

A model for identifying and ranking dangerous accident locations: a case study in Flanders

Peer-reviewed author version

BRIJS, Tom; VAN DEN BOSSCHE, Filip; WETS, Geert & KARLIS, Dimitris (2006) A model for identifying and ranking dangerous accident locations: a case study in Flanders. In: *STATISTICA NEERLANDICA*, 60(4). p. 457-476.

DOI: 10.1111/j.1467-9574.2006.00341.x

Handle: <http://hdl.handle.net/1942/1497>

A Model for Identifying and Ranking Dangerous Accident Locations: A Case-Study in Flanders

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

[†] 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

These days, road safety has become a major concern in most modern societies. In this respect, the determination of road locations that are more dangerous than others (black spots or also called 'sites with promise') can help in better scheduling road safety policies. The present paper proposes a multivariate model to identify and rank sites according to their total expected cost to the society. Bayesian estimation of the model via a Markov Chain Monte Carlo (MCMC) approach is discussed in the paper. To illustrate the proposed model, accident data from 23184 accident locations in Flanders (Belgium) are used and a cost function proposed by the European Transport Safety Council is adopted to illustrate the model. It is shown in the paper that the model produces insightful results that can help policy makers in prioritizing road infrastructure investments.

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

¹Corresponding 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 permanent (when the victim is crippled for life or has died). In Belgium, this total cost to the society of traffic accidents is estimated at 3.72 billion Euro per year (Dieleman, 2000).

Social interest therefore lies mainly in preventing traffic accidents. However, this is not an easy task at all. 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 infrastructure works. 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 severity 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 to obtain better and more accurate estimates of the expected number of accidents. 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 hereby are due to the different observational period for different sites and to the different length of the examined roads. Additionally, data concerning the traffic of each site are needed so as to make fair comparisons. Statistical methods must therefore 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 the 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 MCMC methods. The reason for adopting a Bayesian estimation is twofold. Firstly, although the data seem to contain a wealth of information on different accident types, the number of observations per accident location is very low and hence the choice for a Bayesian analysis is preferable. Secondly, the Bayesian treatment of the model enables us to obtain a probabilistic ranking of the locations. The latter is important to verify the impact of the uncertainty/variability that is present in the data on the ranking of locations.

In order to combine all the data into a single number that will be used for ranking the sites, we will introduce 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 functions (Hauer, 1994), we adopt a well-known cost function proposed by the European Safety Council merely for the purpose of illustration.

The remainder of the paper proceeds as follows. Section 2 briefly reviews the multivariate Poisson distribution and provides the details for the particular version that we are going to use. In section 3 we develop the proposed model and describe in detail the Bayesian estimation for that model. The data are described in section 4. In section 5, we apply the model to the data set and we discuss thoroughly the results. We also provide a criterion for deciding on the selection of the most dangerous sites. Finally, concluding remarks can be found in section 6.

2 Multivariate Poisson distribution

Multivariate count data occur in a wide range of different disciplines, including marketing, accident analysis, economics, epidemiology, and many others. It must be recognized however that, despite the wide range of applications that they can model, only few models for this type of data exist and that the published work is not so large. Among them the multivariate Poisson distribution defined in Johnson et al. (1997) plays an important role, basically as the theoretical tool to construct new models. On the contrary the applications are usually limited to a special case of this distribution limiting the insight provided by the model.

The derivation of multivariate Poisson distributions is based on a general multivariate reduction scheme. Assuming Y_r , $r = 1, \dots, k$, are independent univariate Poisson random variables, i.e. $Y_r \sim \text{Poisson}(\theta_r)$, $r = 1, \dots, k$, then the definition of multivariate Poisson models is made through the vector $\mathbf{Y}' = (Y_1, Y_2, \dots, Y_k)$ and an $m \times k$ matrix \mathbf{A} with zeroes and ones and no duplicate columns. Specifically, the vector $\mathbf{X}' = (X_1, X_2, \dots, X_m)$ defined as $\mathbf{X} = \mathbf{A}\mathbf{Y}$ follows a multivariate Poisson distribution. Note that the elements of \mathbf{X} are dependent as indicated by the structure of the matrix \mathbf{A} .

