Alternative Risk Models for Ranking Locations for Safety Improvement

Luis F. Miranda-Moreno, Liping Fu,
Frank F. Saccomanno, and Aurelie Labbe

Many types of statistical models have been proposed for estimating accident risk in transport networks, ranging from basic Poisson and negative binomial models to more complicated models, such as zero-inflated and hierarchical Bayesian models. However, little systematic effort has been devoted to comparing the performance and practical implications of these models and ranking criteria when they are used for identifying hazardous locations. This research investigates the relative performance of three alternative models: the traditional negative binomial model, the heterogeneous negative binomial model, and the Poisson lognormal model. In particular, this work focuses on the impact of the choice of two alternative prior distributions (i.e., gamma versus lognormal) and the effect of allowing variability in the dispersion parameter on the outcome of the analysis. From each model, two alternative accident estimators are computed by using the conditional mean under both marginal and posterior distributions. A sample of Canadian highway–railway intersections with an accident history of 5 years is used to calibrate and evaluate the three alternative models and the two ranking criteria. It is concluded that the choice of model assumptions and ranking criteria can lead to considerably different lists of hazardous locations.

One of the main tasks in a program for improving safety on a transport network is the identification of a list of hazardous locations (e.g., signalized intersections, road segments, highway–railway intersections) that show evidence of high accident risk. Hazardous locations, referred to as black spots or hot spots, can be defined as locations with high accident frequency or risk when involving both frequency and severity of the accidents. These locations are considered to be the most suitable candidates for engineering inspections and implementation of remedial actions, such as installation of new control devices and improvement of location geometry (*1–3*).

A simple approach to identifying black spots is to rank locations according to the observed number of accidents per vehicle mile or vehicles entered into an intersection, computed for each location without use of data from other locations. This approach has several shortcomings. For example, since accidents occur as rare random events over time, this approach is quite sensitive to random variations. Sites with a high accident frequency within one period may experience low accident frequency in following periods. This approach does not consider that accident frequency may tend to its mean over time—a phenomenon known as regression to the mean (*4*). In addition, this way of ranking locations assumes a linear relationship between the number of accidents and traffic exposure (e.g., vehicle miles traveled or vehicle entries in an intersection), which has been argued in many studies to be nonlinear. Finally, relevant location attributes related to accident occurrences are ignored (*3, 4*).

Instead of ranking dangerous locations by using the observed number of accidents, there has been continuous interest in applying different ranking criteria—posterior mean of accident frequency or posterior expectation of ranks, for example—derived from several random effect or Bayesian models—for example, negative binomial (NB) and hierarchical Bayesian models (*1–3*).

Although the standard Poisson regression (assuming a fixed mean) has been applied for modeling accident events, it has limitations in the presence of overdispersion commonly observed in accident data (*5*). To deal with problems of overdispersion, the Poisson gamma or NB regression model has been widely used (*6*). An advantage of the NB model is that it can capture the unmeasured or unobserved heterogeneities that are due to omitted variables and intrinsic randomness. Both the Poisson and the NB regression model have been extended to deal with the excess of zeros, another possible source of overdispersion in count data. These extensions lead to the zero inflated Poisson (ZIP) and zero inflated negative binomial (ZINB) models, respectively, which also have been used to identify hazardous locations and evaluate countermeasures (*7, 8*). A recent review of these four regression models is available elsewhere (*9*). Other mixed Poisson models have been applied in public health and demography studies (*10–12*). Among these models is the Poisson lognormal model, which is considered in this study.

Among the ranking criteria for the identification of black spots, the conditional mean of accident frequency obtained from the marginal distribution of the NB regression model has been used (*13, 14*). Alternatively, several accident estimators derived from the posterior distribution by using Bayesian analysis have been more often applied for this task. Among them are the posterior mean of accident frequency, the potential of accident reduction, the posterior expectation of ranks, and the probability of being the most dangerous locations (*2, 3, 15*). One of the main advantages of the Bayesian analysis is that it combines the information brought by the accident data with prior knowledge into the posterior distribution (*4, 16*). In this work, the terms "ranking criterion" and "accident estimator" are used indistinctively.

Whereas several alternative statistical models and ranking criteria are available in the literature for identifying hazardous locations on a transport network, there have been few systematic studies on the comparative performance and practical implications of some model assumptions. Many important issues remain to be addressed. What

L. F. Miranda-Moreno, Department of Civil Engineering, 200 University Avenue West, University of Waterloo, Ontario N2L 3G1, Canada, and Instituto Mexicano del Transporte, Querétaro, México. L. Fu and F. F. Saccomanno, Department of Civil Engineering, University of Waterloo, 200 University Avenue West, Waterloo, Ontario N2L 3G1, Canada. A. Labbe, Department of Mathematics and Statistics, Laval University, Quebec City, Quebec G1K 7P4, Canada.

Transportation Research Record: Journal of the Transportation Research Board, No. 1908, Transportation Research Board of the National Academies, Washington, D.C., 2005, pp. 1–8.

is the impact of the use of alternative statistical models for decision making? How important is the assumption about the prior or random effect distribution (e.g., gamma versus lognormal) on the output of the analysis? How significant can be the differences between the ranking criteria derived from marginal or posterior distributions? Which ranking criteria are the most appropriate for the identification of dangerous locations?

The primary goal for this research is to provide empirical evidence about the effect of the use of alternative models and criteria on the ranking of locations for safety improvement. Three alternative models—the NB model, the heterogeneous negative binomial (HNB) model, and the Poisson lognormal model—and two ranking criteria—marginal and posterior mean of accident frequency—are considered in this study. A sample of highway–railway grade crossings located in the Canadian railway network is used as an application environment.

