs as well as the NB-L, but significantly outclass the traditional NB model. Furthermore, the NB-GE model has a simpler modeling framework than the NB-L, which makes its application relatively straight forward.

INTRODUCTION

Crash data are usually characterized by over-dispersion, which means that the sample variance is larger than the sample mean (Lord and Mannering, 2010; Mannering and Bhat, 2014). Depending on the characteristics of the data, the level of dispersion can be so large that traditional statistical models, such as the Poisson-gamma or the Poisson-lognormal, cannot effectively be used for analyzing such datasets (Lord and Geedipally, 2011). The high level of dispersion can be influenced by the proportion of observations that have a value equal to zero as well as data that are characterized by very long tails (observations with a large number of crashes).

To overcome this problem, researchers have proposed different tools for analyzing datasets with a large number of zeros and long tails. They include the zero-inflated (ZI) models (Shankar et al., 1997; Shankar et al., 2001), the Negative Binomial-Lindley (NB-L) model (Geedipally et al., 2012; Hallmark et al., 2013), and the Sichel (SI) model (Zou et al., 2013). So far, the proposed models either suffer from methodological deficiencies (i.e., ZI models), the mathematical manipulation for developing the models can be very complex (i.e., NB-L), or the performance is not as good as other newly introduced three-parameter models (i.e., SI). Lord et al. (2005; 2007) and Geedipally et al. (2012) provide additional discussion about these limitations.

To continue the research on the application of three-parameter models for analyzing datasets with excess zeros, this paper documents the development and application of the Negative Binomial Generalized Exponential (NB-GE) generalized linear model (GLM). This model, which has never been developed before, is based on the recently introduced NB-GE distribution (Aryuyuen and Bodhisuwan, 2013). The distribution and model were evaluated using four different datasets. The NB-GE model was compared to the Poisson, NB and NB-L models. Finally, this paper provides information about the application of NB-GE GLM under a Bayesian framework.

BACKGROUND

This section is divided into two parts and describes the characteristics of the NB-GE distribution and the GLM.

NB-GE Distribution

As the name implies, the NB-GE distribution is a combination of NB and GE distributions. The NB-GE distribution was first introduced by Aryuyuen and Bodhisuwan (2013), where they presented the basic properties of the new distribution such as the mean, variance, skewness and kurtosis. Similar to the NB-L distribution (Lord and Geedipally, 2011), this mixed distribution also has a thick tail and works well when the data contains large number of zeros or is highly dispersed. The traditional NB and NB-exponential distributions are the special cases of the NB-GE distribution.

The probability mass function (pmf) of Negative Binomial distribution is given by

$$P(Y = y, \mu, \phi) = \frac{\Gamma(\phi+y)}{\Gamma(\phi)\Gamma(y+1)}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\mu}{\mu+\phi}\right)^{y} \quad (1)$$

Where,

μ = mean response of the observation; and,

ϕ = inverse of the dispersion parameter α (i.e. $\phi = 1/\alpha$).

The generalized exponential distribution has the probability density function (pdf) as follows (Aryuyuen and Bodhisuwan, 2013):

$$f(Z = z, \alpha, \lambda) = \alpha\lambda(1 - e^{\lambda z})^{\alpha-1} e^{-\lambda z};\ \alpha, \lambda > 0, \mathrm{z} > 0 \tag{2}$$

Where,

α = shape parameter; and,

λ = scale parameter.

The exponential distribution is the special case of the GE distribution (i.e., when $\alpha = 1$).

The moment generating function of the GE distribution is given as (Aryuyuen and Bodhisuwan, 2013):

$$\mathrm{M_z(t)} = \frac{\Gamma(\alpha+1)\Gamma\left(1-\frac{t}{\beta}\right)}{\Gamma\left(\alpha-\frac{t}{\beta}+1\right)} \tag{3}$$

The mean and variance of the GE distribution are given as (Gupta and Kundu, 1999):

$$\mathrm{E(Z)} = \frac{1}{\lambda}(\psi(\alpha+1) - \psi(1)) \tag{4}$$

$$\mathrm{Var(Z)} = \frac{1}{\lambda^2}(\psi'(\alpha+1) - \psi'(1)) \tag{5}$$

Where,

$\psi(.)$ = digamma function; and

$\psi'(.)$= derivative of the digamma function $\psi(.)$.