In the most general form $k = \sum_{j=1}^m \binom{m}{j}$. This form arises if matrix \mathbf{A} has the form $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_m]$, 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 ones, while \mathbf{A}_1 becomes the identity matrix of size $m \times m$. It can be shown that

$$E(\mathbf{X}) = \mathbf{A}\mathbf{M} \quad \text{and} \quad \text{Var}(\mathbf{X}) = \mathbf{A}\mathbf{\Sigma}\mathbf{A}'$$

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

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

Similarly

$$\mathbf{M} = (\theta_1, \theta_2, \dots, \theta_k)'$$

In this paper, we focus on the case where $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2]$, for the analysis of multivariate data sets. This is done in order not to impose too much structure to our data. An 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.

Consider the case of trivariate data. With slightly different notation, assume that $\mathbf{Y}' =$

$(Y_1, Y_2, Y_3, Y_{12}, Y_{13}, Y_{23})$. Define $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2]$, i.e. \mathbf{A} has the form

$$\mathbf{A} = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{array} \right].$$

Then, define $\mathbf{X} = (X_1, X_2, X_3)' = \mathbf{A}\mathbf{Y}$ and thus we have the following representation:

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

where $Y_i \sim \text{Poisson}(\theta_i)$, $i \in \{1, 2, 3\}$ and $Y_{ij} \sim \text{Poisson}(\theta_{ij})$, $i, j \in \{1, 2, 3\}$, $i < j$. 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 $\mathbf{A}\boldsymbol{\mu} = (\theta_1 + \theta_{12} + \theta_{13}, \theta_2 + \theta_{12} + \theta_{23}, \theta_3 + \theta_{13} + \theta_{23})'$ and its variance-covariance matrix is given as

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

The parameters θ_{ij} , $i, j = 1, 2, 3, i < j$, 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.

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

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2, X_3 = x_3) &= P(x_1, x_2, x_3) \\ &= \sum_{k=0}^{s_1} \sum_{r=0}^{s_2} \sum_{s=0}^{s_3} \frac{e^{-\theta_{12}} \theta_{12}^k}{k!} \frac{e^{-\theta_{13}} \theta_{13}^r}{r!} \frac{e^{-\theta_{23}} \theta_{23}^s}{s!} \frac{e^{-\theta_1} \theta_1^{x_1 - k - r}}{(x_1 - k - r)!} \frac{e^{-\theta_2} \theta_2^{x_2 - k - s}}{(x_2 - k - s)!} \frac{e^{-\theta_3} \theta_3^{x_3 - r - s}}{(x_3 - r - s)!} \end{aligned}$$

where $x_1, x_2, x_3 = 0, 1, \dots$, $s_1 = \min(x_1, x_2)$, $s_2 = \min(x_1 - k, x_3)$, $s_3 = \min(x_2 - k, x_3 - r)$. 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 that the marginal distributions are univariate Poisson distributions.

We will base our model on this multivariate Poisson model. 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. Our model 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).

Drawbacks of this model are firstly that it has Poisson marginal distributions and thus it cannot handle overdispersion and secondly that it can model only positive dependence. For more complicated models to improve in both aspects the reader is referred to the papers of Chib and

Winkelmann (2001), Munkin and Trivedi (1999), Karlis and Meligkotsidou (2003), Van Ophem (1999), Cameron et al. (2004), Berkout and Plug (2004). However in our case, the data are in accordance with Poisson marginals and the correlation is positive as it is usually the case for accident data. In addition, our model is much simpler than the other models referred above. So, we believe that this model provides an interesting and realistic device for the purpose of ranking accident sites.

3 The Model

Suppose that the data consist of n different sites. The number of accidents for the i -th site is denoted by V_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 $\phi_i t_i$. 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 ϕ_i , $i = 1, \dots, n$ are the pure accident rate for the i -th site.

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 V_i , they follow a 3-variate Poisson distribution. In other words, the above model allows for different correlations between each pair of variables, which is clearly a realistic assumption in the context of traffic accident injuries. In notational form, we assume that

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

Hence, $\mu_{.i}$ reflects the mean parameter 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.

3.1 A Bayesian approach

Our model has the form

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

The likelihood can be written in the complicated form

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

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 $\boldsymbol{\theta} = (\boldsymbol{\phi}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\mu}_3, \boldsymbol{\lambda}_{12}, \boldsymbol{\lambda}_{13}, \boldsymbol{\lambda}_{23})$, where the vectors represented by boldface letters represent the corresponding parameters for all the observations, i.e. $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n)$ 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 3.3).