MODEL DESCRIPTION

In traffic safety studies, the standard Poisson regression model has been applied for modeling accident data (*6*). This regression model assumes that the number of accidents Y_i occurring over a period at a site i is independently Poisson distributed, that is,

$$Y_i \mid \mu_i \sim \text{Poisson}(\mu_i) \tag{1}$$

where the set of independent observations for the n locations is represented by the vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ with corresponding accident mean $\mu = (\mu_1, \mu_2, \ldots, \mu_n)'$. This model is constrained to the assumption $E(Y_i|\mu_i) = \text{Var}(Y_i|\mu_i) = \mu_i$, where μ_i is commonly defined as an exponential function of a vector of covariates, $\mu_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$, where $\boldsymbol{x}_i' = (1, x_{i1}, \ldots, x_{ik})$ is a vector of covariates and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)$ are regression parameters to be estimated from the data. A shortcoming of this model is that the vector of covariates $\boldsymbol{x}_i$ usually does not explain completely the conditional mean because of omitted exogenous variables or randomness (*5*). For example, because of the complexity of accident events and limitations on relevant information (driver behavior, weather conditions), it is impossible to consider all the factors that affect accident occurrence.

To deal with the problem of overdispersion caused by unmeasured heterogeneities, random variations in the conditional mean of the Poisson model can be captured by introducing a random effect term in a multiplicative way. This leads to the mixed Poisson models, such as the Poisson gamma and Poisson lognormal models (*10, 17*).

Poisson Gamma Model

The Poisson gamma model (also called the NB model) permits the relaxation of the assumption that the variance is equal to the mean by introducing a gamma random effect in a multiplicative way (*17*). Thus, instead of assuming the accident mean to be fixed as in the standard Poisson model, here it is assumed to be random and denoted by θ_i. With this assumption, the Poisson gamma model can be presented as follows (*5*):

$$Y_i \mid \theta_i \sim \text{Poisson}(\theta_i) \tag{2a}$$

$$\theta_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\exp(\epsilon_i) \tag{2b}$$

$$\exp(\epsilon_i) \sim \text{gamma}(\phi, \phi) \tag{2c}$$

where as before $\mu_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$. This model implies that $\exp(\epsilon_i)$ follows a gamma distribution with $E[\exp(\epsilon_i)] = 1$ and consequently $\text{Var}[\exp(\epsilon_i)] = 1/\phi$, which is obtained by specifying that the shape and dispersion parameters (ϕ) of the gamma distribution are equal.

NB Model: Fixed Dispersion Parameter

To obtain the marginal distribution of the NB model, the random effect, $\exp(\epsilon_i)$, is integrated. The resulting marginal is equal to the NB probability function as follows:

$$m(y_i \mid \mu_i, \phi) = \frac{\Gamma(y_i + \phi)}{y_i!\Gamma(\phi)}\left(\frac{\phi}{\mu_i + \phi}\right)^{\phi}\left(\frac{\mu_i}{\mu_i + \phi}\right)^{y_i} \qquad y_i = 0,1,\ldots,n \tag{3}$$

where ϕ is the dispersion parameter and usually is expressed for computing convenience as a function of α, that is, $\phi = 1/\alpha$. As in the Poisson model, $\mu_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$. The conditional mean and variance of the NB marginal distribution are given by

$$\mu_i^{\text{NBF}} = E(Y_i \mid \mu_i, \phi) = \mu_i \tag{4}$$

$$\text{Var}(Y_i \mid \mu_i, \phi) = \mu_i(1 + \mu_i/\phi) \tag{5}$$

where NBF is the NB model with fixed ϕ. As shown in the following section, the log likelihood of this model can be maximized numerically by using the Newton–Raphson algorithm to estimate the model parameters (*17*).

HNB Model: Varying Dispersion Parameter

An alternative parameterization of the traditional NB regression model is to allow observed variability in the dispersion parameter. That is, the dispersion parameter ϕ_i is assumed to vary across locations as a function of covariates such as traffic conditions (*2, 18*). This extension is used to try to structure the unmeasured heterogeneities. For example, two highway–railway intersections with the same number of daily trains located in the same railway corridor could have similar accident patterns, and thus the unmeasured heterogeneities associated with these intersections may be structured in the same way. Modeling the dispersion parameter can increase flexibility and thus precision of accident estimates. The HNB model can be defined as follows (*17, 19*):

$$Y_i \mid \mu_i, \phi_i \sim \text{NB}(\mu_i, \phi_i) \tag{6}$$

where ϕ_i is a function of traffic-flow conditions or other site attributes that modifies the magnitude of the dispersion parameter among sites. In this study, ϕ_i is modeled by using the following link function (*19*):

$$\phi_i = 1/[\gamma_0 \times \exp(\boldsymbol{z}_i'\boldsymbol{\gamma})] \tag{7}$$

where $\boldsymbol{z}_i' = (z_{i1}, \ldots, z_{ik})$ is a vector of covariates representing location attributes of each location (not necessary the same as $\boldsymbol{x}_i'$) and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)'$ is a vector of parameters. The μ_i^{NBV} parameter in Equation 6 denotes the marginal mean of accident from the HNB model, that is,

$$\mu_i^{\text{NBV}} = E(Y_i \mid \beta, \phi_i) = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \tag{8}$$

Notice that Equation 8 is the same as Equation 4, except that the regression parameters $\boldsymbol{\beta}$ are estimated, allowing observed variability in ϕ_i. In road safety studies, the gamma probability density function as a random effect or prior has been widely used. With this choice, a close form of the marginal distribution can be obtained (i.e., NB distribution), yielding computational simplification. However, the gamma distribution may not necessarily be the best one for modeling the unobserved heterogeneities in some accident data sets.

