Use of Mixed Distribution Generalized Linear Models to Quantify Safety Effects of Rural Roadway Features

M. Razaur Rahman Shaon and Xiao Qin

A challenge in modeling crash frequency is an excess of sites with no crashes, few sites with a large number of crashes, or both. When there are excess zeros in the data or when the variance of the response is greater than the mean, the data are overdispersed. Recently, a few promising modeling techniques, such as the negative binomial–Lindley (NB-L) and negative binomial–generalized exponential (NB-GE) mixed distribution generalized linear models (GLMs), have been developed to handle count data overdispersion while keeping the core strength of the NB model. This study expanded the discussion on NB-L and NB-GE GLMs by focusing on their capability for modeling crash data as well as quantifying the safety impact of crash contributing factors. The mixed distribution models along with the conventional NB model were applied to a rural two-lane, two-way highway data set. The results showed that both NB-L and NB-GE GLMs could yield results similar to those of the NB model in addition to having mixed distribution probabilities to account for overdispersion. All modeling approaches successfully estimated the combined effects of lane width and shoulder width and identified the same combination with the optimal safety benefits. Both NB-L and NB-GE can be considered viable alternatives for the NB model if better goodness of fit is desired.

Modeling of crash data is important to highway safety. Crash modeling can be used to identify the contributing factors in crash occurrence, to predict future crashes, or to rank crash hot spots for safety treatments. Crash data are often characterized by extra variance in crash occurrence, which can be attributed to the unaccounted variation across sites (*1*). In statistical terms, when the sample variance is significantly greater than the sample mean, it is called data overdispersion. Failing to account for data overdispersion can lead to biased parameter estimates. Many sites with zero crashes along with a long crash tail can create highly dispersed data. A "long crash tail" refers to a right-skewed crash distribution in which a series of sites with very high crash counts exists in a crash data set with very small probabilities. It was noted in the literature that the traditional models, such as Poisson, negative binomial (NB), and Poisson lognormal, cannot effectively handle highly dispersed data sets because of inherit limitations (*1–3*).

Considering the characteristics of crash count data, researchers have proposed novel statistical methods to address data issues, such as outliers, data heterogeneity, multicollinearity, interactions between variables, and excessive zeros. Lord and Mannering documented many of the issues arising in the crash data set and summarized the regression models proposed by researchers in previous literature (*3*). Mannering and Bhat updated the list of methodologies for analyzing crash data afterward (*2*). Of these methodologies, researchers proposed the zero-inflated Poisson (ZIP) and zero-inflated NB (ZINB) models to overcome data overdispersion with excess zeros (*4–6*). But this kind of modeling assumes that data belong to two states: the zero or safe state and the nonzero state. The zero state is desirable but practically impossible because no site is absolutely crash free. This is an important methodological limitation, although the model provides a better fit for the crash data (*7, 8*). To address some of the criticism associated with zero-inflated models, Malyshkina and Mannering proposed the zero-state Markov switching count model (*9*). This methodology allows individual roadway segments to switch between zero and normal-count states over time. One of the important advantages of this Markov switching approach is that it allows for direct statistical estimation of the specific roadway segment state, whereas traditional zero-inflated models do not. When crash data are characterized by excess zeros with a long tail, the mixed modeling approaches such as the NB-Lindley (NB-L), NB-generalized exponential (NB-GE), and Sichel generalized additive models for location, scale, and shape (GAMLSS) generalized linear regression models (GLMs) usually perform better than traditional NB, ZIP, and ZINB models (*10–12;* Zou et al., unpublished work, 2012). The Sichel GAMLSS is formulated with the Sichel distribution, also known as the Poisson generalized inverse Gaussian distribution, with a four-parameter GAMLSS framework. The main benefit of both NB-L and NB-GE GLM approaches is that they maintain the traditional NB characteristics while handling excess data dispersion.

Recent development of mixed GLMs such as NB-L and NB-GE has encouraged researchers to apply novel and sophisticated statistical methods to model nonnormality in the crash data. The study presented in this paper expanded the discussion of NB-L and NB-GE GLMs as viable modeling alternatives for crash data from the theoretical perspective as well as their ability to quantify the impact of crash contributing factors by applying them for rural two-lane, two-way highways.

METHODOLOGY

Two mixed GLMs, NB-L and NB-GE, are introduced with their implementation procedures because neither modeling technique is readily available in any of the commercial statistical software packages.

Department of Civil and Environmental Engineering, University of Wisconsin–Milwaukee, P.O. Box 784, Milwaukee, WI 53201-0784. Corresponding author: X. Qin, qinx@uwm.edu.

Transportation Research Record: Journal of the Transportation Research Board, No. 2583, Transportation Research Board, Washington, D.C., 2016, pp. 134–141.
DOI: 10.3141/2583-17

NB-L GL Model

Lindley introduced the probability density function for the Lindley distribution (*13*). Because of the popularity of exponential family distributions, the Lindley distribution was overlooked in the previous literature (*14*). The Lindley distribution is a mixture of exponential and gamma distribution. The NB-L distribution is a mixture of NB and Lindley distributions. This mixed distribution works well when the data set contains many zeros or with highly dispersed data. Ghitany et al. showed that Lindley is a better distribution than exponential distribution (*14*). The probability mass function and maximum likelihood estimation equations are referred to in the study conducted by Zamani and Ismail (*15*).

