Comparison of Sichel and Negative Binomial Models in Estimating Empirical Bayes Estimates

By

Yajie Zou*
Ph.D. Candidate, Zachry Department of Civil Engineering
Texas A&M University, 3136 TAMU
College Station, TX 77843-3136
Phone: 936/245-5628, fax: 979/845-6481
E-mail: zouyajie@gmail.com

Dominique Lord, Ph.D.
Associate Professor and Zachry Development Professor I
Zachry Department of Civil Engineering
Texas A&M University, TAMU 3136
College Station, TX 77843-3136
Phone: (979) 458-3949, fax: (979) 845-6481
E-mail: d-lord@tamu.edu

Yunlong Zhang, Ph.D.
Associate Professor, Zachry Department of Civil Engineering
Texas A&M University, 3136 TAMU
College Station, TX 77843-3136
Phone: 979/845-9902, fax: 979/845-6481
E-mail: yzhang@civil.tamu.edu

Yichuan Peng
Ph.D. Candidate, Zachry Department of Civil Engineering
Texas A&M University, 3136 TAMU
College Station, TX 77843-3136
Phone: (979) 458-3949, fax: (979) 845-6481
Email: pycgcx@tamu.edu

November 15, 2012
Word Count: 4,624 + 2,750 (5 tables and 6 figures) = 7,374 words
*Corresponding author
ABSTRACT

Traditionally, transportation safety analysts have used the empirical Bayes (EB) method to improve the estimate of the long-term mean of individual sites and to identify hotspots locations. The EB method combines two different sources of information: (1) the expected number of crashes estimated via crash prediction models, and (2) the observed number of crashes at individual sites. Crash prediction models have extensively been estimated using a negative binomial (NB) modeling framework due to the over-dispersion commonly found in crash data. Recent studies have shown that the Sichel (SI) distribution provides a promising avenue for developing crash prediction models. The objective of this study is to examine the application of the SI model in calculating EB estimates. To accomplish the objective of the study, the SI models with a fixed/varying dispersion term are developed using the crash data collected at 4-lane undivided rural highways in Texas. The important conclusions can be summarized as follows: (1) the selection of the crash prediction model (i.e., the SI or NB model) will affect the value of weight factor used for estimating the EB output; (2) the identification of hazardous sites, using the EB method, can be different when the SI model is used. Finally, a simulation study designed to examine which crash prediction model can better identify the hotspot is recommended as our future research.
INTRODUCTION

Statistical models or crash prediction models have been a very popular method for estimating the safety performance of various transportation elements (1). Due to the over-dispersion observed in crash data, the negative binomial (NB) model is probably the most frequently used statistical model for developing crash prediction models. It has been applied extensively in various types of highway safety studies from the identification of hotspots or hazardous sites to the development of accident modification factors. Up to now, other statistical models for analyzing crash data have been proposed by transportation safety analysts (2). These models, such as the zero-inflated models (3), the Poisson-lognormal (4; 5), the Conway-Maxwell-Poisson (6), the gamma (7), and the negative binomial-Lindley (8), have been used as a substitute to the NB model. More recently, the Sichel distribution has been introduced by Zou et al. (9) for modeling highly dispersed count data. The Sichel (SI) distribution is a compound Poisson distribution, which mixes the Poisson distribution with the generalized inverse Gaussian distribution. Previous studies (10, 11) have shown that the Sichel distribution is particularly useful as a model for highly dispersed count data. Besides the count models, Castro et al. (12) proposed a reformulation of count models as a special case of generalized ordered-response models. The proposed modeling framework can accommodate spatial and temporal dependence and was applied to predict crash frequency at urban intersections.