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

$$\begin{aligned} \phi_i &\sim \text{Gamma}(a_1, b_1) \\ \mu_{1i} &\sim \text{Gamma}(a_2, b_2) \\ \mu_{2i} &\sim \text{Gamma}(a_3, b_3) \\ \mu_{3i} &\sim \text{Gamma}(a_4, b_4) \\ \lambda_{12i} &\sim \text{Gamma}(a_5, b_5) \end{aligned}$$

$$\lambda_{13i} \sim \text{Gamma}(a_6, b_6)$$

$$\lambda_{23i} \sim \text{Gamma}(a_7, b_7)$$

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

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

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

Details for the MCMC follow.

3.2 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 multivariate reduction representation of a multivariate Poisson distribution. In our model, the above idea assumes that there are some latent variables $\delta_{1i}, \delta_{2i}, \delta_{3i}, T_{1i}, T_{2i}, T_{3i}$, each one following independently a Poisson distribution with parameter $\lambda_{12i}, \lambda_{13i}, \lambda_{23i}, \mu_{1i}, \mu_{2i}, \mu_{3i}$, respectively. From them 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 $\delta_{ji}, j = 1, 2, 3$ reflect site characteristics that introduce correlation to the working variables. The data augmentation being used is based on considering the unobservable quantities $\delta_{ji}, j = 1, 2, 3, i = 1, \dots, n$ 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 $\boldsymbol{\kappa} = (\delta_{11}, \dots, \delta_{1n}, \delta_{21}, \dots, \delta_{2n}, \delta_{31}, \dots, \delta_{3n})$ be the unobserved data. Augmenting $\boldsymbol{\kappa}$ to the observed data, the joint posterior of the complete data is of the form

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\kappa} | \text{data}) &= \prod_{i=1}^n \frac{e^{-\phi_i t_i} (\phi_i t_i)^{v_i}}{v_i!} \frac{e^{-\lambda_{12i} v_i} (\lambda_{12i} v_i)^{\delta_{1i}}}{\delta_{1i}!} \frac{e^{-\lambda_{13i} v_i} (\lambda_{13i} v_i)^{\delta_{2i}}}{\delta_{2i}!} \frac{e^{-\lambda_{23i} v_i} (\lambda_{23i} v_i)^{\delta_{3i}}}{\delta_{3i}!} \times \\ &\quad \frac{e^{-\mu_{1i} v_i} (\mu_{1i} v_i)^{y_i - \delta_{1i} - \delta_{2i}}}{(y_i - \delta_{1i} - \delta_{2i})!} \frac{e^{-\mu_{2i} v_i} (\mu_{2i} v_i)^{z_i - \delta_{1i} - \delta_{3i}}}{(z_i - \delta_{1i} - \delta_{3i})!} \frac{e^{-\mu_{3i} v_i} (\mu_{3i} v_i)^{w_i - \delta_{2i} - \delta_{3i}}}{(w_i - \delta_{2i} - \delta_{3i})!} \times \\ &\quad [\Gamma(a_1)]^{-1} \phi_i^{a_1 - 1} b_1^{a_1} \exp(-b_1 \phi_i) [\Gamma(a_2)]^{-1} \mu_{1i}^{a_2 - 1} b_2^{a_2} \exp(-b_2 \mu_{1i}) \times \end{aligned}$$

$$\begin{aligned}
& [\Gamma(a_3)]^{-1} \mu_{2i}^{a_3-1} b_3^{a_3} \exp(-b_3 \mu_{2i}) [\Gamma(a_4)]^{-1} \mu_{3i}^{a_4-1} b_4^{a_4} \exp(-b_4 \mu_{3i}) \times \\
& [\Gamma(a_5)]^{-1} \lambda_{12i}^{a_5-1} b_5^{a_5} \exp(-b_5 \lambda_{12i}) [\Gamma(a_6)]^{-1} \lambda_{13i}^{a_6-1} b_6^{a_6} \exp(-b_6 \lambda_{13i}) \times \\
& [\Gamma(a_7)]^{-1} \lambda_{23i}^{a_7-1} b_7^{a_7} \exp(-b_7 \lambda_{23i})
\end{aligned}$$