Geedipally et al. applied NB-L distribution in the GLM context, where the NB-L distribution can also be reparameterized as follows (*11*):

$$P(Y = y, \mu, \Phi, \theta) = \int \text{NB}(y; \Phi, \varepsilon\mu)\,\text{Lindley}(\varepsilon; \theta)\,d\varepsilon \tag{1}$$

Equation 1 illustrates that $\varepsilon\mu$ is the mean of variable Y following the NB distribution, and ε follows the Lindley distribution.

For modeling crash count data, the NB or Poisson-gamma mixture model (NBGLM) is the most representative statistical method used in previous studies. For modeling the mean response for crash count, the most commonly used functional form of NBGLM is the log-link function. Now, if it is assumed that the crash count follows an NB-L (Φ, p) distribution, the mean response function can be structured as follows:

$$E(Y = y) = \mu \times E(\varepsilon) \tag{2}$$

where

$$\mu = e^{\beta_0 + \sum_{i=1}^{q} \beta_i X}$$

$$E(\varepsilon) = \frac{\theta + 2}{\theta(\theta + 1)}$$

Replacing the value of μ and $E(\varepsilon)$, the mean response function can be written as follows:

$$E(Y) = \left(e^{\beta_0 + \sum_{i=1}^{q} \beta_i X} \right) \times \frac{\theta + 2}{\theta(\theta + 1)} = e^{\left\{\beta_0 + \log\left[\frac{\theta+2}{\theta(\theta+1)}\right]\right\} + \sum_{i=1}^{q} \beta_i X} = e^{\beta_0' + \sum_{i=1}^{q} \beta_i X} \tag{3}$$

where

$$\beta_0' = \beta_0 + \log\left[\frac{\theta + 2}{\theta(\theta + 1)}\right]$$

In the literature, researchers presented the NB-L distribution by using a stochastic representation (*12, 15*). The Lindley distribution is not a standard distribution. It is a mixed distribution of gamma and exponential distribution, which can be written in the following structure (*15*):

$$\varepsilon \sim \frac{1}{1+\theta}\text{gamma}(2, \theta) + \left(1 - \frac{1}{1+\theta}\right)\exp(\theta) \tag{4}$$

The exponential distribution (θ) can also be written as gamma distribution (r, θ). If $r = 1$, then the gamma distribution can be equal to exponential distribution. Rewriting Equation 4 will look like

$$\varepsilon \sim \frac{1}{1+\theta}\text{gamma}(2, \theta) + \left(1 - \frac{1}{1+\theta}\right)\text{gamma}(1, \theta) \tag{5}$$

The Lindley distribution needs to be reparameterized for easy interpretability in the NB-L GLM. The mixture of two gamma distributions can be derived to be rewritten with Bernoulli distribution. Assume a random variable z that follows a Bernoulli distribution. Then the special mixed structure can be written as follows:

$$\varepsilon \sim \frac{1}{1+\theta}\text{gamma}(1+1, \theta) + \left(1 - \frac{1}{1+\theta}\right)\text{gamma}(1+0, \theta)$$

$$\varepsilon \sim \sum \text{gamma}(1+z, \theta)\,\text{Bernoulli}\left(z; \frac{1}{1+\theta}\right) \tag{6}$$

With the newly developed structure for Lindley distribution, the NB-L distribution can be written as following a multilevel hierarchical structure:

$$P(Y = y, \mu, \Phi|\theta) = \text{NB}(y; \Phi, \varepsilon\mu)$$

$$\varepsilon \sim \text{gamma}(\varepsilon; 1+z, \theta)$$

$$Z \sim \text{Bernoulli}\left(z; \frac{1}{1+\theta}\right) \tag{7}$$

In previous studies, researchers used Bayesian interface to implement this model due to the hierarchical structure of NB-L GLM (*11, 16*). In the presented structure, the crash count follows an NB distribution which is conditional on a site-specific frailty term. The site-specific frailty term, ε, was assumed to accommodate additional data heterogeneity in crash data. It is necessary to specify prior distribution for the parameters to obtain the Bayesian estimate. Prior distributions are meant to reflect prior knowledge about the parameters of interest. The site-specific frailty term follows an uninformative prior of gamma distribution. The shape parameter in the gamma distribution also follows a Bernoulli distribution that depends on $z = 1/1 + \theta$. In previous work, Geedipally et al. used a beta prior to define z in Bayesian interface (*11*). But in a study conducted by Hallmark et al., rather than specifying the priors for z, the authors directly specified the prior for θ with a weakly informed prior that follows gamma distribution (*16*). It could be difficult for a user to choose between the two approaches to implement an NB-L model.

To build a generalized model, the NB-L model structure must be developed so that it can perform better than or similar to the NB model with or without preponderant zero crashes or with or without a long crash tail. From the hierarchical structure, the formulation can be seen as adding a site-specific offset term in the log-transformed domain of the mean response of NB distribution. Use of a weakly informative prior may yield a model output in which the parameter estimate for Lindley distribution may have a greater contribution to crash prediction than NB distribution. Markov chain Monte Carlo (MCMC) can also suffer from poor mixing because of a correlation between the intercept and the Lindley term (Equation 3). It has been noted in the literature that if prior information is available, it should be used to formulate the informative priors (*17, 18*). To limit the contribution of the mixed effect from Lindley distribution, a prior

should be used to ensure $E(\varepsilon) = 1$. That is, the first moment of Lindley distribution, $\theta + 2/\theta(\theta + 1) = 1$, which yields to an estimate of θ, is equal to 1.41. For limiting the θ value, Geedipally et al. suggested to use a prior for $1/1 + \theta$ that follows a beta distribution, and the reasonable choice for prior distribution is Beta (n/3, n/2), where n is the total observations.

