Accident Analysis and Prevention 170 (2022) 106638

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Derivation of the Empirical Bayesian method for the Negative Binomial-Lindley generalized linear model with application in traffic safety

Ali Khodadadi [a,*], Ioannis Tsapakis [b], Mohammadali Shirazi [c], Subasish Das [d], Dominique Lord [a]

[a] Texas A&M University, 3136 TAMU, College Station, TX 77843-3136, United States
[b] Texas A&M Transportation Institute, 3500 NW Loop 410, Suite 315 San Antonio, TX 78229, United States
[c] University of Maine, Orono, Maine, 04469, United States
[d] Texas A&M Transportation Institute, 3135 TAMU, College Station, TX 77843, United States

ARTICLE INFO

Keywords:
Empirical Bayesian
Full Bayesian
Negative Binomial-Lindley
Hot spot identification
Expected crash values
Crash prediction model

ABSTRACT

The expected crash frequency is the long-term average crash count for a specific site. It is extensively used to systematically evaluate the crash risk associated with roadway elements. To estimate the expected crashes, the Empirical Bayesian (EB) approach is typically employed. The EB method is a computationally convenient approximation to the Full Bayesian (FB) method, which gained popularity due to its simple interpretation, computational efficiency, and the ability to account for the regression to the mean bias. However, the common EB method used in traffic safety analysis is only applicable when the traditional Negative Binomial (NB) model is used. The NB model, however, is not appropriate when data is highly dispersed, skewed, or has a large number of zero observations. The Negative Binomial-Lindley (NB-L) model is a mixture of the NB and Lindley distributions and has shown superior fit compared to the NB model, especially when the dataset is characterized by excess zero observations. Even though several studies have used the NB-L in developing crash prediction models, the application of the NB-L in other safety-related tasks (*e.g.*, hot spot identification) is largely neglected. This study proposed a framework to develop the EB method for the NB-L model and subsequently estimate the expected crash values.A comparison between the EB and FB estimates was performed to validate the approximation framework in general. The results indicated that the proposed EB framework is able to estimate expected crashes with comparable precision to the FB estimate, but with much less computational cost. In addition, a site ranking analysis using the EB estimates was conducted to validate the proposed approximation method in safety studies. However, it should be noted that any other type of safety analysis that requires access to the expected crashes can benefit from the proposed EB method. This study concluded that the proposed EB framework can properly approximate the underlying FB approach and can reasonably be considered as an alternative to the traditional EB formula derived from the NB model. The results of this study can help to extend the application of the advanced predictive models beyond predicting crashes to other safety-related tasks, with no additional computational efforts.

1. Introduction

The roadway safety management process involves multiple steps that are designed to monitor and reduce crash frequencies on existing roadways (Part, 2010). Of these steps, hot spot identification and safety effectiveness evaluation are two key approaches in safety evaluation and analysis. Hot spot identification identifies sites that can benefit the most from safety treatments. Safety effectiveness evaluation (*e.g.*, before-after analysis, cross-sectional analysis) evaluates how safety has changed because of one or more specific treatments implemented to reduce the crashes. Both analyses require reliable and stable measures to quantitatively evaluate the crash risk associated with a roadway entity in a certain time period. There are three main steps associated with each of the aforementioned analyses. The first step involves developing a crash prediction model (also referred to as crash-frequency model). Crash prediction models are the main tool to predict crash frequencies and identify crash contributing factors. In the second step, the crash prediction models are used to assess the crash risk associated with each

[*] Corresponding author.
E-mail addresses: a.khodadadi1994@tamu.edu (A. Khodadadi), i-tsapakis@tti.tamu.edu (I. Tsapakis), shirazi@maine.edu (M. Shirazi), s-das@tti.tamu.edu (S. Das), dlord@civil.tamu.edu (D. Lord).

https://doi.org/10.1016/j.aap.2022.106638
Received 27 September 2021; Received in revised form 7 March 2022; Accepted 12 March 2022

roadway element. The evaluated crash risk then can be used to determine the likelihood of crash occurrence for each specific site, as a function of site characteristics (*i.e.*, explanatory variables), in a certain time period. The final step involves ranking the sites in decreasing order based on the assessed crash risk (in case of hot spot identification), or determining the efficacy of the countermeasure(s) given the assessed crash risk before and after implementing the treatment.

The Negative Binomial (NB) is the most common statistical model to develop crash prediction models and estimate crash frequencies (Lord et al., 2021; Lord and Mannering, 2010; Mannering and Bhat, 2014). As opposed to the Poisson distribution which assumes the mean and variance of crash observations are equal, the NB distribution allows the variance of the response variable to be greater than the mean by using an additional parameter (referred to as over-dispersion parameter). Although research studies showed that the NB model addresses over dispersion commonly observed in crash data, this model does not necessarily account for issues related to other unique characteristics of crash data. In particular, crash datasets are often characterized by excess zero observations or low sample mean. The NB distribution is not flexible enough to deal with abundance of zero observations in the data (Geedipally et al., 2012). In addition, the NB model will output biased results as the sample mean goes lower (Lord, 2006). Different statistical models have been proposed by safety researchers to overcome limitations of the NB model. Poisson log-normal (Song et al., 2006; Park and Lord, 2007; Khazraee et al., 2018; Shirazi and Lord, 2019), Poisson-generalized inverse Gaussian (Zha et al., 2016; Zou et al., 2013), Conway-Maxwell-Poisson (Lord et al., 2010; Abdella et al., 2019), Semiparametric NB model (Shirazi et al., 2016), Poisson-Tweedie (Debrabant et al., 2018; Saha et al., 2020), Generalized Additive Models (Xie and Zhang, 2008), and Negative Binomial-Lindley (NB-L) (Zamani and Ismail, 2010; Lord and Geedipally, 2011; Geedipally et al., 2012; Shirazi et al., 2017; Shaon et al., 2018; Khodadadi et al., 2021) are just a few examples of advanced count models developed over time to overcome or alleviate the limitations of the NB model. NB-L in particular is the subject of interest in this study. The NB-L model is a mixture of the negative binomial and Lindley distribution. This model was first proposed by Zamani and Ismail (2010), and then used in multiple research fields dealing with sparse count data modeling including crash data analysis (Lord and Geedipally, 2011; Geedipally et al., 2012; Shaon et al., 2018; Khodadadi et al., 2021). The NB-L model offers extra flexibility using the Lindley distribution, resulting in a more powerful tool to fit to crash datasets (Shirazi et al., 2016). In particular, compared to the traditional NB models, the NB-L shows a better fit when a crash dataset contains many zero responses, or exhibits high dispersion, large skewness or long tail (Shirazi et al., 2017).

The three steps mentioned above (*i.e.*, developing crash prediction model, crash risk evaluation, and ranking/before-after analysis based on the evaluated crash risk) have been fully investigated for the well-known NB model. However, despite the superiority of the NB-L, no study has examined the application of the NB-L distribution or its generalized linear model (GLM) beyond the first step (predicting crashes). This study fills this research gap by deriving the equations to estimate the expected crash frequency for the NB-L model based on Full Bayesian (FB) and Empirical Bayesian (EB) framework. Therefore, the primary objectives of this study are to (1) develop an EB framework to calculate the expected crash values for the NB-L models, (2) compare the EB and FB expected values to determine if the proposed EB framework properly approximates the underlying FB paradigm, and (3) test the application of the NB-L and its EB estimates of the expected crashes in other safety-related analyses (site ranking in this study) to ensure the applicability of the proposed framework.

2. Background

The NB-L distribution has been used by researchers in various fields, including safety analysis (crash prediction models). Lord and Geedipally (2011) examined the application of the NB-L distribution in highway safety. They applied both the NB and NB-L distributions to simulated and empirical sparse datasets. They found that the NB-L outperforms the traditional NB distribution. To extend the application of NB-L in safety analysis, Lord et al. (2012) introduced a generalized linear NB-L model (NB-L GLM) to link the crash frequencies to the site characteristics. The regression approach has been employed in numerous transportation-related studies to estimate the relationships between the response variable and influential factors (Safaei et al., 2021b; Rostami et al., 2020; Aman et al., 2021; Aman and Smith-Colin, 2020; Safaei et al., 2021a; Asgharpour et al., 2021). Lord et al. (2012) observed that the NB-L GLM provides a better fit compared to the traditional NB GLM when analyzing a sparse or highly-dispersed dataset. Given the superiority of the NB-L over the traditional count models, different parameterizations of the NB-L model have been proposed, discussed, and applied in the literature. Two-parameters NB-L (Zamani and Ismail, 2010), three-parameters NB-L (Denthet et al., 2016), four-parameters NB-L (Tajuddin et al., 2020), Negative Binomial weighted-Lindley (NB-WLindley) (Khodadadi et al., 2022), and Negative Binomial-Lindley with different variance and dispersion structure (Khodadadi et al., 2021) are a few examples of the more advanced and more complex count models that are recently proposed to provide even greater flexibility to the original NB-L model.

