ELSEVIER

Accident Analysis and Prevention 39 (2007) 1192–1201

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Bayesian multiple testing procedures for hotspot identification

Luis F. Miranda-Moreno [a,b,*], Aurélie Labbe [c,1], Liping Fu [d,2]

[a] *Centre for Data and Analysis in Transportation, Economics Department, Université Laval, Québec G1K7P4, Canada*
[b] *Instituto Mexicano del Transporte, Querétaro, Mexico*
[c] *Mathematics and Statistics, Université Laval, Québec G1K7P4, Canada*
[d] *Civil Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada*

Received 10 February 2006; received in revised form 20 February 2007; accepted 18 March 2007

Abstract

Ranking a group of candidate sites and selecting from it the high-risk locations or hotspots for detailed engineering study and countermeasure evaluation is the first step in a transport safety improvement program. Past studies have however mainly focused on the task of applying appropriate methods for ranking locations, with few focusing on the issue of how to define selection methods or threshold rules for hotspot identification. The primary goal of this paper is to introduce a multiple testing-based approach to the problem of selecting hotspots. Following the recent developments in the literature, two testing procedures are studied under a Bayesian framework: Bayesian test with weights (BTW) and a Bayesian test controlling for the posterior false discovery rate (FDR) or false negative rate (FNR). The hypotheses tests are implemented on the basis of two random effect or Bayesian models, namely, the hierarchical Poisson/Gamma or Negative Binomial model and the hierarchical Poisson/Lognormal model. A dataset of highway–railway grade crossings is used as an application example to illustrate the proposed procedures incorporating both the posterior distribution of accident frequency and the posterior distribution of ranks. Results on the effects of various decision parameters used in hotspot identification procedures are discussed.

Keywords: Hotspot; Multiple testing; Bayesian approach

1. Introduction

Hotspot identification is usually the first step in a safety improvement program, in which sites are first sorted according to one or more ranking criteria, and a subset of sites are then selected as high accident risk locations or hotspots. These locations are considered as the most suitable candidates for further safety inspections and implementation of remedial actions. Most of the past efforts in the literature have been devoted to the problem of determining the appropriate safety measures (e.g., Hauer, 1997; Miaou and Song, 2005; Persaud and Hauer, 1984; Persaud, 1991; Persaud et al., 1999; Hauer, 1997; Cheng and Washington, 2005). These efforts have lead to the general consensus that the safety status or accident risk at a given location should be estimated using random effect or Bayesian models (Schluter et al., 1997; Heydecker and Wu, 2001; Tunaru, 2002; Miranda-Moreno et al., 2005; Miaou and Song, 2005).

In contrast, the issue of what decision rules should be used in selecting hotspots is still widely open. Traditionally two strategies are commonly followed as hotspot selection rules (Schluter et al., 1997; Hauer et al., 2004): (1) selecting sites on the basis of the budget available to conduct safety inspections and implement countermeasures and (2) selecting a list of sites based on some specified cutoff value or threshold of accident risk. The former is probably the most common in practice, where hotspots are selected sequentially from the ranked list until all budgeted resources are exhausted. The latter ensures the selection of a list of locations that are deemed dangerous at some critical level, leaving variable the number of locations to be selected (Higle and Witkowski, 1988; Schluter et al., 1997; Heydecker and Wu, 2001). This second strategy is often the result of local safety policies that stipulate tolerance levels of accident risks. It is the most appropriate when one wants to identify a list of hazardous

* Corresponding author at: Centre for Data and Analysis in Transportation, Economics Department, Université Laval, Québec G1K7P4, Canada. Tel. +1 418 656 2131x13574; fax: +1 418 656 2495.
E-mail addresses: luis-f.miranda-moreno.1@ulaval.ca (L.F. Miranda-Moreno), Aurelie.Labbe@mat.ulaval.ca (A. Labbe), lfu@uwaterloo.ca (L. Fu).
[1] Tel.: +1 418 656 2131x5281.
[2] Tel.: +1 519 888 4567x3984.

0001-4575/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2007.03.008

sites exceeding a certain threshold value, e.g., all sites having at least a 80% of chance to exceed a certain accident rate or number of accidents.

One of the main limitations of the budget-based strategy is that it may result in an unnecessarily large list of sites including a number of locations that are in fact not dangerous, or on the contrary, may produce a short list ignoring locations which are truly at high risk. On the other hand, the shortcoming with the threshold-based strategy is the issue of how to define formally the thresholds so that the resulting decisions are sensible with a minimum chance for errors. Notice that for both strategies an erroneous selection of relatively safe sites as hotspots (false positives) can lead to a significant waste of already limited financial resources. Other the other hand, failing to detect true hotspots (false negatives) may result in inefficient reduction of accidents (Higle and Hecht, 1989; Schluter et al., 1997; Cheng and Washington, 2005). The objective of this paper is to introduce a new strategy called Bayesian testing approach that can be used to formally define decision rules for hotspot identification. This method offers a rigorous way of making decisions based on a hypothesis testing framework, so that the number or rate of wrong decisions in the hotspot selection process can be explicitly minimized or controlled. This multiple testing-based methodology has not been used in transport safety literature, but has been widely applied in areas such as genomics and astronomy (e.g., Benjamini and Hochberg, 1995; Genovese and Wasserman, 2002; Muller et al., 2004; Do et al., 2005; Scott and Berger, 2006).

A literature review on hotspot identification is offered in Section 2. The theoretical definition of the Bayesian multiple testing is presented in Section 3. In Section 4 we define how hypotheses tests can be implemented under the posterior distribution of ranks. The hierarchical modeling framework considered here is introduced in Section 5. A numerical example showing various aspects of model assessment and application of the multiple testing procedures is discussed in Section 6. Finally, Section 7 recapitulates the main conclusions and provides directions for future research.

2. Literature review

A hotspot identification methodology includes essentially two elements: safety measures or criteria for ranking sites of interest and decision rules for selecting hotspots. Past research has mostly focused on developing appropriate measures that can be used to quantify the safety status or risk at individual locations of interest. The simplest risk measure is raw accident rates, such as the number or cost of accidents per vehicle-miles traveled or per vehicle entry. Unfortunately, raw risk estimators have several limitations as discussed in a number of studies (e.g., Hauer, 1997; Miaou and Song, 2005). Specifically, a ranking method relying on raw accident rates may produce large numbers of misclassifications (e.g. selecting relatively safe locations as hotspots or vise versa) due to the random variation of traffic accidents from year to year (Persaud and Hauer, 1984; Persaud, 1986, 1991; Hauer, 1997; Cheng and Washington, 2005).