Generalized Exponential Distribution

The generalized exponential distribution was introduced by Gupta and Kundu as an alternative to three-parameter Weibull distribution and three-parameter gamma distribution (*19*). The generalized exponential distribution can coincide with the exponential distribution when $\alpha = 1$. The probability mass function and the moment generating equation for generalized exponential distribution are given in studies conducted by Gupta and Kundu and Aryuyuen and Bodhisuwan (*19, 20*). By differentiating the logarithm of moment function, the mean and variance for generalized exponential distribution can be obtained. The mean and variance can be written as follows:

$$E(Z) = \frac{1}{\beta}(\Psi(\alpha + 1) - \Psi(1)) \tag{8}$$

$$\text{var}(Z) = \frac{1}{\beta^2}(\Psi'(\alpha + 1) - \Psi'(1)) \tag{9}$$

where $\Psi(.)$ is the digamma function and Ψ' is the derivative of the digamma function.

Introduced by Aryuyuen and Bodhisuwan, the NB-GE distribution is a mixture of NB and GE distributions (*20*). In the GLM context, the NB-GE distribution can also be reparameterized as a mixture of NB and generalized exponential distribution (*10*) and formulated in the following structure:

$$P(X = x, \mu, \Phi, \theta) = \int \text{NB}(x; \Phi, z\mu)\text{GE}(z; \alpha, \beta)dz \tag{10}$$

Similar to the NB-L GLM, Equation 1 illustrates that $z\mu$ is the mean of variable Y following the NB distribution, and z follows the generalized exponential distribution. Now, assuming that the crash count follows an NB-L (Φ, p) distribution, the mean response function can be structured as follows:

$$E(X) = \mu \times E(z) = \left(e^{\beta_0 + \sum_{i=1}^{q} \beta_i X} \right) \times \frac{1}{\beta}(\Psi(\alpha + 1) - \Psi(1)) \tag{11}$$

The NB-GE GLM works similarly to NB-L distribution. The site-specific frailty term was assumed to accommodate data heterogeneity that follows generalized exponential distribution. The implementation of the NB-GE GLM with crash count data can be found only in a study conducted by Vangala et al. (*10*). Unlike Lindley distribution, the generalized exponential family is already available in the OpenBUGS library (*21*). The only point of concern to applying the NB-GE model is to define informative priors for two parameters of generalized exponential distribution. Vangala et al. noted that restricting the parameters α and β between some range can improve the mixing in an MCMC and also elicit the expected value of generalized exponential distribution near 1 (*10*). The authors suggested limiting the value of α and β to between 1 and 3. Use of a uniform prior from 1 to 3 can be reasonable to limit the parameters of generalized exponential distribution.

Both NB-L and NB-GE GLMs have additional features that can accommodate data heterogeneity, and both yield results similar to those of the NB model if data heterogeneity is not a concern. The NB-L and NB-GE models can be considered as random intercept models as the site-specific frailty term varies from site to site. According to the amount of data dispersion, the posterior mean of frailty term can adjust the mean estimate of NB distribution. The additional parameters in NB-L and NB-GE GLMs can provide a better fit for overdispersed data. From a model design perspective, these two features give an advantage to NB-L and NB-GE GLMs in modeling complex crash data.

DATA DESCRIPTION

A data set from the South Dakota Department of Transportation roadway, traffic, and accident database was used to establish the objective of this study. The roadway geometric and traffic features were available in the roadway inventory system (RIS). Multiple event tables from RIS were joined to generate homogeneous segments. The accident data set for 2008 to 2014 was collected from South Dakota accident records system. The crash data were joined with roadway data according to their spatial distance. The focus of this study was rural two-lane, two-way highway segments. The whole data set was divided into training and validation data sets. The training data set consisting of crash data from 2008 to 2012 was used to develop a crash prediction model, and the validation data set consisting of crash counts from 2013 and 2014 was used to compare prediction accuracy between models. The final data set contained 16,828 sites of rural two-lane, two-way highway segments with a cumulative segment length of 6,361.53 mi. A total of 77.8% of the sites experienced no crashes during the 5-year period in the training data set. Descriptive statistics of the variable used for model development are provided in Table 1.

THIRTEEN CONTROLLING CRITERIA

The Green Book recommends safe and efficient practices for the design of roadways on the basis of extensive research and study (*22*). After a technical review of the adopted minimum criteria in the Green Book, FHWA identified 13 criteria needing special attention, referred to as the 13 controlling criteria (*23*). The 13 criteria have substantial importance for operational and safety performance of any highway:

1. Design speed,
2. Lane width,
3. Shoulder width,
4. Bridge width,
5. Horizontal alignment,
6. Superelevation,
7. Vertical alignment,
8. Grade,
9. Stopping sight distance,
10. Cross slope,
11. Lateral offset to obstruction,
12. Structural capacity, and
13. Clearance.

