Empirical Bayes estimates of finite mixture of negative binomial regression models and its application to highway safety

Yajie Zou
Key Laboratory of Road and Traffic Engineering of Ministry of Education
Tongji University, Shanghai 201804, China
Tel: (86)13681865023 Email: yajiezou@hotmail.com

John E. Ash
Department of Civil and Environmental Engineering
University of Washington
Box 352700, Seattle, WA 98195-2700
Tel: (936)245-5628 Email: jeash@uw.edu

Byung-Jung Park*
Department of Transportation Engineering
Myongji University, Korea
Phone: 82-31-330-6499, fax: 82-31-330-2885
E-mail: bjpark@mju.ac.kr

Dominique Lord
Zachry Department of Civil Engineering
Texas A&M University, 3136 TAMU
College Station, TX 77843-3136
Phone: 979/458-3949, fax: 979/845-6481
E-mail: d-lord@tamu.edu

Lingtao Wu
Texas A&M Transportation Institute
Texas A&M University System, 3135 TAMU
College Station, Texas 77843-3135
Phone: (979) 845-7214, fax: (979) 845-6481
E-mail: wulingtao@gmail.com

Revised Version

August 13, 2017

* Corresponding author

ABSTRACT
The empirical Bayes (EB) method is commonly used by transportation safety analysts for conducting different types of safety analyses, such as before-after studies and hotspot analyses. The method has numerous benefits including its relative ease of implementation and accounting for the regression-to-the-mean bias. To date, most implementations of the EB method have been applied using a negative binomial (NB) model, as it can easily accommodate the overdispersion commonly observed in crash data. Recent studies have shown that a generalized finite mixture of NB models with K mixture components (GFMNB-K) can also be used to model crash data subjected to overdispersion and generally offers better statistical performance than the traditional NB model (1-component). So far, nobody has developed how the EB method could be used with finite mixtures of NB models. The main objective of this study is therefore to use a GFMNB-K model in the calculation of EB estimates. Specifically, GFMNB-K models with varying weight parameters are developed to analyze crash data from Indiana and Texas. The main finding shows that the rankings produced by the NB and GFMNB-2 models for hotspot identification are often quite different, and this was especially noticeable with the Texas dataset. Finally, a simulation study designed to examine which model formulation can better identify the hotspot is recommended as our future research.

Keywords: Finite mixture model; varying weight parameters; Empirical Bayes method; negative binomial; vehicle crash data; hotspot identification

1. INTRODUCTION

Due to the high societal and monetary losses associated with motor-vehicle crashes, much work has been done to investigate factors associated with crash frequencies and crash severities (Chen et al., 2016, Zou et al.). In the former case, development of models to predict the crash frequency for a given entity in the transportation network has received a tremendous amount of attention in the safety-related literature. These modeling efforts have improved over time as advances have been made to overcome a variety of "data and methodological issues." One such issue commonly experienced in crash data is overdispersion (i.e., the variance of the crash count data is greater than the mean) (Lord and Mannering, 2010, Mannering and Bhat, 2014). Early efforts to handle overdispersion in crash frequency modeling applied negative binomial (NB) (or Poisson-gamma) models (Maycock and Hall, 1984, Hauer et al., 1988). In fact, the negative binomial model is still commonly used today due to its simplicity in estimation via common statistical software packages and due to its simple relationship between the mean and the variance (Lord and Mannering, 2010, Park et al., 2012). Although the NB model can account for overdispersion in crash data, it is far from the only type of model capable of doing so. A variety of other mixed-Poisson models including the Poisson inverse Gaussian (PIG) model (Zha et al., 2014), the Poisson-lognormal model (Connors et al., 2013, Miranda-Moreno et al., 2005, Aguero-Valverde and Jovanis, 2008), the Poisson-Weibull model (Connors et al., 2013, Cheng et al., 2013), the Negative Binomial-Lindley (Geedipally et al., 2012), the Negative Binomial-Generalized Exponential (Vangala et al., 2015) and Sichel model (Zou et al., 2013a) have been applied in crash frequency modeling tasks where overdispersion was present. Although the aforementioned model specifications can be used to describe heterogeneity in crash data, they have limitations in fully accounting for it, especially when some important crash-related factors (i.e., physiological characteristics of drivers, etc.) are difficult to obtain. In such cases, the finite mixture (or latent class) model is considered as one of the major approaches for characterizing the unobserved heterogeneity in the crash data (Mannering et al., 2016). Recently, the generalized Waring model has also been proposed to characterize unobserved heterogeneity (Peng et al., 2014).

According to Frühwirth-Schnatter (2006), the application of finite mixture models can be traced back to 1943. In transportation, such models (in the form of finite mixture models or latent-class models) have been used in both operational and safety applications (Jun, 2010, Xiong and Mannering, 2013, Eluru et al., 2012, Yasmin et al., 2014, Ding et al., 2015, Zou et al., 2017, Tang et al., 2017). In recent years, finite mixtures of regression models have been used in crash prediction applications. Specifically, Park and Lord (2009) compared finite mixtures of Poisson and negative binomial regression models to a traditional negative binomial regression model in an application involving crash data collected at intersections in Toronto. They noted that the data appeared to arise from two sub-populations (defined primarily by annual average daily traffic (AADT) volumes on minor approaches) and the finite mixture of NB models was able to "provide the nature of the over-dispersion in the data". Park et al. (2014) further investigated the relative performance of the finite mixture model and the NB model in terms of hotspot identification. Zou et al. (2014) studied the effects of varying weight parameters for a finite mixture of NB regression models and found that allowing the weight parameters to vary led to better goodness-of-fit and understanding of the overdispersion present in the crash data. More recently, Park et al. (2016) developed crash modification factors using a finite mixture modeling approach. Some covariates were found to have non-linear relationships with crash frequency, and interactive effects were captured in the finite mixture framework. Overall, the results derived from finite mixture models are more reasonable than that of the commonly used NB model.

In terms of applications of crash prediction models, perhaps the two most common are in hotspot identification and for before-after studies (Persaud et al., 2010). In the former case,

the goal is to identify locations that could benefit from implementation of safety treatments, while in the latter case, the goal is to determine the efficacy of an implemented safety treatment. In order to accomplish either of the aforementioned activities, the empirical Bayesian (EB) method has been the common choice (Hauer, 1997). The EB method combines a site's historical crash record with the expected number of crashes estimated from a safety performance function (SPF) for similar sites, and thus it is relatively insensitive to random variations in crash frequency. The EB method can correct for the regression-to-the mean bias, and it is relatively easy to implement compared to the full Bayesian (FB) approach (Hauer et al., 2002, Zou et al., 2013a). Despite these benefits, the EB method can still be biased in specific cases, and readers are referred to (Lord and Kuo, 2012) for additional information as these issues are outside the scope of this study. When applying the EB method, the occurrence of crashes at a site is usually assumed to be Poisson distributed with rate parameter λ. The distribution of the parameter λ is commonly assumed to follow a gamma distribution, in turn giving rise to Poisson-gamma (i.e., NB) regression models commonly used in the EB method (Hauer, 1992, Hauer, 1997, Cheng and Washington, 2005, Cheng and Washington, 2008). Nonetheless, there is no guarantee that the parameter λ is truly gamma distributed. The EB method has been applied with other mixed-Poisson models, such as the Sichel model, in place of the traditional NB model (Hauer, 1997, Zou et al., 2013a).

In this study, to accommodate the heterogeneity of crash data, the finite mixture of K-component NB regression models with varying weight parameters (GFMNB-K) is used. The primary goal of this paper is to develop the EB method when applying the GFMNB-K model and demonstrate how the EB method can be used for this model with actual crash datasets. Following the procedure in (Zou et al., 2013a), EB estimates obtained from the traditional NB model are compared to those obtained from the finite mixture of NB models. Key components of the investigation include a discussion of the component weights for the GFMNB-K models, a comparison of the predicted crash mean and variance values between the two models (i.e., NB versus GFMNB-K), and an examination of the model performances in terms of hotspot identification.