Sometimes crash risk is quantified by criteria such as short-term crash frequency, crash rate, crash severity, or crash cost (Miaou and Song, 2005; Huang et al., 2009; Guo et al., 2020); however, ignoring the influential crash factors (*e.g.*, Annual Average Daily Traffic, roadway characteristics) could make these methods inefficient. In addition, the uncertainty associated with using the raw crash data could reduce the accuracy of the results, especially for long-term planning processes. The limitations associated with using the historical crash records alone led the researchers and transportation agencies to develop statistical approaches to more accurately predict the crash risk (*i.e.*, expected crash risk); they then used these approaches to rank the sites by the magnitude that their estimated crash risk exceeded the normal crash risk, which is estimated using sites with similar characteristics (Huang et al., 2009). The expected crash frequency is the long-term average crash count for a specific site. Considering the expected number of crashes in hot spot identification can overcome or minimize issues such as the regression-to-the-mean (RTM) bias (Hauer, 1997) or limited sample size (Miaou and Lord, 2003). Furthermore, given that the expected crash frequency uses both observed crash data and the number of crashes estimated from a crash prediction model, it can also account for the fundamentally non-linear relationship between the crash frequency and explanatory variables, the unobserved heterogeneity among the sites (Lord and Mannering, 2010), and the uncertainty associated with parameters of the underlying regression model (Miaou and Lord, 2003). The FB and EB are the two methods that are applied to estimate the expected crash frequencies. The FB method requires access to the hierarchical representation of the underlying predictive model in order to draw random samples from the posterior distribution of the parameters of interest. The hierarchical representation of Bayesian models makes the FB approach more flexible than other methods since it eliminates the need for the closed form representation of the model. Hierarchical models are frequently used in crash data analysis. One of the main advantages of the hierarchical models is the ability to incorporate relevant prior knowledge and common beliefs about the parameters into the modeling process in a natural probabilistic way. The FB method has broadly been used in various safety-related analyses such as estimating crash prediction models, before-and-after studies, and hot spot identification (Guo et al., 2019; Farid et al., 2017; Aguero-Valverde and Jovanis, 2009; Miaou and Song, 2005; Miranda-Moreno et al., 2013; Shirazi et al., 2017; Lan and Persaud, 2011; Persaud et al., 2010; Pu et al., 2020).

Despite the broad applications of the FB approach, this method is often computationally intensive. In particular, for complex models involving a large number of observations and many variables, the FB method can be a time-consuming task due to the integration over the distribution of many parameters. Furthermore, the FB approach requires consideration of a prior distribution on all the unknown parameters. However, finding a well-reasoned and well-defined prior distribution for the problem in hand could be quite challenging. The EB method is a promising alternative to the standard FB paradigm. The EB approach is a special case of the general FB framework when some assumptions are simplified. Unlike the FB method where each parameter is defined as a random variable, the EB approach assumes that the parameters in the highest level of hierarchy are known without any uncertainty (Huang et al., 2009). The EB method could be thought of a computationally convenient approximation to the FB method, and has gained popularity among the safety analysts and transportation agencies due to its simple interpretation, computational efficiency, and the ability to account for the RTM bias (Miaou and Lord, 2003; Huang et al., 2009; Persaud et al., 2010; Khattak et al., 2018; Das et al., 2019). Despite the fact that the EB method is a reliable method to estimate the expected crash risks, it is just an approximation to a more general FB paradigm. First of all, the EB method does not account for the uncertainty embedded in the parameters. Parameters of the crash prediction model are estimated from the observed crash data which are naturally subjected to uncertainty. Ignoring these uncertainties might lead to overestimating the precision and/or less accurate estimates (Miaou and Lord, 2003). Secondly, the EB method might be criticized for a double usage of the data (Huang et al., 2009; Hauer, 1997). In an ideal EB procedure, two sources of data should be used. One source is to develop the crash prediction model and get the predicted crash values, and another is to independently enrich the model with prior knowledge. However, in practice, the safety performance functions (SPF) are obtained from the recorded crash frequency and thus both predicted crash frequency and observed crash frequency are derived from the same source of information. Despite these limitations, the EB expected crash frequency is a good approximation for the expected values derived from the FB method as it still accounts for the RTM, can refine the predicted mean of an entity (Zou et al., 2013), and yields similar estimates as FB estimates with comparable precision but less computational cost. All these confirm that the EB approach is a proper, yet less expensive alternative compared to the FB method.

The EB method has been used for the NB model where the expected crash frequency is defined as a linear combination of the predicted crash frequency (derived from the SPF model) and the observed crash frequency (Hauer et al., 2002). Similarly, the EB procedure proposed in the highway safety manual (HSM) is only applicable when the traditional NB model is being used. As mentioned earlier, different extensions of the NB model have been introduced to deal with problematic characteristics of crash data. As these extensions get more complex and go deeper in the hierarchy, the EB procedure becomes harder to implement as the closed form of such distributions are unavailable or hard to compute analytically. Consequently, there is a clear need to examine the application of the EB method when more advanced models are being used. To this end, some studies attempted to translate the EB framework for more complex models such as Sichel (Zou et al., 2013) or finite mixture NB (Zou et al., 2018).

In terms of ranking procedures, generally, there are two types of ranking approaches, naive ranking and model-based ranking (Huang et al., 2009). The naive ranking method uses the raw crash data to make an ordered list of sites for hot spot identifications. The model-based ranking approaches, however, take the expected crash values as an indication of the crash risk (Huang et al., 2009). The expected crash values are extensively used to sort a list of roadway entities. However, the obtained sorted list could potentially differ among the ranking criteria since they are based on different measures and logics. Several model-based ranking criteria for hot spot identification have been investigated in the literature to better represent the stochastic nature of the crash data. Many studies have recommended ranking sites based on the FB or EB expected value of the Poisson mean, and they concluded that the use of posterior Poisson mean would result in a more reliable and more accurate order compared to the naive ranking criteria (Guo et al., 2019; Lee et al., 2019; Meng et al., 2020; Lan and Persaud, 2011). In the same token, Shen and Louis (1998) discussed that the posterior Poisson mean is an optimal choice when inferences about the expected crashes are of interest. However, it might perform poorly if the rank of the expected crashes is the subject of interest. Consequently, some studies have attempted to directly take uncertainties in rankings into considerations and employed a Bayesian framework in ranking criteria as well (Laird and Louis, 1989; Miaou and Song, 2005; Liu and Sharma, 2018; Shen and Louis, 1998). The posterior ranking criteria (*e.g.*, posterior expected, mode, or median rank) takes all posterior simulations into account for each site's crash risk (not only the posterior mean), and then outputs a ranked list of sites for each simulation run, accordingly. In a comparison study between the EB and FB approach for hot spot identification, Lan and Persaud (2011) explored eight different ranking criteria including posterior expected, posterior mode, and posterior median ranking. The authors concluded that, in general, the posterior rank criteria would perform better than other model-based and naive ranking methods. Similar results were observed in the study done by Laird and Louis (1989). They also concluded that the posterior distribution of a parameter's rank typically carries more information about the true ranking in comparison to the integer rank of that parameter.

3. Methodology

The NB-L distribution is a mixture of the NB and the one-parameter Lindley distribution. The NB-L distribution offers a more flexible structure with more degrees of freedom compared to the traditional NB distribution. Different hierarchical variations of the NB-L distribution have been introduced and analyzed (Geedipally et al., 2012; Zamani and Ismail, 2010; Gomez-Deniz and Calderin-Ojeda, 2017). This study used the original representation developed by Zamani and Ismail (2010) and the generalized linear model proposed by Geedipally et al. (2012) to derive FB and EB procedures. Let Y_i denote the crash frequency following an NB distribution with shape parameter, p_i, and rate (or over dispersion) parameter, ϕ. The hierarchical expression for the NB-L generalized linear model (NB-L GLM) is defined as follows (Geedipally et al., 2012):

$$\begin{aligned} & Y_i \sim NB\ (p_i, \phi); \quad \phi > 0,\ 0 < p_i < 1 \\ & p_i = e^{-\eta_i} \\ & \eta_i \sim Lindley\ (\theta_i) \\ & \\ & \theta_i = \mu_i = e^{\boldsymbol{\beta}\mathbf{X}_i} \\ & \phi \sim \pi_\phi \\ & \boldsymbol{\beta} \sim \pi_\beta \end{aligned} \tag{1}$$

where, θ is the Lindley parameter, $\mathbf{X}_i = (1, X_1, X_2, \ldots X_m)$ is the vector including the contributing variables for site $i, \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)$ is the vector of regression coefficient to be estimated, and π_ϕ and π_β are the prior distribution for ϕ and $\boldsymbol{\beta}$, respectively.

The GLM representation in Eq.(1) was suggested by Geedipally et al. (2012) as an alternative to NB-L GLM, but it has not yet been used for modeling due to its complexity. Note that the above NB-L GLM representation is available in closed-form. Therefore, this representation works well for addressing the objectives of this paper since the EB analysis usually requires the maximum likelihood estimates (MLE) of the parameters, which require access to the closed form formulation of the probability mass function (pmf). In the next section, the derivation of the expected crash values is discussed in detail for both FB and EB methods.