Two of the most important tasks in traffic safety management are network screening to identify sites with potential for safety treatment and the before–after evaluation of the effects of implemented treatments (13). One widely applied approach to help performing both tasks is the popular empirical Bayesian (EB) method. This method can correct for the regression-to-the mean bias (if the site selection bias is not present, see (14)), refines the predicted mean of an entity, and is relative simple to manipulate compared to the fully Bayesian approach (1). As discussed by Hauer (15, 16) and Cheng and Washington (17), the EB method rests on two assumptions: (1) at a given location, crash occurrence obeys the Poisson probability law with mean $\lambda$; (2) the probability distribution of the $\lambda$’s of the population of sites is gamma distributed. On the basis of the above assumptions, the probability that a site selected randomly records $x$ crashes can be approximated by the NB distribution. However, Hauer (16) also pointed out that, although the gamma assumption for the $\lambda$’s is adequate for many examined datasets, this is not a proof that the $\lambda$’s are indeed gamma distributed. Considering the importance of the EB method, there is a need to explore other mixed-Poisson models which might potentially be better used to obtain the EB estimates. Among the mixed-Poisson models, it is found that the NB and SI models both have the quadratic variance-mean relationship (9). Similar to the dispersion parameter of the NB
model, a dispersion term of the SI model is defined to measure the level of dispersion. This dispersion term can be easily used by transportation safety analysts to obtain reliable EB estimates within the SI modeling framework, which is described in the following section. In addition, Gupta and Ong (18) recently examined various mixed Poisson distributions for analyzing long-tailed count data and stated that SI model can provide satisfactory fits in many cases where other models proved inadequate.

The primary objective of this research is to examine the application of the SI model in calculating EB estimates. To accomplish the objectives of this study, both the traditional Sichel (SI) model and the generalized Sichel model (GSI) are considered and compared with the traditional negative binomial (NB) model and the generalized negative binomial model (that is, the NB model with a varying dispersion parameter and it is defined as GNB), respectively. The SI model assumes that the scale and shape parameters of the Sichel distribution are fixed for all locations, while the GSI model assumes that the scale parameter will vary from one site to another. The difference between the SI and NB models is investigated in terms of their effects on EB estimates as well as the identification and ranking of accident-prone sites. In this study, the SI models with a fixed/varying dispersion term are developed using the crash data collected at 4-lane undivided rural highways in Texas.

METHODOLOGY

This section describes the characteristics of the NB and the SI models, as well as the EB Method.

Empirical Bayes Method Based on the Negative Binomial Model

In highway safety, the dispersion parameter of NB models refines the estimates of the predicted mean when the EB method is used. Since the EB estimates can be used to identify hotspots by ranking crash-prone locations and assess the effects of implemented treatments, it is necessary to obtain reliable EB estimates based on an appropriate modeling framework. The selection of crash prediction models (i.e., NB model, Sichel model etc.) for EB estimates may affect the precision of the analysis output. So far, the NB distribution is the most frequently used model by transportation safety analysts for calculating the EB estimates. The NB model has the following model structure: the number of crashes $y$ during some time period is assumed to be Poisson distributed, which is defined by:
\[ p(y \mid \lambda) = \frac{\lambda^y \exp(-\lambda)}{y!} \]  

(1)

where \( \lambda \) = mean response of the observation.

The NB distribution can be viewed as a mixture of Poisson distributions where the Poisson rate is gamma distributed. For the complete derivation of the NB model, the reader is referred to Hilbe (19). The probability density function of the NB is defined as follows:

\[ f(y \mid \mu, \alpha) = \frac{\Gamma(y + 1)}{\Gamma(\frac{1}{\alpha})} \left( \frac{\alpha \mu}{1 + \alpha \mu} \right)^{\frac{1}{\alpha}} \left( \frac{1}{1 + \alpha \mu} \right)^y \]  

(2)

where,

\( y \) = response variable;
\( \mu \) = mean response of the observation; and,
\( \alpha \) = dispersion parameter.

Compared to the Poisson distribution, the NB distribution can allow for over-dispersion. For \( y = 0,1,2,\ldots, \infty \), the mean of \( y \) is \( E[y] = \mu \) and variance is \( \text{VAR}(y) = \mu + \mu^2 \alpha \). If \( \alpha \to 0 \), the crash variance equals the crash mean and the NB distribution converges to the Poisson distribution.