As an alternative, model-based approaches that apply random effect or Bayesian models become more and more popular in safety literature and are widely recommended for the hotspot identification task (Schluter et al., 1997; Heydecker and Wu, 2001; Tunaru, 2002; Miranda-Moreno et al., 2005; Miaou and Song, 2005). Following this approach, many alternative ranking criteria and random effect or Bayesian models have been proposed for hotspot identification. Among the most popular ranking criteria are the posterior mean or rate of accident frequency (Hauer and Persaud, 1984; Higle and Witkowski, 1988; Persaud, 1991; Miranda-Moreno et al., 2005), potential of accident reduction (Persaud et al., 1999; Heydecker and Wu, 2001), posterior probability of being the most dangerous site (Schluter et al., 1997; Tunaru, 2002) and posterior expectation of ranks (Tunaru, 2002; Miaou and Song, 2005). Among the alternative models, we can mention the popular negative binomial, generalized negative binomial, Poisson/Lognormal and hierarchical Bayesian models (Miranda-Moreno et al., 2005). These models have also been extended to account for spatial and/or temporal patterns as well as accident severity (Tunaru, 2002; Miaou and Song, 2005).

A number of studies have also been devoted to the issue of relative performance of Empirical Bayesian methods and other techniques such as the statistical quality control or confidence intervals (e.g., Norden et al., 1956; Laughlin et al., 1975; Hauer and Persaud, 1984; Persaud and Hauer, 1984; Higle and Hecht, 1989; Cheng and Washington, 2005). Some researchers have also proposed to incorporate accident severity or crash costs into risk measures (e.g., Hauer et al., 2004; Geurts et al., 2004; Miranda-Moreno, 2006). Others have classified traffic accidents by specific characteristics such as roadway environment, weather conditions or accident types, such as turning, side-swipe, and rear-end (Flak and Barbaresso, 1982; Sayed et al., 1995). Furthermore, a few researchers have explored the issue of how many years of accident data should be employed in the analysis (May, 1964; Cheng and Washington, 2005).

Research on hotspot selection rules is however noticeably scarce. As mentioned before, an inappropriate selection method may lead to a significant number of misclassifications (false positives or false negatives). This issue has been previously discussed by Higle and Hecht (1989) and Cheng and Washington (2005). Despite its importance some simple selection rules have commonly been adopted (Higle and Witkowski, 1988; Schluter et al., 1997). In this research, we attempt to address this methodological gap by introducing a selection approach that minimizes and/or control the number or proportion of misclassifications explicitly.

3. Decision rules based on Bayesian multiple testing

Considering n random variables $(Y_1, \ldots, Y_n)$ corresponding to n sites under study, where Y_i represents the number of accidents over a given time period at site i ($i = 1, \ldots, n$). We assume that Y_i is distributed according to a probability law with density $f(y_i|\theta_i)$, where θ_i represents the mean number of accidents at site i (parameter of interest). In Bayesian analysis, assuming a distribution with density $\pi(.)$ on θ_i allows the incorporation of

a prior knowledge on the behavior of θ_i. This prior information is combined with the information brought by the sample into the posterior distribution, represented by $p(\theta_i|y_i)$, where y_i represents the observed value of Y_i. The posterior distribution of θ_i is a direct application of Bayes' theorem, and has the following form (Carlin and Louis, 2000),

$$p(\theta_i|y_i) = \frac{f(y_i|\theta_i)\pi(\theta_i)}{m(y_i)} = \frac{f(y_i|\theta_i)\pi(\theta_i)}{\int f(y_i|\theta_i)\pi(\theta_i)\mathrm{d}\theta_i}, \quad (1)$$

where $m(y_i)$ represents the unconditional marginal density function of y_i and $f(y_i|\theta)$ is the observed data likelihood. A detailed discussion on the prior distributions assumed for modeling accident data is provided in Section 5.

A set of hypotheses tests can be defined for each site i to determine whether or not to reject the null hypothesis that a given site is not hotspot, that is,

H_{0i}: $\theta_i \leq k$ (site i is referred as non-hotspot),
H_{1i}: $\theta_i > k$ (site i is referred as hotspot),

where k is a standard value or upper limit of the "acceptable" accident frequency, specified by practitioners according to the application conditions. This critical value may be established in different ways and should reflect particular safety policies, decision-maker's experience, historical information, etc. Obviously, use of a large k-value would indicate that only sites with relatively high accident frequency are to be selected, while use of a small k would result in a large number of sites being targeted as hotspots. In practice, this value can be politically determined, representing the demanded safety level before inspecting and targeting a site. In some cases, it can be simply defined on the basis of the observed average and standard deviation of the number of accidents over the whole population under analysis, that is:

$$k = \bar{y} + sz_0, \quad (2)$$

where, $\bar{y}$ is the observed accident mean using data and s the standard deviation of the observed accident history, and z_0 is a constant representing the number of standard deviations to consider. For a more discussion on the definition of k, we refer to Higle and Witkowski (1988), Schluter et al. (1997) and Heydecker and Wu (2001).

In order to carry on a hypothesis test in a Bayesian framework, a test statistic υ_i can be derived from the posterior distribution of θ_i, such that:

$$\upsilon_i = p(H_{1i}|y_i) = p(\theta_i > k|y_i). \quad (3)$$

Once the test statistic is computed, we reject the null hypothesis H_{0i} for site i, if $\upsilon_i \geq t$, where t is a cutoff or threshold value. Obviously, the number of sites that are detected as hotspots largely depends on the critical threshold t. The following section discusses the challenge underlying the problem of determining the optimal threshold and introduces alternative approaches to address this problem.

3.1. The multiple testing issue

In general, when a hypothesis is tested, two types of errors can occur: (1) type I error, when a safe location is selected as hotspot (false positive) and (2) type II error, when a hotspot is identified as safe site (false negative). In a classical setting, if we test a single hypothesis ($n = 1$), the optimal threshold value is often chosen so that the probability of making a type II error is minimized while the probability of making an type I error is controlled at a reasonable level α. However, when we test multiple hypotheses using a common threshold t, the possible outcomes (over the n tests) may be summarized in Table 1. From these outcomes only the total number of rejected null hypotheses, $D(t)$, can be observed. The variable $V(t)$ is referred to as the number of Type I errors, whereas the variable $T(t)$ represents the number of Type II errors (Dudoit et al., 2004). In this case, instead of choosing the threshold value t in an optimal way *for each test*, we aim to determine the optimal cutoff value *for all of the n tests*.