3.1. Full Bayesian expected values

In the FB paradigm, both parameters and hyper-parameters are assumed to follow pre-defined distributions (*i.e.*, prior distribution) which represent the underlying uncertainty in the parameters. The FB method treats all the parameters as unknown random variables and takes all their uncertainties into account by integrating over the prior distributions. In the Bayesian context, either for EB or FB methods, the posterior predictive distribution is used to estimate the expected crash values. The posterior predictive distribution represents the distribution of the expected data given the observed data and predictive model. The posterior predictive distribution for an expected data point, y_{exp}, given the observed value, y_{obs}, could be written as follows:

$$p\ (y_{exp}|y_{obs}) = \int_{\gamma} p\ (y_{exp}|\gamma, y_{obs})\ p\ (\gamma|y_{obs})\ d\gamma \tag{2}$$

where, $p\ (y_{exp}|\gamma, y_{obs})$ is the likelihood of the expected data given the observed data and model parameters (γ), and $p\ (\gamma|y_{obs})$ is the posterior distribution of the parameters give the observed data. In this section, first we document the derivation of the posterior predictive distribution and the FB expected values for the NB model; then, the same procedure is extended to derive the FB expected crash frequencies for the NB-L model.

The NB distribution itself could be re-parameterized as a continuous mixture of the Poisson and Gamma distributions, where the Poisson mean follows a Gamma distribution. The hierarchical representation of the NB GLM with mean, μ, and over-dispersion parameter, ϕ, could be written as follows:

$$\begin{aligned} &Y_i \sim Poisson\ (\lambda_i) \\ &\lambda_i \sim Gamma\ (\phi, \phi/\mu_i) \\ &\mu_i = e^{\boldsymbol{\beta}\mathbf{X}_i} \end{aligned} \tag{3}$$

$$\begin{aligned} &\boldsymbol{\phi} \sim \pi_{\phi} \\ &\boldsymbol{\beta} \sim \pi_{\boldsymbol{\beta}} \end{aligned}$$

Using the definition of the posterior predictive distribution in Eq.(2), the probability of the expected crashes, y_{exp}, given the observed crash data is as follows:

$$p\ (y_{exp}|y_{obs}) = \int_{\lambda} p\ (y_{exp}|\lambda, y_{obs})\ p\ (\lambda|y_{obs})\ d\lambda \tag{4}$$

where, $p\ (y_{exp}|\lambda) \sim Poisson\ (\lambda)$ and $p\ (\lambda|y_{obs})$ is the posterior distribution of the Poisson mean, λ, which could be written as follows by definition:

$$p\ (\lambda|y_{obs}) = \int_{\phi,\boldsymbol{\beta}} p\ (\lambda|y_{obs}, \boldsymbol{\phi}, \boldsymbol{\beta})\ \pi_{\phi,\boldsymbol{\beta}}\ d(\boldsymbol{\phi}, \boldsymbol{\beta}) \tag{5}$$

Given Eq. (4) and Eq. (5), the Full Bayesian posterior predictive distribution of the NB GLM could be written as follows:

$$p\ (y_{exp}|y_{obs}) = \int_{\lambda} p\ (y_{exp}|\lambda, y_{obs})\ \left(\int_{\phi,\boldsymbol{\beta}} p\ (\lambda|y_{obs}, \boldsymbol{\phi}, \boldsymbol{\beta})\ \pi_{\phi,\boldsymbol{\beta}}\ d(\boldsymbol{\phi}, \boldsymbol{\beta})\right)\ d\lambda \tag{6}$$

Let $\mathscr{P}$ and $\mathscr{G}$ denote the Poisson and Gamma distributions, respectively. Given the definition of the posterior distribution, we know that:

$$p\ (\lambda|y_{obs}, \boldsymbol{\phi}, \boldsymbol{\beta}) \propto \mathscr{P}(y_{obs}|\lambda)\ \ \mathscr{G}(\lambda|\boldsymbol{\phi}, \boldsymbol{\beta}) \tag{7}$$

Therefore, given the fact that the Gamma distribution is a conjugate prior for the Poisson distribution, Eq.(6) could be further simplified as follows:

$$p\ (y_{exp}|y_{obs}) = \int_{\lambda} \mathscr{P}(\lambda)\ \left(\int_{\phi,\boldsymbol{\beta}} \mathscr{G}(y_{obs} + \phi, 1 + \phi \Big/ \mu)\ \pi_{\phi,\boldsymbol{\beta}}\ d(\phi, \boldsymbol{\beta})\right)\ d\lambda \tag{8}$$

Using the FB approach, instead of solving the integral or calculating the closed form representation, we can take the Monte Carlo Markov Chain (MCMC) approach to draw random samples from the posterior predictive distribution. The following steps describe the procedure to draw a random sample from the posterior predictive distribution at each site *i*:

- Draw a random sample from the prior distributions, $\pi_{\boldsymbol{\beta}}$ and π_{ϕ}; then, calculate μ_i;
- Plug in the samples from the previous step in $\mathscr{G}(y + \phi, 1 + \phi/\mu_i)$, and then draw a random sample from the distribution. It gives us a random sample from the λ's posterior distribution;
- Plug in the posterior λ sample from the previous step in $\mathscr{P}(\lambda)$, and then draw a random sample from the distribution. It gives us a random sample from the posterior predictive distribution of the crash frequency at site *i*.

By repeating the hierarchical procedure described above, we can have the necessary samples from the posterior predictive distribution to estimate the expected crash frequency. For this purpose, we can use any measure of centrality (*i.e.*, mode, mean, median) to average out the predictive distribution and achieve the expected value. Note that in the last step, $(y_{exp}|\lambda, y_{obs})$ follows a Poisson distribution. The parameter of the Poisson distribution shows its mean value. Therefore, by drawing random samples and then taking the average, we can find the conditional expectation of λ given the observed data, $E(\lambda|y_{obs})$. This means that if we parameterize the crash frequency as a Poisson mixture model with parameter λ, the posterior predictive distribution would be the same as the posterior distribution of λ. This concept will be useful when developing EB estimates for the NB-L GLM.

Even though the NB-L has been discussed and documented in the literature, no study has yet outlined the derivation of expected crash values for the NB-L GLM. A similar procedure as that used in developing FB for NB GLM is also applicable in the case of NB-L model. Using the NB-L GLM formulation written in Eq.(1), the posterior predictive distribution could be expressed as follows:

$$p\ (y_{exp}|y_{obs}) = \int_{\eta,\phi} p\ (y_{exp}|\eta, \phi, y_{obs}) p\ (\eta, \phi|y_{obs})\ d(\eta, \phi) \tag{9}$$

Putting the full posterior expression of η parameter in Eq.(9), the above formulation could be re-written as follows ($\mathscr{NB}$ denotes the NB distribution):

$$p\ (y_{exp}|y_{obs}) = \int_{\eta,\phi} \mathscr{NB}(e^{-\eta}, \phi)\left(\int_{\boldsymbol{\beta}} p\ (\eta|\boldsymbol{\beta}, y_{obs})\ \pi_{\boldsymbol{\beta}}\ d(\boldsymbol{\beta})\right)\ \pi_{\phi}\ d(\eta, \phi) \tag{10}$$

The Lindley distribution does not have any conjugate prior; hence, the integral above cannot be further simplified. The following procedure should be followed to draw random samples:

- Draw a random sample from each prior distribution, $\pi_{\boldsymbol{\beta}}$ and π_{ϕ}.
- Plug in the sample $\boldsymbol{\beta}$ from the previous step in $p\ (\eta|\boldsymbol{\beta} \mathbf{y}_i)$, and draw a random sample from the distribution. It gives us a random sample from posterior distribution of η.
- Plug in the posterior η sample from the previous step and ϕ from the first step in $\mathscr{NB}(e^{-\eta}, \phi)$, and draw a random sample from the distribution. It gives us a random sample from the posterior predictive of the crash frequency at site *i*.

This procedure could be easily formulated and summarized in statistical software developed for MCMC analysis such as WinBUGS (Lunn et al., 2000), or JAGS (Plummer et al., 2016).

3.2. Empirical Bayesian expected values

As mentioned in the previous section, the main motivation behind the EB method is simplifying the computationally intensive steps of the FB procedure. Unlike the FB method where all the parameters are random variables specified by prior distributions, the EB method does not consider the uncertainty associated with the parameters; instead, the point estimate of the parameters, either maximum likelihood (MLE) or method of moment (MOM) estimates, is used in the highest levels of hierarchy. In the following, the EB approach for the NB model is reviewed, and then the EB approximation for the NB-L model is developed.

The three-step procedure explained for the FB analysis is simplified by some approximations to obtain the expected crash values in the EB paradigm as follows:

Step one – Estimate the parameters of the highest level of hierarchy.