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

$$\begin{aligned}
\delta_{1i} | \cdot & \propto \frac{\lambda_{12i}^{\delta_{1i}}}{\delta_{1i}!(y_i - \delta_{1i})!(z_i - \delta_{1i})!} \left(\frac{1}{\mu_{1i}\mu_{2i}} \right)^{\delta_{1i}} \\
\delta_{2i} | \cdot & \propto \frac{\lambda_{13i}^{\delta_{2i}}}{\delta_{2i}!(y_i - \delta_{2i})!(w_i - \delta_{2i})!} \left(\frac{1}{\mu_{1i}\mu_{3i}} \right)^{\delta_{2i}} \\
\delta_{3i} | \cdot & \propto \frac{\lambda_{23i}^{\delta_{3i}}}{\delta_{3i}!(z_i - \delta_{3i})!(w_i - \delta_{3i})!} \left(\frac{1}{\mu_{2i}\mu_{3i}} \right)^{\delta_{3i}} \\
\phi_i | \cdot & \sim \text{Gamma}(a_1 + v_i, b_1 + t_i), \quad i = 1, \dots, n \\
\mu_{1i} | \cdot & \sim \text{Gamma}(a_2 + y_i - \delta_i, b_2 + v_i), \quad i = 1, \dots, n \\
\mu_{2i} | \cdot & \sim \text{Gamma}(a_3 + z_i - \delta_i, b_3 + v_i), \quad i = 1, \dots, n \\
\mu_{3i} | \cdot & \sim \text{Gamma}(a_4 + w_i - \delta_i, b_4 + v_i), \quad i = 1, \dots, n \\
\lambda_{12i} | \cdot & \sim \text{Gamma}(a_5 + \delta_{1i}, b_5 + v_i), \quad i = 1, \dots, n \\
\lambda_{13i} | \cdot & \sim \text{Gamma}(a_6 + \delta_{2i}, b_6 + v_i), \quad i = 1, \dots, n \\
\lambda_{23i} | \cdot & \sim \text{Gamma}(a_7 + \delta_{3i}, b_7 + v_i), \quad i = 1, \dots, n
\end{aligned}$$

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

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

$x_1, x_2 \in \{0, 1, \dots\}$, $y = 0, \dots, \min(x_1, x_2)$, $\psi > 0$. 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 $P'(k+1) = P'(k) \frac{\rho}{k+1} (x_1 - k)(x_2 - k)$, $k = 0, \dots, s_i - 1$. 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 , $j = 1, \dots, 7$ 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 expected cost can be

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

for some coefficients β_i , $i = 1, 2, 3$ where the three parts 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 using the current values of the parameters and, thus, 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 site i 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. In the sequel we will propose a criterion for selecting the most dangerous sites (perhaps for allocating funds for reconstruction) which is based on such posterior summary measures.

Model checking for the proposed model can be done in the usual way based on predictive densities (see, e.g. Gelfand *et al.*, 1992, Gelfand and Ghosh, 1998) or the currently fashionable deviance information criterion (DIC) (Spiegelhalter *et al.*, 2002). Note also that the model can have an interesting model selection nature by selecting appropriate covariance terms to retain in the model, or as in the multivariate normal model, to check for specific covariance structures. All these are, however, beyond the scope of the present paper. Finally, we adopt an empirical Bayes approach, but examination of the sensitivity of the proposed methodology to the choice of priors is important. In fact, informative priors for site specific parameters can have a large impact on the results by favoring sites against others. In the general setting given in the paper, where all the site-specific parameters have the same priors, the impact on the choice of α_i 's and β_i 's is however small given the large number of sites.

3.3 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.

4 The data

For this study, the official traffic accident records for the region of Flanders (Belgium) have been used for the period 1997-1999. In total 50961 traffic accidents, with at least casualties involved, were registered by the police services in this period at 23184 different road locations on motorways, national and provincial roads². The data consists both of isolated road segments and of intersections. For the road segments, the hectometer stone marker on the road provides the unique position of the accident. For the intersections, all accidents that happen within 50 meters from the center of the intersection are assigned to that intersection.

For each road location in the database, the number of accidents in the given period V_i is counted. Furthermore, a distinction is made according to the severity 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

²It 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.

after the accident), and W_i is the number of heavily injured persons, being every road user who got 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 an 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 .