Poisson Lognormal Model

Instead of assuming a gamma distribution for the random effect $\exp(\epsilon_i)$, a more convenient probability density function can be the lognormal. With this assumption, the Poisson lognormal model can be defined as follows (*10, 17*):

$$Y_i \mid \phi_i \sim \text{Poisson}(\phi_i) \tag{9a}$$

$$\phi_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})\exp(\epsilon_i) \tag{9b}$$

$$\exp(\epsilon_i) \sim \text{lognormal}(0, \sigma^2) \tag{9c}$$

The marginal distribution of this model does not have as close a form as the Poisson gamma model. However, to obtain the maximum likelihood estimates $\hat{\sigma}^2$ and $\hat{\boldsymbol{\beta}}$, we can use several methods, such as the Gauss–Hermite quadrature or the EM algorithm (*12, 17, 20*). The mean and variance of the marginal distribution of this model are given by (*21*)

$$\hat{\mu}_i^{\text{LN}} = \text{E}(Y_i \mid \boldsymbol{\beta}, \sigma) = \exp\left[\boldsymbol{x}_i'\boldsymbol{\beta} + \tfrac{1}{2}\sigma^2\right] \tag{10}$$

$$\text{Var}(Y_i \mid \boldsymbol{\beta}, \sigma) = E(Y_i \mid \boldsymbol{\beta}, \sigma) \times \left\{1 + E(Y_i \mid \boldsymbol{\beta}, \sigma) \times \left[\exp(\sigma^2) - 1\right]\right\} \tag{11}$$

where LN is the Poisson lognormal model. Notice that if $\sigma^2 \to 0$, $E(Y_i \mid \boldsymbol{\beta}, \sigma)$ and $\text{Var}(Y_i \mid \boldsymbol{\beta}, \sigma)$ are reduced to the mean and variance of the Poisson model. Thus, σ can be considered the difference between Poisson and Poisson lognormal models.

This Poisson lognormal model can be a good candidate for modeling accident rates with a heavier-tailed distribution since the lognormal tails are known to be asymptotically heavier than those of the gamma distribution (*20, 22*). For instance, this model can better fit some data than the Poisson gamma under the presence of outliers (*20*). Empirical evidence and other advantages of the Poisson lognormal model were given by Winkelmann (*17*), who compared different models for count data analysis and found that this model better fits a particular data set than the NB model.

Recently, hierarchical Poisson lognormal models have been applied for modeling accident data by using Markov chain Monte Carlo techniques for parameter estimation into the full Bayesian approach (*21, 23*). In this paper, the Poisson lognormal model is applied by using the maximum likelihood approach to estimate the prior parameters (*12, 20*). Here, the Poisson lognormal model is applied to evaluate the impact of the choice of the prior distribution, by comparing this model with the Poisson gamma model.

Parameter Estimation

In the three presented models, parameters of the prior distribution as well as the regression parameters $\boldsymbol{\beta}$ must be specified or estimated. The well-known maximum likelihood method is used here and, in the case of random effects models (or, equivalently, Bayesian models), the likelihood is computed as

$$\ell(\boldsymbol{y} \mid \mu, \rho) = \sum_{i=1}^{n} \ell(y_i \mid \mu_i, \rho), \tag{12}$$

where ρ represents the set of parameters from the prior distribution (for example, in the NB model, $\rho = \phi$) and where

$$\ell(y_i \mid \mu_i, \rho) = \log \int f(y_i \mid \theta_i) g(\theta_i) d\theta_i \tag{13}$$

where $g(\cdot)$ represents the prior distribution applied on θ_i. In practice, numerical methods such as Newton–Raphson are needed to estimate the parameters $\boldsymbol{\beta}$ and ρ (*12, 17, 20*).

ACCIDENT ESTIMATORS BASED ON POSTERIOR MEAN

As mentioned, to incorporate prior information and at the same time combine specific-site attributes (i.e., accident history, traffic volumes, and other location attributes), posterior distribution by using a Bayesian approach has been widely recommended for identifying hazardous locations (*1–3, 15*).

Two main approaches can be distinguished within the class of Bayesian methods: the full Bayes approach and the empirical Bayes (EB) approach. The main difference between these two approaches is in the way the hyperparameters (i.e., parameters from the prior distribution) are determined. In the full Bayes, hyperparameters are determined on the basis of some prior belief on the behavior of the data involved (*16*). However, including prior beliefs about data is challenging and controversial, and this has led many researchers to use the so-called EB approach (*10, 16*).

In the EB approach, the hyperparameters are estimated by using the maximum likelihood technique described in the previous section or any other techniques involving the use of the accident data. This approach has been criticized for implicitly using the data twice. That is, the data are first used to estimate the parameters of the prior distribution, and once these values are determined, the accident history of each location is used to make inferences about the posterior distribution (*4, 16*).

The introduction of a multiplicative random effect, distributed according to a known probability function, is mathematically equivalent to considering that the mean of the Poisson model θ_i follows a specific prior distribution. Thus, the terms "prior" and "random effect" are used indistinctively in this paper.