To obtain parameter estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\phi}$, closed form representation of the NB GLM is essential. The closed form expression of NB GLM is available and could be easily achieved by marginalizing λ variable out:

$$p\ (y|\phi,\boldsymbol{\beta}) = \int_{\lambda} p\ (y|\lambda)\ p\ (\lambda|\phi,\boldsymbol{\beta})\ d\lambda = \frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\mu}{\mu+\phi}\right)^{y} \tag{11}$$

The above expression is the pmf of the NB distribution. $\widehat{\boldsymbol{\beta}}$ and $\widehat{\phi}$ are called the marginal maximum likelihood estimates (MMLE) and could be simply calculated through MLE or MOM approaches.

Step two – Derive the expected value of posterior Poisson mean.

As mentioned before, Gamma is a conjugate prior for Poisson distribution. As a result, the posterior distribution of the Poisson mean, λ, given the data as well as its expected value, is available in closed form (Zou et al., 2018):

$$\begin{aligned} p\left(\lambda\middle|\boldsymbol{\beta},\phi,y\right) &= Gamma\ \left(y+\phi, 1+\phi\middle/\mu\right)\ E\left(\lambda\middle|\boldsymbol{\beta},\phi,y\right) \\ &= \frac{y+\phi}{1+\phi/\mu} = \left(\frac{\mu}{\mu+\phi}\right)y + \left(\frac{\phi}{\mu+\phi}\right)\mu \end{aligned} \tag{12}$$

The above formula (Eq. (13)) is the known EB formula for the expected crash value, which is extensively used in safety analysis (Hauer et al., 2002). Finally, by plugging in $\widehat{\boldsymbol{\beta}}$ and $\widehat{\phi}$ from the previous step in Eq. (13), the EB expected crash frequency for each site is calculated.

The outlined procedure for derivation of the EB estimates for the NB GLM is also applicable in the NB-L GLM. Each step is thoroughly discussed in the following:

Step one – Estimate the parameters of the highest level of hierarchy.

Estimating ϕ and $\boldsymbol{\beta}$ requires the closed form representation of the NB-L GLM. The hierarchical representation of the NB-L model outlined in Eq.(1) can be expressed in closed form by marginalizing η parameter out:

$$p\ (y|\boldsymbol{\beta},\phi) = \int_{\eta} p\ (y|\phi,\eta)\ p\ (\eta|\boldsymbol{\beta})\ d\eta \tag{13}$$

Solving for the above integral would result in the pmf of the NB-L GLM. It follows from the pmf of the NB-L distribution which was developed by (Zamani and Ismail, 2010):

$$p\ (y_i|\phi,\boldsymbol{\beta}) = \frac{e^{2\boldsymbol{\beta}\mathbf{X}_i}}{1+e^{\boldsymbol{\beta}\mathbf{X}_i}}\binom{\phi+y_i-1}{y_i}\sum_{j=0}^{y_i}(-1)^j\binom{y_i}{j}\frac{e^{\boldsymbol{\beta}\mathbf{X}_i}+\phi+j+1}{\left(e^{\boldsymbol{\beta}\mathbf{X}_i}+\phi+j\right)^2} \tag{14}$$

The MLEs, $\widehat{\boldsymbol{\beta}}$ and $\widehat{\phi}$, could then be calculated by maximizing the likelihood (or log-likelihood) function. The log-likelihood function of NB-L GLM is given as follows:

$$ll = \sum_{i=1}^{n}[log\binom{\phi+y_i-1}{y_i} + 2(\boldsymbol{\beta}\mathbf{X}_i) - log(1+e^{\boldsymbol{\beta}\mathbf{X}_i}) \tag{15}$$

$$+log\left(\sum_{j=0}^{y_i}(-1)^j\binom{y_i}{j}\frac{e^{\boldsymbol{\beta}\mathbf{X}_i}+\phi+j+1}{\left(e^{\boldsymbol{\beta}\mathbf{X}_i}+\phi+j\right)^2}\right)]$$

The first partial derivative with respect to the unknown parameters could be written as follows:

$$\frac{\partial ll}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n}\left(2\mathbf{X}_i - \frac{\mathbf{X}_i}{1+e^{\boldsymbol{\beta}\mathbf{X}_i}}\right) + \frac{\sum_{j=0}^{y_i}(-1)^{j+1}\binom{y_i}{j}\frac{\mathbf{X}_i(e^{\boldsymbol{\beta}\mathbf{X}_i}+\phi+j+2)}{(e^{\boldsymbol{\beta}\mathbf{X}_i}+\phi+j)^3}}{\sum_{j=0}^{y_i}(-1)^j\binom{y_i}{j}\frac{e^{\boldsymbol{\beta}\mathbf{X}_i}+\phi+j+1}{(e^{\boldsymbol{\beta}\mathbf{X}_i}+\phi+j)^2}} \tag{16}$$

$$\frac{\partial ll}{\partial\phi} = \frac{\partial}{\partial\phi}\left[\sum_{i=1}^{n}log\binom{\phi+y_i-1}{y_i}\right] + \frac{\sum_{j=0}^{y_i}(-1)^{j+1}\binom{y_i}{j}\frac{e^{\boldsymbol{\beta}\mathbf{X}_i}+\phi+j+2}{(e^{\boldsymbol{\beta}\mathbf{X}_i}+\phi+j)^3}}{\sum_{j=0}^{y_i}(-1)^j\binom{y_i}{j}\frac{e^{\boldsymbol{\beta}\mathbf{X}_i}+\phi+j+1}{(e^{\boldsymbol{\beta}\mathbf{X}_i}+\phi+j)^2}} \tag{17}$$

The first part in Eq.(17) could be re-written as follows (Klugman et al., 2012; Tajuddin et al., 2020):

$$\frac{\partial}{\partial\phi}\sum_{i=1}^{n}log\binom{\phi+y_i-1}{y_i} = \sum_{i=1}^{n}\sum_{m=0}^{y_i-1}\frac{1}{\phi+m} \tag{18}$$

As a result, the partial derivative of the log-likelihood with respect to ϕ is given as:

$$\frac{\partial ll}{\partial\phi} = \sum_{i=1}^{n}\left(\sum_{m=0}^{y_i-1}\frac{1}{\phi+m}\right) + \frac{\sum_{j=0}^{y_i}(-1)^{j+1}\binom{y_i}{j}\frac{e^{\boldsymbol{\beta}\mathbf{X}_i}+\phi+j+2}{(e^{\boldsymbol{\beta}\mathbf{X}_i}+\phi+j)^3}}{\sum_{j=0}^{y_i}(-1)^j\binom{y_i}{j}\frac{e^{\boldsymbol{\beta}\mathbf{X}_i}+\phi+j+1}{(e^{\boldsymbol{\beta}\mathbf{X}_i}+\phi+j)^2}} \tag{19}$$

The above derivative equations could be simultaneously solved using numeric methods (gradient descend, Newton–raphson, etc.) in order to estimate the unknown parameters.

Step two – Derive the expected value of posterior Poisson mean

The hierarchical representation of the NB-L GLM defined in Eq.(1) does not involve the Poisson mean, λ, parameter since λ has already been marginalized out in the definition of the NB distribution. However, we can formulate the NB-L GLM as a functional of λ by breaking down the NB distribution to a mixture of Poisson and Gamma distribution:

$$\begin{aligned} &Y_i \sim Poisson\ (\lambda) \\ &\lambda \sim Gamma\ (\phi, \frac{e^{-\eta_i}}{1-e^{-\eta_i}}) \\ &\eta_i \sim Lindley\ (\theta_i) \\ &\theta_i = \mu_i = e^{\boldsymbol{\beta}\mathbf{X}_i} \end{aligned} \tag{20}$$

Following from the above hierarchical representation, the pmf of the NB-L GLM could be re-written as a function of the Poisson mean, λ:

$$p\ (y|\phi,\boldsymbol{\beta}) = \int_{\lambda} p\ (y|\lambda)\ \left(\int_{\eta} p\ (\lambda|\phi,\eta)\ p\ (\eta|\boldsymbol{\beta})\ d(\eta)\right)\ d\lambda \tag{21}$$

Clearly, the above expression is the pmf of a Poisson mixture distribution with mixing distribution as follows:

$$p\ (\lambda|\phi,\boldsymbol{\beta}) = \int_{\eta} p\ (\lambda|\phi,\eta)\ p\ (\eta|\boldsymbol{\beta})\ d(\eta) \tag{22}$$

Neither the mixing distribution itself, $p\ (\lambda|\phi,\boldsymbol{\beta})$, nor its posterior distribution, $p\ (\lambda|\phi,\boldsymbol{\beta},y)$, can be parameterized in closed-form. Hence, we can't directly calculate the posterior expectation of the Poisson mean, $E(\lambda|\phi,\boldsymbol{\beta},y)$, in the same way we did in the case of the NB model. Instead, we can take advantage of a useful property of the Poisson mixture distributions documented by Karlis and Xekalaki (2005) and Willmot (1986). Suppose Y follows a mixture Poisson distribution with pmf $p\ (x)$. Then, the posterior moments of any order of the Poisson mean, $E\ (\lambda^r|X = x)$, could be calculated as follows:

$$E\ (\lambda^r|X=x) = \frac{p\ (x+r)}{p\ (x)}(x+1)\dots(x+r) \tag{23}$$

where, $p\ (y)$ is the mixed Poisson pmf.