2. METHODOLOGY

The following section provides the derivations for the EB estimates from both the traditional NB and GFMNB-K regression models. Further, the parameter estimation method for the GFMNB-K model is outlined.

2.1. Derivation of EB estimate from NB regression model

Let y_i denote the number of crashes and λ_i denote the crash rate for site i. In the NB regression model, we assume that y_i follows the Poisson distribution with the mean crash rate λ_i and, again λ_i follows the gamma distribution with shape parameter ϕ and rate parameter ϕ/μ_i (where, $\mu_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$). In statistical notation, we have the following:

$$y_i|\lambda_i \sim Poisson(\lambda_i) \tag{1}$$

$$\lambda_i|\phi, \boldsymbol{\beta} \sim Gamma(\phi, \phi/\mu_i) \tag{2}$$

Here, we can see that $E(\lambda_i) = \mu_i$ and $Var(\lambda_i) = {\mu_i}^2/\phi$. Note that ϕ is the shape parameter of the gamma distribution.

In what follows, without loss of generality we omit the subscript i. The joint distribution of (y, λ) has the following probability mass function (pmf):

$$p(y, \lambda) = p(y|\lambda)p(\lambda) \tag{3}$$

$$= \frac{\lambda^y \cdot e^{-\lambda}}{y!} \cdot \frac{(\phi/\mu)^\phi}{\Gamma(\phi)} \cdot \lambda^{\phi-1} \cdot e^{-\lambda(\phi/\mu)} \tag{4}$$

$$= \frac{1}{\Gamma(y+1)} \cdot \frac{(\phi/\mu)^\phi}{\Gamma(\phi)} \cdot \lambda^{y+\phi-1} \cdot e^{-\lambda(1+\phi/\mu)} \tag{5}$$

The marginal distribution of the number of crashes, $p(y)$, is obtained as follows:

$$p(y) = \int_0^\infty p(y|\lambda)p(\lambda)d\lambda \tag{6}$$

$$= \frac{(\phi/\mu)^\phi}{\Gamma(y+1)\Gamma(\phi)} \int_0^\infty \lambda^{y+\phi-1} \cdot e^{-\lambda(1+\phi/\mu)} d\lambda \tag{7}$$

Here, the integrand, $\lambda^{y+\phi-1} \cdot e^{-\lambda(1+\phi/\mu)}$, is a kernel of $\text{Gamma}(y+\phi, 1+\phi/\mu)$ for λ. Thus, the following is true:

$$\frac{(1+\phi/\mu)^{y+\phi}}{\Gamma(y+\phi)} \int_0^\infty \lambda^{y+\phi-1} \cdot e^{-\lambda(1+\phi/\mu)} d\lambda = 1 \tag{8}$$

$$\int_0^\infty \lambda^{y+\phi-1} \cdot e^{-\lambda(1+\phi/\mu)} d\lambda = \frac{\Gamma(y+\phi)}{(1+\phi/\mu)^{y+\phi}} \tag{9}$$

Therefore,

$$p(y) = \frac{(\phi/\mu)^\phi}{\Gamma(y+1)\Gamma(\phi)} \frac{\Gamma(y+\phi)}{(1+\phi/\mu)^{y+\phi}} = \frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)} \left(\frac{\phi}{\mu+\phi}\right)^\phi \left(\frac{\mu}{\mu+\phi}\right)^y \tag{10}$$

The aforementioned expression for $p(y)$ is the pmf for the negative binomial distribution. The marginal mean and variance of y are obtained from the following relationships:

$$E(y) = E\{E(y|\lambda)\} = E(\lambda) = \mu \tag{11}$$

$$Var(y) = E\{Var(y|\lambda)\} + Var\{E(y|\lambda)\} = \mu + \mu^2/\phi \tag{12}$$

Next, the posterior distribution of λ given y is expressed as follows:

$$p(\lambda|y) = \frac{p(y,\lambda)}{p(y)} \propto p(y|\lambda)p(\lambda) \tag{13}$$

$$\propto \lambda^{y+\phi-1} \cdot e^{-\lambda(1+\phi/\mu)} \tag{14}$$

Here, $\lambda^{y+\phi-1} \cdot e^{-\lambda(1+\phi/\mu)}$ is a kernel of $\text{Gamma}(y+\phi, 1+\phi/\mu)$ for λ. Therefore, the following is true:

$$\lambda|y \sim Gamma(y+\phi, 1+\phi/\mu) \tag{15}$$

$$E(\lambda|y) = \frac{y+\phi}{1+\phi/\mu} \tag{16}$$

$$= \left(\frac{\mu}{\mu+\phi}\right) y + \left(\frac{\phi}{\mu+\phi}\right) \mu \tag{17}$$

$$Var(\lambda|y) = \left(\frac{\mu}{\mu+\phi}\right) E(\lambda|y) \tag{18}$$

Ultimately, $E(\lambda|y)$ is the EB estimate for site i with $\frac{\phi}{\mu+\phi}$ as a weighting factor. In the EB method, the parameters ϕ and μ are estimated from existing data through use of the NB regression model. Such procedure is different from that used in the FB method where hyper-prior distributions are set on the aforementioned parameters that themselves depend on a set of higher-level parameters.

2.2. Derivation of EB estimate from GFMNB-K regression model

In this section, the posterior distribution of λ given y under the GFMNB-K is derived. As in the traditional NB regression model, we assume that y follows the Poisson distribution with the mean crash rate λ, and the λ, in this case, follows a K-component finite mixture of gamma distributions as follows:

$$y|\lambda \sim Poisson(\lambda) \tag{19}$$

$$p(\lambda) = \sum_{k=1}^{K} \pi_k p_k(\lambda) \tag{20}$$

where, $p_k(\lambda) \sim Gamma(\phi_k, \phi_k/\mu_k)$, $\pi_k > 0$ and $\sum \pi_k = 1$. Here, we can see that $E(\lambda) = \sum \pi_k \mu_k$ and $Var(\lambda) = \sum \pi_k {\mu_k}^2 (1 + \frac{1}{\phi_k}) - E(\lambda)^2$.

Analogous to Gharib (1995), where two components were considered, the marginal distribution of the number of crashes, $p(y)$ can be derived as follows:

$$p(y) = \int_0^\infty p(y|\lambda) p(\lambda) d\lambda \tag{21}$$

$$= \int_0^\infty \frac{\lambda^y \cdot e^{-\lambda}}{y!} \cdot p(\lambda) d\lambda \tag{22}$$

$$= \sum_{k=1}^{K} \left\{ \pi_k \frac{(\phi_k/\mu_k)^{\phi_k}}{y! \Gamma(\phi_k)} \int_0^\infty \lambda^{y+\phi_k-1} e^{-\lambda(1+\phi_k/\mu_k)} d\lambda \right\} \tag{23}$$

$$= \sum_{k=1}^{K} \left\{ \pi_k \frac{\Gamma(y+\phi_k)}{\Gamma(y+1)\Gamma(\phi_k)} \left(\frac{\phi_k}{\mu_k+\phi_k}\right)^{\phi_k} \left(\frac{\mu_k}{\mu_k+\phi_k}\right)^{y} \right\} \tag{24}$$

Hence, the distribution of $p(y)$ is in the form of GFMNB-K regression model (Park and Lord, 2009). Note that the NB regression model is a special case of GFMNB-K regression model with K=1.