Posterior Mean Under Poisson Gamma Model

Once the gamma distribution is defined as a prior distribution and y_i accidents are observed in a given location, the posterior distribution can be derived on the basis of the Bayes theorem. With the gamma as a prior distribution of θ_i, the posterior distribution of the NB model is also gamma distributed. Thus, the posterior mean under the NB model, denoted by $\hat{\mu}_i^{\text{EBF}}$, can be written as follows (*2, 4*):

$$\hat{\mu}_i^{\text{EBF}} = \left(\frac{1}{1 + \hat{\mu}_i / \hat{\phi}}\right)(\hat{\mu}_i - y_i) + y_i \tag{14}$$

where $\hat{\mu}_i = \exp(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})$ and $\hat{\phi}$ are the maximum likelihood estimates obtained from the NB marginal distribution (Equation 3). Another parameterization of the NB model is to allow variability in ϕ_i, which leads to the HNB model. Thus, to improve the flexibility of the posterior mean of accident frequency, this new arrangement can be applied by computing ϕ_i as a function of some covariates, as presented in Equation 7. To avoid confusion, the posterior mean under the HNB model is denoted by $\hat{\mu}_i^{\text{EBV}}$ and computed as follows:

$$\hat{\mu}_i^{\text{EBV}} = \left[\frac{1}{1+\hat{\mu}_i \times \hat{\gamma}_0 \times \exp(\boldsymbol{z}_i'\hat{\boldsymbol{\gamma}})}\right](\hat{\mu}_i - y_i) + y_i \quad (15)$$

where it can be recalled that the magnitude of $\hat{\phi}_i$ varies from one location to another according to site-specific attributes ($\boldsymbol{z}_i$), such as traffic flows. Heydecker and Wu (*2*) and Miaou and Lord (*18*) considered variability of $\hat{\phi}_i$ by using a similar parameterization.

Approximated Posterior Mean Under Poisson Lognormal Model

For the Poisson lognormal model introduced earlier, the implementation of the EB approach is more complicated because the lognormal is not a conjugate distribution as the gamma for the Poisson model. Therefore, the lognormal posterior mean must be approximated (*10, 11*). The approximate posterior mean introduced by Clayton and Kaldor (*11*) and used also by others (*12*) will be used here and can be written as follows:

$$\hat{\mu}_i^{\text{EBLN}} = \exp\left[\frac{\boldsymbol{x}_i'\hat{\boldsymbol{\beta}} + \hat{\sigma}(y_i + 0.5)\zeta_i - 0.5\hat{\sigma}^2}{1+\hat{\sigma}^2(y_i + 0.5)}\right] \quad (16)$$

where $\hat{\mu}_i^{\text{EBLN}}$ is an approximate posterior mean of the Poisson lognormal model and $\zeta_i = \log(y_i + 0.5)$ given that for values of $y_i = 0$, ζ_i would be undefined. Again, $\hat{\sigma}^2$ and $\hat{\boldsymbol{\beta}}$ are the maximum likelihood estimates obtained from the marginal distribution of the Poisson lognormal model.

COMPARISON OF ALTERNATIVE ACCIDENT ESTIMATORS: CASE STUDY

The previous section described three different models: negative binomial (fixed ϕ), heterogeneous negative binomial (varying ϕ_i), and Poisson lognormal models. For each model, both the conditional mean based on a marginal distribution and the posterior mean of the accident frequency were presented as two alternative ranking criteria for identifying hazardous locations. Therefore, by using the marginal distribution of each model, the following accident estimators can be defined:

- $\hat{\mu}_i^{\text{NBF}}$ = conditional mean of accident frequency based on the marginal distribution of the NB model (fixed ϕ), Equation 4;
- $\hat{\mu}_i^{\text{NBV}}$ = conditional mean of accident frequency based on the marginal distribution of the HNB model (varying ϕ_i), Equation 8; and
- $\hat{\mu}_i^{\text{LN}}$ = conditional mean of accident frequency based on the marginal distribution of the Poisson lognormal model, Equation 10.

In addition, the accident estimators derived from the posterior mean of the accident frequency are denoted by

- $\hat{\mu}_i^{\text{EBF}}$ = posterior mean of accident frequency under the NB model, Equation 14;
- $\hat{\mu}_i^{\text{EBV}}$ = posterior mean of accident frequency under the HNB model, Equation 15; and
- $\hat{\mu}_i^{\text{EBLN}}$ = approximate posterior mean of accident frequency based on the Poisson lognormal model, Equation 16.

The objective for this case study is to compare these alternative accident estimators by using an accident data set obtained from a sample of highway–railway intersections as an application environment. From this will be observed the impact that the model assumptions and ranking criteria previously discussed can have on the identification of hazardous locations.

Data Description

The data set used for this application combines information from two databases provided by Transport Canada and the Canadian Transportation Safety Board. One database consists of a crossing inventory that contains information on approximately 29,500 grade crossings (public and private) located nationally in Canada. The occurrence accident database includes information of car–train accidents recorded for several years. Several groups of attributes are included in the crossing inventory: geographical location, type of warning devices, some geometry features, and road and train traffic volumes (*14*).

A sample of 5,094 highway–railway grade crossings with flashing lights as main warning devices was selected. The other two main groups in the database are crossings with reflectorized signboard and crossings with gates. Splitting the inventory according to the warning devices helps to avoid the problem of correlation among crossing attributes.

For model calibration and ranking criteria comparison, the historical number of accidents from 1996 to 2000 (a 5-year period) were considered. With this consideration, effort was made to exclude significant changes of crossing attributes over time (e.g., traffic conditions).

A brief description of the variables involved in the analysis is presented in Table 1. As a complement, a correlation matrix was estimated to identify high linear correlation among explanatory variables. A small or moderate linear association was found among the attributes involved in the analysis (correlation coefficients less than 0.5).

Model Calibration and Validation

Before the model calibration, the functional form of μ_i is specified for the three regression models. Here, a popular functional form is adopted (*14, 18*):

$$\mu_i = \exp[\beta_0 + \beta_1 \ln(\text{AADT}_i \times T_i) + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}] \text{ or}$$
$$= (\text{AADT}_i \times T_i)^{\beta_1} \exp(\beta_0 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}) \quad (17)$$

where

- AADT_i = average annual daily traffic,
- T_i = number of daily trains, and
- x_i = other geometry attributes of each crossing *i*.