We can use this property to calculate the posterior expectation of the Poisson mean when the NB-L model is being used. In the same way as before, let Y follow the NB-L distribution derived in Eq.(14). Then, the posterior expectation of the Poisson mean could be written as follows:

$$E(\lambda|y,\phi,\boldsymbol{\beta}) = \frac{p\ (y+1|\phi,\boldsymbol{\beta})}{p\ (y|\phi,\boldsymbol{\beta})}\ (y+1) \tag{24}$$

$$= \frac{\frac{e^{2\beta X}}{1+e^{\beta X}} \binom{\phi + y + 1 - 1}{y+1} \sum_{j=0}^{y+1} (-1)^j \binom{y+1}{j} \frac{e^{\beta X}+\phi+j+1}{(e^{\beta X}+\phi+j)^2}}{\frac{e^{2\beta X}}{1+e^{\beta X}} \binom{\phi + y - 1}{y} \sum_{j=0}^{y} (-1)^j \binom{y}{j} \frac{e^{\beta X}+\phi+j+1}{(e^{\beta X}+\phi+j)^2}} \ (y+1) \tag{25}$$

The formula in Eq.(25) can be further summarized as follows:

$$E(\lambda|y, \phi, \boldsymbol{\beta}) = \frac{A(y+1)}{A(y)} \ (y+1) \tag{26}$$

where,

$A(y) = \sum_{j=0}^{y}(-1)^j \binom{y}{j} \frac{e^{\beta X}+\phi+j+1}{(e^{\beta X}+\phi+j)^2} = \sum_{j=0}^{y}(-1)^j \binom{y}{j} \frac{\mu+\phi+j+1}{(\mu+\phi+j)^2}$ Eq.(26) can be used to estimate the posterior mean of λ without knowing any information about the mixing distribution or to solve for the expectation definition itself. The proposed EB formula is comparable with the famous EB formula derived for the NB model indicated in Eq. (13). Finally, we need to incorporate the MLEs, $\widehat{\phi}$ and $\widehat{\boldsymbol{\beta}}$, in Eq.(26). The resulting value is the desired EB expected crash value for the NB-L model. As observed, it does not involve any intensive computation (like the FB method) or solving any complex integral. The next sections describe the dataset used for empirical evaluation of the proposed EB framework as well as the implementation details.

4. Data description

In the previous section, the derivation of the expected crash value using FB and EB frameworks was discussed for the NB-L GLM. In order to examine the developed FB and EB frameworks, this study used two datasets. Both Virginia (2014–2019) and Texas (2014–2019) datasets represent the crash statistics of the non-federal aid system (NFAS) roadways discussed in Khodadadi et al. (2021b) and Das et al. (2021). NFAS roadways are typically characterized by lower volumes and lower crash frequency in comparison with other roadway functional classifications (Khodadadi et al., 2021). Consequently, many NFAS segments experience zero crashes. Further, there are a lot of missing data in many roadway characteristics (*e.g.*, shoulder width) that could potentially be used as predictors. Therefore, a limited number of variables were available to use in the generalized linear modeling framework. However, as the emphasis of this study is to assess and compare the developed framework, using fewer variables is not an issue. The summary statistics of both datasets are provided in Table 1.

5. Modeling results

In this section, the modeling results for both NB and NB-L GLMs are presented, then the proposed EB procedure for NB-L GLM is examined. First, the EB and FB estimates of the expected crashes were compared to generally validate the proposed EB method and show how well the EB estimates mimic the FB estimates. Then, the EB and FB estimates were used in site ranking analysis to determine how similar the produced ranks and identified hot spots were.

Table 1
Summary Statistics of datasets.

Dataset	Variables	Min	Max	Average	Standard Deviation
Texas	Number of crashes	0	15	0.86	1.65
	Average 5-years AADT (vpd)	43	1166	313.8	253
	Segment length (miles)	0.10	4.41	0.96	0.93
Virginia	Number of crashes	0	8	2.01	2.09
	Average 5-years AADT (vpd)	163	5180	694	625
	Segment length (miles)	0.13	5.67	1.35	1.08

5.1. Crash prediction models

The NB and NB-L GLMs were developed for each dataset. For each model, only the AADT and segment length were included as contributing covariates. It should be noted that as the models are developed using the same functional form and compared using the same dataset, therefore, as noted earlier, the omitted variable bias would not be an issue. Also, as discussed before, this study aims to develop EB estimates for the expected crash values and explore whether they approximate the FB estimate properly. Including more variables might enhance the predictive models' performance; however, it will not affect the underlying theoretical framework of deriving EB estimates.

For each GLM, Full Bayesian and maximum likelihood estimates were calculated. We employed the MCMC method using an open-sourced R package, called "RJAGS" (Plummer et al., 2003), to estimate the posterior of parameters. This study assumed a non-informative gamma, and a non-informative normal distribution for the prior distribution of $\boldsymbol{\beta}$'s and ϕ parameters, respectively. In total, three chains and 60,000 iterations were set up to ensure the MCMC convergence. The first 4,000 samples of each chain were discarded. Also, to reduce the potential auto-correlation among the random draws, every third sample of the rest was used for estimations of unknown parameters.

The maximum likelihood estimates are needed in order to develop the EB estimates. Unlike the NB distribution, the NB-L is not a part of natural exponential family distributions. Hence, its log-likelihood function is not strictly concave. Numerical approaches equipped with proper initial values are needed to approach the global maximum point. Due to the significant sensitivity observed among the NB-L partial derivative equations and the initial values, a meta-heuristic genetic algorithm, together with a gradient descent approach, was utilized to ensure convergence to the global maximum point. For this purpose, the "GA" package in R (Scrucca et al., 2013) was used to solve the optimization problem. A total of 200 iterations with 500 initial populations were considered to maximize the objective function. A gradient descant approach was also employed in each iteration of the genetic algorithm to locally search for better estimates and further enhance the maximization process.

The FB estimates and MLEs for the Texas and Virginia datasets for both Nb and NB-L GLMs are summarized in Table 2 and Table 3, respectively. Three performance measures, namely Deviance Information Criteria (DIC), Mean Absolute Deviance (MAD), and Widely Applicable Information Criteria (WAIC) were used for model comparisons (Lord et al., 2021). WAIC was developed by Vehtari et al. (2017) and it appeared to be a robust alternative for DIC in the Bayesian framework (Watanabe and Opper, 2010; Khodadadi et al., 2021). All the performance measures showed that the NB-L models fit the data better than the NB models. These results were expected since both datasets were characterized by a large number of zeros and high skewness (the domain under which the NB-L model performs better than the NB).

As indicated in Table 2 and Table 3, the signs and magnitudes of estimates are different across models. This issue could be accredited to the particular representation of the NB-L model used in this study. As opposed to the NB model where the mean function has a log-linear association with covariates, the mean of the NB-L model has a non-linear relationship with covariates. Consequently, the magnitude and sign of the estimates do not necessarily represent the causal relationship between covariates and crash frequencies. This issue and its associated limitations are covered in the Discussion section further below.

5.2. Expected crash values

Using the above modeling results, we attempted to calculate both FB and EB expected crash values. Estimating the FB expected values in-

Table 2
Modeling results for Virginia dataset.

Variables	NB GLM		NB-L GLM	
	FB estimate (s.d.)	MLE	FB estimate (s.d.)	MLE
Intercept (β_0)	−3.48 (0.22)	−3.47	7.86 (0.36)*	8.27*
Log (AADT) (β_1)	0.51 (0.03)	0.51	−0.51(0.04)*	−0.53*
Length (β_2)	0.57 (0.02)	0.58	−0.60 (0.03)*	−0.59*
ϕ	3.94 (0.42)	3.87	75.07 (14.74)	96.63
DIC	8301		7135	
WAIC	6859		6630	
MAD	0.89		0.86	
Log-likelihood	−3119		−2888	

*The estimates for the NB-L GLM do not carry information regarding the causal association of covariates and crashes (see discussion below).

Table 3
Modeling results for Texas dataset.

Variables	NB GLM		NB-L GLM	
	FB estimate (s.d.)	MLE	FB estimate (s.d.)	MLE
Intercept (β_0)	−5.49 (0.12)	−5.49	8.00 (0.28)*	7.21*
Log (AADT) (β_1)	0.80 (0.02)	0.79	−0.73 (0.02)*	−0.68*
Length (β_2)	0.66 (0.02)	0.66	−0.60 (0.02)*	−0.61*
ϕ	1.07 (0.04)	1.07	17.83 (3.16)	10.71
DIC	26155		22471	
WAIC	21031		20761	
MAD	0.45		0.44	
Log-likelihood	−9023		−8847	

*The estimates for the NB-L GLM do not carry information regarding the causal association of covariates and crashes (see discussion below).