The marginal mean and variance of y are obtained from the following relationships:

$$E(y) = E\{E(y|\lambda)\} = E(\lambda) = \sum_{k=1}^{K} \pi_k \mu_k \tag{25}$$

$$Var(y) = E\{Var(y|\lambda)\} + Var\{E(y|\lambda)\} \tag{26}$$

$$= E(y) + \left\{\sum_{k=1}^{K} \pi_k {\mu_k}^2 (1 + \frac{1}{\phi_k}) - E(y)^2\right\} \quad (27)$$

Next, the posterior distribution of λ given y is expressed as follows:

$$p(\lambda|y) = \frac{p(y,\lambda)}{p(y)} = \frac{p(y|\lambda)p(\lambda)}{p(y)} \quad (28)$$

$$= \sum_{k=1}^{K} \frac{\pi_k p(y|\lambda)}{p(y)} p_k(\lambda) \quad (29)$$

$$= \sum_{k=1}^{K} \frac{\pi_k p_k(y)}{p(y)} \cdot \frac{p(y|\lambda)p_k(\lambda)}{p_k(y)} \quad (30)$$

$$= \sum_{k=1}^{K} w_k(y) \cdot p_k(\lambda|y) \quad (31)$$

where, $p_k(\lambda|y) = \frac{p(y|\lambda)p_k(\lambda)}{p_k(y)}$ is the "*k*-th group posterior" – the posterior distribution if the prior were $p_k(\lambda)$. Additionally, $w_k(y) = \frac{\pi_k p_k(y)}{p(y)}$ is the posterior probability that y was drawn from group k (Muralidharan, 2010).

Hence, the posterior distribution of λ given y is:

$$\lambda|y \sim \sum_{k=1}^{K} w_k(y) \cdot p_k(\lambda|y) \quad (32)$$

$$E(\lambda|y) = \sum_{k=1}^{K} w_k(y) \cdot E_k(\lambda|y) \quad (33)$$

$E(\lambda|y)$ is the EB estimate for site i. In a Poisson-gamma mixture (NB) model, $E_k(\lambda|y) = \left(\frac{\mu_k}{\mu_k+\phi_k}\right) y + \left(\frac{\phi_k}{\mu_k+\phi_k}\right) \mu_k$ as in Eq. 17. Hence, $E(\lambda|y)$ is expressed as follows:

$$E(\lambda|y) = \sum_{k=1}^{K} w_k(y) \left\{\left(\frac{\mu_k}{\mu_k+\phi_k}\right) y + \left(\frac{\phi_k}{\mu_k+\phi_k}\right) \mu_k\right\} \quad (34)$$

$$Var(\lambda|y) = \sum_{k=1}^{K} w_k(y) \left\{E_k(\lambda|y)^2 + \left(\frac{\mu_k}{\mu_k+\phi_k}\right) E_k(\lambda|y)\right\} - E(\lambda|y)^2 \quad (35)$$

where $w_k(y) = \frac{\pi_k p_k(y)}{p(y)}$ and π_k is parameterized as a function of covariates.

2.3. Parameter estimation method

Finite mixture models can be estimated two ways: the maximum likelihood estimation (MLE) method and the Bayesian method. Previously, Park and Lord (Park and Lord, 2009) adopted the Bayesian framework with data augmentation and Markov Chain Monte Carlo (MCMC) techniques to estimate a finite mixture of NB regression models. Maximum likelihood estimation and the Bayesian method both have their advantages and disadvantages (see discussion in (McLachlan and Peel, 2004, Frühwirth-Schnatter, 2006)). For example, the Bayesian method with MCMC techniques can be computationally demanding and complications can arise due to the label switching problem (Park and Lord, 2009, Frühwirth-Schnatter, 2006). The MLE method, on the other hand, is not guaranteed to converge to a global maximum, and thus it requires many different initial points in an effort to achieve the global maximum. In this study, the MLE approach with application of the Expectation Maximization (EM) algorithm was used to estimate the model parameters. When estimating the mixture

models, the solutions of the likelihood function depends on the initial values and each initial value corresponds to its own maximum (i.e., solution) of the likelihood function; that is, multiple local maximum values can be found. In such cases where a known, consistent estimator of Θ (here, Θ denotes the root that maximizes the log likelihood, although in general it could be any parameter/vector of parameters) cannot be found, one may intuitively select the largest of the local maxima found as the root of the likelihood function (McLachlan and Peel, 2004). In order to attempt to achieve the global maximum, the searching process was repeated 20 times using various random initial values, and the optimal root that corresponded to the largest value of likelihood function was selected. A detailed discussion about the parameter estimation method of this method is described in (Zou et al., 2013b).

3. DATA DESCRIPTION

This section introduces the two crash datasets (i.e., Indiana data and Texas data) used in this study.

3.1. Indiana Data

The Indiana dataset contains crash records from 338 rural interstate roadway sections in that state over a five-year period (from year 1995 to 1999). The explanatory variables considered in the modeling results section are provided in Table 1. A total of 5,737 crashes occurred on 218 out of the 338 roadway segments during the five years over which data were collected; the other 120 segments (36%) in the dataset experienced no reported crashes. Interested readers can see Washington et al. (2010) for a complete list of variables in this dataset.

Table 1. Summary statistics of variables in Indiana data

Variables	Minimum	Maximum	Mean(SD†)	Sum
Total number of crashes over five years period	0	329	16.97 (36.30)	5737
Average daily traffic over five years (F)	9442	143422	30237.6 (28776.4)	
Minimum friction reading in the road section over the 5-year period (FR)	15.9	48.2	30.51 (6.67)	
Pavement surface type (1: asphalt, 0: concrete) (PT)	0	1	0.77 (0.42)	
Median width (in feet) (MW)	16	194.7	66.98 (34.17)	
Presence of median barrier (1: present, 0: absent) (BR)	0	1	0.16 (0.37)	
Interior rumble strips (1: present, 0: absent) (RS)	0	1	0.72 (0.45)	
Segment length (in miles) (L)	0.009	11.53	0.89 (1.48)	300.09
Number of lanes (NL)	2	3	2.089(0.28)	
Outside shoulder width (in feet) (OSW)	6.20	21.80	11.28(1.74)	
Speed limit (mph) (SL)	50	65	63.09(3.91)	

NOTE: † Standard deviation.

3.2. Texas Data

The Texas crash dataset contains crash data collected over a five-year (1997-2001) from a total of 1,499 four-lane undivided rural road segments. The data were collected as a part of the National Cooperative Highway Research Program (NCHRP) 17-29 research project (Lord et al., 2008). The minimum segment length was 0.10 mile while the maximum length reached up to 6.28 miles, and the average length was 0.55 mile. Overall, a total of 4,253 crashes were reported to have occurred on 946 of the segments; the other 553 segments (37%) had no reported crashes. Table 2 presents the summary statistics of variables in the Texas dataset.

Table 2. Summary statistics of variables in Texas data

Variables	Minimum	Maximum	Mean(SD†)	Sum
Total number of crashes (five years)	0	97	2.84(5.69)	4253
Average daily traffic over five-year period (F)	42	24800	6613.61 (4010.01)	
Lane width (in feet) (LW)	9.75	16.5	12.57(1.59)	
Total shoulder width (in feet) (SW)	0	40	9.96(8.02)	
Curve density (curves per mile) (CD)	0	18.07	1.43 (2.35)	
Length of segment (in miles) (L)	0.1	6.28	0.55(0.67)	830.49

NOTE: † Standard deviation.

4. MODELING RESULTS

In section 4, the modeling results and the EB estimates of the NB and GFMNB-K models are provided using the two datasets. The GAMLSS.MX package in the software R (Rigby and Stasinopoulos, 2009) was used to estimate the NB and GFMNB-K models.