TABLE 1 Crossing Characteristics and Accident History

Category	Crossing Attributes	Description	Average or Percentage	Minimum	Maximum
Road features	Posted road speed	km per hour	63.5	10	110
	Road type	Collector or arterial = 1, 0 others	37%*	0	1
	Surface material	Asphalt = 1, 0 others	47%*	0	1
Railway features	Max. train speed	Miles per hour	36.6	5	95
	Track number	Number	1.2	1	6
	Track angle	Degrees	68.0	0	90
Traffic volumes	AADT	Average annual daily traffic	2,532.0%	10	57,000
	Daily trains	Trains per day	6.7	1	73
	Exposure	ln(AADT × daily trains)	8.2	2.3	13.7
Accident history	Observed accidents (five-year period)	Accidents per crossing	0.1	0	5
		% of crossings with zero accidents	93.4%	—	—

*% of roads classified as collectors or arterials and % of asphalted roads.

The product of $AADT_i$ and T_i is a measure of traffic exposure and was proposed by Farr (*24*) and Saccomanno et al. (*14*). The functional form given by Equation 17 allows a possible nonlinear relationship between the accident occurrence and traffic flows. Other functional forms have been proposed in the safety traffic literature (*18*). Since there is no information on the time-of-day variations of traffic conditions, it is not possible to consider a more precise measure of exposure.

Once model parameters are estimated by using the accident history of a 5-year period, the accident estimators of each model are computed. The statistical software packages SAS 8.2 and LIMDEP 8.0 are used for model calibration. After all possible combinations of the crossing attributes introduced in Table 1 are tried, the most significant set of covariates is selected for each model (Table 2). The *t*-ratio test was used for the selection process, considering a confidence level of 95%. The coefficients and t-ratio values of the final regression models are given in Table 2.

TABLE 2 Parameter Values and Statistics for Each Model

Regression Model	Variable	Parameter	*t*-ratio	*P*-value
NB	Intercept	−7.965	−20.303	0.000
	Exposure	0.543	14.153	0.000
	Max. train speed	0.011	4.068	0.000
	Road type	0.372	3.303	0.001
	Dispersion (α)	1.385	4.112	0.000
	Log likelihood	−1273.44		
HNB	Intercept	−7.988	−21.244	0.000
	Exposure	0.546	14.712	0.000
	Max. train speed	0.011	4.180	0.000
	Road type	0.340	2.933	0.003
	Constant (γ_0)	2.569	2.775	0.006
	Daily trains (γ_1)	−0.058	−3.541	0.000
	Log likelihood	−1262.23		
Poisson lognormal	Intercept	−8.496	−20.708	0.000
	Exposure	0.545	14.313	0.000
	Max. train speed	0.011	4.092	0.000
	Road type	0.379	3.248	0.001
	Dispersion (σ)	0.992	10.563	0.000
	Log likelihood	−1262.44		
	Vuong test (V)	2.22		

Testing Overdispersion

The next step is to verify the presence of overdispersion in the data by testing the null hypothesis of the inverse dispersion parameter (H_0: $\alpha = 0$) obtained from the NB regression model by applying the *t*-ratio statistic, which is obtained by dividing the estimate $\hat{\alpha}$ by its standard error. Note that the hypothesis $\alpha = 0$ does not make sense mathematically but reflects the case where the variance of $\exp(\epsilon_i)$ is zero, implying that it is considered as being constant. Such a case is then equivalent to the standard Poisson model. Then, by testing the null hypothesis H_0, one is testing if there is enough evidence to assume $\exp(\epsilon_i)$ to be randomly distributed. Alternatively, overdispersion in the data can be identified on the basis of the likelihood ratio (T_{LR}), which is equal to −2 times the difference in the fitted log likelihood of two nested models (*5*).

A value of $\hat{\alpha} = 1.385$ (i.e., $\hat{\phi} = 0.722$) is obtained with a standard error of 0.337 when the NB model is applied (Table 2), which is significantly different from zero at the 95% confidence level, confirming the existence of overdispersion. In addition, the presence of overdispersion can be identified by calculating T_{LR} with the log likelihood values of Poisson and NB regression models. In this case, the log likelihood values of these two models are −1292.44 and −1273.40, respectively. From these values $T_{LR} = 38.0$, which exceeds the 1% critical value of $\chi^2_{.98}(1) = 5.41$. Note that the T_{LR} statistic approximately follows a chi-squared distribution. From these results, overdispersion for unobserved heterogeneity is clearly revealed.

Goodness of Fit

To test the appropriateness of the Poisson lognormal versus the standard Poisson model, use the test statistic Vuong (*V*), which is useful for comparing nonnested regression models (*25*). In this test, if $|V|$ is greater than 1.96 (critical value for a 95% confidence level), it favors the selection of the Poisson lognormal model. With the calibrated Poisson lognormal and Poisson models, the *V* statistic is 2.22, which is greater than the critical value and supports the selection of the second model.

Alternatively, one can compare the goodness of fit of nonnested models calibrated with the same data by using the Akaike information criterion (AIC) (*5*). This criterion is computed as AIC = −2 log likelihood + *k*, where *k* = number of model parameters.

A regression model with a low AIC is preferred. In this study, the values of AIC for the NB, HNB, and Poisson lognormal models are equal to 2550.9, 2528.5, and 2528.9, respectively. Thus, one can conclude that the HNB and Poisson lognormal better fit the data than the traditional NB regression model. Furthermore, the HNB model and Poisson lognormal regression model have a similar quality of goodness of fit.

