A Bayesian Model for Ranking Hazardous Road Sites

Tom Brijs[†1], Dimitris Karlis[‡], Filip Van den Bossche[†] and Geert Wets[†]

[†] Transportation Research Institute
Hasselt University
Wetenschapspark, Gebouw 5
B-3590 Diepenbeek, BELGIUM
email: tom.brijs, filip.vandenbossche, geert.wets@uhasselt.be

[‡] Department of Statistics
Athens University of Economics and Business
76, Patission Str., 10434, Athens, GREECE
email: karlis@aueb.gr

Abstract

Road safety has recently become a major concern in most modern societies. The identification of sites that are more dangerous than others (black spots) can help in better scheduling road safety policies. This paper proposes a methodology for ranking sites according to their level of hazard. The model is innovative in at least two respects. Firstly, it makes use of all relevant information per accident location, including the total number of accidents and the number of fatalities, as well as the number of slight and serious injuries. Secondly, the model includes the use of a cost function to rank the sites with respect to their total expected cost to society. Bayesian estimation for the model via a Markov Chain Monte Carlo (MCMC) approach is proposed. Accident data from 519 intersections in Leuven (Belgium) are used to illustrate the proposed methodology. Furthermore, different cost functions are used in the paper in order to show the impact of the proposed method on the use of different costs per injury type.

Keywords: Gibbs sampling; Markov Chain Monte Carlo; Hierarchical Bayes; Road accidents; Multivariate Poisson distribution;

[1]Corresponding author

1 Introduction

During recent years, road safety has become a major concern for many governments. According to WHO (2004), the social cost of traffic safety in highly developed countries amounts to approximately 2% of annual GDP. Society therefore has an interest in preventing traffic accidents. However, this may not be an easy task. Reducing the number of traffic accidents requires an integrated approach, known as shared responsibility. For example, this can be carried out by improving the active and passive safety of cars, by raising awareness and forcing car drivers to be more careful and by reducing the hazardous condition of roads. The last option involves identifying sites presenting important accident risks so as to make the infrastructure changes needed to reduce the risks at the site. Furthermore, methods that can measure and produce comparable results regarding the risk of each site are of special interest for designing new roads or for enforcing rules. Such rules imply the existence of criteria that judge a specific site to be hazardous. Such criteria can be comparative, i.e. to find the r most hazardous roads, or they might be based on threshold values and hence all the sites that exceed the threshold will be considered as needing change. In practice, these criteria can be combined using relative information about the costs of such repairs.

In this paper, we will concentrate on so-called *black spots*, i.e. dangerous locations where many accidents occur. These situations are, to a great extent, the result of the infrastructure or the way in which it is used. In this study, we will focus on intersections which are classified as black spots after assessing the level of risk, both in terms of the number *and* the gravity of the accidents. Other approaches define black zones (instead of black spots) as spatial concentrations of interdependent high-frequency accident locations (see Flahaut et al., 2003; Thomas, 1996).

From a statistical point of view, we will treat road accidents as random events. As a result, it is impossible to predict the exact circumstances of a single accident. However, in the literature it is commonly assumed that there is an underlying mean accident rate for each individual intersection. In fact, there is a high variation in statistical models in the literature used for analyzing black spot data, but compelling arguments can be found to support the assumption that accident counts follow the Poisson probability law (Lord et al., 2005). In this context, in order to correct for the extra Poisson variation mostly present in accident counts, authors used negative binomial regression models, as for example in Persaud (1990), Hauer (1997) and Abdel-Aty and Radwan (2000). Other authors used generalized Poisson (Kemp, 1973) and logarithmic models (Andreassen and Hoque, 1986). Hauer and Persaud (1987) introduced the Poisson-gamma generalized linear model, allowing the Poisson mean to vary between locations. A comprehensive and elaborate overview of black spot identification techniques is found in Geurts and Wets (2003) and the references therein.

More recently, Bayesian techniques have been used to tackle problems in traffic safety. For instance, Hauer (1986) presented the Empirical Bayes approach as a better estimate of the expected number of accidents, because of the enhanced accuracy of the estimates. Hauer and Persaud (1987) examined the performance of some identification procedures. Empirical Bayes methods were used to estimate proportions of correctly and falsely identified deviant road sections. Belanger (1994) applied Empirical Bayes methods to estimate the safety of four-legged un-signalized intersections. The results were used to identify black spot locations.

However, the use of *hierarchical* Bayes models in traffic safety is less widespread. Schlüter et al. (1997) deal with the problem of selecting a subset of accident sites based on a probability assertion that the worst sites are selected first. They propose different criteria for site selection. To estimate frequency of accidents, a hierarchical Bayes Poisson model has been used. Christiansen et al. (1992) developed a hierarchical Bayes Poisson regression model to estimate and rank accident sites using a modified posterior accident rate estimate as a selection criterion. Davis and Yang (2001) combined hierarchical Bayes methods with an induced exposure model to identify intersections where the crash risk for a subgroup is relatively high. MacNab (2003) adopted a hierarchical Bayes Poisson random effects spatio-temporal methodology to model and analyze accident and injury surveillance data, and Miaou et al. (2003, 2005) used hierarchical Bayes to build model-based risk maps for area-based traffic crashes.

In this paper, we argue that when decisions have to be taken on the investments needed to improve the safety of particular sites, it is important to find a method which examines the risk of the sites in a comparative way; this would allow the higher risk sites to be selected. Problems that hinder this are the different observation periods for different sites and the different lengths of the roads studied. Moreover, data concerning the traffic at each site are needed in order to be able to make fair comparisons. In this context, ranking procedures based on a hierarchical Bayes approach have been proposed. Those methods can deal with the uncertainty and the great variability of the data and produce a probabilistic ranking of those sites. The approach has been used for ranking problems in various application domains, such as educational institutions or hospitals (see, e.g. Goldstein and Spiegelhalter, 1996) as well as in traffic safety (Schlüter et al., 1997). Recently, Tunaru (2002) proposed a hierarchical Bayes approach for ranking accidents sites based on a bivariate Poisson-lognormal distribution. Other interesting work covering the issue of ranking includes that of Bailey and Hewson (2004) who adopted a Bayesian multivariate GLMM to compare the differential performances of several highway authorities based on traffic safety performance indicators.

We expand on this approach by considering a more realistic model for accident behavior taking

into account (1) the number of accidents, (2) the number of fatalities, and (3) the number of slight and serious casualties for a given time period for each site. This is achieved by using a 3-variate Poisson distribution which allows for positive covariance between the variables. The argument for modeling the number of casualties of different types rather than the number of accidents is motivated by the fact that the cost of road accidents to society really depends on the number of people killed and injured rather than on the number of accidents. Thus, any ranking of sites should be based on casualties rather than accidents alone (see also the work of Jones & Jørgensen, 2003, and Bailey & Hewson, 2004).

In order to summarize the data (for a site) by a single number to be used for ranking the sites, we will make use of a cost function that measures the cost of an accident according to the number of fatalities, serious and slight casualties. For the purpose of illustration, we will use two widely differing cost functions. One is proposed by the OECD (Baum and Hohnscheid, 2001); the other is adopted by the Flemish government (see Ministry of Transportation, 2001).

The remainder of the paper proceeds as follows. In section 2 we develop the proposed model. The data are described in section 3. In section 4, we apply the model to the data set and we discuss the results in detail. Finally, concluding remarks can be found in section 5.

