A Bayesian Model for Ranking Hazardous Sites

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

[†] Transportation Research Institute
Limburgs Universitair Centrum
Universitaire Campus, gebouw D
B-3590 Diepenbeek, BELGIUM
email: tom.brijs, filip.vandenbossche, geert.wets@luc.ac.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 determination of sites that are more dangerous than others (black spots) can help in better scheduling road safety policies. The present paper proposes a methodology to rank sites according to their hazardousness. 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, the number of fatalities, as well as the number of both light and severe injuries. Secondly, the model includes the use of a cost function to rank the sites with respect to their total expected cost to the 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 sensitivity of the proposed method on the use of different costs per injury type.

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

[1]Corresponding author

1 Introduction

During recent years, road safety has become a major concern for many governments. Indeed, for most European countries, road accidents constitute a large problem and cost to the society. In the first place, there are the non-material costs associated with road accidents, including the pain, the suffer, the reduced joy of life, and the personal damage in so far it does not affect the wealth but rather the welfare of the victim (Lindenbergh, 1998). Secondly, there are the material costs associated with road accidents, including direct and indirect costs. The direct costs are related to the accident itself, such as administrative costs (e.g. police and emergency services), material damage (e.g. damage to cars, road infrastructure, buildings, etc.), medical costs (e.g. hospital, rehabilitation, prothesis, etc.), and costs related to resulting traffic jams. The indirect costs are caused by the fact that the victim is not able to participate in the economic life for some period, i.e. either temporarily (due to illness) or definite (when the victim has died). In Belgium, this total cost to the society of traffic accidents is estimated at 3.72 billion Euros per year (Dielemann, 2000).

Social interest therefore lies mainly in preventing traffic accidents. However, this is not at all an easy task. In fact, it is well-known that a traffic accident is usually caused by the failure of one or more of a multitude of factors, including the safety condition of the vehicle, the safety condition of the road (and its environment) and finally the safe behavior of the driver (Haddon, 1970). Reducing the number of traffic accidents therefore 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 sensitizing and enforcing car drivers to be more careful and by reducing the hazardousness of roads. The latter involves identifying sites with large accident risk so as to make the necessary infrastructure changes for reducing the risk of the site. Furthermore, methods that can measure and produce comparable results concerning the risk of each site are of special interest for designing new roads or to enforce rules. Such rules imply the existence of criteria that assess that a specific site is hazardous. Such criteria can be comparative, i.e. to find the r, say, most hazardous roads, or they might be based on threshold values and hence all the roads passing the threshold are to be considered for changes. In practice, these criteria can be combined using relative information about the cost of such repairs. But the main goal remains evident, i.e., the need for quantifying the risk of specific sites.

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 being used. Treating black spots is a well-known and frequently used means of improving road safety. In this study, we will focus on intersections, which are classified as black spots after an assessment of the level of risk, both in terms of the number *and* the gravity of the accidents. At some intersections, risk will be higher than what one would expect for a similar location. 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, almost by definition, as random events. In fact, they are indeed the unintentional result of human behavior (OECD, 1997). 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, one can find a high variety in statistical models in the literature for analyzing black spot data, but compelling arguments can be found to support the assumption that accident counts follow the Poisson probability law. In this context, 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 Hauer and Persaud (1987), Hauer (1996), Nassar (1996) and Geurts and Wets (2003).

More recently, Bayesian techniques have been used to tackle problems in traffic safety. Although the problem of hazardous intersection identification has been widely discussed in literature, the interest in Bayesian methods in this domain only originated in the eighties. Ever since, many applications used in some way an Empirical Bayes approach. 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. Hauer (1996) reviewed the development of procedures to identify hazardous locations in general. Vogelesang (1996) gives a comprehensive overview of Empirical Bayes methods in road safety research.

However, the use of *hierarchical* Bayesian 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 accident frequencies, a hierarchical Bayesian Poisson model has been used. Christiansen et al. (1992) developed a hierarchical Bayesian 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. Point and interval estimates of the relative crash risk for older drivers were obtained using the Gibbs sampler.

In this paper, we will argue that when decisions have to be taken so as to spend money for improving the quality of particular sites, it would be interesting to find a method which can examine the risk of the sites in a comparative way and to find the sites with higher risk. Problems that occur

to this direction are due to the different observational period for different sites and to the different length of the examined roads. Moreover, data concerning the traffic of each site are needed so as to make fair comparisons[2]. Statistical methods must account for the sources of this variability. In this context, ranking procedures based on a hierarchical Bayesian approach have been proposed. Those methods can handle the uncertainty and the great variability of the data and produce a probabilistic ranking of those sites. The approach has been applied to ranking problems in various application domains, like 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 an hierarchical Bayesian approach for ranking accidents sites based on a bivariate Poisson-lognormal distribution.

We extend this approach by considering a more realistic model for the accident behavior taking into account (1) the number of accidents, (2) the number of fatalities, and (3) the number of light and severely injured casualties for a given time period for each site. This is done by using a 3-variate Poisson distribution which allows for covariance between the variables. The parameters of the model are estimated via Bayesian estimation facilitated by Markov Chain Monte Carlo (MCMC) methods.

In order to combine all the data into a single number that will 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, heavy and light injured casualties. However, we want to point out that it is not the objective of this paper to propose optimal values for the costs of each type of casualty. Indeed, since there are ethical problems on defining such cost function, the methodology will be provided using a general function. However, for the purpose of illustration, we will use two widely different cost functions, one proposed by Baum and Hohnscheid (2001) and approved by OECD, as well as another that is adopted by the Flemish government (see Ministry of Transportation, 2001).

The remaining 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 thoroughly the results. Finally concluding remarks can be found in section 5.

2 The Model

Suppose that the data consist of n different sites. The number of accidents for the i-th site is denoted by X_i, while the i-th site has been monitored for a time period t_i. We assume that the number of accidents for this site follows a Poisson distribution with parameter $\lambda_i t_i$. Note that according to this definition, t_i is not necessarily the time but it can also incorporate different lengths for the sites and/or different traffic flows. In any case, it is an offset that makes the different sites comparable by cancelling out all other information that may lead to differences. Thus λ_i's, $i = 1, \ldots, n$ are the

[2]Note that traffic flows for each site are important in so far that the focus is on estimating the relative risk of different sites. However, when the focus is on saving the maximum number of lives, then traffic flows are not needed

pure accident rates for the n sites per unit of time, length or traffic intensity.

For each site, we have also the triplets (Y_i, Z_i, W_i) that correspond to the number of fatalities, the number of lightly injured persons and the number of severely injured persons, respectively. 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, instead of the usual model that assumes the same covariance term for all the pairs and has been examined in Tsionas (1999) and Karlis (2003). Our model differs from the model of Johnson et al. (1997), which assumes more (but unrealistic) structure. Derivation details for our model can be found in the appendix.

In the sequel, we call as 3-variate Poisson distribution the joint probability function given by

$$P(y_1, y_2, y_3) = \sum_{k=0}^{s_1} \sum_{r=0}^{s_2} \sum_{s=0}^{s_3} \frac{e^{-\theta_{12}}\theta_{12}^k}{k!} \frac{e^{-\theta_{13}}\theta_{13}^r}{r!} \frac{e^{-\theta_{23}}\theta_{23}^s}{s!} \frac{e^{-\theta_1}\theta_1^{y_1-k-r}}{(y_1-k-r)!} \frac{e^{-\theta_2}\theta_2^{y_2-k-s}}{(y_2-k-s)!} \frac{e^{-\theta_3}\theta_3^{y_3-r-s}}{(y_3-r-s)!}$$

where $s_1 = min(y_1, y_2), s_2 = min(y_1 - k, y_3), s_3 = min(y_2 - k, y_3 - r)$. The above distribution will be denoted as $3 - Poisson(\theta_1, \theta_2, \theta_3, \theta_{12}, \theta_{13}, \theta_{23})$. It can be seen (see Appendix for details) that 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 injuries. To our knowledge, the above distribution has not been used in any application. One can define analogously multivariate Poisson distributions, for details see the Appendix.