4.1. Comparison of EB Estimates from NB and GFMNB-K Models using Indiana Data

The modeling results for two models using the Indiana data are presented in this section. In the modeling process, the segment length was treated as an offset variable. Then, the mean functional form for the standard NB model is adopted as follows:

$$\mu_i = \beta_0 L_i F_i^{\beta_1} e^{\beta_2 * FR_i + \beta_3 * PT_i + \beta_4 * MW_i + \beta_5 * BR_i + \beta_6 * RS_i + \beta_7 * NL_i + \beta_8 * OSW_i + \beta_9 * SL_i} \quad (36)$$

where, μ_i is the expected crash frequency for segment i; L_i is the segment length (miles) of segment i; F_i is the average daily traffic (average over five years) on segment i; FR_i is the minimum friction reading for segment i; PT_i is the pavement surface type for segment i; MW_i is the median width for segment i; BR_i is the presence of median barrier for segment i; RS_i is the presence of interior rumble strips for segment i; NL_i is the number of lane for segment i; OSW_i is the outside shoulder width for segment i; SL_i is the speed limit for segment i; and, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9)'$ is the vector of estimated coefficients.

For the considered explanatory variables, a backward variable selection algorithm was applied to exclude the insignificant variables. After excluding the insignificant variables, the results for the Indiana data using the NB model are provided in Table 3. When estimating the GFMNB-K model, in order to guarantee the convergence criterion for the EM algorithm, only statistically significant independent variables listed in Table 3 are used.

For the GFMNB-K model, the component-wise mean functional form is defined as follows:

$$\mu_{k,i} = \beta_{k,0} L_i F_i^{\beta_{k,1}} e^{\beta_{k,2}*FR_i + \beta_{k,3}*MW_i + \beta_{k,4}*BR_i + \beta_{k,5}*SL_i} \tag{37}$$

where, $\mu_{k,i}$ is the expected crash frequency at segment i for component k; $\boldsymbol{\beta}_k = (\beta_{k,0}, \beta_{k,1}, \beta_{k,2}, \beta_{k,3}, \beta_{k,4}, \beta_{k,5})'$ is the vector of estimated coefficients for component k.

As aforementioned, in the GFMNB-K model, the weight parameters are allowed to vary. For the GFMNB-K models in this study, a linear combination of different available explanatory variables is used to model the weight parameters:

$$\frac{\pi_{ik}}{\pi_{iK}} = e^{\gamma_{0,k}} e^{\gamma_{1,k}*L_i + \gamma_{2,k}*F_i + \gamma_{3,k}*FR_i + \gamma_{4,k}*MW_i + \gamma_{5,k}*BR_i + \gamma_{6,k}*SL_i} \tag{38}$$

where, π_{ik} is the estimated proportion (or weight) of component k for segment i; where, $\boldsymbol{\gamma}_k = (\gamma_{0,k}, \gamma_{1,k}, \gamma_{2,k}, ..., \gamma_{6,k})'$ are the estimated coefficients for component k.

Previously, Zou et al. (2014) applied the GFMNB-K models with different numbers of components to the Indiana data. Based on the reported Bayesian information criterion (BIC) values in their study, the number of components K=2 was selected as optimal for the final model. The goodness-of-fit statistics and modeling results for the GFMNB-2 model are provided in Table 3. Figure 1(a) shows the component weights for each of the two components in the GFMNB-2 model, for each segment in the Indiana dataset. Recall that for each segment, the weight values must sum to unity as shown in Figure 1(a). In addition, one can see from Figure 1(a) that while a large range of weight values for the segments are presented, cases in which full (i.e., 1.00), or nearly full weight, is given to one component in the GFMNB-2 model are most common. Note that for the GFMNB-2 model, the coefficients of variables take reasonable values and correspond to the modeling results from the NB model. The values of the Akaike information criterion (AIC) and Bayesian information criterion indicate that the GFMNB-2 model can provide a better statistical fit than the traditional NB model, which suggest that the crash data may be generated from two distinct sub-populations with different regression coefficients and degrees of over-dispersion, rather than from a single population.

Table 3. Modeling results of NB and GFMNB-2 models using Indiana data

Standard NB							
	Intercept	Ln(F)	FR	MW	BR	SL	$\log(\alpha)^{\dagger}$
Estimate	-6.763	0.69	-0.028	-0.006	-2.972	0.038	-0.117
Std. error	1.527	0.127	0.01	0.002	0.269	0.016	0.126
Number of observations	338						
Log-likelihood	-943.082						
AIC	1900.164						
BIC	1926.925						
GFMNB-2 model							
Component 1	Intercept	Ln(F)	FR	MW	BR	SL	$\log(\alpha)^{\dagger}$
Estimate	-59.957	1.316	-0.018*	-0.004*	-0.234*	0.753	0.236*
Std.error	24.471	0.224	0.018	0.003	0.675	0.375	0.172
Component 2	Intercept	Ln(F)	FR	MW	BR	SL	$\log(\alpha)$
Estimate	-6.389	0.791	-0.018	-0.004	-20.457*	0.010*	-1.976
Std.error	1.089	0.101	0.006	0.002	714.085	0.008	0.192
Estimate of γ_1	Intercept	Segment length	F	FR	MW	BR	SL
	20.919	3.246	-2.16E-05	-0.097	0.0102	5.405	-0.32
Log-likelihood	-846.74						
AIC	1735.48						
BIC	1815.76						

NOTE: † Dispersion parameter $\alpha = 1/\phi$; * Not significant at 5% significance level.

Figure 1(b) shows a comparison of the variance versus the mean predicted by the two different models developed from the Indiana dataset. For small crash mean values (here, less than approximately 25), the GFMNB-2 and traditional NB models yield similar variance. However, as the crash mean increases, the variance of the NB model increases at a much greater rate than that of the GFMNB-2 model. This is partially due to the fact that while the variance for the traditional NB model and GFMNB-2 model are both proportional to the crash mean, the formulation of the variance for the GFMNB-2 model includes a term involving the subtraction of the square of the crash mean that is not present in the variance formulation for the traditional NB model. If a large portion of the weight in the GFMNB-2 model is allocated to the second component (whose inverse dispersion parameter is less than one), one can see how the variance of the GFMNB-2 model could in fact exceed that of the NB model as is

observed for a small number of points in Figure 1(b). Ultimately, increased flexibility in handling overdispersion can be achieved through the use of the GFMNB-2 model compared to the traditional NB model. By estimating a different inverse dispersion parameter for each component in the mixture, a much larger range of variance values can be achieved depending on the values of the inverse dispersion parameters and the weight given to each component.

A crucial part of EB analyses is ranking sites based on their EB estimates of long-range expected crash frequency such that hotspots can be identified. Figure 1(c) shows a plot of rankings as determined by the GFMNB-2 and traditional NB models. Smaller ranking values indicate more hazardous segments based upon EB estimates, with ranking one being the most hazardous site (Park et al., 2014). From the figure, it appears there is some degree of positive correlation between the rankings obtained via the GFMNB-2 and NB models, respectively. Nonetheless, there are still several differences in site ranking values as determined via each of the two models. Table 4 attempts to further quantify these differences. From the table, it can be seen that 35.2% of the segments had rankings that differed by more than ±10 positions when the results of the GFMNB-2 and NB models were compared. In the majority of cases (56.8%), the GFMNB-2 model ranked sites lower than the NB model, and thus may be viewed as slightly more conservative. Although the use of the GFMNB-2 model will affect the weighting factor used in the EB estimates, both models provided comparable rankings (i.e., differences in ranking of less than or equal to ±10 positions) for the majority of segments in this case study.