The dispersion parameter \( \alpha \) of the NB model is very important in calculating the EB estimates. As proposed by Hauer (16), the long term mean for a site \( i \) can be estimated using the EB method:

\[ \hat{\mu}_i = w_i \hat{\mu}_i + (1 - w_i) y_i \]  

(3)

where \( \hat{\mu}_i \) is the EB estimate of the expected number of crashes per year for site \( i \); \( \hat{\mu}_i \) is the estimated number of crashes by crash prediction models for given site \( i \) (estimated using a NB 
model); \( w_i = \frac{1}{1 + \alpha \mu_i} \) is the weight factor estimated as a function of \( \mu_i \) and dispersion parameter \( \alpha \); and \( y_i \) is the observed number of crashes per year at site \( i \).

**Empirical Bayes Method Based on the Sichel Model**

The Sichel distribution (also known as the Poisson-generalized inverse Gaussian distribution) has recently been used for modeling motor vehicle crashes \( (9) \), and it was showed that the Sichel distribution is particularly useful as a model for highly dispersed count data. The formulation of the Sichel distribution, \( SI(\mu, \sigma, \nu) \), is given by,

\[
p(y \mid \mu, \sigma, \nu) = \frac{(\mu/c)^y K_{y+\nu}(\alpha)}{K_{\nu}(1/\sigma)^y!(\alpha\sigma)^{y+\nu}}
\] (4)

where,
- \( y \) = response variable;
- \( \mu \) = mean response of the observation;
- \( \sigma \) = scale parameter;
- \( \nu \) = shape parameter;
- \( \alpha^2 = \sigma^2 + 2\mu(c\sigma)^{-1} \);
- \( c = \frac{K_{\nu+1}(1/\sigma)}{K_{\nu}(1/\sigma)} \); and,
- \( K_{\nu}(t) = \frac{1}{2} \int_0^\infty x^{\nu-1} \exp(-\frac{1}{2}t(x+x^{-1}))dx \) is the modified Bessel function of the third kind.

For \( y = 0, 1, 2, ..., \infty \), the mean of \( y \) is \( E[y] = \mu \) and variance is \( VAR(y) = \mu + \mu^2[2\sigma(\nu+1)/c + 1/c^2 - 1] \). When \( \sigma \rightarrow \infty \) and \( \nu > 0 \), it can be shown that the SI distribution can be reduced to the NB distribution. For more information about the formulation in equation (4) and parameter \( c \), interested readers are referred to Rigby et al. \( (20) \).

In highway safety, the dispersion parameter of NB models takes a central role for
calculating EB estimates \((1)\). As noted by Zou et al. \((9)\), the NB and SI models both have the quadratic variance-mean relationship, that is \(\text{VAR}(y) = \mu + \mu^2 h(\alpha, \sigma, \nu)\) where \(E(y) = \mu\) and \(h(\alpha, \sigma, \nu)\) is a function of the parameters of the mixing distribution. For the NB model, \(h(\alpha, \sigma, \nu) = \alpha\) is defined as the dispersion parameter; on the other hand, for the SI model, \(h(\alpha, \sigma, \nu) = 2\sigma(\nu+1)/\nu + 1/c^2 - 1\) can be viewed as a dispersion term. Similar to the dispersion parameter \(\alpha\) in the NB model, this dispersion term can be also used to measure the level of dispersion. More importantly, the output of the SI model (the dispersion term) can be easily used by practitioners to obtain reliable EB estimates. We developed a SI modeling framework to obtain EB estimates according to the following steps:

1. The weight factor for the SI model is derived as follows. According to Hauer \((16, \text{equation 11.1})\), let \(K\) be the observed number of crashes, and \(\kappa\) be the expected crash count, the EB estimator of \(\kappa\) is:

\[
E(\kappa|K) = wE(\kappa) + (1-w)K
\]

Then the weight factor \(w\) is given as (Hauer \((16\), equation 11.2a)):

\[
w = \frac{1}{1 + \frac{\text{VAR}(\kappa)}{E(\kappa)}}
\]