For our application, we assume that

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

Hence, $\mu_{\cdot i}$ reflects the rate for fatalities, light injuries and severe injuries per accident for the site i, while λ_{ij} are the covariance parameters for each pair of variables.

Note that empirical evidence supports the assumption that there is positive correlation between the three variables Y_i, Z_i, W_i. 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.

Since we have assumed site specific rates for all the variables of interest, it is not easy to proceed with classical estimation methods, as for example with the maximum likelihood method. In order to avoid this overparametrization problem, we will proceed from the Bayesian perspective, which is

the typical procedure for this kind of data. In fact, we will describe an Empirical Bayes approach where the prior parameters will be specified by the data.

2.1 Bayesian Approach

Our model has the form

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

The likelihood can be written in the complicated form

$$\begin{aligned} L(X, Y, Z, W \mid \lambda, \mu_1, \mu_2, \mu_3, \rho) &= \prod_{i=1}^{n} P(y_i, z_i, w_i | x_i) P(x_i) \\ &= \prod_{i=1}^{n} \frac{e^{-\lambda_i t_i}(\lambda_i t_i)^{x_i}}{x_i!} \sum_{k=0}^{s_1} \sum_{r=0}^{s_2} \sum_{s=0}^{s_3} \frac{e^{-\lambda_{12i}x_i}(\lambda_{12i}x_i)^k}{k!} \times \\ &\quad \frac{e^{-\lambda_{13i}x_i}(\lambda_{13i}x_i)^r}{r!} \frac{e^{-\lambda_{23i}x_i}(\lambda_{23i}x_i)^s}{s!} \times \\ &\quad \frac{e^{-\mu_{1i}x_i}(\mu_{1i}x_i)^{y_i-k-r}}{(y_i-k-r)!} \frac{e^{-\mu_{2i}x_i}(\mu_{2i}x_i)^{z_i-k-s}}{(z_i-k-s)!} \frac{e^{-\mu_{3i}x_i}(\mu_{3i}x_i)^{w_i-r-s}}{(w_i-r-s)!} \end{aligned}$$

where $s_1 = min(y_i, z_i), s_2 = min(y_i - k, w_i), s_3 = min(z_i - k, w_i - r)$.

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 with data augmentation will be used in order to explore the posterior distribution of the parameters of interest. A byproduct of this approach is that we can obtain at the same time the posterior distribution of every summary function of the parameters, including ranks. This is exactly the key ingredient of our approach as it enables ranking the sites according to some criteria and/or calculation of the posterior distribution of any cost function.

The vector of parameters can be represented as $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{\lambda}_{12}, \boldsymbol{\lambda}_{13}, \boldsymbol{\lambda}_{23})$, where the vectors represented by boldface letters represent the corresponding parameters for all the observations, i.e. $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$ and similarly for the other vectors.

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. The choice of prior parameters can be based on either diffuse Gamma densities or an Empirical Bayes approach (described in section 2.3).

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

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

$i = 1, \ldots, n$ for all parameters.

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

$$
\begin{aligned}
p(\boldsymbol{\theta} \mid \boldsymbol{X}) &\propto L(X, Y, Z, W \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\
&= \prod_{i=1}^{n} \frac{e^{-\lambda_i t_i} (\lambda_i t_i)^{x_i}}{x_i!} \sum_{k=0}^{s_1} \sum_{r=0}^{s_2} \sum_{s=0}^{s_3} \frac{e^{-\lambda_{12i} x_i} (\lambda_{12i} x_i)^k}{k!} \frac{e^{-\lambda_{13i} x_i} (\lambda_{13i} x_i)^r}{r!} \frac{e^{-\lambda_{23i} x_i} (\lambda_{23i} x_i)^s}{s!} \times \\
&\quad \frac{e^{-\mu_{1i} x_i} (\mu_{1i} x_i)^{y_i - k - r}}{(y_i - k - r)!} \frac{e^{-\mu_{2i} x_i} (\mu_{2i} x_i)^{z_i - k - s}}{(z_i - k - s)!} \frac{e^{-\mu_{3i} x_i} (\mu_{3i} x_i)^{w_i - r - s}}{(w_i - r - s)!} \times \\
&\quad [\Gamma(a_1)]^{-1} \lambda_i^{a_1 - 1} {b_1}^{a_1} exp(-b_1 \lambda_i) [\Gamma(a_2)]^{-1} \mu_{1i}^{a_2 - 1} {b_2}^{a_2} exp(-b_2 \mu_{1i}) \times \\
&\quad [\Gamma(a_3)]^{-1} \mu_{2i}^{a_3 - 1} {b_3}^{a_3} exp(-b_3 \mu_{2i}) [\Gamma(a_4)]^{-1} \mu_{3i}^{a_4 - 1} {b_4}^{a_4} exp(-b_4 \mu_{3i}) \times \\
&\quad [\Gamma(a_5)]^{-1} \lambda_{12i}^{a_5 - 1} {b_5}^{a_5} exp(-b_5 \lambda_{12i}) [\Gamma(a_6)]^{-1} \lambda_{13i}^{a_6 - 1} {b_6}^{a_6} exp(-b_6 \lambda_{13i}) \times \\
&\quad [\Gamma(a_7)]^{-1} \lambda_{23i}^{a_7 - 1} {b_7}^{a_7} exp(-b_7 \lambda_{23i})
\end{aligned}
$$

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 useful form. The unconditional joint density of (Y_i, Z_i, W_i) is not recognizable to belong to any of the known 3-variate discrete distributions. However, marginally, each of the (Y_i, Z_i, W_i) will have a univariate Neyman distribution (see Douglas, 1980). The marginal predictive distribution will be a mixture of Neyman univariate distributions, which again is of unknown form. We omit the details as they are not useful for our scope.

2.2 MCMC details

The key ingredient for constructing the MCMC approach is the data augmentation offered by the multivariate reduction approach that is used to construct the multivariate Poisson distribution. We will make use of the following representation of a multivariate Poisson distribution, known as multivariate reduction (see e.g. Johnson et al., 1997, details in the Appendix).

We start from a series of independent Poisson variables $X_1, \ldots, X_6$ each one following independently a Poisson distribution, i.e, $X_i \sim Poisson(\theta_i), \quad i = 1, \ldots, 6$ and then we create the new variables

$$\begin{aligned} Y_1 &= X_1 + X_4 + X_5, \\ Y_2 &= X_2 + X_4 + X_6, \\ Y_3 &= X_3 + X_5 + X_6 \end{aligned}$$

One can see that X_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 X_5 and X_6. Thus, θ_4 is the covariance parameter between Y_1 and Y_2 and so on. According to the above model, we may have only positive covariances. However, for count data negative covariances are rather rare.

In our model, the above idea assumes that there are some latent variables $\delta_{1i}, \delta_{2i}, \delta_{3i}, T_{1i}, T_{2i}, T_{3i}$ 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. A similar data augmentation has been used by Karlis and Meligkotsidou (2003) for a multivariate Poisson model including regressors.