Decision Implications of Alternative Models

In this section, implications are explored for using alternative models or ranking criteria presented previously, such as choice of the random effect distribution (i.e., gamma versus lognormal), specification of the dispersion parameter (i.e., fixed ϕ versus varying ϕ_i), and the impact of using alternative estimators (i.e., marginal versus posterior means). To achieve this objective, the differences among accident estimators, derived from the same or different model, will be computed on the basis of two measures named percentage deviation and Spearman correlation coefficient.

Percentage Deviation

The percentage deviation can be applied to compare two ranking criteria for the number of locations that are different in the two lists of dangerous locations. For that, a sample of crossings is ranked according to two accident estimators, resulting in two different ranked lists of sites. Then, the number of sites selected from the top of each list is designated as black spots (m), and the percentage of locations that are different in both lists of hazardous sites is calculated. That is, the percentage deviation that results from the comparison of two ranking criteria can be computed as

$$\% \text{ deviation} = 100 \times (1 - s/m) \tag{18}$$

where s is the number of dangerous locations that are common in the two compared lists and m is the total number of dangerous sites selected from the top of a list sorted according to a given criteria (e.g., $\hat{\mu}_i^{EBF}$). A high % deviation is obtained if two ranking criteria generate very different lists of dangerous locations.

To compare the different ranking criteria based on percentage deviation, first specify $\hat{\mu}_i^{EBF}$ as the base ranking criterion, because it is among the most used in practice. Other optimal estimators are compared to this criterion by using the % deviation. Figure 1 shows the percentage deviation for different lists of black spots (m). From this comparison, the following can be observed:

- There are important differences between the posterior mean of the NB model ($\hat{\mu}_i^{EBF}$) and the approximated posterior mean of the Poisson lognormal model ($\hat{\mu}_i^{EBLN}$). Between these two ranking criteria, the percentage deviation varies from 10% to 30%, reaching a maximum value when m is around 300. This result implies that the choice of the prior distribution may lead to significantly different lists of dangerous crossings in this particular analysis.
- The discrepancy between $\hat{\mu}_i^{EBF}$ and $\hat{\mu}_i^{EBV}$ (i.e., posterior mean of the NB model versus posterior mean of the HNB model) is moderated. The deviation varies from 10% to 15% when m is less than 400. This means that between 85% and 90% of the black spots identified with $\hat{\mu}_i^{EBF}$ are the same as the hazardous crossings identified with $\hat{\mu}_i^{EBV}$. As expected, their difference decreases as m increases.

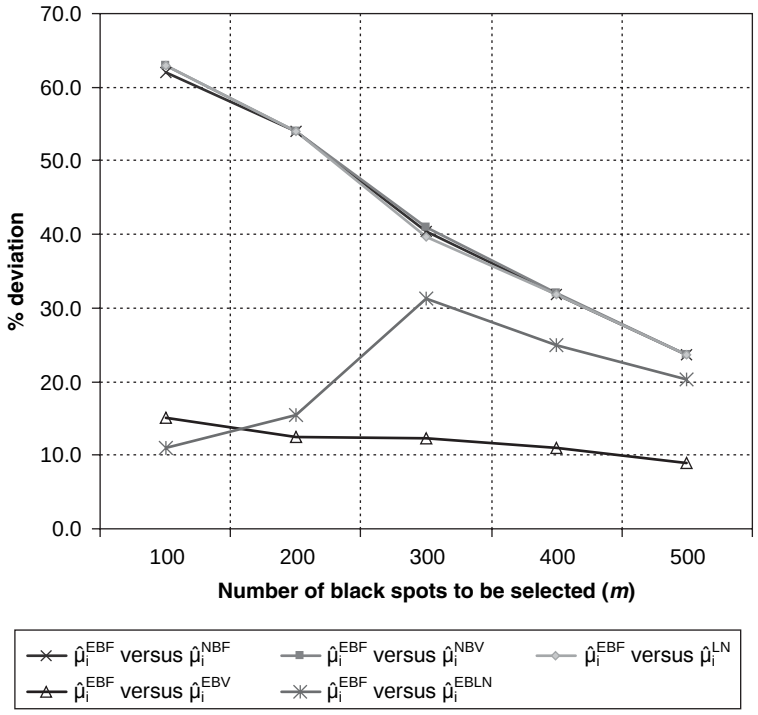

FIGURE 1 Percentage deviation among $\hat{\mu}_i^{EBF}$ and alternative accident estimators.

• Conversely, the differences between $\hat{\mu}_i^{EBF}$ and the conditional mean based on the marginal distribution of each model ($\hat{\mu}_i^{NBF}$, $\hat{\mu}_i^{NBV}$, and $\hat{\mu}_i^{LN}$) are fairly significant. For instance, when one selects a small number of black spots (e.g., $m < 200$ dangerous crossings), the percentage deviation is more than 50% for the three comparisons. However, these differences fall dramatically as the list size of black spots increases. In general, the ranks obtained with marginal means of accidents are significantly different from the ranks obtained with posterior means of accidents. On the other hand, the differences among the three marginal accident estimators appear to be insignificant.

Spearman Correlation Coefficient

The Spearman correlation coefficient is a nonparametric technique that is usually applied to evaluate the degree of linear association between two independent variables (*26*). Here this coefficient is used to measure the correlation between two accident estimators. That is, the coefficient is used to measure the degree of association between two lists of hazardous sites ordered on the basis of two ranking criteria, and it is computed as follows:

$$r = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{m(m^2 - 1)} \tag{19}$$

where d_i is the difference between the two ranks of a specific site i and, as defined previously, m is the number of sites to be selected as hazardous. The value of r can vary from +1 to −1. A value close to +1 suggests that the two ranking criteria are positively linearly related and vice versa. The coefficient r can give some important insights into the dimensions of the shifts in the ranking orders of two lists of hazardous sites. Thus, combining r and percentage deviation gives a better conclusion about the differences obtained with two accident estimators. The coefficient r can be formally tested by using the statistic $t = r/\sqrt{(1 - r^2)/(n - 2)}$ which has a t-distribution with $n - 2$ degrees of freedom. If $|t| > t_{\omega/2}$, the null hypothesis that the correlation coefficient (r) is zero, where ω is the level of significance, is rejected.