To address this issue, Benjamini and Hochberg (1995) proposed the concept of false discovery rate (FDR) for determining optimal thresholds in a multiple testing setting. FDR is defined as the expected proportion of Type I errors among the rejected null hypotheses, and more formally,

$$\begin{aligned} &\text{i.} \quad \mathrm{FDR}(t) = E\left[\frac{V(t)}{D(t)}\right], \text{ if } D(t) > 0 \\ &\text{ii.} \quad \mathrm{FDR}(t) = E[V(t)], \text{ if } D(t) = 0. \end{aligned} \quad (4)$$

Similarly, the false negative rate (FNR) is defined as the expected proportion of Type II errors among the null hypotheses that have been accepted. We refer to Dudoit et al. (2004) for more information on the error rates and the multiple testing procedures in a frequentist setting (based on p-values). For details on Bayesian multiple testing, we refer to Muller et al. (2004), Do et al. (2005), and Scott and Berger (2006), for instance.

3.2. Bayesian hypothesis testing procedures

In this section, we present two testing procedures, each of them providing a control of a specific global error over the n tests performed.

Table 1
Outcomes of multiple tests when using a t-dependent rejection region

	Test result		
	# of H_0 accepted	# of H_0 rejected	
Real state of the sites			
# of truly non-hotspots (H_0)	$U(t)$	$V(t)$	n_0
# of truly hotspots (H_1)	$T(t)$	$S(t)$	n_1
	$n - D(t)$	$D(t)$	n

Notation: n = total number of sites under analysis; n_0 = unknown number of truly non-hotspots; n_1 = unknown number of truly hotspots; $V(t)$ = number of false positives; $T(t)$ = number of false negatives; $S(t)$ = number of unsafe sites correctly classified as hotspots; $U(t)$ = number of safe sites correctly defined as non-hotspots; $D(t)$ = number of rejected null hypotheses (number of sites detected).

3.2.1. Bayesian test with weights (BTW)

Here, the decision of accepting or rejecting the null hypothesis is based on a specific loss function which specifies the possible economic consequences (cost) of making a decision error. The loss function is commonly defined as (Berger, 1985):

$$\begin{cases} 0, & \text{if the decision taken is right, that is, accept } H_{0i} \text{ when } \theta_i \le k \text{ or reject } H_{0i} \text{ when } \theta_i > k \\ c_0, & \text{if we reject } H_{0i} \text{ when } \theta_i \le k \text{ (false positive)}, \\ c_1, & \text{if we accept } H_{0i} \text{ when } \theta_i > k \text{ (false negative)} \end{cases} \tag{5}$$

where c_0 and c_1 represent the losses for making a wrong decision due to a false positive and a false negative error, respectively. It can be shown that the optimal rejection region that minimizes this loss function can be defined on the basis of the posterior probability of the alternative hypothesis for site i, that is,

$$\text{Reject } H_{0i} \text{ if } \upsilon_i = p(H_{1i}|y_i) \ge \frac{c_0}{c_1 + c_0}, \tag{6}$$

where υ_i is the test statistic for site i defined in Eq. (3). Based on this decision rule, the critical threshold t is given by $t_c^* = c_0/[c_1 + c_0]$, where c stands for this Bayesian test with weights (BTW). Note also that c_0 and c_1 might be relatively fixed with respect to each other. For instance, if we specify $c_0 = 2$ and $c_1 = 1$, we obtain $t_c^* = 0.66$. This means that an incorrect decision of selecting a safe site as hotspot has twice the weight or cost of accepting a site as non-hazardous when it is in fact dangerous.

3.2.2. Bayesian test controlling for the posterior FDR or FNR

Alternatively, we can apply multiple testing procedures that provide a direct control of the false discovery and false negative rates defined in Eq. (4). These procedures have the advantage to control the proportion of non-hotspots included in a list (FDR) or to control the proportion of real hotspots that are excluded from a list (FNR). Thus, using the FDR procedure, practitioners may select a list of sites that is expected to contain a pre-specified percentage of non-hotspots, for instance 5% or 10%. Controlling the FNR, practitioners may also select a list such that only 5% of the real hotspots are missed. In a Bayesian setting, these error rates are actually defined as the posterior expected false discovery rate denoted by $\overline{\text{FDR}}(t)$ and the posterior expected false negative rate denoted by $\overline{\text{FNR}}(t)$. Based on the work of Muller et al. (2004), these posterior expected rates are defined as,

$$\begin{aligned} &\text{i.} \quad \overline{\text{FDR}}(t) = E\left[\frac{V(t)}{D(t)+\xi}\middle| y\right] = \frac{\sum d_i(1-\upsilon_i)}{D(t)+\xi}, \\ &\text{ii.} \quad \overline{\text{FNR}}(t) = E\left[\frac{T(t)}{n-D(t)+\xi}\middle| y\right] = \frac{\sum(1-d_i)\upsilon_i}{n-D(t)+\xi}, \end{aligned} \tag{7}$$

where, d_i is referred to as a discovery or negative count, i.e., $d_i = 1$ if H_{0i} is rejected and $d_i = 0$ otherwise. Then, $D(t) = \Sigma d_i$ is the number of rejected null hypothesis or the number of sites declared as hotspots. Furthermore, $V(t) = \Sigma d_i(1-\upsilon_i)$ and $T(t) = \Sigma(1-d_i)\upsilon_i$ denote the posterior expected count of false discoveries and false negatives, respectively. Finally, the additional term ξ avoids division by zero and y is the observed data. Muller et al. (2004) proposed several ways of combining the goals of minimizing false discoveries and false negatives. Here, we consider the bivariate loss function that explicitly acknowledges the two competing goals, which is defined as,

$$L_R = [\overline{\text{FDR}}(t), \overline{\text{FNR}}(t)], \tag{8}$$

where L_R denotes the posterior expected loss function. Under this loss function, the optimal list of sites can be defined as the minimization of $\overline{\text{FNR}}(t)$ subject to $\overline{\text{FDR}}(t) \le \alpha_D$, where α_D denotes the FDR-control level (e.g., 1%, 5%, ...) which is specified by practitioners. Similarly, we can minimize $\overline{\text{FDR}}(t)$ subject to $\overline{\text{FNR}}(t) \le \alpha_N$, where α_N represents the level at which the $\overline{\text{FNR}}(t)$ wants to be controlled. In the first case, when controlling $\overline{\text{FDR}}(t) \le \alpha_D$, the optimal decision takes the form:

$$\text{Reject } H_{0i} \text{ if } \upsilon_i \ge t_D^*, \tag{9}$$

where, the optimal cutoff t_D^* is given by,

$$t_D^* = \min\{t \in [0, 1], \overline{\text{FDR}}(t) \le \alpha_D\}. \tag{10}$$

More details and proofs of these testing procedures can be seen in Muller et al. (2004). For implementation of this multiple testing procedure, we can apply the following steps:

(1) Once H_{0i} has been fixed for each site i, we compute:

$$\upsilon_i = p(H_{1i}|y_i), \tag{11}$$

according to the probability model assumed.