The NB-GE distribution arises by combining the NB and GE distributions. The pmf of NB-GE distribution is given as (Aryuyuen and Bodhisuwan, 2013):

$$f(X = x; r, \alpha, \lambda) = \binom{r+x-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^j \left(\frac{\Gamma(\alpha+1)\Gamma\left(1+\frac{r+j}{\lambda}\right)}{\Gamma\left(\alpha+\frac{r+j}{\lambda}+1\right)} \right);\ \alpha, \lambda > 0 \tag{6}$$

For the moments, skewness, and other details, the reader is referred to Aryuyuen and Bodhisuwan (2013).

NB-GE Generalized Linear Model

The NB-GE distribution is defined as a mixture of NB and GE distributions such that:

$$P(X = x, \mu, \phi, \theta) = \int \mathrm{NB}(x; \phi, z\mu)\mathrm{GE}(z; \alpha, \lambda)\ dz \quad (7)$$

Here, $f(u,;a,b)$ means that f is the distribution of the variable μ, with parameters a and b. The parameter μ is the function of covariates and z follows the GE distribution.

In the context of the NB-GE GLM, the mean response for the number of crashes is estimated by the product of μ and $E(z)$. The parameter μ is assumed to have a log-linear relationship with the covariates and is structured as:

$$\ln(\mu) = \beta_0 + \sum_{i=1}^{q} \beta_i X \quad (8)$$

Where,

X = traffic and geometric variables of a particular site;
βs = regression coefficients to be estimated; and,
q = total number of covariates in the model.

The mean response of the number of crashes is calculated as follows:

$$\mathrm{E(X)} = \mu \times \mathrm{E(Z)} = e^{\sum_{i=1}^{q} \beta i X + \beta 0} \times \frac{1}{\lambda}(\psi(\alpha + 1) - \psi(1)) \quad (9)$$

The crash variance is given by the Equation (11) below:

$$\mathrm{Var(X)} = \mathrm{E(X)} + \left[\mathrm{Var(Z)} + (\mathrm{E(Z)})^2\right]\frac{(1+\phi)}{\phi} - (\mathrm{E(X)})^2 \quad (10)$$

Where E(Z) and Var(Z) are described in the Equations (4) and (5), respectively.

The parameters α, β and ϕ of the NB-GE model can be obtained using the OpenBUGS software (Lunn et al., 2000), since the generalized exponential is already available in the library. For the GLM estimated in this work, three Markov chains were used with a total of 50,000 iterations per chain. The first 40,000 iterations have been discarded. The remaining 10,000 iterations were used for estimating the coefficients.

RESULTS

This part of the paper presents the results related to both the NB-GE distribution and the GLM. For evaluating the NB-GE distribution, its performance is compared with the Poisson, NB and NB-L distributions. For the GLM evaluation, the NB-GE model is compared to the NB and NB-L models.

Examining the NB-GE Distribution

This section describes the comparison analysis among the Poisson, NB, the NB-L and NB-GE distributions. The characteristics of the NB-L distribution and the NB-L model are not described here, but can be found in Lord and Geedipally (2011) and Geedipally et al. (2012), respectively.

The first dataset include single-vehicle fatal crashes that occurred on divided multilane rural highways between 1997 and 2001. The data were collected as a part of NCHRP 17-29 research project titled "*Methodology for estimating the safety performance of multilane rural highways*" (Lord et al., 2008). The data contained 1,721 segments that varied from 0.10 mile to 11.21 miles, with an average equal to 1.01 miles. The sample mean was equal to 0.13. About 89% of the segments had no fatal crash. The second dataset included single-vehicle roadway departure fatal crashes that occurred on 32,672 rural two-lane horizontal curves between 2003 and 2008. The sample mean is equal to 0.14. For this dataset, about 90% of the data experienced no crash during the 5-year period. The goodness-of-fit (GOF) analysis, based on the Chi-Square and log-likelihood, are presented in Tables 1 and 2.