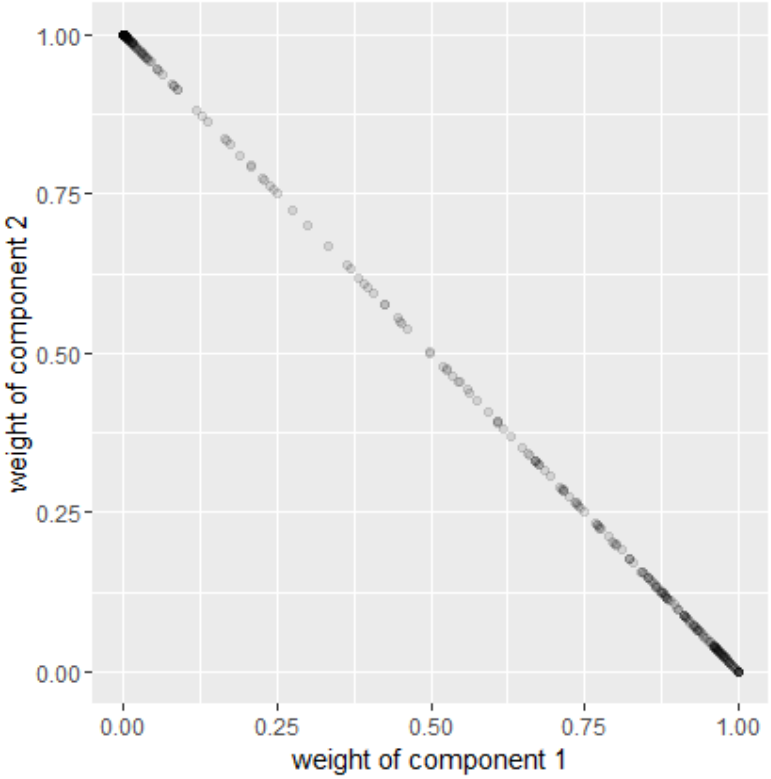

Figure 1(a). Component weights for GFMNB-2 model using Indiana data

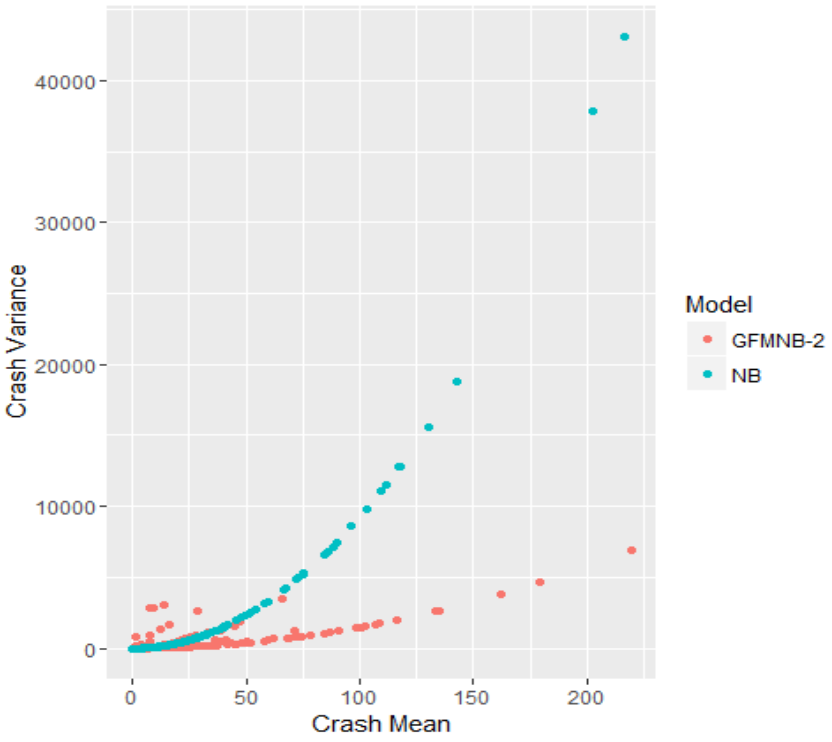

Figure 1(b). Variance versus mean for GFMNB-2 and NB models using Indiana data

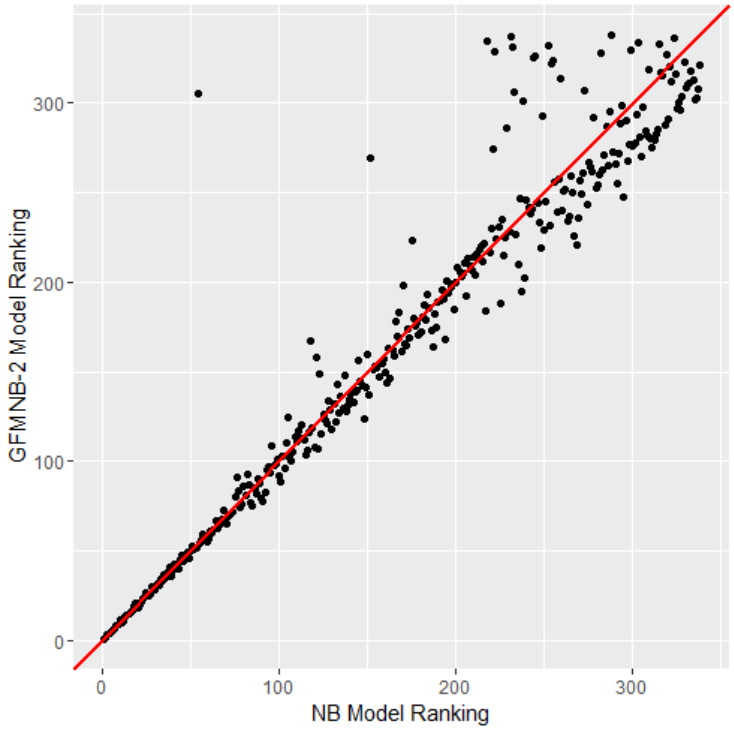

Figure 1(c). Relationship in ranking by GFMNB-2 and NB models (Indiana data)

Table 4. Differences in ranking between GFMNB-2 and NB models using the EB estimates (Indiana data)

Differences in ranking	Difference and percentage
Non-identical ranking	305 (90.2%)
Ranking difference beyond 10 positions	119 (35.2%)
Ranking difference beyond 20 positions	78 (23.1%)

4.2. Comparison of EB estimates from NB and GFMNB models using Texas Data

The following section presents the EB estimates for the NB and GFMNB models that were developed using the Texas dataset. As was the case with the Indiana dataset, the segment length is considered as an offset variable in the models. The mean functional form of the standard NB model is adopted as follows:

$$\mu_i = \beta_0 L_i F_i^{\beta_1} e^{\beta_2 * LW_i + \beta_3 * SW_i + \beta_4 * CD_i} \tag{39}$$

where, μ_i is the expected crash frequency for segment i ; L_i is the segment length (miles) of segment i; F_i is the average daily traffic (average over five years) on segment i; LW_i is the lane width (feet) of segment i; SW_i is the total shoulder width (feet) of segment i; CD_i is the curve density (curves per mile) of segment i; and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'$ are the coefficients to be estimated.

For the GFMNB-K model, the component-wise mean functional form is defined as follows:

$$\mu_{k,i} = \beta_{k,0} L_i F_i^{\beta_{k,1}} e^{\beta_{k,2} * LW_i + \beta_{k,3} * SW_i + \beta_{k,4} * CD_i} \tag{40}$$

where, all variables are defined as aforementioned and the subscript k denotes the kth component. As was the case for the GFMNB-K model developed from the Indiana data, the weight parameters for the GFMNB-K model developed from the Texas data were modeled as a linear combination of the available predictors as follows:

$$\frac{\pi_{ik}}{\pi_{iK}} = e^{\gamma_{0,k}} e^{\gamma_{1,k} * L_i + \gamma_{2,k} * F_i + \gamma_{3,k} * LW_i + \gamma_{4,k} * SW_i + \gamma_{5,k} * CD_i} \tag{41}$$

where, π_{ik} is the estimated proportion (or weight) of component k for segment *i* and $\boldsymbol{\gamma}_k = (\gamma_{0,k}, \gamma_{1,k}, \gamma_{2,k}, ..., \gamma_{5,k})'$ are the estimated coefficients in the weight parameter for component k.