2 The Model

Suppose that data were collected from n different sites. The number of accidents for the i-th site is denoted by X_i. We assume that the number of accidents for this site follows a Poisson distribution with parameter $\phi_i t_i$, where t_i is some known exposure measure depending on the monitoring time, the length of the site and/or the traffic flows. Thus ϕ_i's, $i = 1, \ldots, n$ are the unknown pure accident rates for the n sites.

For each site, we have also the triplet (Y_i, Z_i, W_i). Y_i denotes the number of fatalities (including road users who died in the hospital within 30 days after the accident), W_i is the number of seriously injured people, being every road user who was injured in an accident and whose condition involves an admission for at least 24 hours in the hospital. Every road user who got injured in a car accident, but who does not fit the description of the specification of fatally or seriously injured road user, is included in the third group of slight injuries, denoted by Z_i. We assume that jointly and conditional on the number of accidents X_i, they follow a 3-variate Poisson distribution.

Multivariate extensions of the simple Poisson distribution have been proposed in the literature and since the name has been used for different probability functions, it has caused a lot of confusion. In this paper, we make use of a model that allows for pairwise covariances for each pair of variables

(see Karlis and Meligkotsidou, 2005), instead of the usual model that assumes the same covariance term for all the pairs and has been studied in Tsionas (1999) and Karlis (2003). Our model differs from the model of Johnson et al. (1997), which assumes more (but unrealistic) structure.

In the sequel, we call as 3-variate Poisson distribution the joint probability function of the discrete random variables Y_1, Y_2, Y_3 given by

$$\begin{aligned} & P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) = P(y_1, y_2, y_3) \\ = & \sum_{k=0}^{s_1}\sum_{r=0}^{s_2}\sum_{s=0}^{s_3} \frac{e^{-\theta_{12}}\theta_{12}^k}{k!}\frac{e^{-\theta_{13}}\theta_{13}^r}{r!}\frac{e^{-\theta_{23}}\theta_{23}^s}{s!}\frac{e^{-\theta_1}\theta_1^{y_1-k-r}}{(y_1-k-r)!}\frac{e^{-\theta_2}\theta_2^{y_2-k-s}}{(y_2-k-s)!}\frac{e^{-\theta_3}\theta_3^{y_3-r-s}}{(y_3-r-s)!}, \end{aligned} \tag{1}$$

$y_1, y_2, y_3 = 0, 1, \ldots,$ $s_1 = min(y_1, y_2), s_2 = min(y_1 - k, y_3), s_3 = min(y_2 - k, y_3 - r)$ and $\theta_j > 0$, for $j \in \{1, 2, 3, 12, 13, 23\}$. The above distribution will be denoted as $3-Poisson(\theta_1, \theta_2, \theta_3, \theta_{12}, \theta_{13}, \theta_{23})$. The marginal distributions are univariate Poisson distributions, i.e. $Y_1 \sim Poisson(\theta_1 + \theta_{12} + \theta_{13})$, $Y_2 \sim Poisson(\theta_2 + \theta_{12} + \theta_{23})$, $Y_3 \sim Poisson(\theta_3 + \theta_{13} + \theta_{23})$ and the covariance between Y_i and Y_j is given by the corresponding parameter θ_{ij}. In other words, the above model allows for different correlations between each pair of variables, which is clearly a more realistic assumption in the context of traffic accident analysis. Note that empirical evidence supports the assumption that there is positive correlation between the three variables Y_i, Z_i, W_i, so that there is no need to allow for negative correlation in the model. This is natural since it reflects the severity of the accidents on location i. So, instead of assuming independence between the three variables, by imposing three independent Poisson distributions, we propose a model that takes into account those correlations between the variables, and hence it can model the interdependencies in a more realistic way.

2.1 Bayesian Approach

Our model has the form

$$\begin{aligned} X_i &\sim Poisson(\phi_i t_i) \\ (Y_i, Z_i, W_i) \mid X_i = x_i &\sim 3 - Poisson(\mu_{1i}x_i, \mu_{2i}x_i, \mu_{3i}x_i, \lambda_{12i}x_i, \lambda_{13i}x_i, \lambda_{23i}x_i) \end{aligned}$$

where $\mu_{\cdot i}$ are parameters related to fatalities, slight injuries and serious injuries for the site i, while $\lambda_{\cdot i}$ are the covariance parameters for each pair of variables. The likelihood can be written in the complicated form

$$\begin{aligned} L(X, Y, Z, W \mid \boldsymbol{\theta}) &= \prod_{i=1}^{n} P(y_i, z_i, w_i | x_i) P(x_i) \\ &= \prod_{i=1}^{n} \Big[\frac{e^{-\phi_i t_i}(\phi_i t_i)^{x_i}}{x_i!} \sum_{k=0}^{s_1}\sum_{r=0}^{s_2}\sum_{s=0}^{s_3} \Big\{ \frac{e^{-\lambda_{12i}x_i}(\lambda_{12i}x_i)^k}{k!} \frac{e^{-\lambda_{13i}x_i}(\lambda_{13i}x_i)^r}{r!} \times \end{aligned}$$

$$\frac{e^{-\lambda_{23i}x_i}(\lambda_{23i}x_i)^s}{s!}\frac{e^{-\mu_{1i}x_i}(\mu_{1i}x_i)^{y_i-k-r}}{(y_i-k-r)!}\times$$
$$\frac{e^{-\mu_{2i}x_i}(\mu_{2i}x_i)^{z_i-k-s}}{(z_i-k-s)!}\frac{e^{-\mu_{3i}x_i}(\mu_{3i}x_i)^{w_i-r-s}}{(w_i-r-s)!}\Big\}\Big]$$

where $s_1 = min(y_i, z_i), s_2 = min(y_i - k, w_i), s_3 = min(z_i - k, w_i - r)$ and $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{\lambda}_{12}, \boldsymbol{\lambda}_{13}, \boldsymbol{\lambda}_{23})$ represents the vector of all parameters where the vectors represented by bold letters contain the corresponding parameters for all the observations, i.e. $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)$ and similarly for the other vectors.

Full Bayesian inference is not easy for this likelihood as it involves multiple summations. Therefore, a Markov Chain Monte Carlo (MCMC) technique based on Gibbs sampling will be used to explore the posterior distribution of the parameters of interest.

For each parameter, we will assume a Gamma prior and we also assume that the prior distributions are independent. Thus, the prior distribution for the entire vector of parameters $p(\boldsymbol{\theta})$ will be a product of $7n$ Gamma densities. We describe an hierarchical Bayes approach by specifying hyperpriors for the prior parameters. Alternatively, if subjective knowledge on the topic exists one may elicit prior parameters or one may even consider an Empirical Bayes approach.

More formally, let $x \sim Gamma(a, b)$ denote the Gamma distribution with density $f(x) = x^{a-1}b^a exp(-bx)/\Gamma(a)$. Then, the priors are

$$\begin{array}{lllllllll}
\phi_i & \sim & Gamma(a_1, b_1), & \mu_{1i} & \sim & Gamma(a_2, b_2), & \mu_{2i} & \sim & Gamma(a_3, b_3), \\
\mu_{3i} & \sim & Gamma(a_4, b_4), & \lambda_{12i} & \sim & Gamma(a_5, b_5), & \lambda_{13i} & \sim & Gamma(a_6, b_6), \\
 & & & \lambda_{23i} & \sim & Gamma(a_7, b_7) & & &
\end{array}$$

$i = 1, \ldots, n$ for all parameters. Denote as $\mathbf{a} = (a_1, \ldots, a_7)$ and $\mathbf{b} = (b_1, \ldots, b_7)$ the vectors of Gamma prior parameters. We follow the approach of George *e*t al. (1993) and we define a hyperprior $\pi(\mathbf{a}, \mathbf{b})$ as the product of independent priors, i.e.

$$\pi(\mathbf{a}, \mathbf{b}) = \pi(a_1) \ldots \pi(a_7)\pi(b_1) \ldots \pi(b_7)$$

where $\pi(a_i)$ are exponential distributions with means γ_i and $\pi(b_i)$ are $Gamma(\tau_i, \beta_i)$ distributions.