Table 1. Single-Vehicle Fatal Crashes on Divided Multilane Rural Highways between 1997 and 2001 (Lord and Geedipally, 2011)

Crashes	Observed Frequency	Poisson	NB	NB-L	NB-GE
0	1,532	1,509.2	1,534.4	1,532.9	1,532.6
1	162	198	154.7	158.3	158.9
2	19	13.0	25.8	23.7	23.6
3	6	0.6	4.9	4.6	4.5
4	2	0	1.2	1.4	1.8
Parameters		μ=0.131	μ=0.131 ϕ=0.434	Θ=1532.9 R=1.851	r=1.28 α=1.5 λ=13.569
Chi-square		102.99	2.73	1.68	**1.39**
Log-likelihood		-715.1	-696.1	-695.6	**-694.5**

Bold character shows a better fit.

Table 2. Single-Vehicle Roadway Departure Crashes on Rural Two-Lane Horizontal Curves between 2003 and 2008 (Lord and Geedipally, 2011)

Crashes	Observed Frequency	Poisson	NB	NB-L	NB-GE
0	29,087	28,471.6	29,204.8	29,133.6	29,097.8
1	2,952	3,918.0	2,706	2,855.5	2,908.4
2	464	269.6	567.4	503.1	498.3
3	108	12.4	141.1	120.9	115.9
4	40	0.4	37.8	35.9	34.3
5	9	0.0	10.6	13.1	11
6	5	0.0	3.0	3.3	4.1
7	2	0.0	0.9	3.3	3
8	3	0.0	0.3	0.0	0.9
9	1	0.0	0.1	0.0	0.4
10	1	0.0	0.0	3.3	0.2
Parameters		μ=0.138	μ=0.138 ϕ=0.284	Θ=9.212 R=1.018	r=0.937 α=1.280 λ =8.999
Chi-square		2,297.31	57.47	11.68	**6.38**
Log-likelihood		-14,208.1	-13,557.7	-13,529.8	**-13,525**

Bold character shows a better fit.

The results for both tables show that the NB-GE distribution provide a slightly better fit than the NB-L. As expected, the two three-parameter distributions perform much better than the Poisson and the NB distributions.

Examining the NB-GE GLM

This section describes the modeling results for the NB, NB-L and NB-GE. The models were applied to two datasets. The first dataset contains crash data that was collected over a period of five years at 338 rural interstate sections in Indiana, while the second dataset contains data collected in Michigan in 1996. The summary of the data is given in Tables 3 and 4.

Table 3. Summary Statistics for the Indiana Data (Washington et al., 2011; Lord and Geedipally, 2011)

Variable	Min.	Max.	Average (std. dev)	Total
Number of Crashes (5 years)	0	329	16.97 (36.30)	5737
Average daily traffic over the 5 years (ADT)	9,442	143,422	30,237.6 (28776.4)	--
Minimum friction reading in the road segment over the 5-year period (FRICTION)	15.9	48.2	30.51 (6.67)	--
Pavement surface type (1 if asphalt, 0 if concrete) (PAVEMENT)	0	1	0.77 (0.42)	--
Median width (in feet) (MW)	16	194.7	66.98 (34.17)	--
Presence of median barrier (1 if present, 0 if absent) (BARRIER)	0	1	0.16 (0.37)	--
Interior rumble strips (RUMBLE)	0	1	0.72 (0.45)	--
Segment length (in miles) (L)	0.009	11.53	0.89 (1.48)	300.09

Table 4. Summary Statistics for the Michigan Data (1996) (Lord and Geedipally, 2011)

	Min.	Max.	Average (std. dev)	Total
Number of Crashes (1 year)	0	61	0.68 (1.77)	23168
Annual average daily traffic (AADT)	160	20,994	4,507.5 (3280.6)	--
Segment length (L) (miles)	0.001	54.54	0.18 (0.58)	6212
Shoulder width (in feet) (SW)	0	24	16.94 (5.26)	--
Lane width (in feet) (LW)	8	15	11.22 (0.78)	--
Speed limit (SPEED) (mph)	25	55	52.47 (6.39)	--

The modelling results for the Indiana and Michigan data are summarized in Tables 5 and 6, respectively. The models were assessed using the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002), Mean Absolute Deviance (MAD) (Oh et al., 2003) and the Mean Predicted Square Error (MPSE) (Oh et al., 2003*)*. In both tables, the intercept for the NB-GE needs to be adjusted with the mean of the GE distribution shown in Equation (10), (i.e., $\frac{1}{\lambda}(\psi(\alpha+1)-\psi(1))$) to directly compare it with the intercept of the NB model. This is similar to the adjustment needed for the intercept of the NB-L model. For more details about the adjustment for the NB-L model, the reader is referred to Geedipally et al. (2012). The DIC was calculated based on the characteristics described in Geedipally et al. (2013). Other than the intercept, all coefficients estimated by different models are almost the same.