Based on these data, the Flemish government has recently identified a list of 800 dangerous road locations, so called black spots, to be considered for road infrastructure works to reduce the number of accidents. However, since their selection is not based on a statistical modeling procedure, it is interesting to compare them with the results obtained from the model suggested above.

5 Results

Before discussing the results of the proposed model, some remarks about the data set should be added that are relevant to the put the results in the right perspective. First, since data is available only for road locations where accidents happened, all results should be interpreted conditional on the 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 spatial 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 modeling approach followed in this paper.

Finally, $t_i = 1$ for all $i = 1, \dots, n$ locations, since all accident sites are comparable in length, the time periods of the data are the same and we do not possess traffic flow information.

5.1 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 accidents data set are the empirical Bayes estimates $a_1 = 0.85$, $a_2 = 0.00$, $a_3 = 1.18$, $a_4 = 0.02$, $a_5 = 0.11$, $a_6 = 0.01$, $a_7 = 1.00$ and $b_1 = 0.07$.

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

posterior density ³.

5.2 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 will adopt 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 (e.g. what are the quantities that have to be measured in order to estimate the cost for a severely injured person?). Therefore, the cost function, which is obtained from calculations made by the European Transport Safety Council (ETSC), is merely used for illustrative purposes and does not reflect any ethical or political statement by the authors.

More specifically, the cost function from the ETSC values the loss of a human life as the result of an accident equal to 6.9 heavy casualties and 22.8 light casualties. Consequently, the expected total cost of accidents for a particular road location can be calculated using the following cost function

$$c_i = E(C_i) = 22.8E(Y_i) + 3.3E(W_i) + E(Z_i)$$

The cost function is based on economic arguments and includes all the expenses related to a death or an injury. Using this cost function, individual sites can be ranked. Let the vector $\mathbf{c}^{(j)}$ with elements $c_i^{(j)}$, $i = 1, \dots, n$ contain the expected costs for each site at the j -th MCMC iteration. Then, one can assign a rank to each site according to its cost value and transform the vector $\mathbf{c}^{(j)}$ into a vector $\mathbf{R}^{(j)}$ which contains the ranks for all the sites. The posterior distribution for the ranks of each site can then easily be constructed as explained in section 3.2.

Figure 1 about here

Figure 1 shows the posterior mean ranks for a random selection of 100 sites out of the total of 23184 sites⁴. In fact, for each site one can see the posterior mean rank together with its \pm posterior standard deviation to show the variability present in estimating the ranks. From the traffic safety point of view, figure 1 shows that some locations on the left tend to be ranked higher (i.e. more dangerous), whereas several other locations exist with similar mean ranks indicating that these locations are of the same hazard. It is also apparent that the variability is much larger for the sites with a smaller mean rank. This is due to the fact that since it appears that there are several

³Details about convergence properties are available on request, but we omit them to save space.

⁴We chose to show the mean ranks of only 100 sites selected at random to improve the readability of the graph.

locations that are of similar hazard, their permutation during the MCMC iterations leads to a higher variance.

In other words, the graph shows that there is a small group of locations that are ranked as very dangerous during the entire MCMC chain, but that there is also a much larger group of locations for which it is difficult to say if one location is more dangerous than the other, i.e., for which the ranking over the MCMC iterations is subject to large perturbations. It would therefore be interesting to construct a procedure to decide which locations should be considered to be a candidate for the 800 most dangerous locations, as selected by the Flemish government.

5.3 A criterion for selecting sites

One of the advantages of the MCMC method is that it allows for exploring posterior distributions of certain functions related to the parameters of interest.

Denote by $\theta^{(j)}$ the vector of parameters at the j -th iteration of the MCMC. From this vector we are able to calculate the expected cost for each site. We may proceed further by deriving the probability for each site of being one of the r worst sites. To help the exposition and provide generality suppose that we have n sites. Since the smaller the cost the better the site, a site is bad if the cost is high. Suppose that we need to estimate the probability that the site i belongs to the r worst sites. This implies that its cost is among the r highest and hence its rank is larger than $n - r$ (since the ranking procedure gives ranks from the smaller cost to the larger). Then the estimated probability is calculated as

$$P_r(i) = \frac{\sum_{j=1}^B I(R_i^{(j)} > n - r)}{B}$$

where $I(A)$ is the indicator function and B the number of iterations kept after the burn-in.