Zou et al. (2013b) previously studied development of the GFMNB-K on the Texas data and determined that use of K=2 mixture components was appropriate based on the modeling results. Parameter estimates and associated standard errors, as well as goodness-of-fit statistics for both the traditional NB and GFMNB-2 models are presented in Table 5. Brief examination of the coefficients shows that the signs seem plausible. Note that a rather limited selection of explanatory variables are included in the model. Some other explanatory variables (e.g., truck percentage, number of days with wet surface, etc.) cannot be accessed because of the age of the Texas data. Previous studies (Washington et al., 2010) indicate that omitting such critical variables can lead to biased parameter estimates. It should be also pointed out that as recently discussed in Wu et al. (2015), only very important missing variables affect the modeling results of regression models.

Figure 2(a) shows the values of the component weights for both components in the GFMNB-2 model; each point corresponds to one segment in the Texas dataset. Unlike the GFMNB-2 model developed from the Indiana data, the model developed from the Texas data has a lesser proportion of sites with more balanced weightings (i.e., scenarios in which neither

of the component weights is near unity). This can be visualized in Figure 2(a) by noting the middle portion of the line has comparatively sparse point density compared to the line in Figure 1(a). Thus, it appears that many segments in the Texas data are better described by one component in the NB mixture. Table 5 also shows goodness-of-fit statistics for each of the two models developed from the Texas data. Both AIC and BIC values indicate that the GFMNB-2 model provides better statistical fit when compared to the traditional NB model.

Table 5. Modeling results for NB and GFMNB-2 models using the Texas data

Standard NB						
	Intercept	Ln(F)	LW	SW	CD	$\log(\alpha)^{\dagger}$
Estimate	-7.949	0.975	-0.053	-0.010	0.067	-0.939
Std. error	0.406	0.044	0.017	0.003	0.012	0.091
Number of observations	1499					
Log-likelihood	-2561.38					
AIC	5134.77					
BIC	5166.65					
GFMNB-2 model						
Component 1	Intercept	Ln(F)	LW	SW	CD	$\log(\alpha)^{\dagger}$
Estimate	-5.090	0.740	-0.139	0.004*	0.218	-1.509
Std.error	0.525	0.054	0.023	0.005	0.026	0.141
Component 2	Intercept	Ln(F)	LW	SW	CD	$\log(\alpha)$
Estimate	-10.554	1.210	-0.012*	-0.009*	0.039	-0.677
Std.error	0.650	0.070	0.023	0.005	0.015	0.119
Estimate of γ_1	Intercept	Segment length	F	LW	SW	CD
	45.183	-279.644	0.008	-0.999	6.594	12.701
Log-likelihood	-2510.145					
AIC	5056.01					
BIC	5151.64					

NOTE: † Dispersion parameter $\alpha = 1/\phi$; * Not significant at 5% significance level.

The predicted crash mean and variance were also compared for the GFMNB-2 and NB models developed from the Texas dataset, and Figure 2(b) shows a plot of these values. As was observed with the Indiana dataset, variance values are similar for both models at small values of the crash mean (here, less than approximately 10). Additionally, as the crash mean increases, the rate at which the variance grows for the NB model exceeds that of the GFMNB-2 model in the majority of cases. This effect can partially be explained by the fact the variance formulation for the GFMNB-2 model involves a weighted sum of terms that decreases as the inverse dispersion parameter grows. Thus, segments with a high proportion of weight allocated to Component 1 of the GFMNB-2 (i.e., the component for which the value of the inverse dispersion parameter is greater than for that of the NB model) may have smaller variance than was estimated by the NB model for a given mean value of crashes. Additionally, the inclusion of a difference term involving the square of the mean value of crashes for a given site also leads to decreased variance for the GFMNB-2 model when compared to the NB model in many cases. For the selection of sites at which the estimated variance from the GFMNB-2 model exceeded that from the NB model, one plausible explanation would be that in such cases, a high proportion of the weight in the mixture model was allocated to Component 2 of the GFMNB-

2 model (i.e., the component whose inverse dispersion parameter value was smaller than that estimated with the NB model). As was observed for the plot of variance versus mean crash values for the Indiana data, the use of the GMFNB-2 allows for more flexible modeling of overdispersion as the variance can be affected by a range of inverse dispersion parameter values located between those estimated for each of the components of the mixture.

Comparison of the GFMNB-2 and NB models' hotspot ranking behavior, based upon EB estimates, was also performed using the Texas dataset. Similar to Figure 1(c), which showed results for the Indiana data, the plot of GFMNB-2 versus NB rankings in Figure 2(c) shows a positive trend. Despite this positive correlation, there were substantial differences in rankings between the two models. A total of 71.5% of the ranking values differed by more than ±10 positions. Further, more than one quarter of segments (27.0% to be exact) had ranking values that differed by more than ±50 positions between the two models. Overall, ranking values as determined from the EB estimates using the GFMNB-2 model were lower than those from the EB estimates developed using the traditional NB model in the majority of cases (57.1%). Hence, as was observed when considering the Indiana dataset, the EB estimates obtained through use of the GFMNB-2 model on the Texas dataset were more conservative than those obtained via the NB model. That said, the number of large discrepancies in rankings (i.e., more than ±50 positions difference) suggests that use of the GFMNB-2 model can substantially affect weight values used to calculate EB estimates when compared to a traditional NB model. Note that the model developed from the Texas data contains only a few explanatory variables. Thus, it may be possible that as a result of including more explanatory variables in the model, the ranking difference between the NB and GFMNB-2 models could be further reduced.