Table 5. Modeling Results for the Indiana Data.

Variable	NB		NB-L		NB-GE	
	Value	Std. dev	Value	Std. dev	Value	Std. dev
INTERCEPT (β_0)	-4.779	0.979	-3.739	1.115	-3.233	0.423
Ln(ADT) (β_1)	0.7219	0.091	0.630	0.106	0.570	0.067
FRICTION (β_2)	-0.02774	0.008	-0.02746	0.011	-0.0284	0.010
PAVEMENT (β_3)	0.4613	0.135	0.4327	0.217	0.481	0.158
MW (β_4)	-0.00497	0.001	-0.00616	0.002	-0.00651	0.002
BARRIER (β_5)	-3.195	0.234	-3.238	0.326	-3.240	0.324
RUMBLE (β_6)	-0.4047	0.131	-0.3976	0.213	-0.349	0.154
α[1]	0.934	0.118	0.238	0.083	2.339	0.427
λ					1.526	0.284
DIC	1,900		**1,701**		1,784	
MAD[2]	6.91		**6.89**		7.04	
MSPE[3]	206.76		**195.54**		202.93	

[1] α cannot be compared directly between the models; [2] $MAD = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|$;

[3] $MSPE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$

Bold character shows a better fit.

The cumulative residual plot for Indiana data is given in Figure 1. For this residual plot, the values have been adjusted so that the final value is zero. It can be observed that the NB-GE model seems to perform slightly better than NB-L model. Also, both NB-GE and NB-L seem to fit the data better than the NB model. It was also observed that when cumulative residuals are not adjusted, the NB-L and NB-GE models predict the total number crashes for all sites closer than that of the NB model.