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

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\kappa} \mid data) &= \prod_{i=1}^{n} \frac{e^{-\lambda_i t_i}(\lambda_i t_i)^{x_i}}{x_i!} \frac{e^{-\lambda_{12i} x_i}(\lambda_{12i} x_i)^{\delta_{1i}}}{\delta_{1i}!} \frac{e^{-\lambda_{13i} x_i}(\lambda_{13i} x_i)^{\delta_{2i}}}{\delta_{2i}!} \frac{e^{-\lambda_{23i} x_i}(\lambda_{23i} x_i)^{\delta_{3i}}}{\delta_{3i}!} \times \\ & \frac{e^{-\mu_{1i} x_i}(\mu_{1i} x_i)^{y_i - \delta_{1i} - \delta_{2i}}}{(y_i - \delta_{1i} - \delta_{2i})!} \frac{e^{-\mu_{2i} x_i}(\mu_{2i} x_i)^{z_i - \delta_{1i} - \delta_{3i}}}{(z_i - \delta_{1i} - \delta_{3i})!} \frac{e^{-\mu_{3i} x_i}(\mu_{3i} x_i)^{w_i - \delta_{2i} - \delta_{3i}}}{(w_i - \delta_{2i} - \delta_{3i})!} \times \\ & [\Gamma(a_1)]^{-1} \lambda_i^{a_1 - 1} {b_1}^{a_1} exp(-b_1 \lambda_i) [\Gamma(a_2)]^{-1} \mu_{1i}^{a_2 - 1} {b_2}^{a_2} exp(-b_2 \mu_{1i}) \times \\ & [\Gamma(a_3)]^{-1} \mu_{2i}^{a_3 - 1} {b_3}^{a_3} exp(-b_3 \mu_{2i}) [\Gamma(a_4)]^{-1} \mu_{3i}^{a_4 - 1} {b_4}^{a_4} exp(-b_4 \mu_{3i}) \times \\ & [\Gamma(a_5)]^{-1} \lambda_{12i}^{a_5 - 1} {b_5}^{a_5} exp(-b_5 \lambda_{12i}) [\Gamma(a_6)]^{-1} \lambda_{13i}^{a_6 - 1} {b_6}^{a_6} exp(-b_6 \lambda_{13i}) \times \\ & [\Gamma(a_7)]^{-1} \lambda_{23i}^{a_7 - 1} {b_7}^{a_7} exp(-b_7 \lambda_{23i}) \end{aligned}$$

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

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

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

Simulation from the Gamma conditionals is straightforward, however, simulation from the posterior density of δ_{ji}, $j = 1, 2, 3$ is not easy. Yet, a simple table look-up method suffices since in each case δ_{ji} 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

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

$x_1, x_2 \in \{0, 1, \ldots\}$, $y = 0, \ldots, min(x_1, x_2)$, $\psi > 0$. This is of the same form as our conditionals. Since the required probabilities are 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 $P'(k+1) = P'(k)\frac{\rho}{k+1}(x_1 - k)(x_2 - k)$, $k = 0, \ldots, s_i - 1$. Then, rescale the probabilities in order to sum to 1 and one obtains the conditional probabilities needed for the simulation via table look-up.

The choice of the hyperparameters a_j, b_j, $j = 1, \ldots, 7$ can be either diffuse priors in order to reflect our ignorance or they can be obtained in an Empirical Bayes way from the data. For practical reasons, it is advocated to use informative priors for λ_{ji}, because diffuse priors can have serious effects on the convergence properties of the chain. Especially, for small counts the chain can be trapped in 0 values for the pseudoparameters λ_{ji}.

One can see that the Bayesian estimation is split in two parts. The first part, involves only estimation of Poisson parameters and this can be easily accomplished via standard conjugate analysis. The second part, involves Bayesian estimation for a multivariate Poisson distribution.

MCMC offers the opportunity to derive the posterior distribution of any function of the parameters. For 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 a large impact as it measures the hazard of a site taking into account all these aspects.

A simple form of this cost can be

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

for some coefficients β_i, $i = 1, 2, 3$ where the three parts corresponds to expected cost of fatalities, light injuries and severe 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 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. Thus, it allows for comparing different sites taking into account the randomness in collecting and reporting the data.

2.3 Empirical Bayes

Proceeding in an Empirical Bayes spirit, the parameters of the prior distribution have to be obtained from the data. Using the derivation of the model, we can proceed by using the results concerning Empirical Bayes estimation for a simple Poisson model derived by Gaver and O'Muircheartaigh (1987). Alternatively, one can use the joint probability function to derive moment based Empirical Bayes estimates for the prior distributions.

Since we have 14 prior parameters, we need 14 equations from the data in order to get those prior values via moment matching. In order both to simplify the problem but, at the same time, to use some prior information useful for implementation reasons, we elicited only 11 prior parameters from the data while we put $b_5 = b_6 = b_7 = 1$. Those parameters affect the prior of the covariance terms and thus, in order to avoid trapping the chain in zero values while we generate the posterior distributions of δ_i, we use a rather small variance for the priors associated with the covariance parameters.

For the remaining parameters, we used the marginal means and variances (8 equations) as well as the three covariances. This system is sufficient to provide values for all the prior parameters. Note that values for the other three prior parameters $b_5 = b_6 = b_7$ can be elicited in this way, but we believe that this is an unnecessary complication. Note also that in cases where the system of 11 equations does not have solutions in the admissible range (the prior parameters as being parameters of the Gamma density ought to be positive) we tried to satisfy the mean relationships so as the priors to have the correct means.

3 The data

For this study, the official Belgian traffic accident records for the city of Leuven (Belgium) have been used. The Federal Government organizes the systematic registration of accidents involving

injuries. The objective is the gathering of interesting data that allows the analysis of accidents to get better insight into traffic safety problems on both the federal and regional level. Data is systematically registered for each traffic accident involving fatalities or injured victims on public roads in Belgium [3].

The motivations to investigate accidents at intersections for a small university city are twofold. First of all, since Leuven is a university town, many students live and travel in the inner city. The manifest presence of this subgroup results in very specific traffic patterns. Second, Leuven is not only a university city, but also a center of economic activity that brings along an ever-increasing stream of commuter traffic. For these reasons, the local government is working on a mobility plan. The objective is to extend basic mobility and to preserve and strengthen the accessibility for all means of transport. At the same time, car traffic should be controlled and the safety level should increase, together with the quality of life and environment. To achieve these targets, a safe infrastructure should be provided. In fact, a selection of dangerous road sections will be made and investments will be done according to a priority list (for details, see "Mobiliteitsplan voor de Stad Leuven" (in Dutch), 2002, *http://www.leuven.be*).

This study is based on the data set of traffic accidents for the years 1991 to 1998 on intersections in the city of Leuven. The intersections can be split into three groups, according to their localization. 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 are leaving the inner city. Also smaller intersections are to 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, 2323 accidents at 519 intersections were identified, with accident counts ranging from 1 to 62. For each intersection, the number of accidents in the given period X_i is counted. Furthermore, a distinction is made according to the gravity of the accidents. For each intersection 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 heavily injured persons, being every road user who got injured in a crash 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 to whom the specification of fatally or heavily injured road user does not apply, is counted in the third group of light injuries, denoted by Z_i.

Some remarks about the data set should be added, however. First, since data is available only for intersections where accidents happened, all results should be interpreted conditional on the

[3]It should be mentioned, however, that underreporting may affect the results. This is caused by the fact that some people involved in an accident fail to call the police and by the fact that not all accident forms are sent to the National Institute of Statistics. The latter is especially true for accidents involving only one road user, light accidents and accidents involving weak road users, like pedestrians.

occurrence of accidents. This is also the reason why no explanatory variables are used, because they would not be generally significant. Second, abstraction is made of the order of the accidents over the years. Third, 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 set, it is certainly useful and instructive to illustrate the modelling approach followed in this paper.

For 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 the same and we do not possess traffic flow information. The influence of different traffic flows on the results will be discussed later in the next section.

4 Results

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

4.1 Computational Details

A first problem in the data was the fact that counts related to fatalities and severe injuries were rather small. The variance to the mean ratio was slightly smaller than 1 indicating that their marginal distribution is not overdispersed relative to the Poisson distribution. This caused a problem in moment matching for deriving the prior parameters. In order to proceed in such cases, we set the overdispersion parameter of the Gamma prior equal to 1, i.e. $b_i = 1$, $i = 2, 3, 4, 5, 6, 7$. The other prior parameters for the ring dataset are the Empirical Bayes estimates $a_1 = 0.856$, $a_2 = 0.00245$, $a_3 = 1.181$, $a_4 = 0.02687$, $a_5 = 0.1135$, $a_6 = 0.01537$, $a_7 = 1.0087$ and $b_1 = 0.07042359$.