The weight factor \(w\) is a function of mean and variance of \(\kappa\) and is always a number between 0 and 1. So far, the above two equations do not rest on any assumption about the distribution of \(\kappa\). If \(\kappa\) is gamma distributed, then the resulting \(K\) follows a NB distribution. If we let \(\kappa\) take a generalized inverse Gaussian distribution, then \(K\) follows a SI distribution. The generalized inverse Gaussian \(\text{GIG}(\mu, \sigma, \nu)\) is defined as (Equation 7 is obtained from a reparameterization of equation 10.39 of Rigby and Stasinopoulos \((21)\)):

\[
f(\kappa|\mu, \sigma, \nu) = \left(\frac{c}{\mu}\right)^\nu \left[\frac{\kappa^{\nu-1}}{2H_v\left(\frac{1}{\sigma}\right)}\right] \exp\left[-\frac{1}{2\sigma}\left(\frac{c\kappa}{\mu} + \frac{\mu}{c\kappa}\right)\right]
\]
where $\mu > 0$, $\sigma > 0$, $-\infty < \nu < \infty$, $c = \frac{H_{\nu+1}(1/\sigma)}{H_{\nu}(1/\sigma)}$ and $H_{\nu}(t) = \frac{1}{2} \int_{0}^{\infty} x^{\nu-1} \exp(-\frac{1}{2} t(x+x^{-1})) dx$. The mean of $\kappa$ is $E[\kappa] = \mu$ and variance is $VAR(\kappa) = \mu^2 [2\sigma(\nu+1)/c + 1/c^2 - 1]$.

Substituting for $E[\kappa]$ and $VAR(\kappa)$ from above into the equation (6), we obtain the function to estimate weight factor:

$$w = \frac{1}{1 + \frac{VAR(\kappa)}{E(\kappa)}} = \frac{1}{1 + \mu [2\sigma(\nu+1)/c + 1/c^2 - 1]} = \frac{1}{1 + E(\kappa)h(\sigma, \nu)}$$

(8)

where $h(\sigma, \nu)$ is the dispersion term for the SI model.

2. Within the SI modeling framework, the long term mean for a site $i$ using the EB method is given by:

$$\hat{\mu}_i = w_i \hat{\mu}_i + (1-w_i) y_i$$

(9)

where $\hat{\mu}_i$ is the EB estimate of the expected number of crashes per year for site $i$; $\mu_i$ is the estimated number of crashes by crash prediction models for given site $i$ (estimated using a SI model); $w_i = \frac{1}{1 + h(\sigma, \nu) \hat{\mu}_i}$ is the weight factor estimated as a function of $\hat{\mu}_i$ and $h(\sigma, \nu) = 2\sigma(\nu+1)/c + 1/c^2 - 1$; and $y_i$ is the observed number of crashes per year at site $i$. Thus, the dispersion term in the SI model can be easily used for estimating the weight factor in the EB method.

**DATA DESCRIPTION**

The crash dataset used in this study was collected at 4-lane undivided rural segments in Texas. This dataset contains crash data collected on 1499 undivided rural segments over a five-year period from 1997 to 2001. This dataset has been used extensively in other research (9; 22). During the
study period, 553 out of the 1,499 (37%) segments did not have any reported crashes, and a total of 4,253 crashes occurred on 946 segments. Table 1 provides the summary statistics for individual road segments in the Texas data.

**TABLE 1 Summary Statistics of Characteristics for Individual Road Segments in the Texas Data**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean(SD†)</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of crashes (5 years)</td>
<td>0</td>
<td>97</td>
<td>2.84(5.69)</td>
<td>4253</td>
</tr>
<tr>
<td>Average daily traffic over the 5 years (F)</td>
<td>42</td>
<td>24800</td>
<td>6613.61 (4010.01)</td>
<td></td>
</tr>
<tr>
<td>Lane Width (LW)</td>
<td>9.75</td>
<td>16.5</td>
<td>12.57(1.59)</td>
<td></td>
</tr>
<tr>
<td>Total Shoulder Width (SW)</td>
<td>0</td>
<td>40</td>
<td>9.96(8.02)</td>
<td></td>
</tr>
<tr>
<td>Curve Density (CD)</td>
<td>0</td>
<td>18.07</td>
<td>1.43 (2.35)</td>
<td></td>
</tr>
<tr>
<td>Segment Length (L) (miles)</td>
<td>0.1</td>
<td>6.28</td>
<td>0.55(0.67)</td>
<td>830.49</td>
</tr>
</tbody>
</table>

† SD = Standard Deviation.