Let $\boldsymbol{X}$ denote the totality of the data. Using these priors, the posterior takes the form of

$$p(\boldsymbol{\theta} \mid \boldsymbol{X}) \propto L(X, Y, Z, W \mid \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \mathbf{a}, \mathbf{b})\pi(\mathbf{a}, \mathbf{b})$$

The predictive distribution can be found by

$$P(X, Y, Z, W) = \int_{\boldsymbol{\theta}} L(X, Y, Z, W \mid \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

which is not of a convenient form. Note that $p(\boldsymbol{\theta}) = \int_{\mathbf{a},\mathbf{b}} p(\boldsymbol{\theta}|\mathbf{a}, \mathbf{b})\pi(\mathbf{a}, \mathbf{b})\mathbf{dadb}$. It may be shown that since the distribution of $(Y_i, Z_i, W_i) \mid X_i$ is a 3-variate Poisson distribution, integrating out

the X_i leads to a 3-variate joint density conditional on θ; marginally, each of (Y_i, Z_i, W_i) will have a univariate Neyman type A distribution (see Douglas, 1980). However, the marginal predictive distribution will be a mixture of Neyman univariate distributions, which is of unknown form. Practically, predictive distributions may easily be estimated from the MCMC output.

2.2 MCMC details

The key ingredient for constructing the MCMC approach is the data augmentation offered by the following representation of a multivariate Poisson distribution, known as multivariate reduction (see e.g. Johnson et al., 1997 and Joe, 1997, section 4.6). We start from a series of independent Poisson variables $\Delta_1, \ldots, \Delta_6$ each one following independently a Poisson distribution, i.e, $\Delta_i \sim Poisson(\xi_i),\ \ i = 1, \ldots, 6$ and then we create the new variables $Y_1 = \Delta_1 + \Delta_4 + \Delta_5$, $Y_2 = \Delta_2 + \Delta_4 + \Delta_6$ and $Y_3 = \Delta_3 + \Delta_5 + \Delta_6$. One can see that Δ_4 appears in both Y_1 and Y_2 and thus it is the term that measures the covariance of Y_1 and Y_2. A similar interpretation holds for Δ_5 and Δ_6. Thus, ξ_4 is the covariance parameter between Y_1 and Y_2 and so on.

In our model, the above idea assumes that there are some latent variables $\delta_{1i}, \delta_{2i}, \delta_{3i}, T_{1i}, T_{2i}, T_{3i}$, where $T_{ji} \mid X_i, \boldsymbol{\theta} \sim Poisson(\mu_{ji} x_i)$ for $j = 1, 2, 3$, $\delta_{1i} \mid X_i, \boldsymbol{\theta} \sim Poisson(\lambda_{12i} x_i)$, $\delta_{2i} \mid X_i, \boldsymbol{\theta} \sim Poisson(\lambda_{13i} x_i)$, $\delta_{3i} \mid X_i, \boldsymbol{\theta} \sim Poisson(\lambda_{23i} x_i)$, from which we construct the working variables $Y_i = T_{1i} + \delta_{1i} + \delta_{2i}, Z_i = T_{2i} + \delta_{1i} + \delta_{3i}, W_i = T_{3i} + \delta_{2i} + \delta_{3i}$. The variables δ_{ji}, $j = 1, 2, 3$ reflect site characteristics that introduce correlation to the working variables. The data augmentation being used is based on considering the unobservable quantities δ_{ji}, $j = 1, 2, 3$, $i = 1, \ldots, n$ as parameters and then to proceed by updating their values according to their posterior distribution. For the other parameters, one may use the standard Gamma conjugate priors to facilitate the computations.

Let $\boldsymbol{\kappa} = (\delta_{11}, \ldots, \delta_{1n}, \delta_{21}, \ldots, \delta_{2n}, \delta_{31}, \ldots, \delta_{3n})$ be the unobserved data. Augmenting parameters $\boldsymbol{\theta}$ by $\boldsymbol{\kappa}$, the joint posterior of all the parameters is of the form

$$
\begin{aligned}
p(\boldsymbol{\theta}, \boldsymbol{\kappa} | data) \quad \propto \quad & \prod_{i=1}^{n} \frac{e^{-\phi_i t_i} (\phi_i t_i)^{x_i}}{x_i!} \frac{e^{-\lambda_{12i} x_i} (\lambda_{12i} x_i)^{\delta_{1i}}}{\delta_{1i}!} \frac{e^{-\lambda_{13i} x_i} (\lambda_{13i} x_i)^{\delta_{2i}}}{\delta_{2i}!} \frac{e^{-\lambda_{23i} x_i} (\lambda_{23i} x_i)^{\delta_{3i}}}{\delta_{3i}!} \times \\
& \frac{e^{-\mu_{1i} x_i} (\mu_{1i} x_i)^{y_i - \delta_{1i} - \delta_{2i}}}{(y_i - \delta_{1i} - \delta_{2i})!} \frac{e^{-\mu_{2i} x_i} (\mu_{2i} x_i)^{z_i - \delta_{1i} - \delta_{3i}}}{(z_i - \delta_{1i} - \delta_{3i})!} \frac{e^{-\mu_{3i} x_i} (\mu_{3i} x_i)^{w_i - \delta_{2i} - \delta_{3i}}}{(w_i - \delta_{2i} - \delta_{3i})!} p(\boldsymbol{\theta} \mid \mathbf{a}, \mathbf{b}) \pi(\mathbf{a}, \mathbf{b})
\end{aligned}
$$

Now, the conditional posteriors can be derived ($\cdot$ denotes the remaining parameters) as

$$
\begin{aligned}
\delta_{1i} \mid \cdot \quad & \propto \quad \frac{\lambda_{12i}^{\delta_{1i}}}{\delta_{1i}!(y_i - \delta_{1i} - \delta_{2i})!(z_i - \delta_{1i} - \delta_{3i})!} \left(\frac{1}{x_i \mu_{1i} \mu_{2i}} \right)^{\delta_{1i}}, \quad \delta_{1i} = 0, \ldots, min(y_i - \delta_{2i}, z_i - \delta_{3i}), \\
\delta_{2i} \mid \cdot \quad & \propto \quad \frac{\lambda_{13i}^{\delta_{2i}}}{\delta_{2i}!(y_i - \delta_{1i} - \delta_{2i})!(w_i - \delta_{2i} - \delta_{3i})!} \left(\frac{1}{x_i \mu_{1i} \mu_{3i}} \right)^{\delta_{2i}}, \quad \delta_{2i} = 0, \ldots, min(y_i - \delta_{1i}, w_i - \delta_{3i}), \\
\delta_{3i} \mid \cdot \quad & \propto \quad \frac{\lambda_{23i}^{\delta_{3i}}}{\delta_{3i}!(z_i - \delta_{1i} - \delta_{3i})!(w_i - \delta_{2i} - \delta_{3i})!} \left(\frac{1}{x_i \mu_{2i} \mu_{3i}} \right)^{\delta_{3i}}, \quad \delta_{2i} = 0, \ldots, min(z_i - \delta_{1i}, w_i - \delta_{2i}),
\end{aligned}
$$