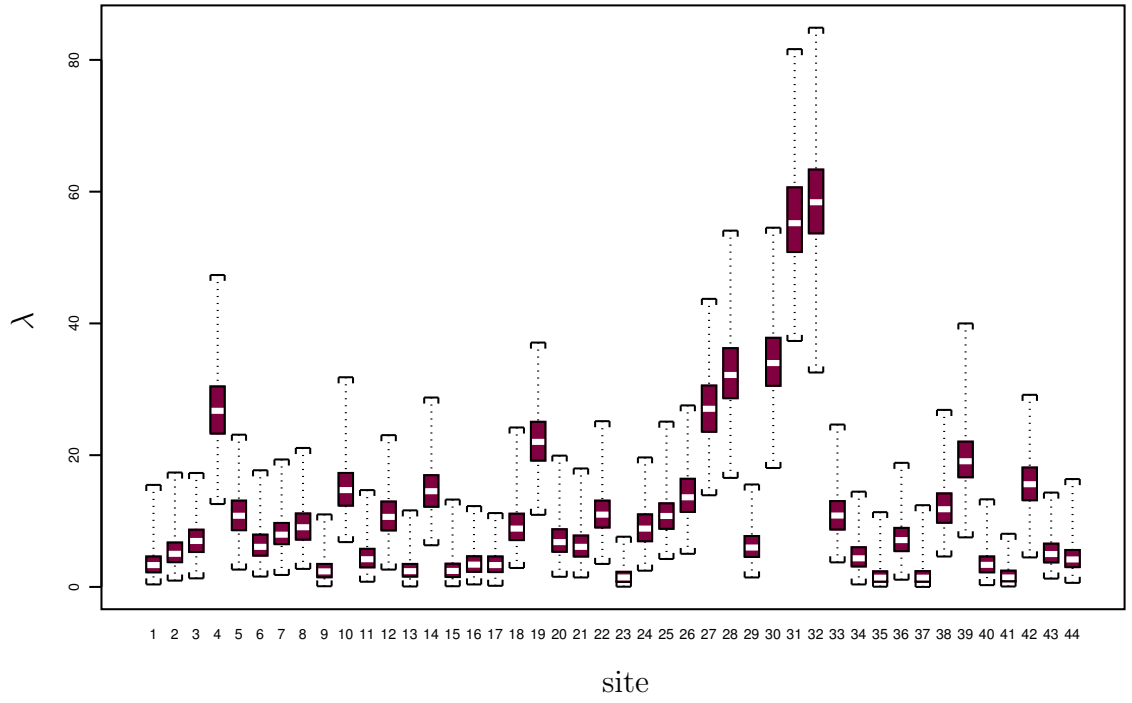
We run the MCMC for a burn-in period of 1000 iterations and then we sampled every 10th value. From the autocorrelation plots, no interesting autocorrelations existed. We found that the chain converged easily and that the sampled values are indeed independent draws from the target posterior density [4].

[4]Details about convergence properties are available on request, but we omit them to save space.

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, in figure 1 one can see the posterior distributions of λ_i's (i.e. the mean number of accidents) for all 44 sites on the ring. From the figure, it is clear that there are sites (4, 27, 28, 30, 31 and 32) with much higher accident rates than others.

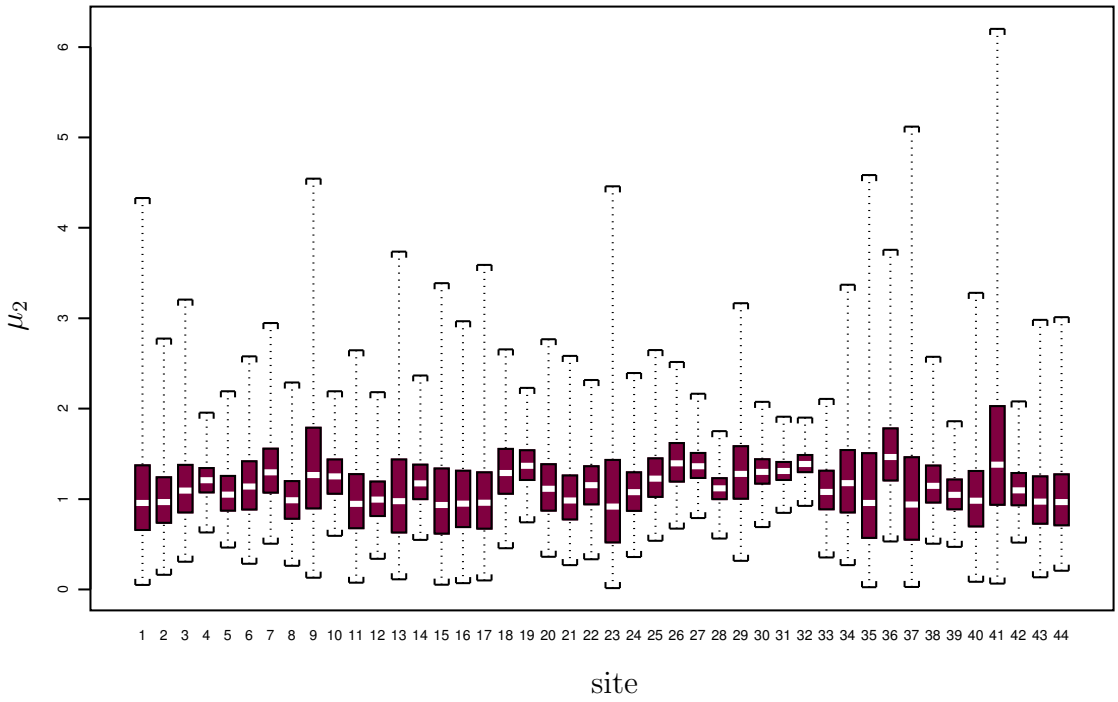
Figure 1: Boxplots of the posterior densities for λ for all 44 sites on the ring of Leuven



More interesting conclusions can be found in Figure 2. In fact, this figure depicts the posterior distribution for the parameters μ_2, which is the rate of light injuries per accident for each site. One can observe that the rate is relatively the same across all sites, which implies that the light injuries rate is homogeneous across those intersections. Of course, the observed values are much different as they refer to different number of accidents.

One of the advantages of MCMC is that we can obtain posterior distributions for any function of the parameters. So, in figure 3 one can see the posterior distribution of μ_2/μ_3, i.e. the ratio of light injuries versus severe injuries. In fact, we have plotted the numbers in logarithmic scale. It is interesting to see that the majority of the sites have a homogeneous behavior, but that there are some sites for which the ratio is much smaller. This implies that the ratios of light to severe injuries are smaller and thus those differences reflect differences on the kind of accidents at those sites. For example, the fact that the severe injuries are higher may imply more severe collisions on

Figure 2: Boxplots of the posterior densities for μ_2 for all 44 sites on the ring of Leuven



those intersections.