**MODELING RESULTS**

This section provides the EB estimates for the NB and SI models and is divided into two parts. In the first part, the NB and SI models were estimated by using the fixed dispersion parameter and fixed dispersion term, respectively. In the second part, the GNB and GSI models were developed by using different parameterizations for a varying dispersion parameter and a varying scale parameter. In this study, all models were estimated using gamlss package in the software R.

**Comparison of Effects of the NB and SI models on EB Estimates**

The modeling results for the Texas data are provided in this section. The mean functional form is adopted as follows:

$$
\mu_i = \beta_0 L_i F_i e^{\beta_2 \times LW_i + \beta_3 \times SW_i + \beta_4 \times CD_i}
$$

(10)

where $\mu_i$ is the estimated numbers of crashes at segment i; $L_i$ is the segment length in miles for segment i; $F_i$ is the flow (average daily traffic over five years) traveling on segment i; $SW_i$ is
the total shoulder width in feet for segment $i$; $CD_i$ is the curve density (curves per mile) for segment $i$; and $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'$ are the estimated coefficients.

We applied the NB and SI models to the Texas data. The modeling results are provided in Table 2. The coefficients for both models have the plausible values. The goodness-of-fit statistics (Log-likelihood, Akaike information criterion (AIC) and Bayesian information criterion (BIC)) indicate that the SI model can work better than the NB model.

**TABLE 2 Modeling Results for NB and SI Models with the Texas Data**

<table>
<thead>
<tr>
<th>Estimates</th>
<th>NB</th>
<th>SI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept $\ln(\beta_0)$</td>
<td>-7.9489</td>
<td>-7.9984</td>
</tr>
<tr>
<td>$\ln$ (Average daily traffic) $\beta_1$</td>
<td>0.9749</td>
<td>0.9926</td>
</tr>
<tr>
<td>Lane Width $\beta_2$</td>
<td>-0.0533</td>
<td>-0.0600</td>
</tr>
<tr>
<td>Total Shoulder Width $\beta_3$</td>
<td>-0.0100</td>
<td>-0.0100</td>
</tr>
<tr>
<td>Curve Density $\beta_4$</td>
<td>0.0675</td>
<td>0.0627</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.3907</td>
<td>-</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>-</td>
<td>1.6181</td>
</tr>
<tr>
<td>$\nu$</td>
<td>-</td>
<td>1.012</td>
</tr>
<tr>
<td>Log-likelihood</td>
<td>2561.39</td>
<td>2550.23</td>
</tr>
<tr>
<td>AIC</td>
<td>5134.77</td>
<td>5114.45</td>
</tr>
<tr>
<td>BIC</td>
<td>5166.65</td>
<td>5151.64</td>
</tr>
</tbody>
</table>

Note: - = not applicable.

* SE = Standard Error.

Figure 1 shows the comparison of variance predicted by the NB and SI models. It can be observed that, when the crash mean is less than 10, both models can yield almost the same variance. Since the variance of the predicted number of crashes is proportional to the crash mean, the difference in variance between two models is more apparent for sites with high predicted crash mean. Overall, the figure demonstrates that the variance of the NB model increases at a slower rate than that of the SI model as the number of predicted accidents increases. This can be explained by the fact that the dispersion term, $h(\sigma, \nu) = 2\sigma(\nu + 1)/c + 1/c^2 - 1$, of the SI model is estimated as a function of the scale parameter $\sigma$ and the shape parameter $\nu$. Thus, the dispersion term of the SI model is more flexible than the dispersion parameter of the NB model.
and can allow higher dispersion. Since this dataset is highly dispersed, the crash variance might be better captured by the SI model than by the NB model, especially at larger crash mean values.