It is interesting that the above probabilities allow for an heuristic rule for selecting the worst sites. Suppose that all sites have the same characteristics. This way we expect that for all the sites the required probabilities will be exactly the same as any differences will be merely random perturbations. In that case, we expect that this probability will be equal to r/n for each site. Sites with probability above this limit reveal a deviation from the argument about equal sites. Of course, due to random perturbations, some probabilities will be larger even in the case of equal sites. The situation resembles the case of the scree plot in Principal Components Analysis for selecting the number of components. To facilitate further this approach we may calculate some confidence intervals for the probabilities by replicating the procedure for a number of times. This will reveal sites with probability above the limit in a more rigorous basis reducing the effect of random perturbations.

For the Flemish data set, we have $r = 800$. From the 3000 sampled values during MCMC, we split them in 30 batches of 100 replications each, and we calculated the probabilities as above for each batch. This results with 30 values for each probability. In Figure 2 and 3, we plot the

probability calculated from all 3000 sampled values together with its smaller and larger values from the 30 replications. This provides an indication about the existing variability. Recall that since we deal with estimated probabilities one may construct quick and rough confidence intervals for probabilities using as an estimate of the variance of the probability p , i.e., the quantity $Bp(1 - p)$ where B is the replication size or use the fact that the criterion is a kind of a sample mean and thus based on asymptotic result to find a consistent estimate of its variance. In Figure 2 and 3, the horizontal line represents the probability r/n i.e. the probability of each site in the case of pure noise. For our case this equals $800/23184$.

Figure 2 and 3 about here

Applying this procedure to the Flemish accident data set leads to the following interesting results. From figure 2, which shows all 23184 locations, it can be seen that most of the locations share a similar probability close to zero, indicating that they can be excluded as being candidates for the 800 worst accident locations in Flanders. Only a limited number of sites (located at the left of the figure) show a high probability of belonging to the 800 worst sites. Indeed, when looking more closely at the left part of the figure 2, which is depicted in figure 3, some interesting conclusions can be drawn.

First of all, figure 3 shows that although there is a small group of very dangerous locations having a high probability (i.e. much above the baseline), there are at least 1500 locations that could potentially belong to the 800 most dangerous locations to be considered for a safety audit, because the lower limit of their confidence interval falls above the given baseline. This is an extremely important finding since it shows that the current selection of 800 dangerous locations is highly debatable. Indeed, apart from the small group of obviously dangerous locations with high probabilities, there is a large group of locations (with probabilities around 0.2) that are more or less equally dangerous and this shows that a large part of the current selection of 800 locations by the Flemish government is somewhat subject to randomness. In other words, a large part of the selected locations could equally well be interchanged with other locations that are equally dangerous.

Secondly, one can identify two jumps in figure 3, hereby creating three 'clusters' of locations. The first cluster of locations are those having a very high probability of belonging to the 800 worst sites, i.e., say $p > 0.4$. These locations (some 250) are extremely dangerous and should in any case deserve the attention of the traffic safety engineers. Their confidence intervals are also very small, which supports their severeness. The second group is much larger and shows a rather constant probability of about $p = 0.4$ with much larger confidence intervals. For this group, consisting of about 1250 locations, it is much harder to say which of them should be considered as belonging to the 800 most dangerous locations in Flanders. Indeed, the lower limit of their confidence intervals is still above the baseline, which makes them almost equal candidates for the 800 worst sites.

Finally, there is a third group of locations (the rest) with probabilities $p \leq 0.1$, for which the probability of belonging to the 800 worst sites is extremely small. Furthermore, the estimated

confidence limits are quite small, which corroborates the assumption that it is very unlikely that one of these locations would be a valid candidate to belong to the 800 worst sites in Flanders.

6 Concluding Remarks

In this paper, a Bayesian procedure using an MCMC was developed for ranking accident locations in Flanders, Belgium. The procedure takes into account not only the number of fatalities, but also the number of injuries (severe and light) and combines this information by means of a cost function in order to rank the sites.