4.3 Ranking sites using a cost function

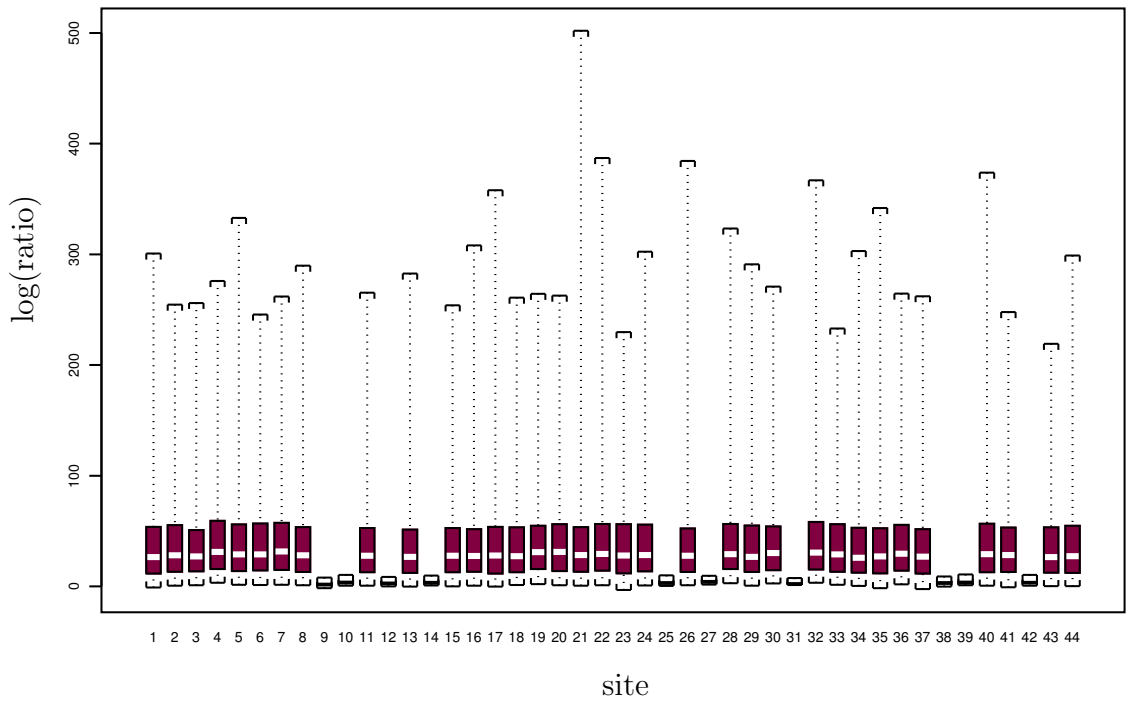
As mentioned in the introduction, one of the strong points of the proposed methodology is the ability to rank accident sites based on a combination of criteria, i.e. the number of fatalities, heavy and light injuries for each site, instead of using only one of them. However, in order to combine the information contained in those three variables, we need a cost function that in some sense assigns a cost to each variable, i.e. assigns a weight to each type of injury.

Once again, 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?) or economic arguments (what are the quantities that have to be measured in order to estimate the cost for a severely injured person?). For illustrative purposes, 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 (2001) and has been approved by the Organization for Economic Cooperation and Development (OECD). It measures the cost of accidents in a particular site by the following cost function

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

Figure 3: Boxplots of the posterior densities for μ_2/μ_3 for all 44 sites on the ring of Leuven in logscale

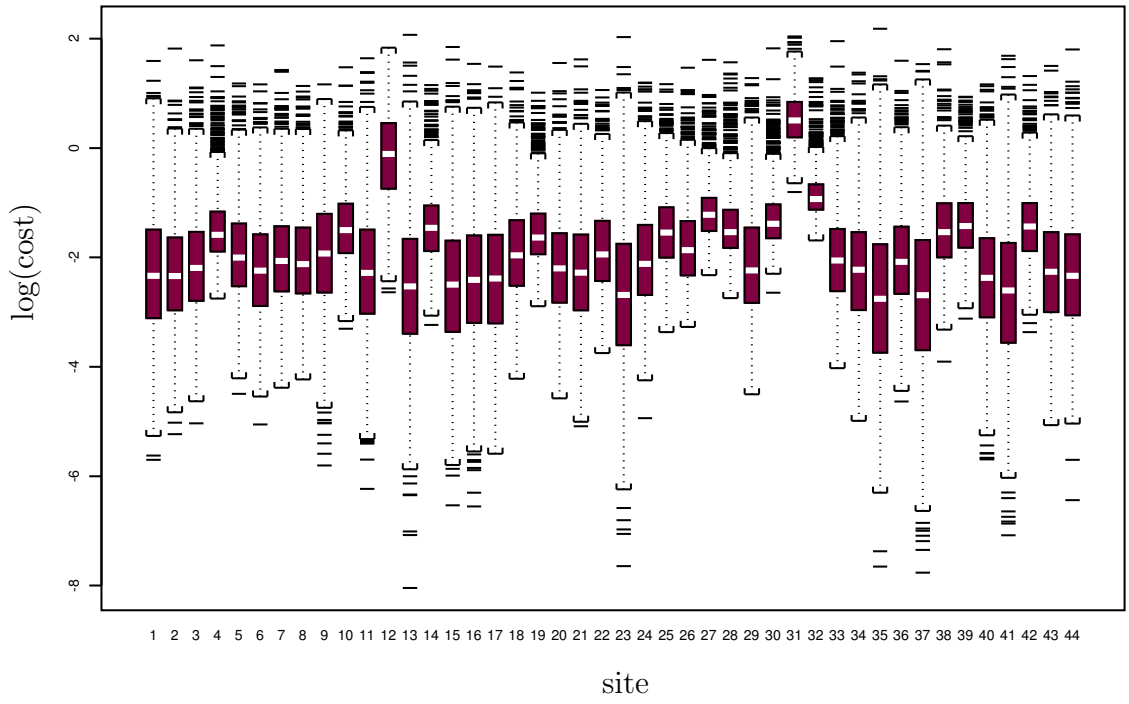


in million Euros. The function is based on economic arguments and includes all the expenses related to a death or an injury. It is interesting to show how much different are the weights assigned to each part of the equation. In fact, under this cost setting, a death is 14 times more expensive (and hence more important for the calculations) compared to a severe injury.

Figure 4 depicts the posterior distribution of the cost for each site taking into account all available parameters. We have used log-scale in order to improve the quality of the graph. The posterior distributions have very large right tails. For some sites the cost is clearly much larger than for other sites, e.g. for intersections 12, 31 and 32.

Using the above 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. In figure 5 one can see those posterior rankings. Sites 12, 31 and 32 are ranked as the worst and they differ very much from the rest. However, there are some interesting points that must be mentioned. There are a lot of sites with similar rankings, the variability for those sites is quite large. In fact, these are sites of similar behavior and the differences are due to random variability. From the graph, one can identify some sites that show a resistant difference by the others but the ranking of

Figure 4: Boxplot of the posterior densities for the cost, using the cost function of Baum(2001)in logscale



the remaining is just a random perturbation and thus of no interest. The latter is very important when the decision to be made is based on the relative ranking (perhaps the mean ranking) of those sites and there is only a budget for a predetermined number of sites. Then, there is a danger that some sites are not different but the decision can be made on ground of random variability at the iterations. Concluding, the approach proposed offers the opportunity to examine whether there are sites that are significantly worse than others.

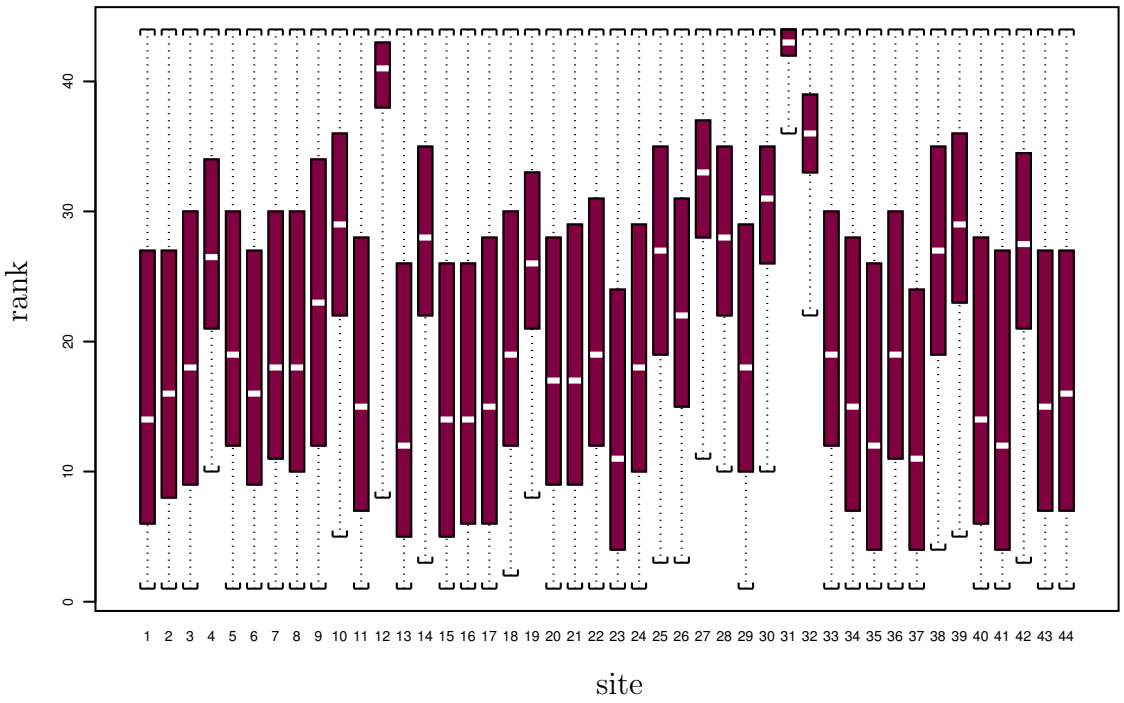
Another interesting case is the cost function adopted by the Flemish government in Belgium (Ministry of Transportation, 2001). This function has the form

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

which in fact assigns weights $(5, 3, 1)$ to deaths, severe and light injuries respectively. This cost function is somewhat different from the previous in so far that 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). The differences with the function of Baum are clear. This function clearly downweights the deaths. The rational for this is that, as a result of the definition of the different injury types (see section 3), deaths and heavy injuries are more closely related than light injuries. The ranking based on this function also can be seen in figure 6.