$$
\begin{aligned}
\phi_i \mid \cdot &\sim Gamma(a_1 + x_i, b_1 + t_i), \quad i = 1, \ldots, n \\
\mu_{1i} \mid \cdot &\sim Gamma(a_2 + y_i - \delta_{1i} - \delta_{2i}, b_2 + x_i), \quad i = 1, \ldots, n \\
\mu_{2i} \mid \cdot &\sim Gamma(a_3 + z_i - \delta_{1i} - \delta_{3i}, b_3 + x_i), \quad i = 1, \ldots, n \\
\mu_{3i} \mid \cdot &\sim Gamma(a_4 + w_i - \delta_{2i} - \delta_{3i}, b_4 + x_i), \quad i = 1, \ldots, n \\
\lambda_{12i} \mid \cdot &\sim Gamma(a_5 + \delta_{1i}, b_5 + x_i), \quad i = 1, \ldots, n \\
\lambda_{13i} \mid \cdot &\sim Gamma(a_6 + \delta_{2i}, b_6 + x_i), \quad i = 1, \ldots, n \\
\lambda_{23i} \mid \cdot &\sim Gamma(a_7 + \delta_{3i}, b_7 + x_i), \quad i = 1, \ldots, n
\end{aligned}
$$

For the prior parameters, the conditional posteriors are of the form

$$
\begin{aligned}
a_j \mid \cdot &\propto \left[\frac{b_j^{a_j}}{\Gamma(a_j)}\right]^n (\Omega_j')^{a_j} \gamma_j^{-1} \exp(-a_j/\gamma_j), \quad j = 1, \ldots, 7 \\
b_j \mid \cdot &\sim Gamma(\tau_j + na_j, \beta_j + \Omega_j), \quad j = 1, \ldots, 7
\end{aligned}
$$

where Ω_j equals $\sum \phi_i, \sum \mu_{1i}, \sum \mu_{2i}, \sum \mu_{3i}, \sum \lambda_{12i}, \sum \lambda_{13i}, \sum \lambda_{23i}$ respectively for $j = 1, \ldots, 7$ and similarly Ω_j' equals $\prod \phi_i, \prod \mu_{1i}, \prod \mu_{2i}, \prod \mu_{3i}, \prod \lambda_{12i}, \prod \lambda_{13i}, \prod \lambda_{23i}$ respectively for $j = 1, \ldots, 7$.

Simulation from the Gamma conditionals is straightforward. Simulation from the posteriors for a_j is easy since the densities are log-concave (see, George *et* al., 1993) and thus the adaptive rejection sampling algorithm of Gilks and Wild (1992) can be applied. Simulation from the posterior density of δ_{ji}, $j = 1, 2, 3$ is not easy. Yet, a simple table look-up method (e.g. Devroye, 1986) suffices since in each case it can take only finite values from 0 to s. Suppose the general case where we want to simulate a random variable from a distribution with probability function

$$
P(Y = y \mid \psi, x_1, x_2) \propto \frac{\psi^y}{y!(x_1 - y)!(x_2 - y)!},
$$

where $x_1, x_2 \in \{0, 1, \ldots, \}$, $y = 0, \ldots, min(x_1, x_2)$, and $\psi > 0$. This is of the same form as our conditionals. Since the required probabilities are defined in a finite range, they can be computed via a recursive scheme. The scheme is as follows: since the calculation of the normalizing constant is not trivial, start with $P'(0) = 1$ and then use the relationship $P'(y+1) = P'(y)\frac{\psi}{y+1}(x_1 - y)(x_2 - y)$, $y = 0, \ldots, s_i - 1$. Then, rescale the probabilities in order to sum to 1 and one obtains the probabilities needed for the simulation via table look-up.

MCMC offers the opportunity of deriving the posterior distribution of any function of the parameters. In our case, the function of interest is the expected cost C_i for the i-th site. For decision purposes this cost, measured as a function of the expected accidents and fatalities and/or injuries, can have an important impact as it measures the hazard of a site taking into account all these aspects.

A simple form of this cost may be

$$C_i = \beta_1(\mu_{1i} + \lambda_{12i} + \lambda_{13i})\phi_i + \beta_2(\mu_{2i} + \lambda_{12i} + \lambda_{23i})\phi_i + \beta_3(\mu_{3i} + \lambda_{13i} + \lambda_{23i})\phi_i$$

for some coefficients β_i, $i = 1, 2, 3$ where the three parts correspond to expected cost of fatalities, slight injuries and serious injuries correspondingly. At each iteration of the chain, the values of the costs can be calculated and their posterior distributions can be obtained. The costs can then be used to rank the sites according to their expected total cost to the society.

So, if $r_i^{(j)}$ denotes the rank of the i-th site at the j-th iteration, then one can construct the posterior distributions of the ranks as well, or any posterior summary of them. In other words, if the criterion for taking corrective actions is to allocate funds to the most dangerous sites, the posterior mean ranks offer such a classification. Otherwise, if the criterion is based on whether the expected cost is above a given threshold, then the posterior distribution of the costs are of interest. In both cases, the results of the analysis can be used for decision making. Perhaps, the most important contribution of such a ranking is the fact that we take into account the uncertainty for the ranking since it is not based on deterministic criteria.

3 The data

For the purposes of this study we have used the official traffic accidents on intersections for the city of Leuven (Belgium) for the years 1991 to 1998. The intersections can be split into three groups, according to their location. The inner city is characterized by some star-shaped arterial roads, and other smaller roads, that are mostly of the same type. The ring road is a larger secondary road with some very large intersections, where the arterial roads lead out of the inner city. Smaller intersections can also be found on the ring road, typically having no traffic lights. The road network outside the ring road is quite diverse. There are some built-up areas, secondary roads to surrounding cities and approaches to and exits from the major highways.

In total, 2,323 accidents at 519 intersections were identified, with accident counts ranging from 1 to 62. Some remarks about the data should be added at this point. First, since data are available only for intersections where accidents happened, all results should be interpreted conditional on the occurrence of accidents (we will study the impact of the omission of zero accident sites on the ranking results in section 4.6). This is also the reason why no explanatory variables are used, because they would not be generally significant. Second, the model does not consider spatial correlations among intersections. In fact, one could argue that neighboring sites might have an influence on the safety between each other. Distances and geographical neighborhood should be

measured in order to take correlations into account. This complex extension is not worked out in this paper. Although these restrictions might limit somewhat the practical use of the data, it is certainly useful and instructive to illustrate the modeling approach followed in this paper.

For the purposes of this study, $t_i = 1$ for all $i = 1, \ldots, n$, since all accident sites are intersections (so there is no different segment length per site), the time periods of the data are identical and we do not have traffic flow information. The influence of different traffic flows on the results will, however, be discussed later in the next section.

4 Results

For the data set concerning Leuven, we applied the proposed methodology to both the entire data set of 519 intersections and to a smaller data set concerning the 44 intersections on the ring road. The latter intersections are the most dangerous since the speed on the ring (70 km/h) is usually much higher and thus the accidents more serious. For the sake of improving the presentation, we will use both data sets. The smaller data set will be used to illustrate the approach with respect to the quantities of interest, while both data sets will be used for elaborating the ranking procedure.

4.1 Computational Details

We ran the MCMC for 6000 iterations and used the first 1000 as a burn-in period. For the remaining 5000 iterations we sampled every 10th value to remove autocorrelation. From the autocorrelation plots, no interesting autocorrelations existed. We found that the chain converged easily and that the sampled values are indeed independent draws from the target posterior density.