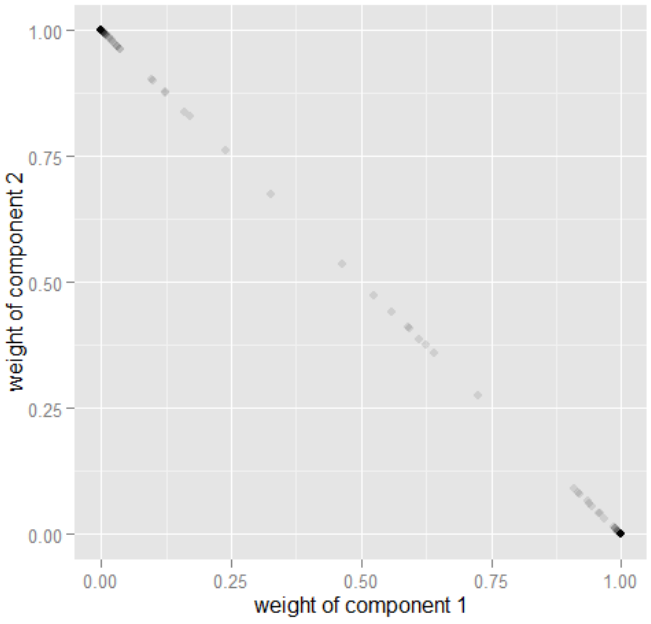

Figure 2(a). Component weights for GFMNB-2 model using Texas data

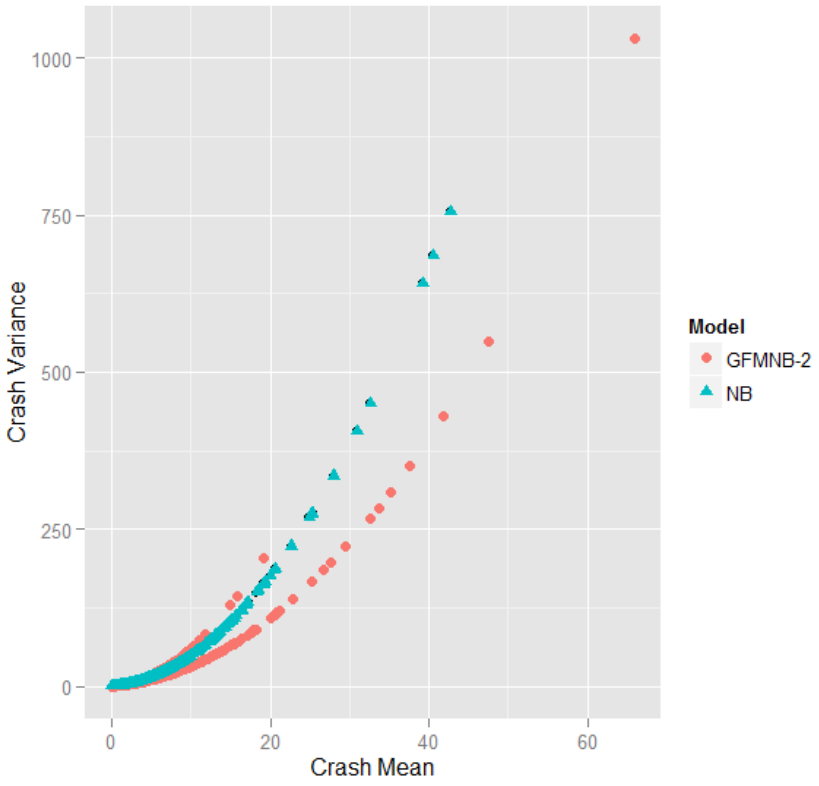

Figure 2(b). Variance versus mean for GFMNB-2 and NB models using Texas data

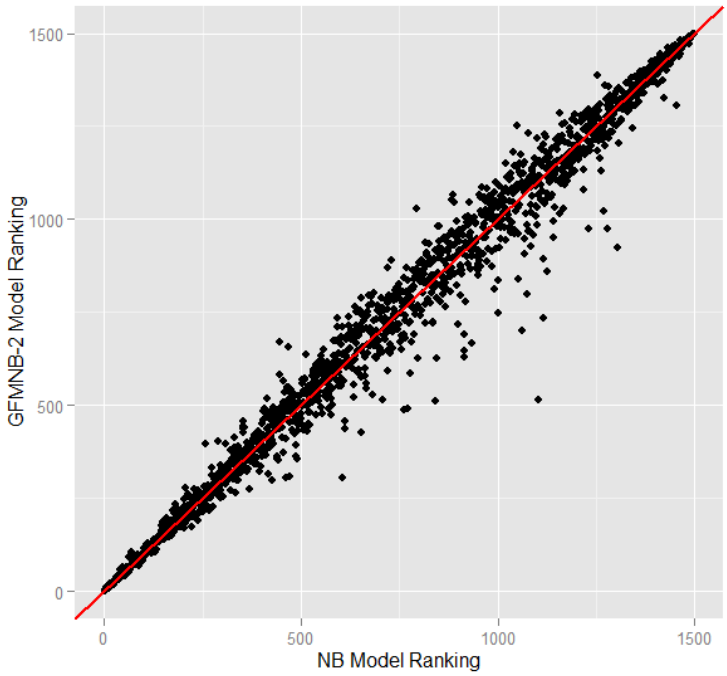

Figure 2(c). Relationship in ranking by GFMNB-2 and NB models (Texas data)

Table 6. Differences in ranking between GFMNB-2 and NB models using the EB estimates (Texas data)

Differences in ranking	Difference and percentage
Non-identical ranking	1466 (97.8%)
Ranking difference beyond 10 positions	1072 (71.5%)
Ranking difference beyond 50 positions	405 (27.0%)

5. DISCUSSION

Interesting modeling results from the NB and GFMNB-2 models warrant further discussion. First, this study used the observed crash data collected from highway segments in Indiana and Texas to compare the performance of two different EB formulations. Although working with the observed crash data has its benefits, one major drawback is that it is nearly impossible to know the true value of safety at a particular site (Miranda-Moreno et al., 2005). This problem can be overcome if one uses simulated data. In such cases, actual performance of the EB method using the different model formulations could be compared based upon how the sites were ranked for the hotspot analysis. Thus, one future research effort may involve simulation studies to validate whether or not a GFMNB-K model outperforms a traditional NB model for the EB method in terms of hotspot identification (Wu et al., 2014).

Second, in current-generation safety studies, finite mixture models and random-parameters models are two primary methods used to capture the unobserved heterogeneity in crash data (Mannering et al., 2016). Compared to the proposed finite mixture model (or latent class model), the random parameters models can allow parameters to differ across observations to characterize the unobserved heterogeneity. In the random parameters formulation, the explanatory variables can have varying influences on the response variables (Hensher and Greene, 2003). As discussed by Xiong and Mannering (2013), a finite mixture model assumes a restrictive homogeneity within the sub-population and the random parameters models assumes a continuous distribution. To overcome these model constraints, the finite mixture random parameters approach should be considered to not only account for group-specific heterogeneity in the population but also allow for the individual level heterogeneity within each group (Xiong and Mannering, 2013). In the future, both random parameters NB models and finite mixture random parameters NB models should be further applied with the EB method.

6. SUMMARY AND CONCLUSIONS

For transportation safety analysts, the EB method has become a popular tool due to its relative ease of implementation compared to other approaches and its use of combined data sources (i.e., observed and predicted crash counts) to help ensure quality estimates. This study developed the EB method when applying a GFMNB-2 model and demonstrated how this new formulation of the EB method can be applied to actual crash datasets. Both of the NB and GFMNB-2 model formulations were developed from two commonly used crash datasets collected in Indiana and Texas, respectively. For the traditional NB models, a fixed dispersion parameter was considered. For the GFMNB-2 models, the weight parameters were estimated as a linear combination of the predictor variables. The main conclusion of the study can be summarized as follows: although the site rankings by EB estimates may appear correlated between the NB and GFMNB-2 models, there are often large discrepancies. This was especially noticeable in the Texas dataset where 27.0% of the sites had ranking values that differed by more than 50 positions when the rankings based upon the EB estimates using the NB and

GFMNB-2 models were compared. The unobserved heterogeneity in the crash data may explain the observed difference in ranking. In this study, given that the accuracy of the EB technique depends on how we select similar sites (that ultimately provide the data needed for SPF development), the different hotspot identification results obtained from NB and GFMNB-2 models suggest that the EB estimates from the GFMNB-2 model may possibly improve the identification accuracy, especially when the NB model is misspecified for the heterogeneous crash data. Although the proposed method is not yet practice ready, transportation safety analysts may use the GFMNB-K model to calculate the EB estimates and avoid manually identifying similar groups within the heterogeneous crash dataset, a task that could prove difficult as the underlying subpopulations in the data are unknown. In the future, a simulation study should be conducted to further examine which crash prediction model (i.e. GFMNB-K or NB model) can better identify hotspots.

ACKNOWLEDGEMENTS

This research is sponsored jointly by the National Natural Science Foundation of China (grant no. 51608386) and Shanghai Sailing Program (grant no. 16YF1411900). The paper benefitted from the input of reviewers and their comments are greatly appreciated.

REFERENCES

AGUERO-VALVERDE, J. & JOVANIS, P. P. 2008. Analysis of road crash frequency with spatial models. *Transportation Research Record: Journal of the Transportation Research Board,* 2061**,** 55-63.

CHEN, C., ZHANG, G., QIAN, Z., TAREFDER, R. A. & TIAN, Z. 2016. Investigating driver injury severity patterns in rollover crashes using support vector machine models. *Accident Analysis & Prevention,* 90**,** 128-139.

CHENG, L., GEEDIPALLY, S. R. & LORD, D. 2013. The Poisson–Weibull generalized linear model for analyzing motor vehicle crash data. *Safety Science,* 54**,** 38-42.

CHENG, W. & WASHINGTON, S. 2008. New criteria for evaluating methods of identifying hot spots. *Transportation Research Record: Journal of the Transportation Research Board***,** 76-85.

CHENG, W. & WASHINGTON, S. P. 2005. Experimental evaluation of hotspot identification methods. *Accident Analysis & Prevention,* 37**,** 870-881.