An interesting observation when comparing figures 5 and 6 is that some sites are recognized as

Figure 5: Boxplots of the ranking the sites, using the cost function of Baum(2001)



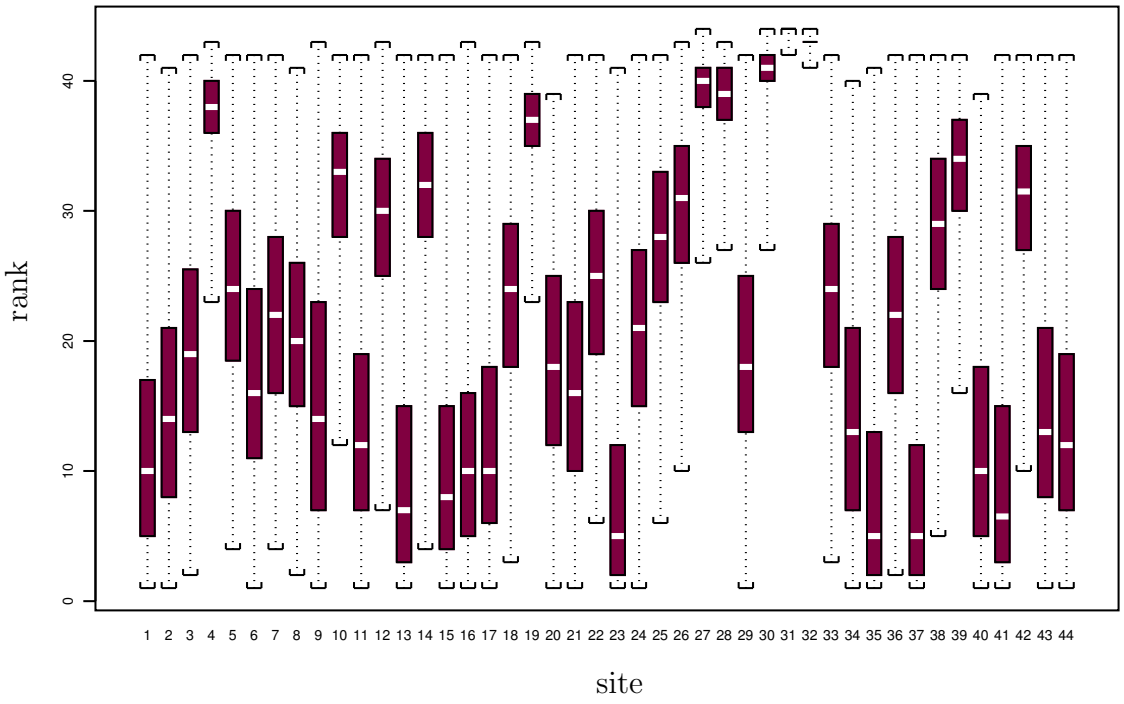
dangerous by both approaches. In fact, the variability on the Flemish cost function is somewhat lower and offers an easier ranking of the sites, but it is apparent that the sites are almost ranked in a similar way. We will pursue more this issue in the next subsection when examining the entire dataset.

4.4 Sensitivity analysis of cost parameters

Both cost functions, introduced in the previous section, enable to put different (absolute or relative) weights on each accident type. As a result, we expect the ranking of intersections to be somewhat different according to the weights 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 large budgets to remedy the, say r, most dangerous intersections. In general, it sounds reasonable that the results will not coincide. However, if only the first r most dangerous sites are to be identified, we expect the methods to agree more or less on the same sites. In other words, if some sites are inherently dangerous, we expect them to be identified regardless of the cost function being used.

For example, in figure 7 the ranks for each site on the ring are plotted against each other for both cost functions, from 1 (most dangerous) to 44 (least dangerous). At least two conclusions

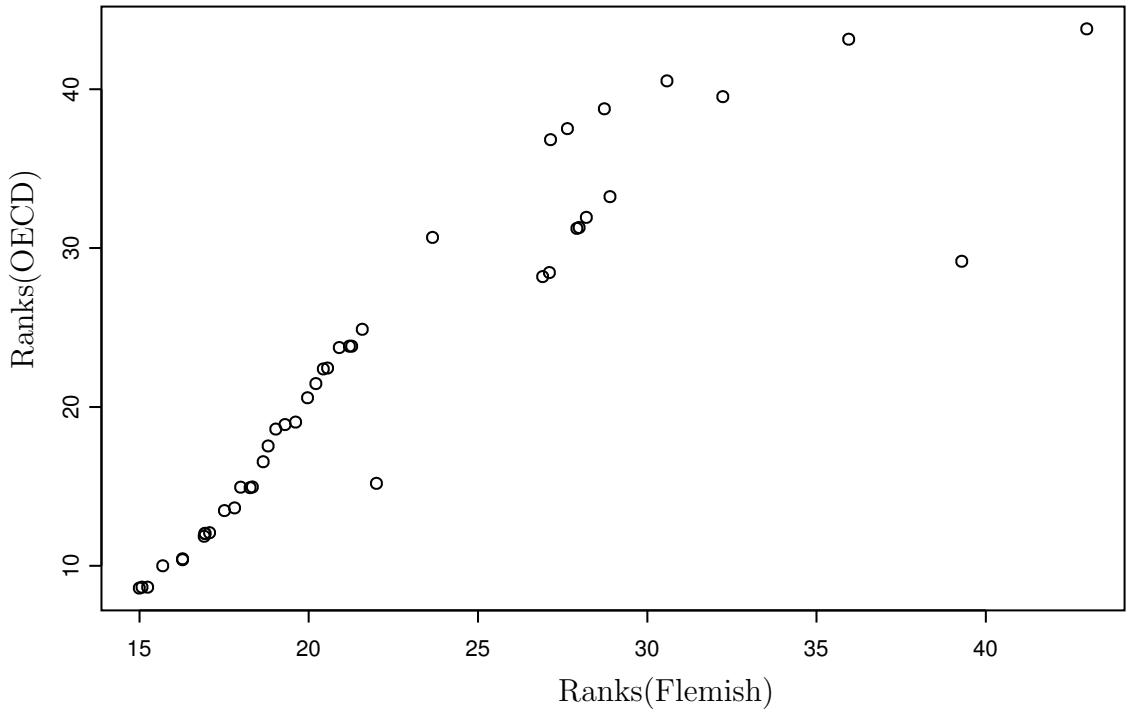
Figure 6: Boxplots of the ranking the sites, using the Flemish approach



can be drawn from this figure. Firstly, if both approaches would yield the same rankings, we would intuitively expect all points to lie on a straight line. This is not the case. However, this can be explained by the different variability of the rankings. Yet, in the case of a perfect match between the rankings, the two rankings must create a monotone curve, which in some sense can be seen in figure 7. Secondly, figure 7 shows that the most dangerous intersections (situated in the lower left corner of the graph) coincide more than the least dangerous intersections (situated in the upper right corner of the graph), indicating that there is much more agreement between both cost functions towards the identification of the top-most dangerous sites, instead of less dangerous sites.