![Variance versus mean for NB and SI models.](image)

**FIGURE 1** Variance versus mean for NB and SI models.

To compare the effects of the NB and SI models on EB estimates, the proposed two models are applied to the Texas data. As seen in Figure 2 (a) and (b), first, both models have similar relationships between the weight factor and the model estimates, and the weight factor is inversely related to model estimate $E(y_i)$. The higher the model estimates, the smaller the weight factors, and vice versa. Second, when the model estimate $E(y_i)$ is small, both models can yield almost the same weight factor. However, when the model estimate increases, the SI model gradually produces a lower value for the weight factor than those of the NB model. This is because the dispersion term of the Sichel model ($h(\sigma, \nu) = 0.58$) is larger than the dispersion parameter of the NB model ($\alpha = 0.39$). As a result, compared to the NB model, the SI model
will put more weight on the observed number of crashes than the model estimates for the accident-prone sites when calculating the EB estimates. This might be an advantage for the SI model when analyzing the highly dispersed crash data. Since for the accident-prone sites, the model estimate can possibly underestimate the number of crashes. For example, for the sites with the observed number of crashes larger than 10, the average model estimate produced by the NB model is 14.76, while the average observed number of crashes is 18.76. Thus, for these accident-prone sites, it might be appropriate to put more weight on the observed number of crashes in obtaining the EB estimates to reflect the actual safety performance.
Figure 3 illustrates the relationship between the hotspot identification lists ranked by the NB model and by the SI model. Smaller values in the ranking imply more hazardous segments in terms of EB estimates (1). A preliminary examination of the figure seems to indicate a positive association between the NB model ranking and the SI model ranking. However, a further look shows that there is some variation in the ranking between the two models. As shown in Table 3, a detailed comparison of the ranking between two models suggests that the ranking differs significantly. For example, 18.8% of the observations had a ±50 position difference in ranking between the NB and SI models (note that the number of road segments in the Texas data is 1499). Overall, the analysis in this section suggests that the SI model can affect the value of weight factor used for estimating the EB output. Consequently, the SI model will influence the identification of hazardous sites.
FIGURE 3 Relationships in ranking by NB and SI models.

TABLE 3 Differences in Ranking between NB and SI Models using the EB Estimates

<table>
<thead>
<tr>
<th>Differences in ranking</th>
<th>Difference and percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-identical ranking</td>
<td>1411 (94.13%)</td>
</tr>
<tr>
<td>Ranking difference beyond 10 positions</td>
<td>926 (61.77%)</td>
</tr>
<tr>
<td>Ranking difference beyond 50 positions</td>
<td>282 (18.81%)</td>
</tr>
</tbody>
</table>

There are 1499 road segments in the Texas data.

**Comparison of Effects of the GNB and GSI models on EB Estimates**

Recent studies in transportation safety have shown that the dispersion parameter of NB models can be potentially dependent upon the explanatory variables and NB models with a varying dispersion parameter can provide better statistical fitting performance (23) or help describing the characteristics of the dispersion (24). Moreover, the varying dispersion parameter has been
shown to influence EB estimates (1). Thus, the effects of GNB and GSI models on EB estimates are also compared.

For the Texas data, the selection of the functional form will affect the value of weight factor used for estimating the EB output. Thus, we considered different functional forms for estimating the dispersion parameter $\alpha_i$ and the scale parameter $\sigma_i$. The functional forms are summarized as follows.

Dispersion parameter of GNB model:

Model 1: $\alpha_i = \gamma_0 L_i$  
Model 2: $\alpha_i = \gamma_0 / L_i$ 
Model 3: $\alpha_i = \gamma_0 L_i^{\gamma_1}$  

Scale parameter of GSI model:

Model 1: $\sigma_i = \gamma_0 L_i$  
Model 2: $\sigma_i = \gamma_0 / L_i$  
Model 3: $\sigma_i = \gamma_0 L_i^{\gamma_1}$  

where $\alpha_i$ is the estimated dispersion parameter at segment $i$; $\sigma_i$ is the estimated scale parameter at segment $i$; $L_i$ is the segment length in miles for segment $i$; and $\gamma = (\gamma_0, \gamma_1)'$ are the estimated coefficients.