For the parameters of the hyperprior we selected $\gamma_j = \tau_j = \beta_j = 1$ for $j = 1, \ldots, 7$. Note that we have used several other choices for examining the sensitivity of the approach and we did not find important differences. For this reason, we report the results based on these values.

4.2 Posterior densities for the parameters of interest

The most interesting aspect of this MCMC approach is the fact that one can obtain posterior summaries for several quantities of interest by running a simple chain. For example, Figure 1a shows the posterior distribution of ϕ_i's for all 44 sites on the ring. It is clear that sites 4, 27, 28, 30, 31 and 32 have higher accident rates than the other sites.

Figure 1 about here

More interesting conclusions can be found in Figure 1b, which depicts the posterior distribution for the parameters $\mu_2 + \lambda_{12} + \lambda_{23}$, which is the rate of slight injuries per accident for each site. We observe that the rate is relatively the same across all sites, which suggests that the slight injuries rate may be homogeneous across those intersections. The observed values are more different because they refer to different numbers of accidents.

4.3 Ranking sites using a cost function

As mentioned in the introduction, one of the strong points of the proposed methodology is the ability to rank accident sites based on a combination of criteria, i.e. the number of fatalities, serious and slight injuries for each site. However, in order to combine the information contained in those three variables, we need a cost function that assigns a cost to each variable, i.e. assigns a weight to each type of injury. We want to stress that assigning costs to different injury types is a rather controversial issue for a variety of reasons, including ethical arguments (e.g. can we assign a cost to a human life?) and economic arguments (what are the quantities that have to be measured in order to estimate the cost of a seriously injured person?). For purposes of illustration, we will use two different cost functions in terms of the weights assigned to each injury type, mainly in order to allow for a sensitivity analysis of the proposed methodology.

The first cost function was proposed by Baum and Høhnscheid (2001) and has been approved by the Organization for Economic Cooperation and Development (OECD). It measures the cost of accidents (in millions of Euro) at a particular site by the following cost function

$$C_i = E(Y_i) + 0.075E(W_i) + 0.0035E(Z_i)$$

The function is based on economic arguments and includes all the expenses related to a fatality or an injury. The difference in the weights assigned to each part of the equation is interesting. For example, under this cost setting, a fatality is weighted 14 times more expensive (and hence more important for the calculations) compared to a serious injury.

Another interesting cost function is adopted by the Flemish government in Belgium (Ministry of Transportation, 2001). This function takes the form

$$C_i = 5E(Y_i) + 3E(W_i) + E(Z_i)$$

which in fact assigns weights $(5, 3, 1)$ to deaths, serious and slight injuries respectively. This cost function is somewhat different from the previous in so much as it does not result in a total cost figure (i.e. an amount), but returns an overall score based on the scores for each injury type (i.e. a plain number). This function clearly undervalues the fatalities. The rationale for this is that, as

a result of the definition of the different injury types (see section 2), fatalities and serious injuries are more closely related than slight injuries.

Using either cost function, individual sites can be ranked. Let the vector $\boldsymbol{c}^r$ contain the costs for each site at the r-th MCMC iteration. Then, one can assign a rank to each site according to its cost value and transform the vector $\boldsymbol{c}^r$ to a vector $\boldsymbol{R}^r$ which contains the ranks for all the sites. The posterior distribution for the rank of each site can then easily be constructed. Figure 2 depicts the posterior distribution of the cost and the ranking for each site taking into account all available parameters for the two different cost functions. We have used log-scale for the costs in order to improve the quality of the graph. Note also that the costs are not comparable as they are in different scales. The posterior distributions have very large right tails. For some sites the cost is clearly greater than for other sites, e.g. for intersections 12, 31 and 32, although the results are more pronounced in the case of the Flemish cost function than the OECD cost function.

Figure 2 about here

With respect to the rankings, Figure 2 shows that, except for some sites that are ranked as dangerous in both cost functions (e.g. 4, 10, 12, 31, 32), there are a lot of sites with similar rankings. Moreover, although the variability of the results for the Flemish cost function is somewhat lower and thus enables an easier ranking of the sites, the overall variability is quite large. In other words, many sites show a similar behavior and the differences are due to random perturbations and thus of no interest. The latter is very important when investment decisions are based on the relative ranking (e.g. the mean ranking) of those sites and there is only a budget for a predetermined number of sites. In this case, there is a danger that some sites are not different in terms of their risk and that an investment decision is made on grounds of random variability. To conclude, the approach proposed offers the opportunity to examine whether there are sites that are significantly worse than others.

4.4 Sensitivity analysis of cost parameters

Both cost functions give different (absolute or relative) weight to each accident type. As a result, we have already demonstrated that the ranking of intersections will be somewhat different depending on the weight assigned to each accident type. Road safety decision makers will therefore be highly interested in the sensitivity of those rankings with respect to the parameter choices being made. Indeed, if different parameter choices result in totally different rankings, then policy makers should evaluate carefully the impact of their decisions before allocating important budgets to remedy the, say r, most dangerous intersections. In general, it sounds reasonable that the results will not

coincide perfectly. However, if it is only the first r most dangerous sites that are to be identified, we expect the methods to agree more or less on the same sites.

Figure 3 shows, for the entire data set of 519 intersections, the percentage agreement between the two approaches, i.e. the percentage of sites that appear by both approaches in the list of the r most dangerous sites, as a function of r based on their mean rank. Figure 3 shows that there is a quite heavy discrepancy between both cost functions relating to the top-10 most dangerous intersections (only 70 to 80% of the most dangerous sites are considered identical by both methods), except for the fact that both cost functions agree about the most dangerous site overall. One should be careful, however, in interpreting such a graph since the percentage of discrepancy is naturally higher for small values of r than for larger values. When r increases above 60, the parity between both functions increases again. This graph clearly shows that policy makers should be careful in selecting the right value for r. When r is set too low, dangerous sites as identified by both cost functions, will not be dealt with. In contrast, when r is set too high, some sites will be selected as dangerous although they are classified as dangerous by one cost function and not dangerous or less dangerous by another.

Figure 3 about here

4.5 Goodness of fit

In order to demonstrate the goodness-of-fit of the proposed model, the predictive distributions for the data were obtained from the MCMC output. Samples from the predictive distributions for each site, were obtained by simulating from the conditional density of the data using the current draw of parameters for each site from the Gibbs sampler. In particular, we constructed 95% credible intervals for each data point based on the sample quantiles and examined whether the observed values belonged to this interval.

Figure 4 presents the predictive intervals for the 44 sites at the ring of Leuven for the number of accidents, the number of slightly injured and the number of seriously injured. We skip the same plot for the number of fatalities as this variable takes very small values and the plot is not informative. One can see that the model predicts the observed values quite well. However, in some sense, this was expected since one parameter for each data point was used and hence there is some overfitting.

Figure 4 about here

Furthermore, in order to check the fit of models one may use a variety of predictive p-values given by $p_i = P\left[D_i(y^{rep}, \theta) > D_i(y, \theta)\right]$, where D_i is an appropriate statistic and y^{rep} are data

generated from the predictive distribution and y the observed data. In order to check the fit of the marginal distributions we have adopted χ^2 discrepancy quantities, as proposed by Gelman *et al.* (1995, p.172). Small values of the p-value reflect the implausibility of the data under the model and hence the lack of fit of the model to the data. The p-values for the total number of accidents, fatalities, severely injured and slightly injured were 0.208, 0.753, 0.452 and 0.241 respectively. Therefore the goodness of fit is satisfactory.