In order to validate this assumption, we used the entire dataset from Leuven, including all the 519 sites to rank them according to both cost functions. For instance, assume that one is interested in finding the r most dangerous intersections. Figure 8 shows the percentage agreement between the ranking of 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 (only plotted up to $r = 300$ for clarity). Figure 8 shows that, for the dataset of Leuven, both cost functions disagree quite heavily on the top-10 most dangerous intersections (only 50 to 60% of the most dangerous sites are considered identical by both methods).

Figure 7: Ranks using the two different cost functions



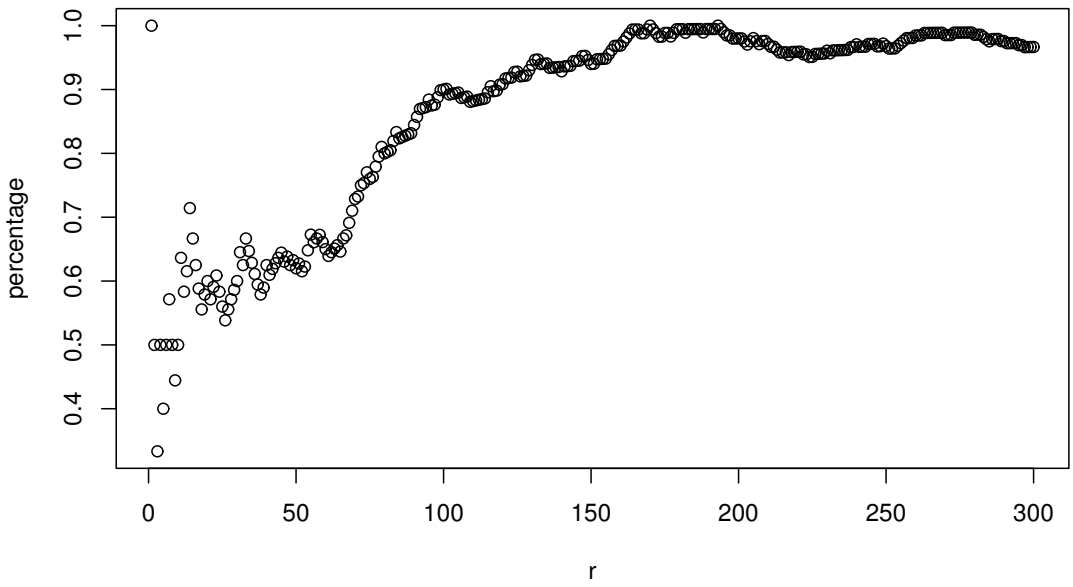
However, both cost functions agree on the most dangerous site overall[5]. When r increases above 60, the agreement between both functions increases again and reaches almost 100% for values of $r > 160$ up to 519. This graph clearly shows that policy makers should be careful in selecting the right value for r, i.e. the number of dangerous sites to allocate money. Indeed, when r is set too low, dangerous sites as identified by both cost functions, will be left untouched. 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 as not or less dangerous by another.

4.5 What about covariates?

In this paper, we did not deal with covariates in our model. In some sense, this may sound strange. However, we have decided not to include covariates for two reasons. Firstly, our scope was on 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 happened. Therefore, in no case we would have a balanced design to include covariate effects.

[5]One should be careful, however, in interpreting such graph since the percentage disagreement is naturally higher for small values of r than for larger values. For example, if only one out of three sites for a value of $r = 3$ is different, this will result in an agreement of only 66%.

Figure 8: Percentage of agreement for ranking the r most hazard sites using the two different cost functions



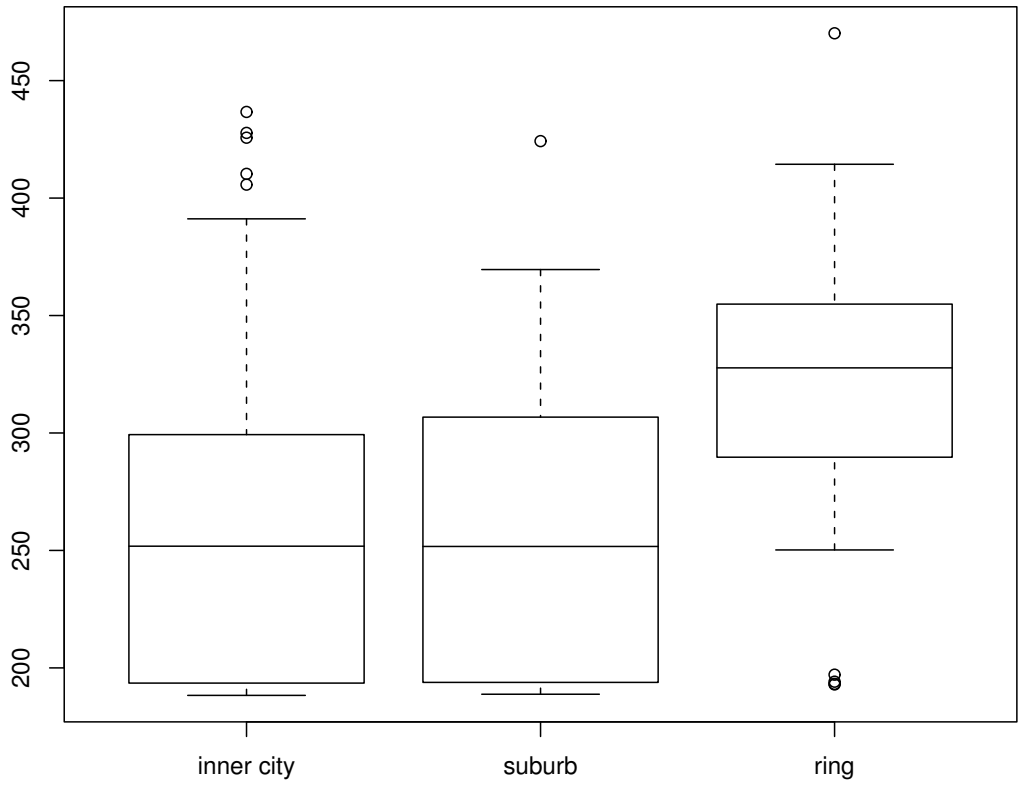
Nevertheless, since it is known both from the literature and from practice that the most dangerous intersections are often situated on ring roads, in figure 9 we have plotted on the Y-axis the mean rankings based on the OECD cost function for the sites with respect to their locations. One can see that indeed the sites situated on the ring road of Leuven are much more dangerous (i.e. their mean ranking is higher), perhaps due to the larger speed of the vehicles on the ring compared with 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, especially 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, this problem can have tremendous impact on the society, not only because it can reduce the accidents on a particular site but, at the same time, one may allocate budgets that could be given to another site, more dangerous, in fact.

In the present paper, we developed a hierarchical Bayesian procedure for ranking sites. The

Figure 9: Boxplots of the mean rankings of sites, with respect the position of sites



procedure takes into account not only the fatalities, but also the injuries (severe and light) and combines this information by means of a cost function in order to rank the sites.

The choice of this cost function, however, is not the purpose of the present paper. In fact, we think that this issue is rather controversial. However, we used two different cost functions for illustrative purposes and to perform a sensitivity analysis towards the results. The first cost function is based on economic arguments and adopts absolute monetary values to express the cost to the society of each accident type. The second cost function is based on pragmatic reasons for decision taking and assigns relative weights to each accident type in order to prioritize budgets to the most dangerous sites. In both cases, our approach can incorporate this kind of information while ranking a site.

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 decision making, one can see whether the chosen sites are really the most dangerous or there are other sites with almost similar characteristics.

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 (see e.g. Tsionas, 1999) that assumes the same covariance for each pair.

6 Acknowledgements

This work is done while Dimitris Karlis visited the Transportation Research Institute at the Limburgs Universitair Centrum in Diepenbeek, Belgium. Furthermore, work on this subject has been supported by a grant given by the Flemish Policy Research Centre on Traffic Safety.