TABLE 1 Variable Summary Statistics of Training Data Set

Variable	Description	Mean	Minimum	Maximum	SD
Crash	Crash count	0.619	0	88	2.382
AADT	Annual average daily traffic over 5 years	920.797	45	21,396	913.721
Seg_Length	Segment length in miles	0.378	0.01	16.494	1.015
Lane_Wid	Lane width in feet	12.956	9	30	2.109
Shoulder_Wid	Average shoulder width in feet	3.052	0	15	2.564
Speed_Limi	Speed limit	57.342	20	65	10.653
R_Mile	Radius of curvature in miles	0.082	0.01	1.084	0.186
Hcur_Len	Horizontal curve length	0.050	0.01	1.025	0.114
Grade	Grade percentage	1.197	3	15	1.977
Rumble_Strip	Yes (71.1%) No (28.9%)				
Curve_Flag	Yes (8.0%) No (92.0%)				
Grade_Flag	Yes (4.5%) No (95.5%)				

NOTE: Categorical variables are presented as percentage of total highway miles.

AASHTO evaluated the contribution of the 13 controlling criteria for geometric design (*24*). The safety effects for the controlling criteria on rural two-lane highways provided in this study were carefully reviewed. This study considered whether the mixed distribution models with complex mathematical structure can effectively identify the statistically significant association between crash frequency and any of these variables.

For choosing the statistically significant variables for model development, the correlation matrix was investigated for the data set, followed by the NB model that was developed as benchmarks. According to the correlation matrix results and the NB model output, speed limit, grade percentage, horizontal curve length, and rumble strip variables were discarded because no statistically significant relationship was found between crash count and these variables. Of all available explanatory variables, some of the 13 controlling criteria were found not to be statistically significant in the final model. This finding underscores the fundamental difference in the concept of nominal safety (whether a roadway, design alternative, or design element meets minimum design criteria) and substantive safety (actual or expected safety performance of a highway, usually measured by crash data).

To investigate the design exceptions and their impacts on a rural two-lane highway, the lane width, the shoulder width, and their interaction were explored. Both lane width and shoulder width variables were divided into four levels as recommended in the *Highway Capacity Manual* (*25*), and the level designations are provided in Table 2.

RESULTS AND DISCUSSION

Table 3 summarizes the results of the final model for NB, NB-L, and NB-GE modeling approaches. The segment length variable was considered as an offset. To check statistical significance of parameter estimates, the 95% credible interval for each parameter estimate was reviewed. The 95% credible interval of each parameter estimate does not include zero. A total of three Markov chains were used in each model estimation process. For each chain, 30,000 iterations were used. The first 15,000 iterations were used as burn-in samples for estimating the model parameters (these were discarded). The Gelman–Rubin convergence statistics (G-R statistics) were reviewed to verify the model convergence. Mitra and Washington recommended that the convergence is achieved when the G-R statistic is less than 1.2 (*1*). For assessing overall model prediction accuracy of applied models in a Bayesian framework, Krnjajić and Draper suggested to evaluate the log scoring criterion (*26–28*). The formulation of a cross-validated version of log score (LS_{CV}) is presented in the following equation:

$$\mathrm{LS}_{\mathrm{CV}}\left(M_j \middle| y\mathcal{B}\right) = \frac{1}{n}\sum_{i=1}^{n} \log\, p\left(y_i \middle| y_{-i} M_j \mathcal{B}\right) \tag{12}$$

where

y_{-i} = y-vector (response variable) with ith observation omitted,
M_j = model specification, and
$\mathcal{B}$ = set of propositions (true–false statements) summarizing background information.

TABLE 2 Lane Width and Shoulder Width Level and Designations

Level	Level Designation	Site Count
Lane Width		
≤10 ft	A	996
(10, 11] ft	B	886
(11, 12] ft	C	8,253
>12 ft	D	6,693
Shoulder Width		
[0, 2] ft	P	7,808
(2, 4] ft	Q	4,509
(4, 6] ft	R	2,822
>6 ft	S	1,689

TABLE 3 Final Model Results for Rural Two-Lane Highways

	NB		NB-L		NB-GE	
Variable	Mean	SD	Mean	SD	Mean	SD
Intercept	–3.12	0.236	–2.394	0.337	–2.294	0.215
log(AADT)	0.7229	0.031	0.618	0.038	0.663	0.022
R_mile	–1.578	0.144	–1.591	0.192	–1.551	0.155
Curve_Flag	–0.889	0.064	–0.894	0.092	–0.879	0.07
Lane_Wid:B	–0.489	0.129	–0.516	0.184	–0.517	0.141
Lane_Wid:C	–0.366	0.081	–0.343	0.147	–0.358	0.102
Lane_Wid:D	–0.364	0.079	–0.307	0.142	–0.357	0.098
SW:Q	–1.703	0.406	–1.403	0.253	–1.361	0.481
SW:R	–0.585	0.196	–0.245	0.211	–0.547	0.235
SW:S	–0.512	0.781	–0.329	0.688	–0.852	0.345
LW_B:SW_Q	1.615	0.646	1.335	0.595	1.272	0.70
LW_B:SW_R	0.913	0.364	0.647	0.411	0.884	0.389
LW_B:SW_S	0.539	0.868	0.487	0.804	1.014	0.517
LW_C:SW_Q	1.587	0.407	1.311	0.26	1.244	0.485
LW_C:SW_R	0.602	0.199	0.318	0.216	0.582	0.247
LW_C:SW_S	0.495	0.787	0.404	0.678	0.892	0.34
LW_D:SW_Q	1.636	0.408	1.354	0.254	1.322	0.481
LW_D:SW_R	0.549	0.194	0.296	0.224	0.549	0.237
LW_D:SW_S	0.51	0.788	0.382	0.486	0.898	0.343
Inverse–dispersion	2.239	0.134			8.929	0.816
Dispersion			0.104	0.003		
Θ			1.498	0.029		
Alpha					2.888	0.097
Lambda					2.828	0.167
DIC	22,090		21,090		21,550	
Dbar	22,070		19,010		20,060	
pD	19.8		2,080		1,488	
LS_{FS}	–1.85		–1.79		–1.82	

NOTE: Dispersion and inverse–dispersion are not directly comparable because of different specifications in model structure.