In addition, Table 1 shows the posterior summaries for the hyperparameters a_j, b_j. While the hyperparameters themselves do not have a simple interpretation, their ratio a_j/b_j corresponds to the mean of the parameters of the model. For example a_1/b_1 corresponds to the mean of the accidents rate and it is close to the observed value 4.47. A similar interpretation can be given for the other hyperparameters according to the model specification. One can verify that the ratios are quite close to the observed values they describe.

Table 1 about here

Another aspect with respect to model fit must also be emphasized. Usually, accident data are truncated in the sense that we have data for sites where accidents have occurred. This implies that perhaps there are some more sites where accidents have not occurred. For our case, the period covered is very long (8 years) and thus we expect that only a limited number of such sites have been omitted. In addition, if one calculates the probability of zero based on the posterior mean, the derived probability is close to 0.01, which is small. For this reason, we did not consider the truncated model, which would be much more complicated in nature. The next section examines the effect (if any) of the absence of zero-accident locations on the ranking results.

4.6 Impact on ranking of absence of zero accident sites

In order to examine the effect of removing from the data sites without any accidents, the following experiment was carried out. We inflated our data by adding 200 sites without accidents (and thus without fatalities or injuries). The aim is to examine whether the inclusion of such sites would also inflate the ranking. We adopted the ranking using the OECD costs. Now, the model was calculated again but with $n+m$ observations, where $n = 519$ the original data and $m = 200$ the inflation data. It turns out that, for the inflation data, their ranking is very different from the other sites. In fact, their mean rank is much smaller and they are ranked as a clearly separated group of non-dangerous sites. But more important, we are interested in the impact on the ranking of the non-zero accident sites, and on the ranking of the most dangerous sites in particular. After verifying the relative rank positions of the most dangerous sites, it turns out that these are not affected by the inclusion of

zero-accident sites. Thus, from a practical point of view, our approach can identify the dangerous locations even if sites with no accidents are not present in the data.

4.7 What about covariates?

In this paper, we did not include covariates in our model for two reasons. Firstly, our interest was in the ranking of the intersections. The use of covariate information would imply that we take into account the differences due to these covariates and thus the rankings would not be useful anymore. Secondly, all intersections included in the model are conditional on the fact that at least one accident occurred. Therefore, in no case we would have a balanced design to include covariate effects. Nevertheless, from a transportation point of view, we expect the mean rank of a site to be related to the type of location (type of road) where the site is situated. Indeed, it turns out that the sites located on the ring road of Leuven are the most dangerous (have the highest rank), perhaps due to the higher speed of the vehicles on the ring compared to the speed in built-up areas. Intersections situated in the suburbs or in the inner city are ranked lower and there are no interesting differences between them. In fact, it is known by traffic policy makers in Leuven that those intersections on the ring road are more dangerous, especially those where the ring intersects with some major roads connecting the city center with the suburbs outside the ring.

5 Concluding Remarks

The problem of ranking sites or identifying black spots is perhaps a difficult one. This is especially the case since accidents are rare events and thus the observed data are not necessarily a good indication, i.e. they are merely draws from an underlying density distribution. From the point of view of policy making, a solution to this problem can have a substantial impact on society, not only because it can reduce the accidents on a particular site but, at the same time, it can prevent budgets to be allocated to the wrong sites.

In the present paper, we developed a hierarchical Bayes procedure for ranking sites. The procedure takes into account not only fatalities, but also injuries (serious and slight) and combines this information by means of a cost function in order to rank the sites.

Perhaps, the most interesting insight offered by our model is that it does not only rank the sites but it also takes into account the variability of this ranking. Hence, for the purpose of decision making, one can see whether the chosen sites are really the most dangerous or if there are other sites with almost similar characteristics. Note, however, that the use of mean rank implicitly implies a

specific structure for the ranks and thus any other rank summary could also be used (see, Marden, 1995).

From the methodological point of view, the model suggested in the present paper is based on a 3-variate Poisson distribution with different covariances for each pair of variables. This approach is rather new in the literature and this model is more realistic than the common covariance model that assumes the same covariance for each pair.

Finally, we recall that the proposed model was based on the Poisson assumption. This can be limiting for some applications as it cannot adequately describe extreme events that may occur in accident analysis, as for example when an accident produces more than one fatality. This is a limitation of the proposed model, but more advanced models may be constructed in a similar manner.

Acknowledgements

This work was carried out during a visit by Dimitris Karlis to the Transportation Research Institute at Hasselt University in Diepenbeek, Belgium. Furthermore, work on this subject has been supported by a grant given by the Flemish Government to the Flemish Research Center for Traffic Safety. The authors would like to thank anonymous referees for their helpful comments that improved the paper.

References

Abdel-Aty, M.A., Radwan, A.E., 2000. Modelling traffic accident occurrence and involvement. Accident Analysis and Prevention 32(5), 633-642.

Andreassen, D.C., Hoque, M.M., 1986. Intersection accident frequencies. Traffic Engineering and Control 27(10), 514-517.

Bailey, T.C., Hewson, P.J., 2004. Simultaneous modelling of multiple traffic safety performance indicators by using a multivariate generalized linear mixed model. Journal of the Royal Statistical Society A 167(3), 501-517.

Baum, H., Høhnscheid, K.J., 2001. Measuring the road accident costs. Economic Evaluation of Road Traffic Safety Measures: Round Table No. 117, OECD Publications.

Belanger, C., 1994. Estimation of Safety of Four-legged Unsignalized Intersections. Transportation Research Record 1467, 23-29.

Christiansen, C.L., Morris, C.N., Pendleton, O.J., 1992. A Hierarchical Poisson Model with Beta Adjustments for Traffic Accident Analysis. Center for Statistical Sciences Technical Report 103, University of Texas at Austin.

Davis, G.A., Yang, S., 2001. Bayesian Identification of High-risk Intersections for Older Drivers via Gibbs Sampling. Transportation Research Record 1746, 84-89.

Douglas, J.B., 1980. Analysis with Standard Contagious Distributions. Statistical Distributions in Scientific Work Series 4. International Cooperative Publishing House, Fairland, Maryland USA.

Devroye, L. (1986). *Non-Uniform Random Variate Generation.* New-York: Springer-Verlag

Flahaut, B., Mouchart, M., San Martin, E., Thomas, I., 2003. The local spatial autocorrelation and the kernel method for identifying black zones: a comparative approach. Accident Analysis and Prevention, 35 (6) 991-1004.

George, E.I., Makov, U.E., and Smith, A.F.M., 1993. Conjugate likelihood distributions. Scandinavian Journal of Statistics, 20, 147-156.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis.* First edition. Chapman and Hall.

Geurts, K., Wets, G., 2003. Black Spot Analysis Methods: Literature Review. Doc.nr. RA-2003-07. Flemish Research Center for Traffic Safety, Diepenbeek, Belgium.

Gilks, W.R. and Wild, P. 1992. Adaptive rejection sampling for Gibbs sampling. Journal of the Royal Statistical Society, Series C, 41, 337-348.

Goldstein H., Spiegelhalter, D.J., 1996. League tables and their limitations: Statistical Issues in comparisons of institutional performance (with discussion). Journal of the Royal Statistical Society A 159, 385-443.

Hauer, E., 1986. On the Estimation of the Expected Number of Accidents. Accident Analysis and Prevention 18(1), 1-12.

Hauer, E., 1997. Observational before-after studies in road safety, Pergamon, Oxford.

Hauer, E., Persaud, B., 1987. How to estimate the safety of rail-highway grade crossing and the effects of warning devices. Transportation Research Record 1114, 131-140.