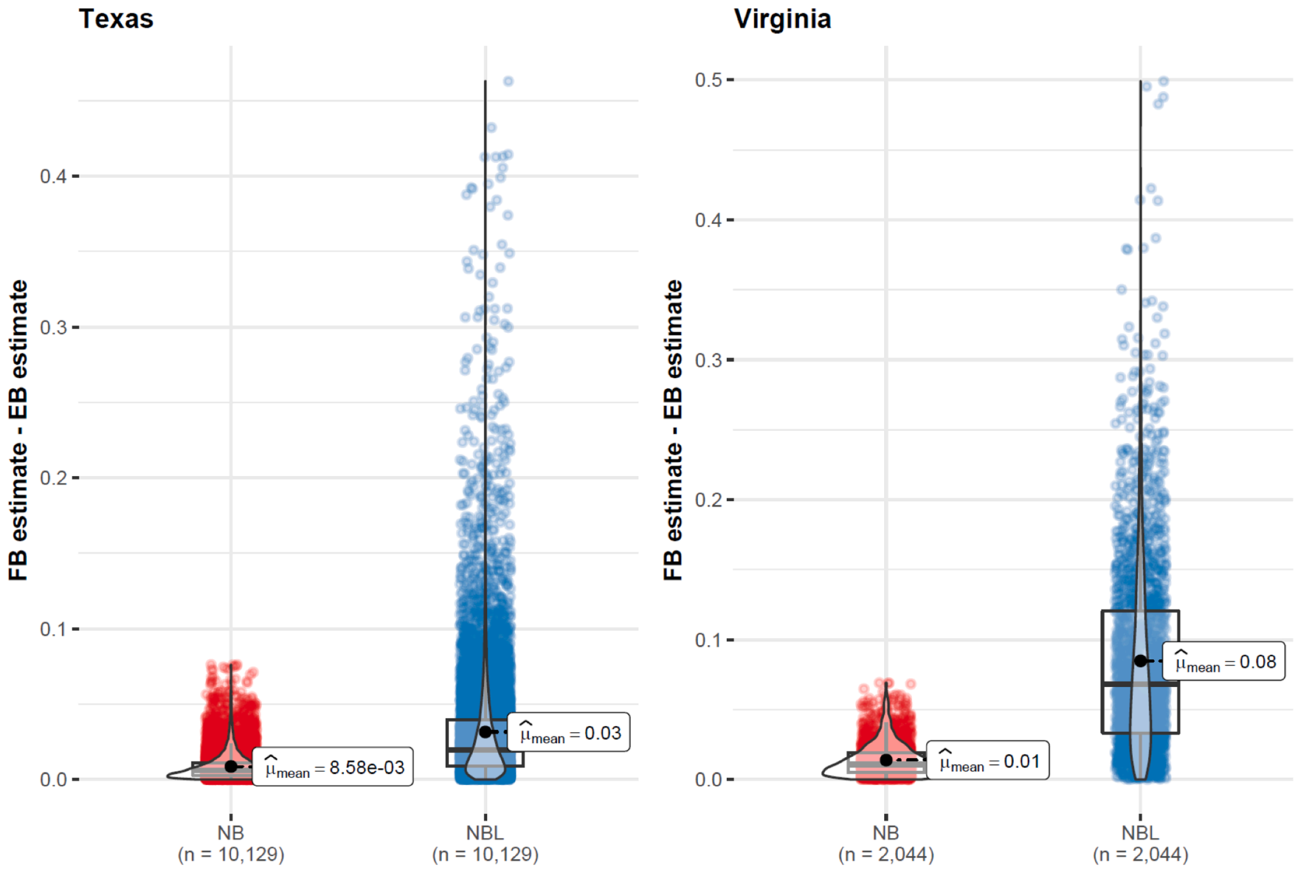

Fig. 1. The absolute difference between the EB and FB estimates of the expected crashes.

volves multiple random sampling steps, which are doable using any software developed for MCMC analysis. However, the MCMC analysis requires the model tree defined using the standard distributions. The Lindley distribution is not a standard distribution but can be re-parameterized as a two-component gamma mixture (Zamani and Ismail, 2010):

$$\epsilon \sim \text{Lindley}(\theta) \equiv \frac{1}{1+\theta}\ \text{Gamma}(2,\theta) + \frac{\theta}{\theta+1}\ \text{Gamma}(1,\theta) \tag{27}$$

To calculate the EB expected crashes, we merely plugged in the MLEs (*i.e.*, $\widehat{\phi}, \widehat{\beta}$) from Table 2 and Table 3 in the EB formula developed in Eq. (25). The EB expected crash value for site i would be as follows:

$$E(\lambda_i|y_i) = \frac{A(y_i+1)}{A(y_i)}(y_i+1) \tag{28}$$

where, $A(y_i) = \sum_{j=0}^{y_i}(-1)^j \binom{y_i}{j} \frac{e^{\widehat{\beta X}} + \widehat{\phi} + j + 1}{(e^{\widehat{\beta \mathbf{X}_i}} + \widehat{\phi} + j)^2}$

Results from both methods were observed and compared to see whether EB estimates properly approximate the FB estimates. The absolute difference between the estimates for the NB and NB-L models are plotted in Fig. 1. These violin plots show the extent to which the FB expected values and their approximated EB counterparts are different. As seen, the difference between the mean of the expected values are quite small for both models; this indicates that the proposed EB formula for the NB-L GLM can accurately approximate the FB procedure, thus avoiding the demanding MCMC analysis.

However, the difference between the EB and FB expected values were larger for the NB-L in comparison to those of the NB model. This issue is fully covered in the discussion section. Even though the difference between the EB and FB estimates seem relatively larger for the NB-L models, the authors observed that the relative differences will not exceed 25% for sites with crash experiences. This indicates that the EB and FB estimates are close, and it is anticipated that the EB estimates will be adequate for safety applications.

5.3. Application in site ranking

The empirical results in the previous section showed that the EB estimates of the expected crash values are appropriate alternatives to the FB estimates. Both the absolute and relative differences between the EB and FB estimates of expected crashes were small, indicating that the EB estimates well approximate their FB counterparts. However, sometimes the order of the expected crash values is of interest rather than the magnitude itself. In site ranking studies, the aim is to sort the study sites in decreasing order of their evaluated crash risk (expected crash value). As a result, this study examined the application of the proposed EB framework for the NB-L GLM in site ranking.

In order to assess how a ranking criterion performs, a reference ranking is needed as the basis for the comparison. Several studies found Posterior ranking criteria to be a better alternative to integer ranking when the model is implemented as a Bayesian framework (Laird and Louis, 1989; Miaou and Song, 2005; Liu and Sharma, 2018; Shen and Louis, 1998). Laird and Louis (1989) observed that the posterior distribution of ranks carries more information than the integer rank that is usually assigned to the parameter mean. The Posterior ranking criterion takes into account all posterior simulations for each site's crash risk (not only the posterior mean), and results in a ranking for each simulation run. Eventually, by taking the average of the simulated ranks for each site, the posterior expected rank is achieved. This study assumed the posterior expected ranks as the reference ranks in order to compare the ranking criteria produced by FB and EB expected crash values for the NB-L model. The first 10, 20, 50, 100, 200, and 500 top ranked sites were identified for each ranking criterion. Table 4 shows the number of sites that appeared in both the ranking criteria being evaluated and the reference ranking (posterior expected ranks). As seen in both datasets, the hot spots identified by the EB approach are quite similar to those identified by the FB approach. Similarity of the ranks indicate that the proposed EB approach can be a proper alternative to the FB approach not only in crash prediction, but also in hotspot identification. Site ranking is the only safety analysis evaluated in this study; however, any other type of safety studies that use the long-term crash mean can benefit from the proposed EB method.

Table 4
Site ranking results.

Dataset	Risk Evaluation Criteria	Unordered Ranked Groups					
		1–10	1–20	1–50	1–100	1–200	1–500
Texas	NB-L FB	10/10 *	20/20	50/50	100/100	200/200	500/500
	NB-L EB	9/10	20/20	46/50	97/100	190/200	486/500
Virginia	NB-L FB	10/10	19/20	49/50	99/100	196/200	498/500
	NB-L EB	8/10	15/20	49/50	95/100	196/200	492/500

*Posterior expected ranking is the reference ranking criteria.

6. Discussion

Modeling results from previous works and the current study indicated that compared to the traditional NB model, the NB-L model provides a superior fit when analyzing crash datasets with excess zero observations. Consequently, the NB-L has a better performance in evaluating the crash risk associated with each site. However, the NB-L expected values were only available using the FB approach, which could be difficult to compute for large datasets. This study introduced the EB framework to approximate the FB estimates of the NB-L model. Results from the previous section showed that the proposed EB framework for the NB-L can properly approximate the underlying FB procedure, so it can be used for analyses that require expected crash values (*e.g.*, site ranking, before-after analysis). Some interesting findings, results, and limitations are discussed below.