For example, by using this correlation coefficient, one can explore the degree of association between $\hat{\mu}_i^{EBF}$ and $\hat{\mu}_i^{EBV}$ (i.e., fixed ϕ versus varying ϕ_i). Also, one can see the impact of the choice of prior (i.e., gamma versus lognormal) by estimating the degree of association between $\hat{\mu}_i^{EBF}$ and $\hat{\mu}_i^{EBL}$. The results of the Spearman correlation coefficient are presented in Figure 2 for different dimensions of m.

From Figure 2 it can be seen that the degree of correlation between $\hat{\mu}_i^{EBF}$ and $\hat{\mu}_i^{EBV}$ is high (more than 80%) when the list size is greater than 300 black spots. However, when the list size is less than 300 black spots, the correlation is moderate (less than 60%). In this case, allowing variability in the dispersion parameter can produce considerable changes in the positions of the locations into a list of black spots.

In addition, the degree of association between $\hat{\mu}_i^{EBF}$ and $\hat{\mu}_i^{EBL}$ is lower than that between $\hat{\mu}_i^{EBF}$ and $\hat{\mu}_i^{EBV}$, especially when m is in the range of 200 to 300 sites. Thus, it can again be seen that the choice of the prior distribution has an important effect on the ranks of the sample of crossings used in this study.

CONCLUSIONS AND FUTURE RESEARCH

This research investigated the relative performance and decision implications of three alternative risk models and two ranking criteria in the context of identification of locations for safety improvements. An accident data set of Canadian highway–railway intersections was used to calibrate and evaluate the different models and ranking criteria. Some of the main conclusions are summarized as follows:

• The Poisson lognormal model has been introduced as an alternative to the NB model into the context of the empirical Bayes approach.

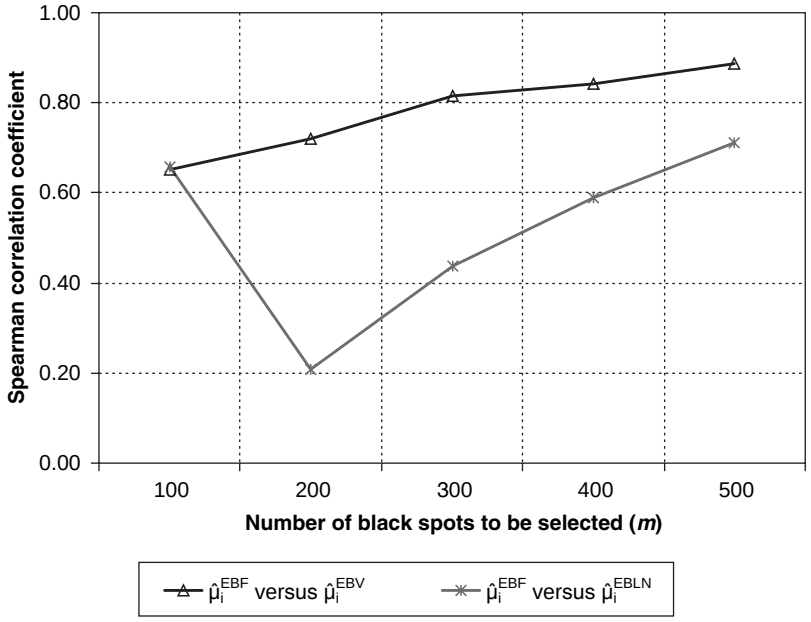

FIGURE 2 Degree of association among $\hat{\mu}_i^{EBF}$ and two alternative posterior means, $\hat{\mu}_i^{EBV}$ and $\hat{\mu}_i^{EBLN}$.

A comparison of these two models was done to observe the effect that the choice of the prior distribution can have on the identification of hazardous sites. In this case study, the choice of the prior could lead to considerably different lists of hazardous locations.

- Allowing variability in the dispersion parameter can add flexibility to the posterior mean of the NB model and thus improve both goodness of fit and accuracy of accident estimates. In the case study, both the Poisson lognormal model and the HNB model better fit the data than did the traditional NB model.
- For a given model, use of the expected accident frequency under the marginal distribution yields strikingly different lists of hazardous locations than use of the posterior mean as ranking criterion. As observed in previous studies, the use of the posterior distribution may be more appropriate than the use of the marginal distribution of a given model for the identification of dangerous sites.

Finally, it should be noted that this research represents the first step toward the goal of developing a set of useful guidelines that can be used by practitioners to select the most appropriate tools for identifying high-risk locations. Many important issues remain to be addressed before this goal can be achieved:

- The conclusions obtained from this study are limited to the data set used in the analysis. Further investigations that use more data sets from different transportation facilities are necessary to substantiate the findings and generalize these preliminary conclusions. For instance, it is possible that the choice of lognormal distribution could be advantageous when modeling accident data sets that contain locations with a relatively high number of accidents. In such cases, the long tail of the lognormal distribution makes it a good competitor of the gamma distribution.
- This research covers only a limited number of ranking criteria. Thus, it is important to investigate the impact of use of other ranking criteria, such as posterior expectation of ranks and probability of being the most dangerous location.
- Because of the rarity of accident events, it is difficult to know the true value or a reliable estimate of the safety status of a given location on the basis of limited observation history. This property suggests that it is impossible to obtain the absolute performance of a model based on accident history. This research has attempted to answer only the question of the relative differences between the alternative models and ranking criteria. Some researchers have suggested using simulated data to quantify the absolute performance of alternative models (*9*). This is an area for future research.
- Previous work suggests that the ranking of locations may be sensitive to accident severity. It is also interesting to study different accident types and various classes of accident severity.