(2) When controlling for the FDR, we compute t_D^* according to Eq. (10), where $\overline{\text{FDR}}(t)$ is given by Eq. (7-i). In practice, t_D^* can be determined by computing $\overline{\text{FDR}}(t)$ for a finite number of values of t, chosen in [0,1], for instance, $t = 0.99$, 0.98, ... Then, t_D^* will correspond to the minimum t such that $\overline{\text{FDR}}(t) \le \alpha_D$. These two steps are very easy to implement and more details are provided in Miranda-Moreno (2006).

When the aim is to control the proportion of hotspots excluded from a list, decision makers may control the $\overline{\text{FNR}}(t)$ and minimize the $\overline{\text{FDR}}(t)$. For that, the multiple testing procedure described above can be repeated in a similar fashion, with the difference that now the optimal cutoff t_N^* is computed as,

$$t_N^* = \min\{t \in [0, 1], \overline{\text{FNR}}(t) \le \alpha_N\}, \tag{12}$$

where α_N denotes the level at which the $\overline{\text{FNR}}(t)$ is to be controlled.

4. Hypotheses tests based on posterior distribution of ranks

In our previous discussion, we have shown that Bayesian hypotheses tests may be formulated on the basis of the posterior distribution of θ_i. The decision as to whether or not a site should

be considered as a hotspot could also be made on the basis of its relative rank as compared to other sites. The rank of a site i (R_i) under the safety measure θ_i is defined as follows (Rao, 2003),

$$R_i = \sum_{j=1}^{n} I(\theta_i \geq \theta_j), \tag{13}$$

where $I(\cdot)$ is an indicator function and the smallest θ_i has rank 1. In other words, the greatest ranks correspond to the most hazardous sites. We define $\mathbf{R} = (R_1, R_1, \ldots, R_n)'$ as the rank vector for a sample of n sites. Similar to the process of identifying hotspots based on θ_i described above, hotspots can be identified on the basis of the ranks of the sites by testing the following hypotheses,

H_{0i}: $R_i \leq q$ (site i is referred to as non-hotspot),
H_{1i}: $R_i > q$ (site i is referred as hotspot),

where q is a standard or upper limit rank specified by the decision-makers. For instance, we can define q as a certain proportion of n, that is, $q = \gamma \times n$, where γ is a percentage, e.g., 70%, 80%, etc. This hypothesis test can be used when the interest focuses on the identification of sites with ranks greater than a certain percentile value. Once H_{0i} is established, the test statistic $\upsilon_i = p(H_{1i}|y_i)$ is computed under the posterior distribution of R_i, that is, $\upsilon_i = p(R_i > q|y_i)$. The optimal cutoff t^* is then determined according to any of the multiple testing procedures defined in the previous section.

In the present context, one of the advantages of the multiple testing methods discussed above is that they can be used under any modeling settings currently available in the road safety literature. For instance, one can utilize an Empirical Bayes approach implemented via the negative binomial model or a full Bayesian approach applied via hierarchical Bayes models defined in the following section.

5. Hierarchical Bayes models for accident data

Hierarchical Bayes models have been extensively applied for modeling traffic accident data. These models can deal with problems of over-dispersion due to unobserved heterogeneities, and allow the user to incorporate site-specific attributes and complex variations, e.g., time and/or space patterns in the data. In the present context, one major advantage is the flexibility in terms of distributions that are offered, and the possibility to obtain posterior densities of the parameters of interest. The most popular model in road safety is the Negative Binomial or Poisson/Gamma model. This model is widely used in part due to its computational advantages. However, less attention has been devoted to alternative models such as the Poisson/Lognormal model which may be more suitable for modeling accident rates with a heavier-tailed distribution than the Gamma (Winkelmann, 2003; Miranda-Moreno et al., 2005). These two models are defined next in a hierarchical Bayes framework.

5.1. Model framework

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 $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)'$. Here, we start by assuming that for each site i, the number of accidents over a period of time (Y_i) follows an independent Poisson distribution with its mean (θ_i) defined by a regression model as follows (Winkelmann, 2003):

$$\theta_i = \exp(\mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i). \tag{14}$$

where $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})'$ is a vector of covariates representing site-specific attributes, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)'$ is a vector of unknown regression parameters, and ε_i represents the model error for site i. This expression (14) can be written as,

(i) $\theta_i = \mu_i \vartheta_i$,

where $\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$ and $\vartheta_i = \exp(\varepsilon_i)$. Commonly, ϑ_i is viewed as an unmeasured heterogeneity due to omitted exogenous covariates and randomness. The hierarchical structure of the model is the following:

$$\begin{aligned} &\text{(i)} \quad Y_i|\theta_i \overset{\text{iid}}{\sim} \text{Poisson}(\theta_i), \\ &\text{(ii)} \quad \vartheta_i|\eta \sim \pi_\vartheta(\eta) \text{ and } \eta \sim \pi_\eta(.) \\ &\text{(iii)} \quad f(\boldsymbol{\beta}) \propto 1 \end{aligned} \tag{15}$$

This means that a prior distribution π_ϑ is assumed on the error parameter ϑ_i, which depends on another parameter η, with prior π_η. In addition, $f(\boldsymbol{\beta})$ denotes the prior on the regression parameters $\boldsymbol{\beta}$, which is commonly assumed to be flat (normal) or diffuse. A flat prior indicates that a mean equal to 0 and a very large variance is assumed on each of the regression parameters, e.g. $\beta_j \sim N(0,1000)$. Furthermore, the parameters $\boldsymbol{\beta}$ and η, are mutually independent. According to the specification of the priors π_ϑ and π_η, two alternative models are defined as follows.