Joe, H. 1997. Multivariate Models and Dependence Concepts. Chapman and Hall, London.

Johnson, N., Kotz, S., Balakrishnan, N., 1997. Discrete Multivariate Distributions, Wiley, New York.

Jones, A.P., Jørgensen, S.H., 2003. The use of multilevel models for the prediction of road accident outcomes. Accident Analysis and Prevention 35, 59-69.

Karlis, D., 2003. An EM algorithm for multivariate Poisson distribution and related models. Journal of Applied Statistics 30, 63-77.

Karlis, D., Meligkotsidou, L., 2005. Multivariate Poisson Regression with Full Covariance Structure. Statistics and Computing, 15, 255-265.

Kemp, C.D., 1973. An elementary ambiguity in accident theory. Accident Analysis and Prevention 5(4), 371-373.

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis and Prevention 37(1), 35-46.

MacNab, Y.C., 2003. A Bayesian hierarchical model for accident and injury surveillance. Accident Analysis and Prevention 35(1), 91-102.

Marden, J.I., 1995 *Analyzing and Modelling Rank Data*, Chapman and Hall.

Miaou, S.-P., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. Accident Analysis and Prevention 37(4), 699-720.

Miaou, S.-P, Song, J.J., Mallick, B.K., 2003. Roadway traffic crash mapping: a space-time modeling approach. Journal of Transportation and Statistics 6(1), 33-57.

Ministry of Transportation, 2001. Design mobility plan Flanders, Belgium, available at http://viwc.lin.vlaanderen.be/mobiliteit .

Persaud, B., 1990. Black spot identification and treatment evaluation. The Research and Development Branch, Ontario, Ministry of Transportation.

Schlüter, 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, 293-316.

Thomas, I., 1996. Spatial data aggregation: exploratory analysis of road accidents. Accident Analysis and Prevention 28, 251-264.

Tsionas, E.G., 1999. Bayesian analysis of the multivariate Poisson distribution. Communications in Statistics - Theory and Methods 28, 431-451.

Tunaru, R., 2002. Hierarchical Bayesian Models for Multiple Count Data. Austrian Journal of Statistics 31, 221-229.

WHO (2004). World report on road traffic injury prevention. Available at: http://www.who.int/world-health-day/2004/infomaterials/world_report/en/index.html

	a_1	a_2	a_3	a_4	a_5	a_6	a_7
mean	1.175	0.193	22.500	0.771	0.002	0.006	0.007
st.dev.	0.084	0.121	4.177	0.233	0.010	0.018	0.024

	b_1	b_2	b_3	b_4	b_5	b_6	b_7
mean	0.149	3.983	1.868	2.615	3.096	3.468	3.439
st.dev.	0.022	15.865	3.489	6.839	9.586	12.024	11.827

	a_1/b_1	a_2/b_2	a_3/b_3	a_4/b_4	a_5/b_5	a_6/b_6	a_7/b_7
mean	4.478	0.007	1.194	0.035	0.001	0.003	0.001
st.dev.	0.215	0.002	0.028	0.005	0.024	0.069	0.001

Table 1: Posterior summaries for the hyper-parameters a_j, b_j and their ratio a_j/b_j for $j = 1, \ldots, 7$

Figure 1: Boxplots of the posterior densities for ϕ_i's and the rate of slight injuries per accident for all 44 sites on the ring of Leuven

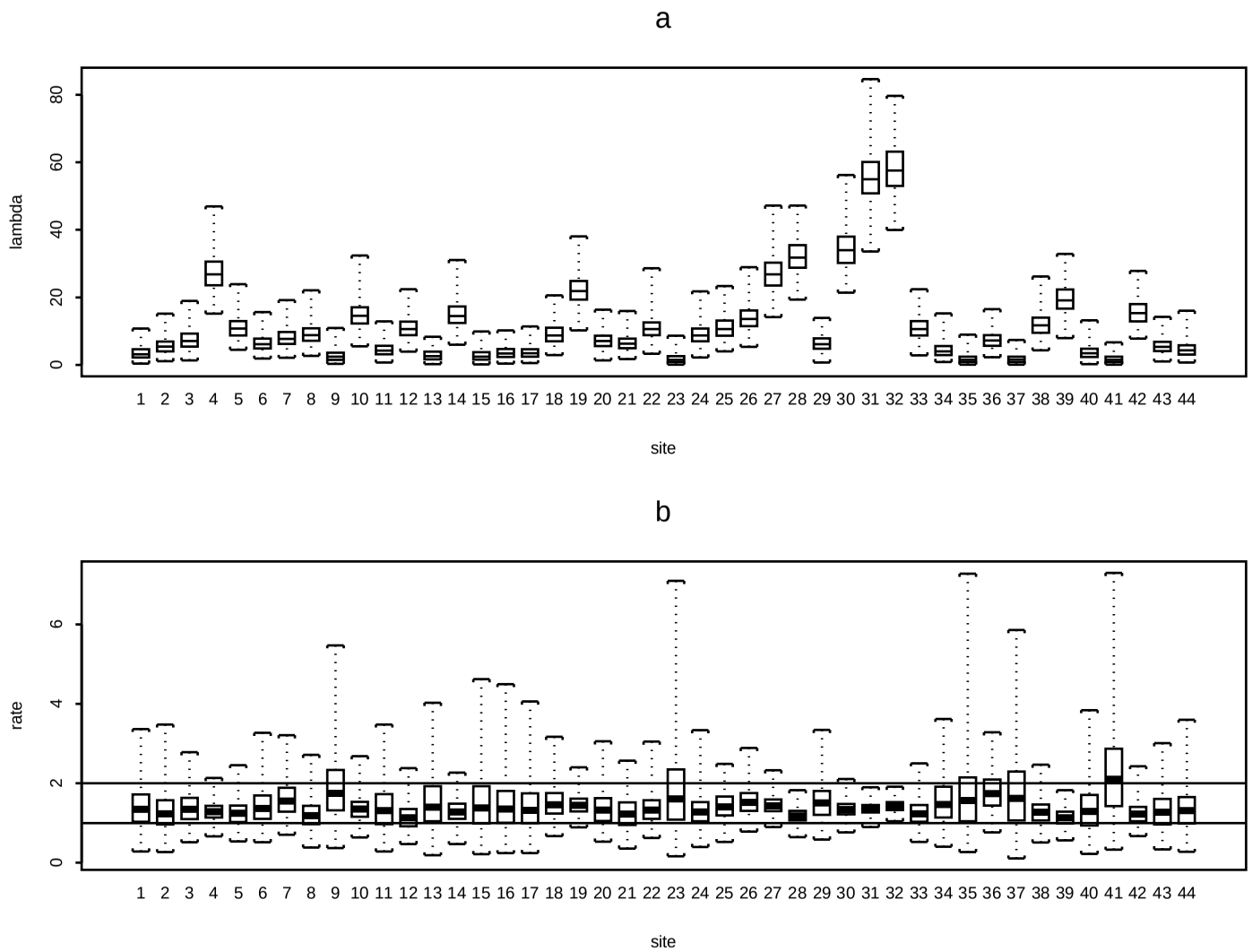

Figure 2: Posterior distributions for the log(cost) (1st row) and the ranks (2nd row) using both cost functions