For a large sample size, LS_{CV} can be computationally expensive as it requires n separate MCMC runs after each observation is omitted. In the setting in which predictive distribution is not available in closed form, it can also be computationally expensive (*28*). Draper and Krnjajić suggested another form of log scoring that omits the leave-one-out idea, called the full sample log score (LS_{FS}) (*28*), which is formulated in Equation 13. The model with the higher log score is considered to have better prediction accuracy and better small sample model discrimination ability.

$$LS_{FS}\left(M_j \middle| y\mathcal{B}\right) = \frac{1}{n}\sum_{i=1}^{n} \log p\left(y_i \middle| y\ M_j\mathcal{B}\right) \tag{13}$$

The coefficient estimates in Table 3 show that (*a*) the intercept is quite different among three models because a site-specific frailty term was considered as an additional offset in both NB-L and NB-GE modeling, (*b*) the main effect estimates have the same sign and the mean value are similar, and (*c*) the interactions have rather different mean values. The estimated coefficient for lane width shows that with the increase in lane width, the reduction in crash occurrence gets smaller. The same trend is observed for shoulder width. For considering the combined effect of lane width and shoulder width, the combined mean was calculated from the main and interaction terms, as shown in Table 4.

The results in Table 4 consider a baseline of lane width less than or equal to 10 ft and shoulder width of [0, 2] feet. The combined interaction coefficients suggest there is a reduction in crash occurrence with the increase in lane width from 10 ft and shoulder width from [0, 2] feet. To better understand the effect of the lane width–shoulder width combination, pseudoelasticity was computed. The pseudoelasticity is the percent increase in the crash frequency caused by the change in indicator variable, which can be formulated by Equation 14 (*29*):

$$E_{x_{ik}}^{\lambda_i} = \frac{\exp(\beta_k) - 1}{\exp(\beta_k)} \tag{14}$$

where β_k is the estimated model coefficient for indicator variable x_{ik}.

The estimated pseudoelasticity for the combination of lane width and shoulder width is presented in Figure 1.

TABLE 4 Combined Lane Width–Shoulder Width Interaction Coefficients

Shoulder Width	Lane Width		
	B: (10, 11] ft	C: (11, 12] ft	D: >12 ft
NB			
Q: (2, 4] ft	–0.578	–0.482	–0.431
R: (4, 6] ft	–0.162	–0.349	–0.400
S: >6 ft	–0.463	–0.383	–0.366
NB-L			
Q: (2, 4] ft	–0.584	–0.435	–0.356
R: (4, 6] ft	–0.114	–0.270	–0.256
S: >6 ft	–0.357	–0.267	–0.254
NB-GE			
Q: (2, 4] ft	–0.606	–0.475	–0.397
R: (4, 6] ft	–0.181	–0.323	–0.356
S: >6 ft	–0.355	–0.318	–0.312

NOTE: Combined interaction coefficient = coefficient for lane width level + coefficient for shoulder width level + interaction coefficient between lane width and shoulder width.

In Figure 1, the pseudoelasticity estimates are shown with a bar chart for each modeling approach. The pseudoelasticity estimates of the interaction effect suggest that a combination of a lane width of (10, 11] feet and shoulder width of (2, 4] feet has the maximum reduction in crash occurrence. Another important point is that with the same lane width, an increase in shoulder width to the next level of (4, 6] feet has the lowest reduction in crash occurrence within all interactions. Increasing lane width and shoulder width for a rural two-lane highway may not add safety benefits. In this study, the results suggest that a roadway with 11-ft lane width and 4-ft shoulder width has optimal safety benefits. Despite different coefficient trends between models, all three modeling approaches can identify the same lane width–shoulder width combination for optimal benefits.

A comparison of the findings with previous literature showed that the results conform with the study conducted by Lee et al. that found that the logarithm of crash rate was the highest for 12-ft lanes and lower for lane width less than or greater than 12 ft (*30*). Qin et al. found that there is an increase in single-vehicle crash occurrence with the increase in lane width on Michigan rural two-lane highways (*5*). In previous studies, the authors found that crash frequency usually drops with an increase in lane width (*31–33*), plausibly because of a larger separation between vehicles in adjacent lanes. However, the larger separation resulting from wider lanes may make drivers feel safe and thus increase their speed. Hauer suggested that a larger separation between vehicles tends to increase