CONNORS, R. D., MAHER, M., WOOD, A., MOUNTAIN, L. & ROPKINS, K. 2013. Methodology for fitting and updating predictive accident models with trend. *Accident Analysis & Prevention,* 56**,** 82-94.

DING, C., MA, X., WANG, Y. & WANG, Y. 2015. Exploring the influential factors in incident clearance time: Disentangling causation from self-selection bias. *Accident Analysis & Prevention,* 85**,** 58-65.

ELURU, N., BAGHERI, M., MIRANDA-MORENO, L. F. & FU, L. 2012. A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings. *Accident Analysis & Prevention,* 47**,** 119-127.

FRÜHWIRTH-SCHNATTER, S. 2006. *Finite mixture and Markov switching models*, Springer Science & Business Media.

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**,** 258-265.

GHARIB, M. 1995. Two characterisations of a gamma mixture distribution. *Bulletin of the*

Australian Mathematical Society, 52**,** 353-358.
HAUER, E. 1992. Empirical Bayes approach to the estimation of "unsafety": the multivariate regression method. *Accident Analysis & Prevention,* 24**,** 457-477.
HAUER, E. 1997. *Observational Before/After Studies in Road Safety. Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety*.
HAUER, E., HARWOOD, D., COUNCIL, F. & GRIFFITH, M. 2002. Estimating safety by the empirical Bayes method: a tutorial. *Transportation Research Record: Journal of the Transportation Research Board***,** 126-131.
HAUER, E., NG, J. C. & LOVELL, J. 1988. *Estimation of safety at signalized intersections (with discussion and closure)*.
HENSHER, D. A. & GREENE, W. H. 2003. The mixed logit model: the state of practice. *Transportation,* 30**,** 133-176.
JUN, J. 2010. Understanding the variability of speed distributions under mixed traffic conditions caused by holiday traffic. *Transportation Research Part C-Emerging Technologies,* 18**,** 599-610.
LORD, D., GEEDIPALLY, S. R., PERSAUD, B. N., WASHINGTON, S. P., VAN SCHALKWYK, I., IVAN, J. N., LYON, C. & JONSSON, T. 2008. Methodology to predict the safety performance of rural multilane highways.
LORD, D. & KUO, P.-F. 2012. Examining the effects of site selection criteria for evaluating the effectiveness of traffic safety countermeasures. *Accident Analysis & Prevention,* 47**,** 52-63.
LORD, D. & MANNERING, F. 2010. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice,* 44**,** 291-305.
MANNERING, F. L. & BHAT, C. R. 2014. Analytic methods in accident research: Methodological frontier and future directions. *Analytic methods in accident research,* 1**,** 1-22.
MANNERING, F. L., SHANKAR, V. & BHAT, C. R. 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. *Analytic Methods in Accident Research,* 11**,** 1-16.
MAYCOCK, G. & HALL, R. 1984. *Accidents at 4-arm roundabouts*.
MCLACHLAN, G. & PEEL, D. 2004. *Finite mixture models*, John Wiley & Sons.
MIRANDA-MORENO, L., FU, L., SACCOMANNO, F. & LABBE, A. 2005. Alternative risk models for ranking locations for safety improvement. *Transportation Research Record: Journal of the Transportation Research Board***,** 1-8.
MURALIDHARAN, O. 2010. An empirical Bayes mixture method for effect size and false discovery rate estimation. *The Annals of Applied Statistics***,** 422-438.
PARK, B.-J. & LORD, D. 2009. Application of finite mixture models for vehicle crash data analysis. *Accident Analysis & Prevention,* 41**,** 683-691.
PARK, B.-J., LORD, D. & LEE, C. 2014. Finite mixture modeling for vehicle crash data with application to hotspot identification. *Accident Analysis & Prevention,* 71**,** 319-326.
PARK, B.-J., LORD, D. & WU, L. 2016. Finite mixture modeling approach for developing crash modification factors in highway safety analysis. *Accident Analysis & Prevention,* 97**,** 274-287.
PARK, E. S., CARLSON, P. J., PORTER, R. J. & ANDERSEN, C. K. 2012. Safety effects of wider edge lines on rural, two-lane highways. *Accident Analysis & Prevention,* 48**,** 317-325.
PENG, Y., LORD, D. & ZOU, Y. 2014. Applying the Generalized Waring model for investigating sources of variance in motor vehicle crash analysis. *Accident Analysis & Prevention,* 73**,** 20-26.

PERSAUD, B., LAN, B., LYON, C. & BHIM, R. 2010. Comparison of empirical Bayes and full Bayes approaches for before–after road safety evaluations. *Accident Analysis & Prevention,* 42**,** 38-43.

RIGBY, R. & STASINOPOULOS, D. 2009. A flexible regression approach using GAMLSS in R. *London Metropolitan University, London*.

TANG, J., LIU, F., ZOU, Y., ZHANG, W. & WANG, Y. 2017. An Improved Fuzzy Neural Network for Traffic Speed Prediction Considering Periodic Characteristic. *IEEE Transactions on Intelligent Transportation Systems,* PP**,** 1-11.

VANGALA, P., LORD, D. & GEEDIPALLY, S. R. 2015. Exploring the application of the Negative Binomial–Generalized Exponential model for analyzing traffic crash data with excess zeros. *Analytic Methods in Accident Research,* 7**,** 29-36.

WASHINGTON, S. P., KARLAFTIS, M. G. & MANNERING, F. L. 2010. *Statistical and econometric methods for transportation data analysis*, CRC press.

WU, L., LORD, D. & ZOU, Y. 2015. Validation of crash modification factors derived from cross-sectional studies with regression models. *Transportation Research Record: Journal of the Transportation Research Board***,** 88-96.

WU, L., ZOU, Y. & LORD, D. 2014. Comparison of Sichel and Negative Binomial Models in Hot Spot Identification. *Transportation Research Record: Journal of the Transportation Research Board,* 2460**,** 107-116.

XIONG, Y. & MANNERING, F. L. 2013. The heterogeneous effects of guardian supervision on adolescent driver-injury severities: A finite-mixture random-parameters approach. *Transportation Research Part B: Methodological,* 49**,** 39-54.

YASMIN, S., ELURU, N., BHAT, C. R. & TAY, R. 2014. A latent segmentation based generalized ordered logit model to examine factors influencing driver injury severity. *Analytic methods in accident research,* 1**,** 23-38.

ZHA, L., LORD, D. & ZOU, Y. 2014. The Poisson Inverse Gaussian (PIG) Generalized Linear Regression Model for Analyzing Motor Vehicle Crash Data. *Journal of Transportation Safety & Security***,** 00-00.

ZOU, Y., LORD, D., ZHANG, Y. & PENG, Y. 2013a. Comparison of Sichel and Negative Binomial Models in Estimating Empirical Bayes Estimates. *Transportation Research Record: Journal of the Transportation Research Board,* 2392**,** 11-21.

ZOU, Y., TANG, J., WU, L., HENRICKSON, K. & WANG, Y. Quantile analysis of factors influencing the time taken to clear road traffic incidents. *Proceedings of the Institution of Civil Engineers - Transport,* 0**,** 1-9.

ZOU, Y., YANG, H., ZHANG, Y., TANG, J. & ZHANG, W. 2017. Mixture modeling of freeway speed and headway data using multivariate skew-t distributions. *Transportmetrica A: Transport Science***,** 1-28.

ZOU, Y., ZHANG, Y. & LORD, D. 2013b. Application of finite mixture of negative binomial regression models with varying weight parameters for vehicle crash data analysis. *Accident Analysis & Prevention,* 50**,** 1042-1051.

ZOU, Y., ZHANG, Y. & LORD, D. 2014. Analyzing different functional forms of the varying weight parameter for finite mixture of negative binomial regression models. *Analytic methods in accident research,* 1**,** 39-52.