5.2. Hierarchical Poisson/Gamma model

Based on the model framework defined in (15), the hierarchical Bayes (HB) version of the Poisson/Gamma model can be written as (Carlin and Louis, 2000; Rao, 2003),

$$\begin{aligned} &\text{(i)} \quad \vartheta_i|\phi, \delta \sim \text{Gamma}(\phi, \delta) \\ &\text{(ii)} \quad \phi \sim \text{Exponential}(c) \\ &\text{(iii)} \quad \delta \sim \text{Gamma}(a, b) \end{aligned} \tag{16}$$

where ϕ and δ are the shape and scale parameters of the Gamma distribution. The hyper-parameters a, b and c have fixed values. By specifying $\phi = \delta$, the hierarchical Negative Binomial (HNB) model is derived with the following assumptions,

$$\begin{aligned} &\vartheta_i|\phi \sim \text{Gamma}(\phi, \phi) \\ &\phi \sim \text{Exponential}(c) \end{aligned} \tag{17}$$

Notice that for a given value of ϕ, ϑ_i follows a Gamma distribution with $E[\vartheta_i] = 1$ and $\text{Var}[\vartheta_i] = 1/\phi$.

5.3. Hierarchical Poisson/Lognormal model

Instead of assuming a Gamma prior for ϑ_i, one can assume an alternative probability density function such as the Lognormal. With this assumption, the hierarchical Poisson/Lognormal (HPL) model is written as follows (Rao, 2003):

$$\begin{aligned} &\text{(i)} \quad \varepsilon_i = \log(\vartheta_i)|\sigma^2 \sim \text{Normal}(0, \sigma^2), \text{ and} \\ &\text{(ii)} \quad \sigma^{-2} \sim \text{Gamma}(a, b). \end{aligned} \tag{18}$$

As in the previous models, the hyper-prior parameters a and b have fixed values and are specified by modelers.

These hierarchical Bayes models are argued to be more flexible than the traditional two-stage Poisson models since more stages of randomness can be considered (Carlin and Louis, 2000; Rao, 2003). These models have been also extended to allow spatial variations for modeling geographical dependence such as the conditional auto-regressive or CAR models, and to the multivariate case (Miaou and Song, 2005).

5.4. Bayesian inference and model selection

When working with hierarchical Bayesian models, posterior distributions are not tractable algebraically in many cases, as is the case for the models considered here. This problem can be solved by generating a large number of samples from the posterior distribution using Markov Chain Monte-Carlo (MCMC) algorithms. From these samples, posterior quantities of interest are computed for the model parameters (Carlin and Louis, 2000). In this paper, MCMC algorithms implemented in WinBUGS, such as Gibbs and Metropolis-Hastings, are utilized for this purpose. A test statistic υ_i for each site is also computed using the MCMC samples in order to apply the multiple testing procedures described previously.

To compare two alternative models (e.g., HNB versus HPL model), the popular Deviance Information Criterion (DIC) proposed by Spiegelhalter et al. (2002) is utilized. This criterion is very popular for Bayesian model selection and is based on the posterior distribution of the *deviance* statistic defined as $D_\theta = -2\log f(y|\boldsymbol{\theta})$. Formally, it may be written as,

$$\text{DIC} = \bar{D} + p_\text{D} = 2\bar{D} - D(\bar{\boldsymbol{\theta}}), \tag{19}$$

where $p_\text{D} = \bar{D} - D(\bar{\boldsymbol{\theta}})$ and $\bar{D}$ is the posterior mean of the *deviance* statistic, i.e., $\bar{D} = E[-2\log f(y|\boldsymbol{\theta})]$. In addition, $D(\bar{\boldsymbol{\theta}}) = -2\log f(y|\bar{\boldsymbol{\theta}})]$, which means that the *deviance* is obtained by substituting on D_θ the posterior mean of $\boldsymbol{\theta}$ denoted by $\bar{\boldsymbol{\theta}}$. The model with the smallest DIC is defined as the model that would best fit the data.

6. A case study

We start this section by describing the decision process that practitioners may follow in order to select a subgroup of candidate sites based on any of the Bayesian tests defined previously. This decision process is summarized in Fig. 1, where the first step is the model selection (for instance, HNB or HPL model), which can be done based on the DIC defined previously. As a second step, the hypothesis test is defined according to the choice of posterior distribution, $p(\theta_i|y_i)$ or $p(R_i|y_i)$. As part of this step, the critical limit values k and q must first be specified in order to compute the test statistic υ_i for each site under analysis. Finally, a Bayesian test is selected depending on the type of error rate that is to be controlled. If the Bayesian test with weights (BTW) is applied, the weights of c_0 and c_1 must be designated by practitioners. If the FDR or FNR test is applied, the levels of control α_D or α_N must be defined.

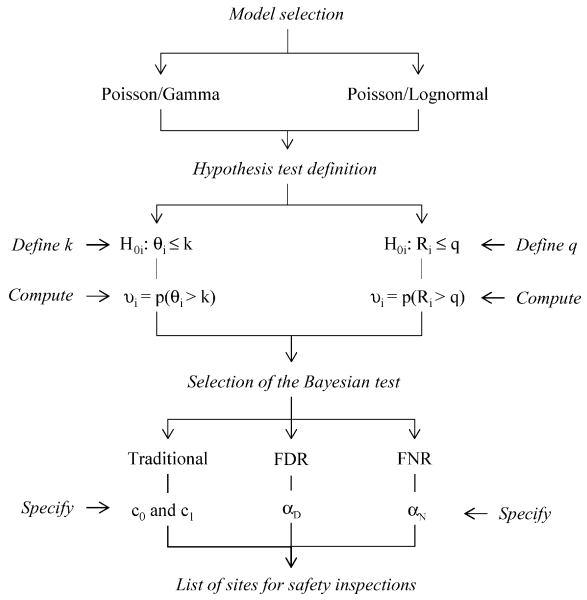

Fig. 1. Decision process for identifying hazardous locations.

In this section, the application of the testing procedures discussed above is illustrated using empirical data. A sample of highway–railway intersections in Canada is considered as an application environment. This is extracted from two databases provided by Transport Canada and the Transportation Safety Board of Canada. One database consists of a grade crossing inventory, including several crossing features such as types of warning device, geometric characteristics and traffic conditions, e.g., posted road speed, maximum train speed, average annual daily traffic, average daily trains, etc. The second database provides information on car-train accidents for several years. These databases have been used by the same authors in previous work (Miranda-Moreno et al., 2005). Here, we consider a sample of public crossings with gates as main warning devices, which comprises approximately 1773 grade crossings, as well as the history of accidents of the period 1997–2001 (5 years of accident information). According to our previous work, we consider as a measure of traffic exposure the natural logarithm of the product of average daily trains and average annual daily traffic (AADT).