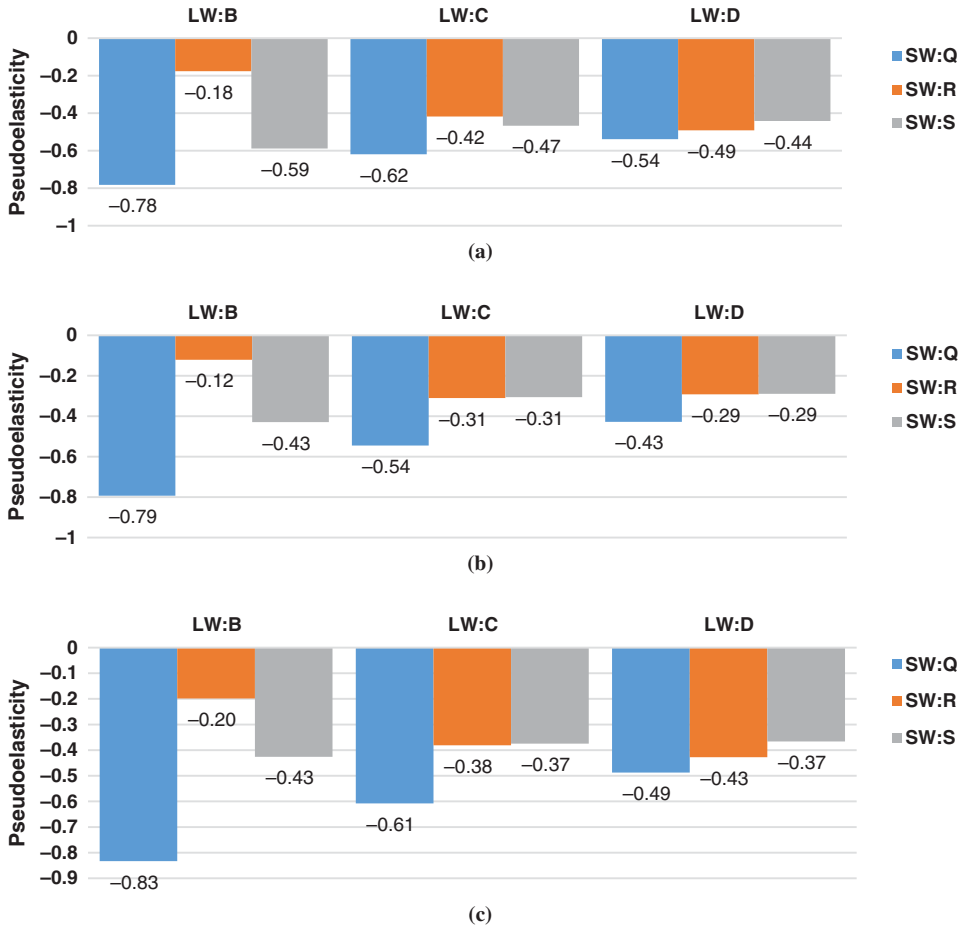

FIGURE 1 Comparison of pseudo-elasticity between models: (*a*) NB, (*b*) NB-L, and (*c*) NB-GE.

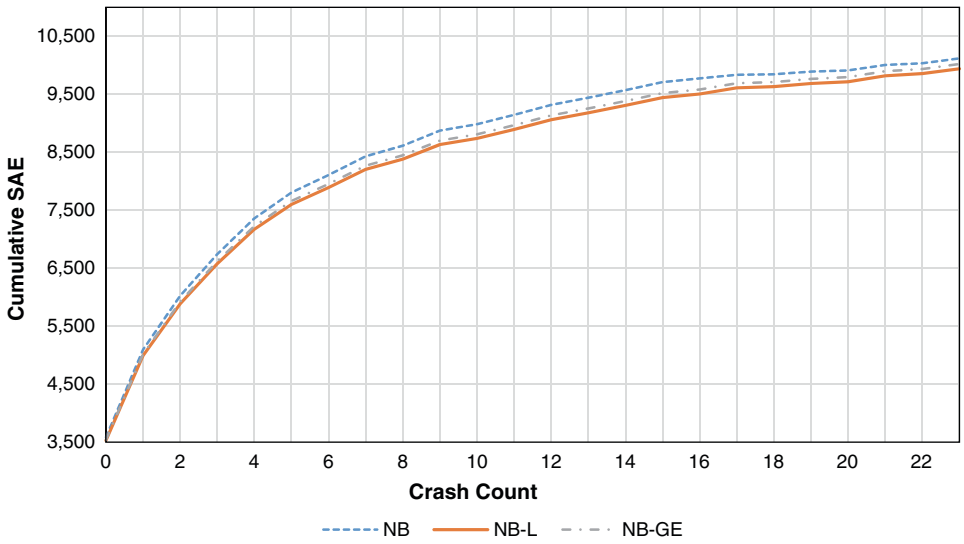

FIGURE 2 Cumulative sum of squares error plot with crash count.

speed and reduce spacing between vehicles, which may contribute to higher crash occurrences (*34*). Driver behavior influences, complicates, and sometimes compromises the safety outcome of the countermeasures, as reflected in an FHWA-sponsored study in which the safety effectiveness of a lane width and shoulder width combination on rural two-lane undivided roads was evaluated (*35*). The authors compared the safety effectiveness of various allocations of total paved width into lane width and shoulder width by computing the corresponding crash modification factor. The authors concluded that there is no definitive trend in the relationship between crash frequency and the lane width–shoulder width combination.