Table 4 presents the modeling results of the GNB and GSI models, respectively. The goodness-of-fit statistics indicate that Model 3 fits the data better than the other two models for the GNB and GSI models. Previously, Geedipally et al. (25) examined different functional forms for the dispersion parameter using the Texas data and found Model 3 to be the best model. Overall, the GSI model provides a slightly better fit than the GNB model.
TABLE 4 Modeling Results for GNB and GSI Models with the Texas Data

<table>
<thead>
<tr>
<th></th>
<th>GNB Model</th>
<th></th>
<th>GSI Model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
<td>Model 3</td>
<td></td>
</tr>
<tr>
<td>Estimates</td>
<td>Value</td>
<td>SE</td>
<td>Value</td>
<td>SE</td>
</tr>
<tr>
<td>Intercept ln(β₀)</td>
<td>-8.4812</td>
<td>0.4002</td>
<td>-7.2909</td>
<td>0.3858</td>
</tr>
<tr>
<td>Ln(Average daily traffic) β₁</td>
<td>1.0184</td>
<td>0.0429</td>
<td>0.9234</td>
<td>0.0411</td>
</tr>
<tr>
<td>Lane Width β₂</td>
<td>-0.0429</td>
<td>0.0115</td>
<td>-0.07</td>
<td>0.0161</td>
</tr>
<tr>
<td>Total Shoulder Width β₃</td>
<td>-0.0086</td>
<td>0.003</td>
<td>-0.0112</td>
<td>0.0032</td>
</tr>
<tr>
<td>Curve Density β₄</td>
<td>0.062</td>
<td>0.0105</td>
<td>0.0943</td>
<td>0.013</td>
</tr>
<tr>
<td>Intercept ln(γ₀)</td>
<td>-1.059</td>
<td>0.0935</td>
<td>-1.491</td>
<td>0.0717</td>
</tr>
<tr>
<td>Segment Length γ₁</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-0.4901</td>
</tr>
<tr>
<td>Log-likelihood</td>
<td>2634.04</td>
<td>2563.91</td>
<td>2548.2</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>5280.07</td>
<td>5139.81</td>
<td>5110.41</td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>5311.95</td>
<td>5171.69</td>
<td>5147.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Model 1</td>
<td>Model 2</td>
<td>Model 3</td>
<td></td>
</tr>
<tr>
<td>Estimates</td>
<td>Value</td>
<td>SE</td>
<td>Value</td>
<td>SE</td>
</tr>
<tr>
<td>Intercept ln(β₀)</td>
<td>-7.9994</td>
<td>0.3868</td>
<td>-7.7736</td>
<td>0.3837</td>
</tr>
<tr>
<td>Ln(Average daily traffic) β₁</td>
<td>0.9931</td>
<td>0.0404</td>
<td>0.9696</td>
<td>0.039</td>
</tr>
<tr>
<td>Lane Width β₂</td>
<td>-0.0602</td>
<td>0.0173</td>
<td>-0.0618</td>
<td>0.0182</td>
</tr>
<tr>
<td>Total Shoulder Width β₃</td>
<td>-0.01</td>
<td>0.0033</td>
<td>-0.0102</td>
<td>0.0035</td>
</tr>
<tr>
<td>Curve Density β₄</td>
<td>0.0627</td>
<td>0.0121</td>
<td>0.0716</td>
<td>0.0126</td>
</tr>
<tr>
<td>Intercept ln(γ₀)</td>
<td>35.19</td>
<td>0.001</td>
<td>-0.6084</td>
<td>0.1869</td>
</tr>
<tr>
<td>Segment Length γ₁</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-0.557</td>
</tr>
<tr>
<td>ν</td>
<td>-3.643</td>
<td>0.2565</td>
<td>-2.678</td>
<td>0.2661</td>
</tr>
<tr>
<td>Log-likelihood</td>
<td>2550.23</td>
<td>2542.47</td>
<td>2540.23</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>5114.47</td>
<td>5098.93</td>
<td>5096.46</td>
<td></td>
</tr>
<tr>
<td>BIC</td>
<td>5151.64</td>
<td>5136.12</td>
<td>5138.96</td>
<td></td>
</tr>
</tbody>
</table>

Note: - = not applicable.

Henceforth, Model 3 is adopted as the functional form for estimating the dispersion parameter and the scale parameter. Figure 4 shows the comparison of variance predicted by the GNB and GSI models. It can be observed that, when the crash mean is less than 20, both models can yield almost the same variance. For higher crash means, the GSI model predicted a slightly higher variance than the GNB model. Overall, the figure illustrates that the difference in variance
between the GNB and GSI models is not large. Moreover, it can be concluded that the variance of the GSI model (or GNB model) increases at a slower rate than that of the SI model (or NB model) as the number of crash mean increases. As discussed by EI-Basyouny and Sayed (23), this is probably because the GNB and GSI models have more parameters and consequently lower variability.