REFERENCES

1. Persaud, B., C. Lyon, and T. Nguyen. Empirical Bayes Procedure for Ranking Sites for Safety Investigation by Potential for Safety Improvement. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1665,* TRB, National Research Council, Washington, D.C., 1999, pp. 7–12.
2. Heydecker, B. G., and J. Wu. Identification of Sites for Accident Remedial Work by Bayesian Statistical Methods: An Example of Uncertain Inference. *Advances in Engineering Software,* Vol. 32, 2001, pp. 859–869.
3. Miaou, S.-P., and J. J. Song. Bayesian Ranking of Sites for Engineering Safety Improvements: Decision Parameter, Treatability, Statistical Criterion, and Cost Function. Presented at 83rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2004.
4. Hauer, E. *Observational Before-After Studies in Road Safety: Estimating the Effect of Highway and Traffic Engineering Measures on Road Safety.* Elsevier Science Publishers, Amsterdam, 1997.
5. Cameron, A. C., and P. K. Trivedi. *Regression Analysis of Count Data.* Cambridge University Press, Cambridge, England, 1998.
6. Miaou, S.-P. The Relationship Between Truck Accidents and Geometric Design of Road Sections: Poisson Versus Negative Binomial Regressions. *Accident Analysis and Prevention,* Vol. 26, No. 4, 1994, pp. 471–482.
7. Shankar, V., J. Milton, and F. Mannering. Modeling Accident Frequency as Zero-Altered Probability Process: An Empirical Inquiry. *Accident Analysis and Prevention,* Vol. 29, No. 6, 1997, pp. 829–837.
8. Miranda-Moreno, L. F., and L. Fu. A Comparative Study of Alternative Risk Estimators for Ranking Highway-Rail Grade Crossings for Safety Improvement. Presented at 13th Pan-American Conference on Traffic and Transportation Engineering, New York, 2004.
9. Lord, D., S. P. Washington, and J. N. Ivan. Statistical Challenges with Modeling Motor Vehicle Crashes: Understanding Implications of Alternative Approaches. Presented at 83rd Annual Meeting of the Transportation Research Board, Washington, D.C., 2004.
10. Rao, J. *Small Area Estimation.* John Wiley and Sons, New York, 2003.
11. Clayton, D., and J. Kaldor. Empirical Bayes Estimates of Age-Standardized Relative Risk for Use in Disease Mapping. *Biometrics,* Vol. 43, No. 3, 1987, pp. 671–681.
12. Meza, J. Empirical Bayes Estimation Smoothing of Relative Risk in Disease Mapping. *Journal of Statistical Planning and Inference,* Vol. 112, 2003, pp. 43–62.
13. Saccomanno, F., R. Grossi, D. Greco, and A. Mehmood. Identifying Black Spots Along Highway SS107 in Southern Italy Using Two Models. *Journal of Transportation Engineering,* Nov. 2001, pp. 515–552.
14. Saccomanno, F. F., L. Fu, and L. F. Miranda-Moreno. Risk-Based Model for Identifying Highway-Rail Grade Crossing Blackspots. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1862,* Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 127–135.
15. Schluter P. J., J. J. Deely, and A. J. Nicholson. Ranking and Selecting Motor Vehicle Accident Sites by Using a Hierarchical Bayesian Model. *The Statistician,* Vol. 46, No. 3, 1997, pp. 293–316.
16. Carlin, B., and T. Louis. *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall, London, 2000.
17. Winkelmann, R. *Econometric Analysis of Count Data.* Springer-Verlag, Berlin, 2003.
18. Miaou, S.-P., and D. Lord. Modeling Traffic Crash-Flow Relationships for Intersections: Dispersion Parameter, Functional Form, and Bayes Versus Empirical Bayes Methods. In *Transportation Research Record: Journal of the Transportation Research Board, No. 1840,* Transportation Research Board of the National Academies, Washington, D.C., 2003, pp. 31–40.
19. Greene, W. LIMDEP, Version 8.0. 2002. www.limdep.com.
20. Sohn, S. Y. A Comparative Study of Four Estimators for Analyzing the Random Event Rate of the Poisson Process. *Journal of Statistical Computation and Simulation,* Vol. 49, 1994, pp. 1–10.
21. Tunaru, R. Hierarchical Bayesian Models for Multiple Count Data. *Austrian Journal of Statistics,* Vol. 31, No. 2–3, 2002, pp. 221–229.
22. Kaas, R., and O. Heseelager. Ordering Claim Size Distributions and Mixed Poisson Probabilities. *Insurances: Mathematics and Economics,* Vol. 17, 1995, pp. 193–201.
23. Miaou, S.-P., J. J. Song, and B. K. Mallick. Roadway Traffic Crash Mapping: A Space-Time Modeling Approach. *Journal of Transportation and Statistics,* Vol. 6, No. 1, 2003, pp. 33–57.
24. Farr, E. *Summary of the DOT-Rail-Highway Crossing Resource Allocation Procedure: Revised,* Federal Railway Administration, U.S. Department of Transportation, 1987.
25. Washington, S. P., M. G. Karlaftis, and F. L. Mannering. *Statistical and Econometric Methods for Transportation Data Analysis.* Chapman & Hall/CRC, Washington, D.C., 2003.
26. Mansfield, E. *Statistics for Business and Economics.* Norton, New York, 1983.

The Statistical Methodology and Statistical Computer Software in Transportation Research Committee sponsored publication of this paper.