6.1. Model calibration and selection

For model calibration, the hyper-parameter c of the HNB model might be fixed in advance. In this study, we follow a more reasonable approach by taking advantage of the dispersion parameter estimate ($\hat{\phi}$) obtained by maximizing the NB marginal

likelihood. That is, we first calibrate the NB model to the studied dataset which yields a dispersion parameter of $\hat{\phi} = 0.64$. Based on the fact that the expectation of the Exponential hyper-prior for ϕ is $1/c$, we can then assume that $E(\phi|c) = 1/c = 0.64$, from which $c = 1.56$. For the HPL model, a vague or non-informative hyper-prior is assumed for σ^{-2} with both parameters a and b, equal to 0.001. This suggests a mean equal to 1 and a very large variance. We refer to Washington and Oh (2006) for more discussions on how to incorporate prior knowledge under a Bayesian framework.

Once the hyper-parameters are fixed, posterior distributions of model parameters are sampled using WinBUGS 1.4. In this study, 6000 simulation iterations were carried out for each parameter of interest, using the first 1000 samples as burn-in iterations. In order to select the crossing attributes to be included in the final model, we first obtained the posterior expected values of all the regression coefficients, along with their standard deviations and 95% confidence intervals. From those, we only selected the attributes whose regression coefficients did not contain 0 in the 95% confidence intervals, i.e., regression parameters significantly different from zero at the 95% confidence level. These crossing attributes are: (1) road type represented as a binary variable (road type = 1 for arterials or collectors, and 0 otherwise), (2) posted road speed, and (3) traffic exposure computed as a function of daily traffic (AADT) and daily trains.

The posterior summary of the regression coefficients $\boldsymbol{\beta}$ along with the dispersion parameters ϕ and σ^2 were computed for both the HNB and HPL models. The results are presented in Table 2. The posterior mean of each regression coefficient, expect β_0, is positive, making sense from a safety point of view and confirming the results obtained in our previous works (Miranda-Moreno et al., 2005). In addition, from the DIC results presented in the same table, we can see that a better fit to the data was obtained when applying the HNB model. The DIC value computed with the HNB model is smaller than the one obtained with the HPL model.

6.2. Some results

The implementation of multiple testing methods involves different decision parameters which can have a significant impact on the size as well as composition of the final hotspot list. In this section, we present some results of a sensitivity study to illustrate the effects of these parameters, including:

a. type of error rates: BTW, FDR or FNR,
b. posterior distributions: θ_i versus R_i,
c. model choice: HNB versus HPL model,
d. control levels: α_D and α_N.

For this analysis, the following values were assumed for k and q:

- $k = 0.18$ for H_{0i}: $\theta_i \leq k$. This k-value is fixed according to Eq. (2), where the observed average number of accidents ($\bar{y}$) in the dataset studied is approximately 0.12, the standard deviation (s) is 0.4 and we consider that $z_0 = 1.5$ standard deviations. This standard value is fixed using the empirical information about accident frequency, independent of the model assumptions.
- $q = 1418$ for H_{0i}: $R_i \leq q$. This q-value corresponds to a $\gamma = 80$th percentile of the ranks where $n = 1773$ locations, that is, $q = 0.8 \times 1773$.

6.2.1. Effect of different error rates

To illustrate the effect of the use of alternative error rates, we consider the following control levels and weights under the two hypotheses defined before and the HNB model:

- A control level $\alpha_D = 10\%$, when using the false discovery rate procedure, that is, the list selected is expected to contain at most 10% of non-hazardous sites,
- A control level $\alpha_N = 10\%$, when applying false negative rate method, that is, at most 10% of non-hotspots are expected to be hazardous,
- Two sets of weights when using the Bayesian test with weights (BTW): $c_0 = 2$ and $c_1 = 1$ (i.e., a false discovery is twice as costly as a false negative) and $c_0 = 1$, $c_1 = 2$ (i.e., a false negative is twice as costly as a false positive).

These results are given in Table 3. As expected, the hotspot list obtained when limiting the FDR is less than the list identified by controlling the FNR. While aiming at controlling FDR has the advantage of having a smaller list of hotspots, it suffers from a higher FNR. Conversely, the approach of controlling FNR ensures that the true dangerous sites are not missed, but it does

Table 2
Summary of model calibration results

Model	Parameters	Posterior mean	S.E.	95% Conf. interval
HNB	Intercept (β_0)	−6.429	0.717	(−7.764, −4.955)
	Road type (β_1)	0.499	0.164	(0.171, 0.815)
	Posted road speed (β_2)	0.011	0.005	(0.001, 0.021)
	Traffic exposure (β_3)	0.323	0.054	(0.214, 0.429)
	ϕ	0.691	0.237	(0.381, 1.332)
	Goodness of fit measure: DIC = 1191.25			
HPL	Intercept (β_0)	−7.041	0.718	(−8.316, −5.657)
	Road type (β_1)	0.506	0.172	(0.167, 0.842)
	Posted road speed (β_2)	0.011	0.005	(0.001, 0.022)
	Traffic exposure (β_3)	0.327	0.049	(0.228, 0.421)
	σ^2	1.034	0.338	(0.596, 2.012)
	Goodness of fit measure: DIC = 1250.40			

Table 3
Number of hotspots identified with different Bayesian tests (Using the HNB model)

Testing method	For H_{1i}: $\theta > 0.18$		For H_{1i}: $R_i > 1418$	
	Threshold	Number of intersections	Threshold	Number of intersections
$\overline{\text{FDR}}(t) \leq 10\%$	$t_D^* = 0.80$	36	$t_D^* = 0.80$	37
BTW, $c_0 = 2$, $c_0 = 1$	$t_c^* = 0.67$	93	$t_c^* = 0.67$	95
BTW, $c_0 = 1$, $c_1 = 2$	$t_c^* = 0.33$	282	$t_c^* = 0.33$	283
$\overline{\text{FNR}}(t) \leq 10\%$	$t_N^* = 0.27$	464	$t_N^* = 0.27$	474

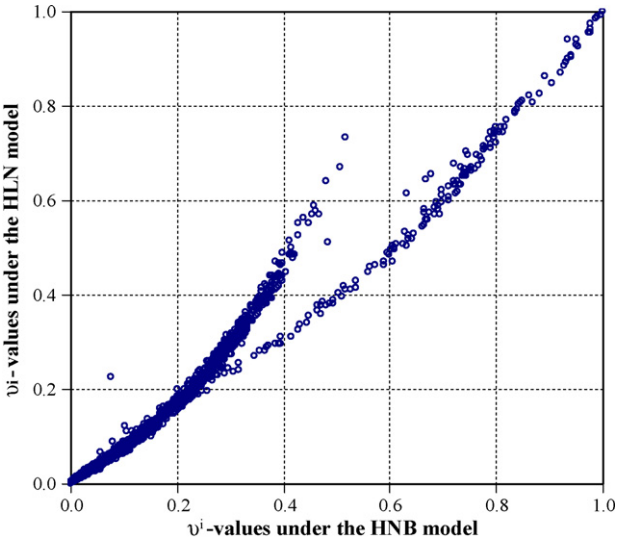

Fig. 2. Comparison of υ_i-values under H_{1i}: $\theta > 0.18$: HNB vs. HLN model.

so at the cost of defining safe sites as hotspots. Furthermore, it can be noted that for a reasonable range of values of c_0 and c_1, the results from the BTW were bounded by the ones obtained with the FDR and FNR procedures. This suggests that the BTW can be used as an alternative to achieve a balance between FDR and FNR.