As indicated in Table 2 and Table 3, neither the signs nor the magnitudes of the estimated coefficients are comparable between the NB and NB-L models. This issue can be attributed to the way the mean function is structured. The NB model in this study is parameterized by its mean, μ, and overdispersion parameter, ϕ. The mean function is assumed to have a log-linear relationship with the site characteristics, (*i.e.*, $\mathbf{X}$) through the regression coefficients (*i.e.*, $\boldsymbol{\beta}$). As a result, the coefficients are directly related to the mean crashes so, their magnitude and sign carry information about how and to what extent each covariate affects the crash frequencies. However, the NB-L model is parameterized differently. Geedipally et al. (2012) introduced two different parameterizations for the NB-L GLM. The first one uses the NB formulated by mean and overdispersion parameter where each site-specific mean value is multiplied by an adjustment factor (or as indicated in the original paper, frailty term), ϵ:

$$Y \sim NB\ (y; \epsilon\mu, \phi)\epsilon \sim \text{Lindley}\ (\theta)\mu = e^{\beta \mathbf{X}} \tag{29}$$

This parameterization is easy to interpret, and given that $E(Y) = \mu$, the regression coefficients are directly related to the mean response. This representation, however, is not available in closed form and hence, cannot be used in the EB framework proposed in this study. Instead, we used the second parameterization, which is available in closed form:

$$Y \sim NB\ (y; p, \phi) - ln\ (p) \sim Lindley\ (\theta)\theta = e^{\beta \mathbf{X}} \tag{30}$$

This parameterization, which follows the original NB-L parameterization discussed in Zamani and Ismail (2010), links the site characteristics to the Lindley parameter, θ, not the mean. The mean of this parameterization can be written as follows (Zamani and Ismail, 2010):

$$E(Y) = \phi\left(\frac{\theta^3}{(\theta+1)(\theta-1)^2-1}\right) \tag{31}$$

As seen in the above definition, the mean response is a non-linear and non-invertible function of regression coefficients and overdispersion parameter. Consequently, the signs and magnitudes of the regression coefficients in the second parameterization do not necessarily show the causal relationship between the covariates and crash frequency. To put it concisely, the proposed EB framework and the NB-L representation we utilized in this study (indicated in Eq. 30) are only applicable when the expected crash values are of interest. The expected crashes can then be used in various safety-related studies such as hot spot identification or before-after analysis. However, if the goal is to determine the underlying relationship between the crash frequencies and contributing factors, the other parameterization of the NB-L model indicated in Eq.(29) should be employed (see Khodadadi et al. (2021b, 2012, 2019, 2016)).

In addition, we observed that the log-likelihood function of the NB-L model behaves unpredictably when large-valued parameters or large inputs are involved. This issue can be attributed to the summation term existing in the NB-L closed-form expression, $\sum_{j=0}^{y}\binom{y}{j}(-1)^{j}\frac{e^{\beta\mathrm{X}}+\phi+j+1}{(e^{\beta\mathrm{X}}+\phi+j)^{2}}$. This summation part results from the following substitution, $(1-e^{-\lambda})^{y} = \sum_{j=0}^{y}\binom{y}{j}(-1)^{j}e^{-\lambda j}$, which was used in Zamani and Ismail (2010),Khodadadi et al. (2021a) to derive the pmf of the NB-L distribution. This part outputs small negative values when large y's or large-valued parameters are input. Consequently, proper initial values are required when maximizing the likelihood function to avoid large estimates and negative likelihoods, and ensure valid estimates and inferences.

Furthermore, a larger difference between the FB and EB expected values was observed in the NB-L compared to the NB model. This issue can be attributed to two reasons. First, the NB-L likelihood is not strictly concave and hence, the global optimization is not possible or very difficult to get. Numerical approaches are needed to approach to the global maximum point as much as possible which eventually lead to a range of local maxima and a range of estimates. Unlike in the case of the NB model where a single set of MLEs achieve the global maximum, in the NB-L model, a range of local estimates would be achieved whose accuracy depend of the initial values. Therefore, the MLEs and the FB estimates are quite different for the NB-L model (see Table 2). This difference between the MLEs and FB estimate will result in the different EB and FB estimates of the expected values. Another potential reason could be the bias-variance trade-off. Due to the flexible structure of the NB-L model, it tends to output less-biased and hence, high-variance results. High variability of the NB-L model is mirrored in high variance of the expected values.

Aside from the high-variability of the NB-L model, the absolute and relative differences showed negligible values. Similarity of the ranking from the EB and FB estimates also confirmed that the EB estimates will be adequate for safety applications. The proposed framework will be specially useful in situations where the traditional NB models are not flexible enough (*e.g.*, abundance of zeros in the data or high skewness), or output biased results (*e.g.*, data with low sample mean).

7. Summary and conclusions

Even though there is rich literature on the advanced predictive models in traffic safety, little has been done to extend their application to other roadway safety tasks. The NB-L model has been proposed for sparse count data modeling and, as indicated in several studies, provides superior performance compared to the common NB model used in traffic safety. However, its application has not been examined in other roadway safety tasks.

Expected crash values estimated from the crash prediction models are the main evaluation tool in safety analysis and represent the long-term risk associated with a roadway entity. This study proposed an EB framework to approximate the underlying FB method and derive the expected crash values for the NB-L model. The derived expected crashes can be used in various safety-related studies (*e.g.*, hot spot identification, before-after analysis). The results showed that the proposed EB framework is able to estimate expected crashes with comparable precision to the FB estimate but with much lower computational costs. The proposed framework was further examined in site ranking analysis. We observed that ranks produced by the EB estimates were similar to those of FB estimates, indicating that the proposed framework can be safely employed in other highway safety tasks such as hot spot identification analysis.

The EB approach introduced in this study can be utilized in any type of analysis that requires access to the expected crash values. The resulting EB expected crashes take advantage of the probabilistic structure of the FB paradigm while avoiding its time-consuming computational efforts. For future studies, a similar framework as introduced in this study can be used to develop an EB method for other advanced predictive models in traffic safety. Also, further work should be performed to validate the application of the framework in other safety-related analyses such as before-after analysis.

Author contribution statement

The authors confirm contribution to the paper as follows: study conception and design: Ali Khodadadi and Dominique Lord; data collection: Ali Khodadadi, Ioannis Tsapakis, and Subasish Das; analysis and interpretation of results: Ali Khodadadi, Mohammadali Shirazi; draft manuscript preparation: Ali Khodadadi, Mohammadali Shirazi, Ioannis Tsapakis, Subasish Das, and Dominique Lord. All authors reviewed the results and approved the final version of the manuscript.

Funding sources

The work was completed as part of SAFE-D University Transportation Center project Use of Disruptive Technologies to Support Safety Analysis and Meet New Federal Requirements. The study was partly funded by the A.P. and Florence Wiley Faculty Fellow.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

aaa

References

Abdella, G.M., Kim, J., Al-Khalifa, K.N., Hamouda, A.M., 2019. Penalized conway-maxwell-poisson regression for modelling dispersed discrete data: The case study of motor vehicle crash frequency. Saf. Sci. 120, 157–163.

Aguero-Valverde, J., Jovanis, P.P., 2009. Bayesian multivariate poisson lognormal models for crash severity modeling and site ranking. Transp. Res. Rec. 2136, 82–91.

Aman, J.J., Smith-Colin, J., Zhang, W., 2021. Listen to e-scooter riders: Mining rider satisfaction factors from app store reviews. Transp. Res. Part D: Transp. Environ. 95, 102856.

Aman, J.J.C., Smith-Colin, J., 2020. Transit deserts: Equity analysis of public transit accessibility. J. Transp. Geogr. 89, 102869.

Asgharpour, S., Javadinasr, M., Bayati, Z., et al., 2021. Investigating severity of motorcycle-involved crashes in a developing country. Presented at 101th Annual Meeting of the Transportation Research Board, Washington, D.C., 2022.

Darzian Rostami, A., Katthe, A., Sohrabi, A., Jahangiri, A., 2020. Predicting critical bicycle-vehicle conflicts at signalized intersections. J. Adv. Transp., 2020.

Das, S., Bibeka, A., Sun, X., Zhou, H.G., Jalayer, M., 2019. Elderly pedestrian fatal crash-related contributing factors: applying empirical bayes geometric mean method. Transp. Res. Rec. 2673, 254–263.

Das, S., Tsapakis, I., Khodadadi, A., 2021. Safety performance functions for low-volume rural minor collector two-lane roadways. IATSS Res.

Debrabant, B., Halekoh, U., Bonat, W.H., Hansen, D.L., Hjelmborg, J., Lauritsen, J., 2018. Identifying traffic accident black spots with poisson-tweedie models. Acc. Anal. Prevent. 111, 147–154.

Denthet, S., Thongteeraparp, A., Bodhisuwan, W., 2016. Mixed distribution of negative binomial and two-parameter lindley distributions. In: 2016 12th International Conference on Mathematics, Statistics, and Their Applications (ICMSA), IEEE, pp. 104–107.

Farid, A., Abdel-Aty, M., Lee, J., Eluru, N., 2017. Application of bayesian informative priors to enhance the transferability of safety performance functions. J. Saf. Res. 62, 155–161.

Geedipally, S.R., Lord, D., Dhavala, S.S., 2012. The negative binomial-lindley generalized linear model: Characteristics and application using crash data. Acc. Anal. Prevent. 45, 258–265.

Gomez-Deniz, E., Calderin-Ojeda, E., 2017. An alternative representation of the negative binomial-lindley distribution. new results and applications. arXiv preprint arXiv: 1703.04812.