Appendix

The derivation of multivariate Poisson distributions is based on a general multivariate reduction scheme. Assuming Y_r, $r = 1, ..., k$, are independent univariate Poisson random variables, i.e. $Y_r \sim Poisson(\theta_r)$, $r = 1, ..., k$, then the definition of multivariate Poisson models is made through the vector $\boldsymbol{Y}' = (Y_1, Y_2, ..., Y_k)$ and an $m \times k$ matrix $\boldsymbol{A}$ with zeroes and ones and no duplicate columns. Specifically, the vector $\boldsymbol{X}' = (X_1, X_2, ..., X_m)$ defined as $\boldsymbol{X} = \boldsymbol{AY}$ follows a multivariate Poisson distribution. Note that, due to the reproductive property of the Poisson distribution, one may allow $\boldsymbol{A}$ to take any positive integer value. Without loss of generality we restrict $\boldsymbol{A}$ to take only 0 or 1 values.

Each element of $\boldsymbol{X}$ can be expressed as a linear combination of the variables Y_i, $i = 1, ..., k$, with coefficients zero and one. In this framework, the variability of the random vector $\boldsymbol{X}$, which has the m-variate Poisson distribution, is explained through the variability of k independent univariate Poisson random variables. Note that the elements of $\boldsymbol{X}$ are dependent as indicated by the structure of the matrix $\boldsymbol{A}$.

The most general form of the multivariate Poisson distribution arises if the matrix $\mathbf{A}$ has the form $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2, \dots \boldsymbol{A}_m]$, where $\mathbf{A}_j$, $j = 1, \dots, m$ is a sub-matrix of dimensions $m \times \binom{m}{j}$, each column of $\mathbf{A}_j$ has exactly j ones and $(m - j)$ zeroes and no duplicate columns exist. Thus, $\mathbf{A}_m$ is the column vector of $\mathbf{1}$s, while $\mathbf{A_1}$ becomes the identity matrix of size $m \times m$.

The reduced models for m variables derived from $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_m]$ are frequently used in the literature and the resulting distributions are commonly referred to as the multivariate Poisson distributions (see e.g. Tsionas, 2001; Karlis, 2003) This class of models is the only one used in practice even though the theoretical treatment of the model has already been described (e.g. Mahamunulu, 1967). In this paper, we regard a more complicated structure of the matrix $\boldsymbol{A}$. We focus on the case where $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2]$, for the analysis of multivariate data sets. This is done in order not to impose too much structure to our data.

The above definition of the multivariate Poisson distribution provides a straightforward generalization of the univariate case. Not only does each element of $\boldsymbol{X}$ marginally follow a univariate Poisson distribution, but also the parameters of the joint distribution of $X_1, X_2, ..., X_m$ have an obvious interpretation, naturally extended from the univariate case.

For the general model we have

$$E(X) = \mathbf{AM}$$

and

$$Var(X) = \mathbf{A\Sigma A}'$$

where $\mathbf{M}$ and $\mathbf{\Sigma}$ are the mean vector and the variance covariance matrix for the variables $Y_0, Y_1, \dots, Y_k$ respectively. $\mathbf{\Sigma}$ is diagonal because of the independence of Y_i's and has the form

$$\mathbf{\Sigma} = diag(\theta_1, \theta_2, \dots, \theta_m)$$

Similarly

$$\mathbf{M} = (\theta_1, \theta_2, \ldots, \theta_m)'$$

Another interesting feature of this model is that it allows for covariance terms separately for each pair of variables and thus it can be considered as a counterpart of the multivariate normal distributions suitable for multivariate count data.

For the case of the trivariate Poisson model defined by the matrix $\boldsymbol{A} = [\boldsymbol{A}_1, \boldsymbol{A}_2]$ it takes the form

$$\begin{aligned} X_1 &= Y_1 + Y_{12} + Y_{13} \\ X_2 &= Y_2 + Y_{12} + Y_{23} \\ X_3 &= Y_3 + Y_{13} + Y_{23} \end{aligned} \tag{1}$$

where $Y_i \sim Poisson(\theta_i)$, $i \in \{1, 2, 3\}$ and $Y_{ij} \sim Poisson(\theta_{ij})$, $i, j \in \{1, 2, 3\}$, $i < j$. Now, the random variables X_1, X_2, X_3 follow jointly a trivariate Poisson distribution with parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_{12}, \theta_{13}, \theta_{23})'$. The mean vector of this distribution is $\mathbf{AM}$=$(\theta_1 + \theta_{12} + \theta_{13}, \theta_2 + \theta_{12} + \theta_{23}, \theta_3 + \theta_{13} + \theta_{23})'$ and its variance-covariance matrix is given as

$$\mathbf{A\Sigma A}' = \begin{bmatrix} \theta_1 + \theta_{12} + \theta_{13} & \theta_{12} & \theta_{13} \\ \theta_{12} & \theta_2 + \theta_{12} + \theta_{23} & \theta_{23} \\ \theta_{13} & \theta_{23} & \theta_3 + \theta_{13} + \theta_{23} \end{bmatrix}.$$

The parameters θ_{ij}, $i, j = 1, 2, 3, i \neq j$, have the straightforward interpretation of being the covariances between the variables X_i and X_j and, thus, we refer to them as the covariance parameters. The parameters θ_i, $i = 1, 2, 3$, appear only at the marginal means and we refer to them as the mean parameters. The model with the common covariance term can be obtained by setting $\mathbf{AM}$=$(\theta_1 + \theta_0, \theta_2 + \theta_0, \theta_3 + \theta_0)'$ and each covariance term equal to θ_0. The mean and the variance-covariance matrix for the m-variate Poisson distribution ($m > 3$) are defined in an analogous manner. It is clear that this model is more flexible and reasonable for real applications than the one with common covariance.

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.

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

Bureau of Transport Economics, 2001. The Black Spot Program: An Evaluation of the First Three Years, Australia, (http://www.dotars.gov.au/btre/docs/r104/htm/contents.htm)

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.

Dielemann, R., 2000. (in Dutch:) Huidige ontwikkelingen van het verkeersveiligheidbeleid, Doc.nr. 00-12n-7/12/00. BIVV, Brussels, Belgium.

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

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, in press.

Gaver, D., O'Muircheartaigh, I.G., 1987. Robust Empirical Bayes Analysis of Event Rates. Technometrics 29, 1-15.

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

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.

Haddon, W., 1970. A logical framework for categorizing highway safety phenomena and activity. The Journal of Trauma 12, 193-207.

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

Hauer, E., 1996. Identification of "Sites With Promise". Transportation Research Record 975, 54-60.

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

Hauer, E., Persaud, B., 1984. Problem of Identifying Hazardous Locations Using Accident Data. Transportation Research Record 975, 36-43.

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.

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

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

Karlis, D., Meligkotsidou, L., 2003. Multivariate Poisson Regression with Full Covariance Structure, research report.

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

Lindenbergh, S.D., 1998. (in Dutch:) Smartengeld, Ph.D. Dissertation, Leiden University, The Netherlands.

Mahamunulu, D.M., 1967. A note on regression in the multivariate Poisson distribution. Journal of the American Statistical Association 62, 251-258.

Mobiliteitsplan voor de Stad Leuven, 2002. (in Dutch), http://www.leuven.be.

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

Nassar, S., 1996. Integrated Road Accident Risk Model, Unpublished Ph.D. Dissertation, Waterloo, Ontario, Canada.

OECD, 1997. Road safety principles and models: review of descriptive, predictive, risk and accident consequence models. OCDE Road Transport Research OCDE/GD(97)153, Paris.

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.

Tsionas, E.G., 2001. Bayesian multivariate Poisson regression. Communications in Statistics - Theory and Methods 30, 243-255.

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

Vogelesang, A.W., 1996. Bayesian Methods in Road Safety Research: an Overview. Institute for Road Safety Research (SWOV), Leidschendam, The Netherlands.