6.2.2. *Posterior distribution effect, θ_i versus R_i*

A fair comparison of the performance of the posterior of ranks versus the posterior distribution of θ_i may not be easy, as we are dealing with two totally different types of hypotheses. For instance, in testing H_{0i}: $\theta_i \leq k$ and H_{0i}: $R_i \leq q$, the resulting list largely depends on the upper limit values k and q, chosen for both hypotheses. However, it may happen in some situations that the two tests produce very similar lists. An example of that can be seen in Table 3, with $k = 0.18$ and $q = 1418$, where the choice was made to yield lists of the same number of sites.

6.2.3. *Impact of the model type*

The effect of model choice is illustrated in Fig. 2, where the υ_i-values of the two alternative models are compared under the same hypothesis, H_{1i}: $\theta_i \leq 0.18$. In this figure, we can observe relevant differences in the υ_i-values coming from two alternative models, specifically for υ_i-values between 0.2 and 0.6. In this range, the HPL tends to produce larger υ_i-values than the HPL. This happens to be for crossings with very large traffic exposure but no accident history. In this particular case, the output seems to be relatively sensitive to the hyper-prior assumptions on the dispersion parameter (Miranda-Moreno, 2006). Obviously, the model choice may have a significant effect on the hotspot list size. For instance, under the same hypothesis test we see in Fig. 3a that the lists identified with the FDR test and HNB model can be up to 30% larger than those detected with the HPL model. An important impact can also be observed when applying the FNR test (Fig. 3b).

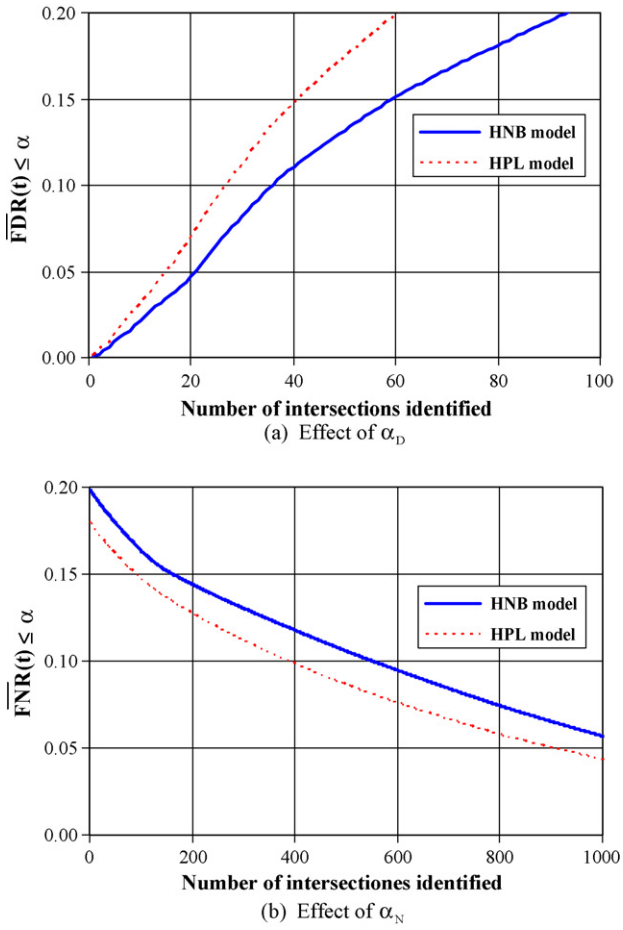

Fig. 3. Influence of the control levels (α_D and α_N) on the hotspot list size.

6.2.4. *Effects of α_D and α_N*

The results shown in Fig. 3a and b also indicate that the critical level of controls α_D and α_N have a significant impact on the size of the hotspot list. Obviously, the number of hotspots identified under either error rate is a function of the control levels. Very small values of α_D produce very small hotspot lists. Conversely, very small values of α_N produce very large hotspot lists. This result is expected, signifying the importance of selecting an appropriate critical error rate before a hotspot identification process is initiated.

In summary, practitioners should be informed that the number of sites that are considered to be hotspots depends on many factors, including:

- The minimum acceptable level of risk that is used in the hypothesis tests (i.e., k or q). For example, H_{0i}: $\theta_i \leq 0.12$ would gives larger lists than H_{0i}: $\theta_i \leq 0.18$.
- The type of error rates used in controlling the overall decision error (i.e., FDR or FNR). FNR gives larger list than FDR.
- Level of control on the error rates. For example, controlling $\overline{\text{FDR}}(t) \leq 5\%$ gives a smaller list than controlling $\overline{\text{FDR}}(t) \leq 10\%$, and $\overline{\text{FNR}}(t) \leq 10\%$ gives a smaller list than $\overline{\text{FNR}}(t) \leq 5\%$.

7. Conclusions

Two Bayesian testing procedures have been introduced for hotspot identification. These procedures can be used not only for ranking a group of locations but also for selecting a list of sites for further engineering safety inspections with a target error rate. They afford practitioners the opportunity to account for uncertainty in model parameters and safety measures while minimizing the false discovery rate (FDR) or false negative rate (FNR) in the selection process.

In particular, the FDR and FNR based procedures are convenient when we want to control the expected rates of false positives or false negatives among the subgroup of locations designated as hotspots. These methods are argued to be more powerful than competing ones, having a natural interpretation and being straightforward in their application. When using the Bayesian test with weights, practitioners need to designate the weights of c_0 and c_1. Under a safety improvement program with large budget constrains, c_0 should be assigned a larger value than c_1. Moreover, when we want to minimize simultaneously both error rates, comparable values should be given to c_0 and c_1. In addition, from a social perspective, we should ideally concentrate on the control of FNR in order to minimize the number of hotspots excluded from a list for further safety inspections. Furthermore, the FDR criterion is economically interpretable since the posterior expected FDR gives the expected proportion of the investment that could be wasted on false leads.