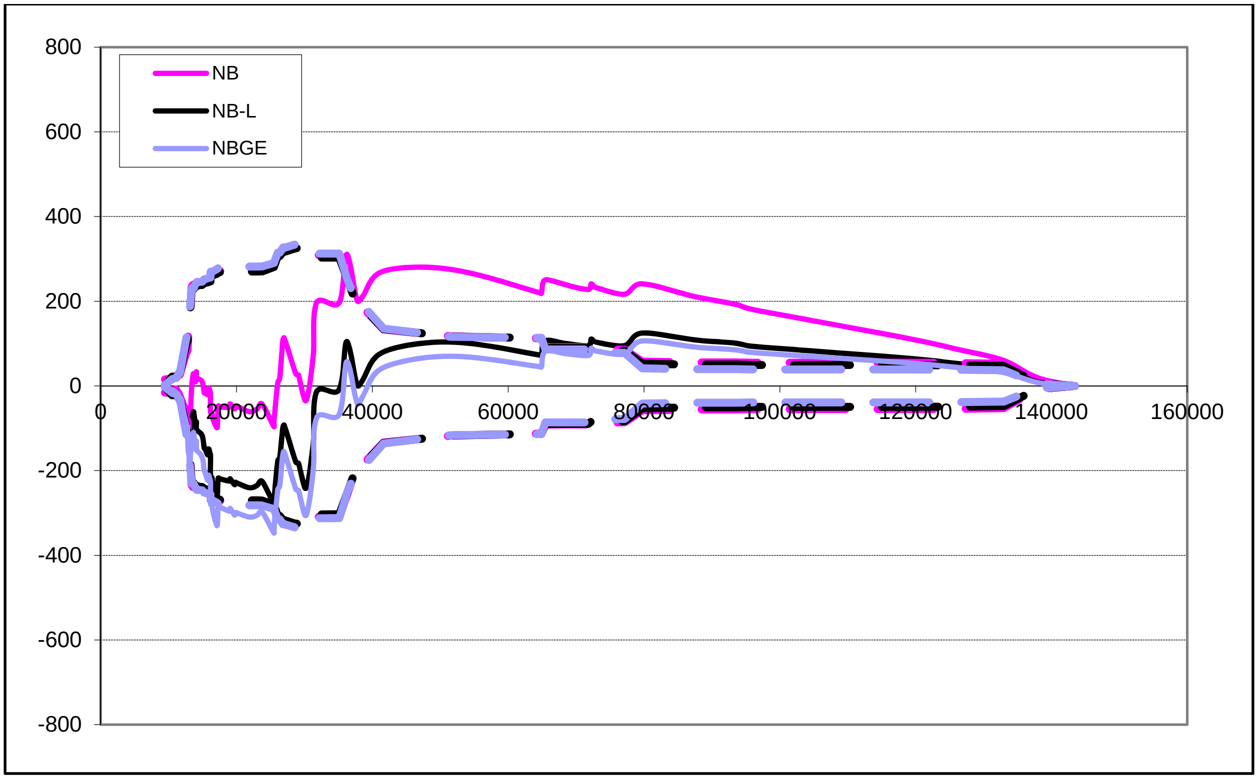

Figure 1. Cumulative Residual Plot for Indiana Data

Table 6. Modeling Results for the Michigan Data.

Variable	NB		NB-L		NB-GE	
	Value	Std. dev	Value	Std. dev	Value	Std. dev
INTERCEPT (β_0)	-3.412	0.239	-3.2607	0.193	-1.58	0.1724
Ln(AADT) (β_1)	0.4267	0.014	0.4243	0.015	0.4212	0.011948
L (β_2)	0.9571	0.009	0.9615	0.009	0.9705	0.0087
SW (β_3)	*-0.00009*	*0.002*	*-0.0003*	*0.002*	*0.0007*	*0.0023*
LW (β_4)	0.0589	0.013	0.0508	0.011	0.0374	0.0105
SPEED (β_5)	0.0098	0.002	0.0091	0.002	0.0071	0.0019
α^1	0.5727	0.019	0.1024	0.002	2.829	0.1462
λ					7.18	1.457
DIC	59,354		**56,046**		57,670	
MAD[2]	0.651		**0.648**		0.6498	
MSPE[3]	**2.831**		2.884		3.089	

Italic character shows that the value is not statistically significant at the 5% level.
Bold character shows a better fit.

[1] α cannot be compared directly between the models; [2] $MAD = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right|$;

[3] $MSPE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$

Figure 2 shows the CURE plots for Michigan data. From the adjusted cumulative residuals, it seems to show that the NB model performs slightly better than the NB-GE and NB-L models. However, when the total number of crashes for all sites is considered, the NB-L model performs much better than the NB model followed by the NB-GE model. The difference between the predicted and observed values -322.1, -232.3, and -3.9 for the NB, NB-GE and NB-L, respectively. Furthermore, the cumulative curve is almost in the negative region for the entire range of the traffic flow values. Recall that the difference for the Zero-Inflated NB was 281.1 for the same dataset (Geedipally et al., 2012).