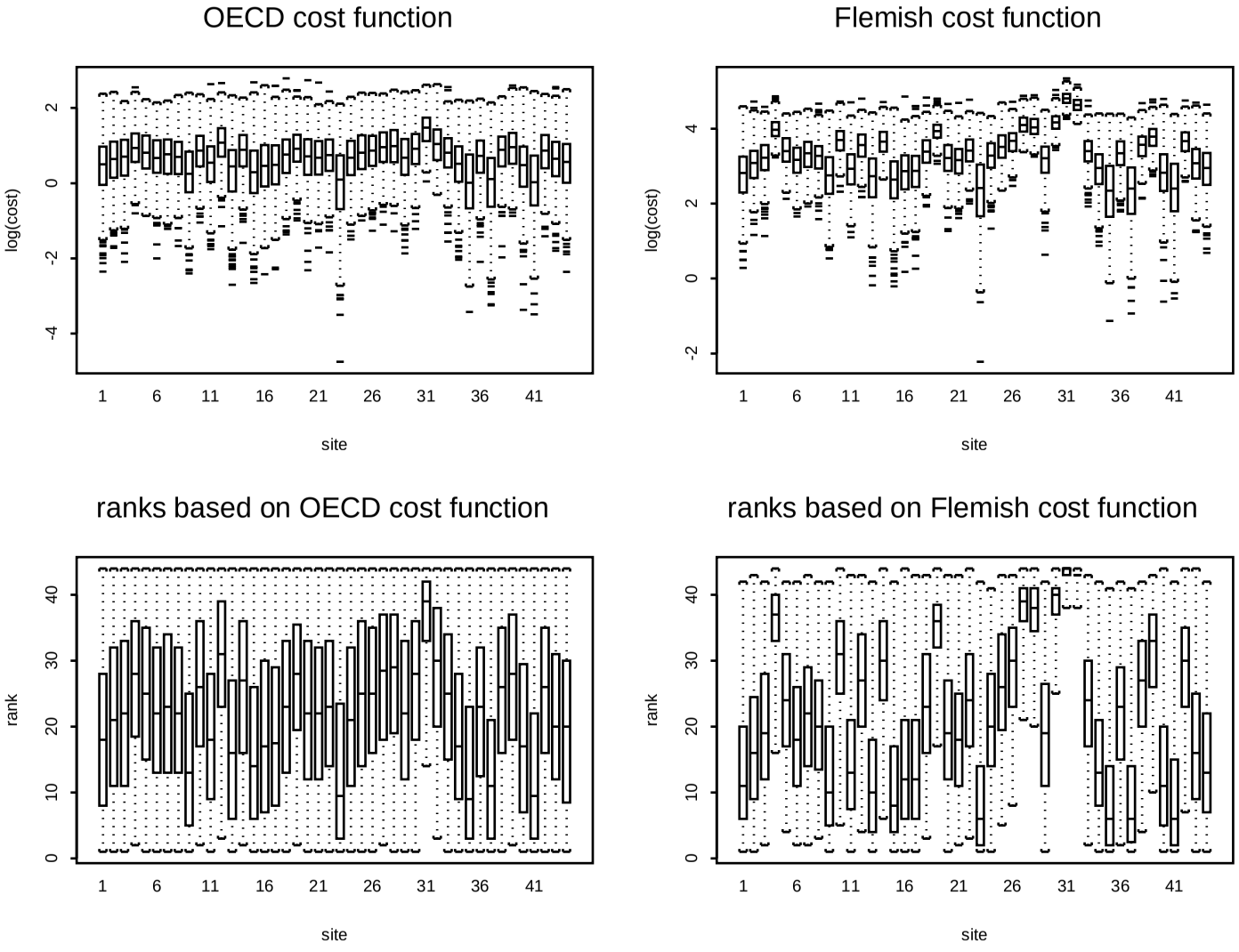

Figure 3: Percentage of agreement for ranking the r most hazard sites using the two different cost functions based on mean ranks

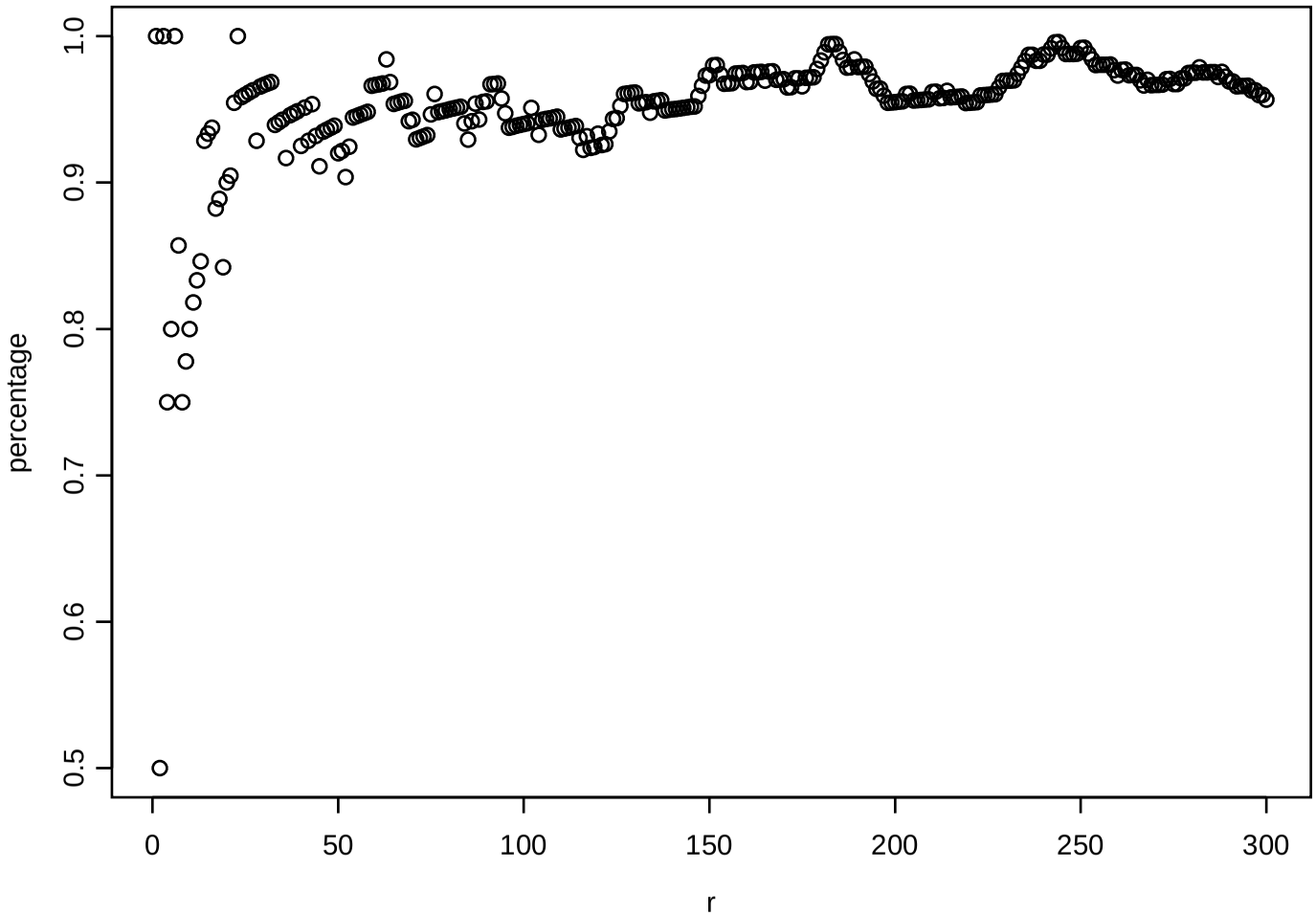

Figure 4: 95% predictive intervals based on the MCMC replications. The points represent the observed values