The results provided in Table 3 contain a model comparison that uses the deviance information criterion (DIC). Geedipally et al. noted that the model parameterization can influence the estimation of DIC value, and comparisons with DIC should be done only between models that have similar parameterization (*36*). The authors also recommended that both NB-L and NB-GE models be compared with the NB model for similar parameterization (*10, 11, 36*). The comparison of DIC value between models shows that the NB-L model performs better than the traditional NB and NB-GE models. The DIC value consists of two components: (*a*) measures of how well the model fits the data (Dbar) and (*b*) a measure of model complexity (pD). The Dbar estimates illustrates that NB-L has superior data fitting than the NB-GE and traditional NB models. The pD measure explains the magnitude of complexity for estimating the parameters. The mix distributions add a significant amount of effective numbers of parameters while implementing the model. A comparison of pD shows that the NB-L GLM has the maximum effective number of parameters among all GLMs. Comparing LS_{FS} shows a similar trend to DIC. Krnjajić and Draper showed that for a fixed-effect modeling approach, DIC and log scores are negatively correlated and motivation of LS coincides with the goal of DIC (*26*). In comparisons of random-effects models, LS may have a different trend from DIC.

For the posterior mean of coefficient to predict crashes, the overall model prediction accuracy must be checked. The validation data set was used to compare prediction accuracy between models. The validation data set consists of crashes that occurred in 2013 and 2014 on the rural two-lane, two-way state highway system in South Dakota. Figure 2 shows the cumulative sum of absolute error (SAE) for NB, NB-L, and NB-GE modeling approaches with the validation data set. In this plot, the *x*-axis represents crash count and the *y*-axis represents cumulative SAE for each crash count.

Figure 2 shows that for both NB-L and NB-GE models, the cumulative SAE was smaller than NB until a certain level of crash counts. With high crash counts, the cumulative SAE value stabilized to the total SAE, which is almost similar for all three modeling approaches.

CONCLUSION

Improving crash prediction accuracy is a challenge for transportation professionals. This study investigated the implementation procedure of sophisticated and complex mathematical structured NB-L and NB-GE GLMs with a two-lane, two-way rural highway crash data set. The model results were compared with the NB model. It was found that both NB-L and NB-GE GLMs not only maintained the strength of NB distribution but also accounted for data overdispersion with excessive number of zeros. A comparison of DIC and LS_{FS} showed that both NB-L and NB-GE provide better statistical goodness of fit than the NB model. The cumulative SAE comparison also showed that both NB-L and NB-GE have a smaller SAE.

A discussion of estimated coefficients for lane width–shoulder width interactions among models underscored the important concept of substantive safety as well as its application to the design exceptions. The coefficient estimates suggested that increasing lane width or shoulder width on rural two-lane highways might not add safety benefits given an optimal combination of lane width and shoulder width. Although the NB model is adequate for modeling overdispersed data under many circumstances, both NB-L and NB-GE can be considered as viable alternatives if there are preponderant zeros in the data and better goodness of fit is desired.

REFERENCES

1. Mitra, S., and S. Washington. On the Nature of Over-Dispersion in Motor Vehicle Crash Prediction Models. *Accident Analysis and Prevention,* Vol. 39, No. 3, 2007, pp. 459–468.