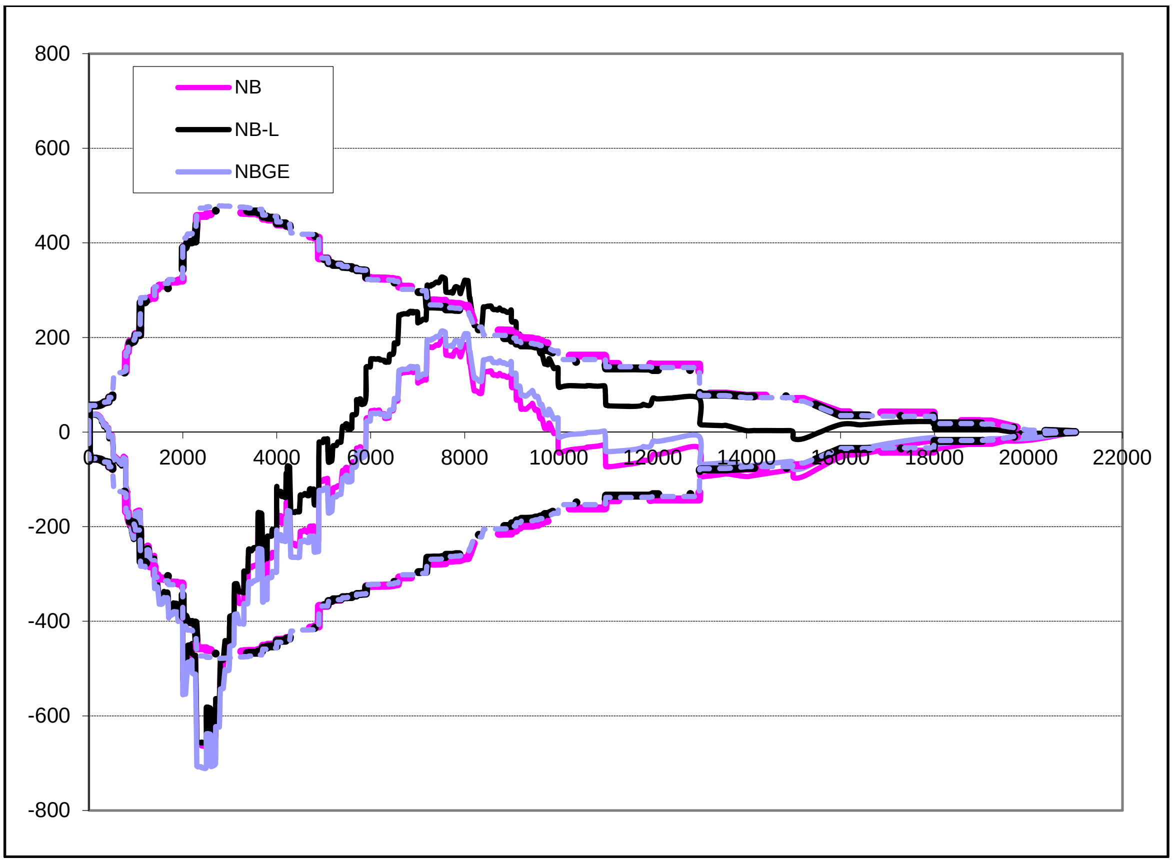

Figure 2. Cumulative Residual Plot for Michigan Data

The results in these two tables show that the NB-GE model provides as good results as for the NB-L, significantly outclass the NB model. Although not shown here, it performs much better than the zero-inflated NB model used on the same data (Lord and Geedipally, 2011). Interestingly, the standard errors for the NB-GE are smaller than those from the NB-L for both datasets. This means that the NB-GE may be capturing a little more of the variance than the NB-L.

The parameterization of the NB-GE used in this work is slightly different than the parameterization described in Aryuyuen and Bodhisuwan (2013), but is similar to the one proposed for the NB-L in Geedipally et al. (2012). When a GLM is considered with the original formulation proposed by

Aryuyuen and Bodhisuwan (2013), the mean response is a non-linear, non-invertible function of the covariates and the parameters, which makes it difficult to characterize the predicted response. On the contrary, the parameterization proposed in this paper is easily interpretable. To understand more about this parameterization, the reader is referred to Geedipally et al. (2012).

Despite the nice interpretability offered by this characterization, MCMC chains could still suffer from poor mixing. This often results from the fact that the $GE(\alpha, \lambda)$ distribution behaves quite differently for α <=1 and α >1. Moreover, the mean of the GE tends to infinity as α increases for a given fixed value of λ. This problem can be mitigated by restricting the parameter α between some range. In this paper, a prior of the form α ~ uniform (1, 3) was used. In addition, when the α and λ of the GE distribution were varied and the α was plotted against the density f(x), the following observations were made: 1) for $\alpha < 0.5$, f(x) is concentrated around 0; 2) for $\alpha > 1$, f(x) has different support depending on the scale parameter, is unimodal. Thus, the range (1, 3) for α and (1, 2) for λ was considered reasonable.