![FIGURE 4 Variance versus mean for GNB and GSI models.](image)

We also compare the effects of the GNB and GSI models on EB estimates. As seen in Figure 5 (a) and (b), both models have similar relationships between the weight factor and the model estimates. Although both models can yield almost the same weight factor, it can be observed that the GSI model usually produces a slightly lower value for the weight factor than that of the GNB model. From Figure 5 (a), it can be seen that the crash prediction model (i.e., NB or SI model) with a varying dispersion parameter/term will affect the weight factor. Moreover, adding a varying dispersion parameter/term in the NB or SI model will have a similar effect on the weight factor.
(a) Weight Factor by GNB model

(b) Weight Factor by GSI Model
FIGURE 5 Weight factors produced by GNB and GSI models (a) relationships between the weight factor and the model estimates (b) associations of weight factors between GNB and GSI models.

Figure 6 illustrates the relationship between the hotspot identification lists ranked by the GNB model and by the GSI model. Although there is some variation in the ranking between two models, Figure 6 shows a stronger positive association between the GNB model ranking and the GSI model ranking than the association between the NB model ranking and the SI model ranking (as illustrated in Figure 3). As shown in Table 5, a detailed comparison of the ranking between the GNB and GSI models demonstrates that there are some differences in ranking. Overall, the analysis in this section suggests that the GSI model can affect the value of weight factor used for estimating the EB output. Consequently, the GSI model will influence the identification of hazardous sites as well.
TABLE 5 Differences in Ranking between GNB and GSI Models using the EB Estimates

<table>
<thead>
<tr>
<th>Differences in ranking</th>
<th>Difference and percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-identical ranking</td>
<td>1356 (90.46%)</td>
</tr>
<tr>
<td>Ranking difference beyond 10 positions</td>
<td>643 (42.89%)</td>
</tr>
<tr>
<td>Ranking difference beyond 50 positions</td>
<td>58 (3.86%)</td>
</tr>
</tbody>
</table>

There are 1499 road segments in the Texas data.

SUMMARY AND CONCLUSIONS

This paper has documented the difference in obtaining the EB estimates by using the SI and NB models. The NB model is the most commonly used crash prediction model. Recently, the SI model was introduced for crash data modeling. The SI model has three different parameters and a dispersion term is defined to measure the level of dispersion as well as obtain the EB estimates. The SI model is more flexible and can handle the observed high dispersion more efficiently than the NB model while maintaining the same logical properties. Moreover, although not included in the paper (due to space limitations), we compared the predictive abilities of the NB and SI models. The results suggest that the SI model provides better prediction performance for the Texas data.

The objective of this study is to examine and compare the effects of the SI and NB models on EB estimates. To accomplish the study objectives, the NB and SI models with fixed/varying dispersion parameter/term are developed using the Texas data. The important conclusions can be summarized as follows. (1) The selection of the crash prediction model (i.e., the SI or NB model) will affect the value of weight factor used for estimating the EB output. And the SI model usually put more weight on the observed number of crashes than the model estimates for the crash-prone sites when calculating the EB estimates. (2) The identification of hazardous sites, using the EB method, can be affected when the SI model is used. The difference in weight factor should play an important role for the observed difference in ranking. Similar conclusions can be made for the GSI and GNB models, although the difference in ranking between two models is less significant.

Note that only a few explanatory variables are included in the mean functional form for the NB and SI models. Due to the age of the datasets, we are not able to get access to other important explanatory variables. However, considering the mean structure used in this study, it is possible that if a well-defined mean functional form is used, the difference in ranking between NB and SI models could be slightly different. Thus, more work needs to be done to verify this.
assumption. In addition, due to the rarity of accident events, it is difficult to know the real safety value of a given location (4). In contrast, as indicated by Miranda-Moreno et al. (26), when working with simulated data we have the advantage of starting with the true safety state of each site and hence, can establish the ones that are truly hotspots. Thus, a simulation study designed to examine which crash prediction model (i.e. SI or NB model) can better identify the hotspot is the subject of our future research.

REFERENCES