2. Mannering, F.L., and C.R. Bhat. Analytic Methods in Accident Research: Methodological Frontier and Future Directions. *Analytic Methods in Accident Research,* Vol. 1, 2014, pp. 1–22.
3. Lord, D., and F. Mannering. The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. *Transportation Research Part A,* Vol. 44, No. 5, 2010, pp. 291–305.
4. Shankar, V., J. Milton, and F. Mannering. Modeling Accident Frequencies as Zero-Altered Probability Processes: An Empirical Inquiry. *Accident Analysis and Prevention,* Vol. 29, No. 6, 1997, pp. 829–837.
5. Qin, X., J.N. Ivan, and N. Ravishanker. Selecting Exposure Measures in Crash Rate Prediction for Two-Lane Highway Segments. *Accident Analysis and Prevention,* Vol. 36, No. 2, 2004, pp. 183–191.
6. Kumara, S.S., and H.C. Chin. Application of Poisson Underreporting Model to Examine Crash Frequencies at Signalized Three-Legged Intersections. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1908,* Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 46–50.
7. Warton, D.I. Many Zeros Does Not Mean Zero Inflation: Comparing the Goodness-of-Fit of Parametric Models to Multivariate Abundance Data. *Environmetrics,* Vol. 16, No. 3, 2005, pp. 275–289.
8. Lord, D., S.P. Washington, and J.N. Ivan. Poisson, Poisson-Gamma and Zero-Inflated Regression Models of Motor Vehicle Crashes: Balancing Statistical Fit and Theory. *Accident Analysis and Prevention,* Vol. 37, 2005, pp. 35–46.
9. Malyshkina, N.V., and F.L. Mannering. Zero-State Markov Switching Count-Data Models: An Empirical Assessment. *Accident Analysis and Prevention,* Vol. 42, No. 1, 2010, pp. 122–130.
10. Vangala, P., D. Lord, and S.R. Geedipally. An Application of the Negative Binomial-Generalized Exponential Model for Analyzing Traffic Crash Data with Excess Zeros. Presented at 94th Annual Meeting of the Transportation Research Board, Washington, D.C., 2015.
11. Geedipally, S.R., D. Lord, and S.S. Dhavala. The Negative Binomial-Lindley Generalized Linear Model: Characteristics and Application Using Crash Data. *Accident Analysis and Prevention,* Vol. 45, 2012, pp. 258–265.
12. Lord, D., and S.R. Geedipally. The Negative Binomial–Lindley Distribution as a Tool for Analyzing Crash Data Characterized by a Large Amount of Zeros. *Accident Analysis and Prevention,* Vol. 43, No. 5, 2011, pp. 1738–1742.
13. Lindley, D.V. Fiducial Distributions and Bayes' Theorem. *Journal of the Royal Statistical Society, Series B (Methodological),* 1958, pp. 102–107.
14. Ghitany, M., B. Atieh, and S. Nadarajah. Lindley Distribution and Its Application. *Mathematics and Computers in Simulation,* Vol. 78, No. 4, 2008, pp. 493–506.
15. Zamani, H., and N. Ismail. Negative Binomial-Lindley Distribution and Its Application. *Journal of Mathematics and Statistics,* Vol. 6, No. 1, 2010, pp. 4–9.
16. Hallmark, S.L., Y. Qiu, M. Pawlovitch, and T.J. McDonald. Assessing the Safety Impacts of Paved Shoulders. *Journal of Transportation Safety and Security,* Vol. 5, No. 2, 2013, pp. 131–147.
17. Bedrick, E.J., R. Christensen, and W. Johnson. A New Perspective on Priors for Generalized Linear Models. *Journal of the American Statistical Association,* Vol. 91, No. 436, 1996, pp. 1450–1460.
18. Schlüter, P., J. Deely, and A. Nicholson. Ranking and Selecting Motor Vehicle Accident Sites by Using a Hierarchical Bayesian Model. *Journal of the Royal Statistical Society: Series D (The Statistician),* Vol. 46, No. 3, 1997, pp. 293–316.
19. Gupta, R.D., and D. Kundu. Theory and Methods: Generalized Exponential Distributions. *Australian and New Zealand Journal of Statistics,* Vol. 41, No. 2, 1999, pp. 173–188.
20. Aryuyuen, S., and W. Bodhisuwan. The Negative Binomial-Generalized Exponential (NB-GE) Distribution. *Applied Mathematical Sciences,* Vol. 7, No. 22, 2013, pp. 1093–1105.
21. Spiegelhalter, D., A. Thomas, N. Best, and D. Lunn. *OpenBUGS User Manual, Version 3.0. 2.* MRC Biostatistics Unit, Cambridge, United Kingdom, 2007.
22. *Policy on Geometric Design of Highways and Streets,* Vol. 3. AASHTO, Washington, D.C., 2001.
23. Stein, W.J., and T.R. Neuman. *Mitigation Strategies for Design Exceptions.* FHWA, U.S. Department of Transportation, 2007.
24. Harwood, D.W., J.M. Hutton, C. Fees, K.M. Bauer, A. Glen, and H. Ouren. *NCHRP Report 783: Evaluation of the 13 Controlling Criteria for Geometric Design.* Transportation Research Board of the National Academies, Washington, D.C., 2014.
25. *Highway Capacity Manual.* TRB, National Research Council, Washington, D.C., 2000.
26. Krnjajić, M., and D. Draper. Bayesian Model Comparison: Log Scores and DIC. *Statistics and Probability Letters,* Vol. 88, 2014, pp. 9–14.
27. Draper, D. Bayesian Model Specification: Heuristics and Examples. In *Bayesian Theory and Applications,* Oxford University Press, Oxford, United Kingdom, 2013, pp. 409–431.
28. Draper, D., and M. Krnjajić. Calibration Results for Bayesian Model Specification. *Bayesian Analysis,* Vol. 1, No. 1, 2010, pp. 1–43.
29. Washington, S.P., M.G. Karlaftis, and F.L. Mannering. *Statistical and Econometric Methods for Transportation Data Analysis.* CRC Press, Boca Raton, Fla., 2010.
30. Lee, C., M. Abdel-Aty, J. Park, and J.-H. Wang. Development of Crash Modification Factors for Changing Lane Width on Roadway Segments Using Generalized Nonlinear Models. *Accident Analysis and Prevention,* Vol. 76, 2015, pp. 83–91.
31. Bonneson, J.A., D. Lord, K.H. Zimmerman, K. Fitzpatrick, and M.P. Pratt. *Development of Tools for Evaluating the Safety Implications of Highway Design Decisions.* Technical report. FHWA, U.S. Department of Transportation, 2007.
32. Haleem, K., A. Gan, and J. Lu. Using Multivariate Adaptive Regression Splines (MARS) to Develop Crash Modification Factors for Urban Freeway Interchange Influence Areas. *Accident Analysis and Prevention,* Vol. 55, 2013, pp. 12–21.
33. Park, E.S., P.J. Carlson, R.J. Porter, and C.K. Andersen. Safety Effects of Wider Edge Lines on Rural, Two-Lane Highways. *Accident Analysis and Prevention,* Vol. 48, 2012, pp. 317–325.
34. Hauer, E. *Lane Width and Safety.* 2000. http://ca.geocities.com/hauer@rogers.com/Pubs/Lanewidth.pdf.
35. Gross, F., P.P. Jovanis, K.A. Eccles, and K.-Y. Chen. *Safety Evaluation of Lane and Shoulder Width Combinations on Rural, Two-Lane, Undivided Roads.* FHWA, U.S. Department of Transportation, 2009.
36. Geedipally, S.R., D. Lord, and S.S. Dhavala. A Caution About Using Deviance Information Criterion While Modeling Traffic Crashes. *Safety Science,* Vol. 62, 2014, pp. 495–498.

The Standing Committee on Statistical Methods peer-reviewed this paper.