Since the GE distribution currently exists in the "OpenBUGS" software, the NB-GE model estimation can be easily implemented. However, the computational time for MCMC runs was slightly longer than the NB model because it involves an additional parameter compared to the NB model. However, the difference in computational times between the two models was not very large.

SUMMARY AND CONCLUSION

This paper has described the development and application of the NB-GE distribution and GLM for analyzing crash data characterized by a large tails and/or an excess of zeros. The NB-GE was compared to the Poisson, NB and the recently introduced three-parameter model NB-L using a total of four datasets. The analyses were carried using four different datasets and the distribution and GLM were estimated in R and OpenBUGS.

The study results show that the NB-GE and NB-L distributions can fit such data sets better than the NB distribution. The same scenario was noted for the GLM. The NB-GE is also easier to implement than the NB-L, which may make this model more useful. Further work is nonetheless still needed to determine when the NB-GE (and NB-L) are needed depending on the characteristics of the data (e.g., level of dispersion, skewness, proportion of zeros, etc.), as suggested by Lord and Geedipally (2014).

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Soma S. Dhavala for providing valuable information about coding the NB-GE in OpenBUGS; Dr. Xiao Qin from South Dakota State University for providing us with the Michigan data and Dr. Fred Mannering for providing the Indiana data.

REFERENCES

Aryuyuen. S., Bodhisuwan, W., 2013. Negative Binomial-Generalized Exponential (NB-GE) Distribution. Applied Mathematical Sciences 7 (22), 1093-1105.

Geedipally, S.R., Lord, D., Dhavala, S.S., 2012. The Negative Binomial-Lindley Generalized Linear Model: Characteristics and Application using Crash Data. Accident Analysis & Prevention 45 (2), 258-265.

Geedipally, S.R., Lord, D., Dhavala, S.S., 2013. A caution about using the Deviance Information Criterion while modeling traffic crashes. Safety Science 62, 495-498.

Gupta, R.D., Kundu, D., 1999. Generalized exponential distributions, Austral. New Zealand J. Statist., 41, 173–188.

Hallmark, S.L., Qiu, Y., Pawlovitch, M., McDonald, T.J., 2013. Assessing the Safety Impacts of Paved Shoulders. Journal of Transportation Safety and Security 5 (2), 131-147.

Lord, D., Geedipally, S.R., 2011. The Negative Binomial-Lindley Distribution as a Tool for Analyzing Crash Data Characterized by a Large Amount of Zeros. Accident Analysis & Prevention 43 (5), 1738-1742.

Lord, D., Geedipally, S.R., 2014. Models for Analyzing Count Data with a Lot of Zeros. Annual Joint Statistical Meeting, Session 589, American Statistical Association, Boston, MA.

Lord, D., Geedipally, S.R., Persaud, B.N., Washington, S.P., van Schalkwyk, I., Ivan, J.N., Lyon, C., Jonsson, T., 2008. Methodology for estimating the safety performance of multilane rural highways. NCHRP Web-Only Document 126, National Cooperation Highway Research Program, Washington, D.C.

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

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

Lord, D., Washington, S.P., Ivan, J.N., 2007. Further Notes on the Application of Zero Inflated Models in Highway Safety. Accident Analysis & Prevention 39 (1), 53-57.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. WinBUGS -- a Bayesian modelling framework: concepts, structure, and extensibility. Statistics and Computing 10, 325-337.

Mannering, F., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Analytic Methods in Accident Research 1, 1-22.

Oh, J., Lyon, C., Washington, S.P., Persaud, B.N., Bared, J., 2003. Validation of the FHWA Crash Models for Rural Intersections: Lessons Learned. Transportation Research Record 1840, 41-49.

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

Shankar, V.N., Ulfarsson, G.F., Pendyala, R.M., Nebergal, M.B., 2003. Modeling crashes involving pedestrians and motorized traffic. Safety Science 41 (7), 627-640.

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

Washington, S., Karlaftis, M.G., Mannering, F., 2011. Statistical and Econometric Methods for Transportation Data Analysis. Second edition, Chapman and Hall/CRC, Boca Raton, FL.

Zou, Y., Lord, D., Zhang, Y., Peng, Y, 2013. Comparison of Sichel and negative binomial models in estimating empirical Bayes estimates. Transportation Research Record 2392, 11-21.