Guo, X., Wu, L., Lord, D., 2020. Generalized criteria for evaluating hotspot identification methods. Acc. Anal. Prevent. 145, 105684.

Guo, X., Wu, L., Zou, Y., Fawcett, L., 2019. Comparative analysis of empirical bayes and bayesian hierarchical models in hotspot identification. Transp. Res. Rec. 2673, 111–121.

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.W., Council, F.M., Griffith, M.S., 2002. Estimating safety by the empirical bayes method: a tutorial. Transp. Res. Rec. 1784, 126–131.

Huang, H., Chin, H.C., Haque, M.M., 2009. Empirical evaluation of alternative approaches in identifying crash hot spots: Naive ranking, empirical bayes, full bayes methods. Transp. Res. Rec. 2103, 32–41.

Karlis, D., Xekalaki, E., 2005. Mixed poisson distributions. International Statistical Review/Revue Internationale de Statistique 35–58.

Khattak, Z.H., Magalotti, M.J., Fontaine, M.D., 2018. Estimating safety effects of adaptive signal control technology using the empirical bayes method. J. Saf. Res. 64, 121–128.

Khazraee, S.H., Johnson, V., Lord, D., 2018. Bayesian poisson hierarchical models for crash data analysis: Investigating the impact of model choice on site-specific predictions. Acc. Anal. Prevent. 117, 181–195.

Khodadadi, A., Tsapakis, I., Das, S., Lord, D., Li, Y., 2021. Application of different negative binomial parameterizations to develop safety performance functions for non-federal aid system roads. Acc. Anal. Prevent. 156, 106103.

Khodadadi, A., Shirazi, M., Geedipaly, S., Lord, D., 2022. Evaluating alternative variations of negative binomial-lindley distribution for modeling crash data. Transportmetrica A: Transport Science.

Klugman, S.A., Panjer, H.H., Willmot, G.E., 2012. Loss models: from data to decisions, vol. 715. John Wiley & Sons.

Laird, N.M., Louis, T.A., 1989. Empirical bayes ranking methods. J. Educ. Stat. 14, 29–46.

Lan, B., Persaud, B., 2011. Fully bayesian approach to investigate and evaluate ranking criteria for black spot identification. Transp. Res. Rec. 2237, 117–125.

Lee, A.S., Lin, W.H., Gill, G.S., Cheng, W., 2019. An enhanced empirical bayesian method for identifying road hotspots and predicting number of crashes. J. Transp. Saf. Secur. 11, 562–578.

Liu, C., Sharma, A., 2018. Using the multivariate spatio-temporal bayesian model to analyze traffic crashes by severity. Analyt. Methods Acc. Res. 17, 14–31.

Lord, D., 2006. Modeling motor vehicle crashes using poisson-gamma models: Examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. Acc. Anal. Prevent. 38, 751–766.

Lord, D., Geedipally, S.R., 2011. The negative binomial–lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. Acc. Anal. Prevent. 43, 1738–1742.

Lord, D., Geedipally, S.R., Guikema, S.D., 2010. Extension of the application of conway-maxwell-poisson models: analyzing traffic crash data exhibiting underdispersion. Risk Analysis: Int. J. 30, 1268–1276.

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

Lord, D., Park, B.J., Model, P.G., 2012. Negative binomial regression models and estimation methods, Probability Density and Likelihood Functions. Texas A&M University, Korea Transport Institute, pp. 1–15.

Lord, D., Qin, X., Geedipally, S.R., 2021. Highway Safety Analytics and Modeling. Elsevier B.V., Amsterdam, The Netherlands.

Lunn, D.J., Thomas, A., Best, N., Spiegelhalter, D., 2000. Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. Stat. Comput. 10, 325–337.

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.

Meng, Y., Wu, L., Ma, C., Guo, X., Wang, X., 2020. A comparative analysis of intersection hotspot identification: Fixed vs. varying dispersion parameters in negative binomial models. J. Transp. Saf. Secur. 1–18.

Miaou, S.P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and bayes versus empirical bayes methods. Transp. Res. Rec. 1840, 31–40.

Miaou, S.P., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. Acc. Anal. Prevent. 37, 699–720.

Miranda-Moreno, L.F., Heydari, S., Lord, D., Fu, L., 2013. Bayesian road safety analysis: Incorporation of past evidence and effect of hyper-prior choice. J. Saf. Res. 46, 31–40.

Park, E.S., Lord, D., 2007. Multivariate poisson-lognormal models for jointly modeling crash frequency by severity. Transp. Res. Rec. 2019, 1–6.

Part, D., 2010. Highway safety manual. American Association of State Highway and Transportation Officials: Washington, DC, USA.

Persaud, B., Lan, B., Lyon, C., Bhim, R., 2010. Comparison of empirical bayes and full bayes approaches for before–after road safety evaluations. Acc. Anal. Prevent. 42, 38–43.

Plummer, M., et al., 2003. Jags: A program for analysis of bayesian graphical models using gibbs sampling, in. In: Proceedings of the 3rd international workshop on distributed statistical computing, Vienna, Austria, pp. 1–10.

Plummer, M., et al., 2016. rjags: Bayesian graphical models using mcmc. R package version 4.

Pu, Z., Li, Z., Jiang, Y., Wang, Y., 2020. Full bayesian before-after analysis of safety effects of variable speed limit system. In: IEEE transactions on intelligent transportation systems.

Safaei, B., Safaei, N., Masoud, A., Seyedekrami, S., 2021a. Weighing criteria and prioritizing strategies to reduce motorcycle-related injuries using combination of fuzzy topsis and ahp methods. Adv. Transp. Stud. 54.

Safaei, N., Zhou, C., Safaei, B., Masoud, A., 2021b. Gasoline prices and their relationship to the number of fatal crashes on us roads. Transp. Eng. 4, 100053.

Saha, D., Alluri, P., Dumbaugh, E., Gan, A., 2020. Application of the poisson-tweedie distribution in analyzing crash frequency data. Acc. Anal. Prevent. 137, 105456.

Scrucca, L., et al., 2013. Ga: a package for genetic algorithms in r. J. Stat. Softw. 53, 1–37.

Shaon, M.R.R., Qin, X., Shirazi, M., Lord, D., Geedipally, S.R., 2018. Developing a random parameters negative binomial-lindley model to analyze highly over-dispersed crash count data. Analytic methods in accident research 18, 33–44.

Shen, W., Louis, T.A., 1998. Triple-goal estimates in two-stage hierarchical models. J. R. Stat. Soc.: Ser. B (Statistical Methodology) 60, 455–471.

Shirazi, M., Dhavala, S.S., Lord, D., Geedipally, S.R., 2017. A methodology to design heuristics for model selection based on the characteristics of data: Application to investigate when the negative binomial lindley (nb-l) is preferred over the negative binomial (nb). Acc. Anal. Prevent. 107, 186–194.

Shirazi, M., Lord, D., 2019. Characteristics-based heuristics to select a logical distribution between the poisson-gamma and the poisson-lognormal for crash data modelling. Transportmetrica A: Transp. Sci. 15, 1791–1803.

Shirazi, M., Lord, D., Dhavala, S.S., Geedipally, S.R., 2016. A semiparametric negative binomial generalized linear model for modeling over-dispersed count data with a heavy tail: Characteristics and applications to crash data. Acc. Anal. Prevent. 91, 10–18.

Song, J.J., Ghosh, M., Miaou, S., Mallick, B., 2006. Bayesian multivariate spatial models for roadway traffic crash mapping. J. Multivariate Anal. 97, 246–273.

Tajuddin, R.R.M., Ismail, N., Ibrahim, K., Bakar, S.A.A., 2020. A four-parameter negative binomial-lindley distribution for modeling over and underdispersed count data with excess zeros. Commun. Stat.-Theory Methods 1–13.

Vehtari, A., Gelman, A., Gabry, J., 2017. Practical bayesian model evaluation using leave-one-out cross-validation and waic. Stat. Comput. 27, 1413–1432.

Watanabe, S., Opper, M., 2010. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. 11.

Willmot, G., 1986. Mixed compound poisson distributions. ASTIN Bull.: J. IAA 16, S59–S79.

Xie, Y., Zhang, Y., 2008. Crash frequency analysis with generalized additive models. Transp. Res. Rec. 2061, 39–45.

Zamani, H., Ismail, N., 2010. Negative binomial-lindley distribution and its application. J. Math. Stat. 6, 4–9.

Zha, L., Lord, D., Zou, Y., 2016. The poisson inverse gaussian (pig) generalized linear regression model for analyzing motor vehicle crash data. J. Transp. Saf. Secur. 8, 18–35.

Zou, Y., Ash, J.E., Park, B.J., Lord, D., Wu, L., 2018. Empirical bayes estimates of finite mixture of negative binomial regression models and its application to highway safety. J. Appl. Stat. 45, 1652–1669.

Zou, Y., Lord, D., Zhang, Y., Peng, Y., 2013. Comparison of sichel and negative binomial models in estimating empirical bayes estimates. Transp. Res. Rec. 2392, 11–21.