In this paper, we have also shown how both the posterior distribution of accident frequency and posterior distribution of ranks can be used to carry on hypothesis tests. Mainly, the choice of one approach over the other should be motivated by the practitioner's interest. If the interest is the relative ranking of locations, inference based on the posterior distribution of ranks is more efficient than posterior distribution of the mean number of accidents. However, when the group of locations under consideration (n) is relatively large, the computation of the posterior distribution of ranks can be more challenging. A simulation study designed to compare these alternative ranking methods is the subject of our current research. Note that when using empirical accident data, we do not have the ground truth information (a priori) about which sites are truly dangerous. In contrast, 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. By comparing the expected versus observed (true) error rates we can then assess the method's power to detect the true hotspots (Miranda-Moreno, 2006). This approach of using simulated data to validate a new methodology or model has been widely used in the field of statistics.

This research has adopted a hierarchical Poisson framework that has widely been recommended for modeling accident data with substantial sources of variations due to unobserved heterogeneity and space-time trends. Our case study indicated that the model choice may have an important impact on the results of the hotspot selection process. This can be due to the low accident frequency of the studied dataset and the hyper-prior assumptions on the dispersion parameter (the use of diffuse priors). Given that the model choice may be relevant in such cases, we recommend that several models should be applied to the same dataset, and a goodness-of-fit analysis performed in order to select the best model. The performance of alternative hyper-prior assumptions, models and safety measures is also part of our current research.

References

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rates: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B (57), 289–300.

Berger, J.O., 1985. Statistical Decision Theory and Bayesian Analysis. Springer Series in Statistics.

Carlin, B., Louis, T., 2000. Bayes and Empirical Bayes Methods for Data Analysis. Chapman and Hall/CRC Press, London.

Cheng, W., Washington, S., 2005. Experimental evaluation of hotspot identification methods. Accid. Anal. Prevent. 37, 870–881.

Do, K., Müller, P., Tang, F., 2005. A Bayesian mixture model for differential gene expression. J. R. Stat. Soc. C (54), 627–644.

Dudoit, S.M., van der Laan, Pollard, K., 2004. Multiple testing. Part I. Single-step procedures for control of general type I error rates. Stat. Appl. Genet. Mol. Biol. (3), 1, Article 13.

Flak, M.A., Barbaresso, J.C., 1982. Use of computerized roadway information system in safety analyses. Transport. Res. Record 844, 50–55.

Genovese, C., Wasserman, L., 2002. Operating characteristics and extensions of the FDR procedure. J. R. Stat. Soc. B (64), 499–517.

Geurts, K., Wets, G., Brijs, T., Vanhoof, K., 2004. Identification and ranking of black spots: sensitivity analysis. Transport. Res. Record 1897, 34–42.

Hauer, E., Persaud, B.N., 1984. Problem of identifying hazardous locations using accident data. Transport. Res. Record 975, 36–43.

Hauer, E., 1997. Observational Before-After Studies in Road Safety. Pergamon, Tarrytown, NY.

Hauer, E., Kononov, J., Allery, B.K., Griffith, M.S., 2004. Screening the road network for sites with promise. Transport. Res. Record 1784, 27–32.

Higle, J.L., Witkowski, J.M., 1988. Bayesian identification of hazardous sites. Transport. Res. Record 1185, 24–35.

Higle, J.L., Hecht, M.B., 1989. A comparison of techniques for the identification of hazardous locations. Transport. Res. Record 1238, 10–19.

Heydecker, B.G., Wu, J., 2001. Identification of sites for accident remedial work by Bayesian statistical methods: an example of uncertain inference. Adv. Eng. Software (32), 859–869.

Laughlin, J.C., Haefner, L.E., Hall, J.W., Clough, D.R., 1975. Methods for Evaluating Highway Safety Improvements. NCHRP Report 162: TRB, National Research Council, Washington, DC.

May, J.F., 1964. A determination of accident prone location. Traffic Eng. 34 (5), 21–27.

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

Miranda-Moreno, L.F., Fu, L., Saccomanno, F., Labbe, A., 2005. Alternative risk models for ranking locations for safety improvement. Transport. Res. Record 1908, 1–13.

Miranda-Moreno, L.F., 2006. Statistical Models and Methods for the Identification of Hazardous Locations for Safety Improvements. Ph.D. Thesis. Department of Civil Engineering, University of Waterloo.

Muller, P., Parmigiani, G., Robert, C., Rousseau, J., 2004. Optimal sample size for multiple testing: the case of gene expression microarrays. J. Am. Stat. Assoc. (99), 990–1001.

Norden, M., Orlansky, J., Jacobs, H., 1956. Application of statistical quality-control techniques to analysis of highway-accident data. Highway Res. Board 117, 17–31.

Persaud, B.N., Hauer, E., 1984. Comparison of two methods for de-biasing before and after accident studies. Transport. Res. Record 975, 43–49.

Persaud, B.N., 1986. Safety migration, the influence of traffic volumes, and other issues in evaluating safety effectiveness. Transport. Res. Record 1086, 33–41.

Persaud, B.N., 1991. Estimating accident potential of Ontario road sections. Transport. Res. Record 1327, 47–53.

Persaud, B.N., Lyon, C., Nguyen, T., 1999. Empirical Bayes procedure for ranking sites for safety investigation by potential for safety improvement. Transport. Res. Record 1665, 7–12.

Rao, J.N.K., 2003. Small Area Estimation. Wiley, New York.

Sayed, T., Abdelwahab, W., Navin, F., 1995. Identifying accident-prone locations using fuzzy pattern recognition. J. Transport. Eng. (ASCE) 121 (4), 352–358.

Schluter, P.J., Deely, J.J., Nicholson, A.J., 1997. Ranking and selecting motor vehicle accident sites by using a hierarchical Bayesian model. The Statistician 46 (3), 293–316.

Scott, J.G., Berger, J.O., 2006. An exploration of aspects of Bayesian multiple testing. J. Stat. Plann. Inference 136 (7), 2144–2162.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. B (64), 583–640.

Tunaru, R., 2002. Hierarchical Bayesian models for multiple count data. Aust. J. Stat. 31 (3), 221–229.

Washington, S., Oh, J., 2006. Bayesian methodology incorporating expert judgment for ranking countermeasure effectiveness under uncertainty: example applied to at grade railroad crossings in Korea. Accid. Anal. Prevent. 38, 346–356.

Winkelmann, R., 2003. Econometric Analysis of Count Data. Springer, Germany.