From the methodological point of view, the model suggested in the present paper is based on a 3-variate Poisson distribution. The model assumes that the covariances vary across sites. Yet, there is little evidence for this assumption and hence, naturally, one may assume constant covariances across sites by not allowing the covariances in (2) to depend on the v_i 's. This would reduce the complexity of the model as one needs only to update three covariance parameters rather than $3n$. In fact, running this reduced model, we found that this assumption had negligible (if any) effect on our results. Yet, in order to provide a general framework, we kept the more general model with varying covariances.

From the traffic safety point of view, the most interesting insight offered by our model is that it does not only rank the sites but that 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 whether there are other sites with almost similar characteristics. It is important to note, however, that this paper does not provide a cost-benefit analysis of road infrastructure investments. To this end, one would need the cost of alternative road infrastructure investments together with their respective accident modification factors (AMF's). Indeed, per location, a number of alternatives may be available (e.g. decrease speed, construct a median barrier, etc.) to increase safety with different respective costs and effectiveness. Since this information is lacking, one could say that this paper proposes the optimal ranking given an unlimited available budget.

7 Acknowledgements

Part of this work was 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 Government to the Flemish Policy Research Center for Traffic Safety.

References

- Abdel-Aty, M.A. and Radwan, A.E. (2000) Modelling traffic accident occurrence and involvement. *Accident Analysis and Prevention*, **32(5)**, 633-642.
- Andreassen, D.C. and Hoque, M.M. (1986) Intersection accident frequencies. *Traffic Engineering and Control*, **27(10)**, 514-517.
- 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.
- Berkhout, P. and Plug, E. (2004). A bivariate Poisson count data model using conditional probabilities. *Statistica Neerlandica*, 58, 349–364
- Cameron, A.C., Li, T. , Trivedi, P.K. and Zimmer, D.M. (2004) Modelling the differences in counted outcomes using bivariate copula models: with application to mismeasured counts. *to appear in Journal of Econometrics*
- Chib, S and Winkelmann, R. (2001). Markov Chain Monte Carlo Analysis of Correlated Count data. *Journal of Business and Economic Statistics*, **19**, 428-435.
- Christiansen, C.L., Morris, C.N. and 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. and Yang, S. (2001) Bayesian Identification of High-risk Intersections for Older Drivers via Gibbs Sampling. *Transportation Research Record*, **1746**, 84-89.
- Dieleman, R. (2000) (in Dutch:) Huidige ontwikkelingen van het verkeersveiligheidsbeleid, *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. and 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.
- Gaver, D. and O’Muircheartaigh, I.G. (1987) Robust empirical Bayes analysis of event rates. *Technometrics*, **29**, 1-15.

- Gelfand, A.E., Dey, D.K. and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. In *Bayesian Statistics 4 (Edited by J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith)*, 147-167. Oxford University Press, Oxford.
- Gelfand, A.E. and Ghosh, S.K. (1998). Model Choice: A Minimum Posterior Predictive Loss Approach. *Biometrika*, 85, 1–13.
- Geurts, K. and Wets, G. (2003) Black Spot Analysis Methods: Literature Review. *Doc.nr. RA-2003-07*. Flemish Research Center for Traffic Safety, Diepenbeek, Belgium.
- Goldstein H. and 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. (1994) Can one estimate the value of life or is it better to be dead than stuck in traffic? *Transportation Research, Series A*, **28**, 109-118.
- 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. and Persaud, B. (1984) Problem of Identifying Hazardous Locations Using Accident Data. *Transportation Research Record*, **975**, 36-43.
- Hauer, E. and 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. and 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. and 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.

- Lindenberg, S.D. (1998) (in Dutch:) Smartengeld, *Ph.D. Dissertation*, Leiden University, The Netherlands.
- Munkin, M.K. and Trivedi, P.K (1999) Simulated maximum likelihood estimation of multivariate mixed-Poisson regression models, with application *Econometrics Journal*, 2, pages 29-48
- 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. and Nicholson, A.J. (1997) Ranking and selecting motor vehicle accident sites by using a hierarchical Bayesian model. *The Statistician*, **46**, 293-316.
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P. and van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit (with discussion). *Journal of the Royal Statistical Society B*, **64**, 583–639.
- 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.
- van Ophem, H. (1999) A general method to estimate correlated discrete random variables. *Econometric Theory*, 15, 228–237
- Vogesang, A.W. (1996) *Bayesian Methods in Road Safety Research: an Overview*. Institute for Road Safety Research (SWOV), Leidschendam, The Netherlands.

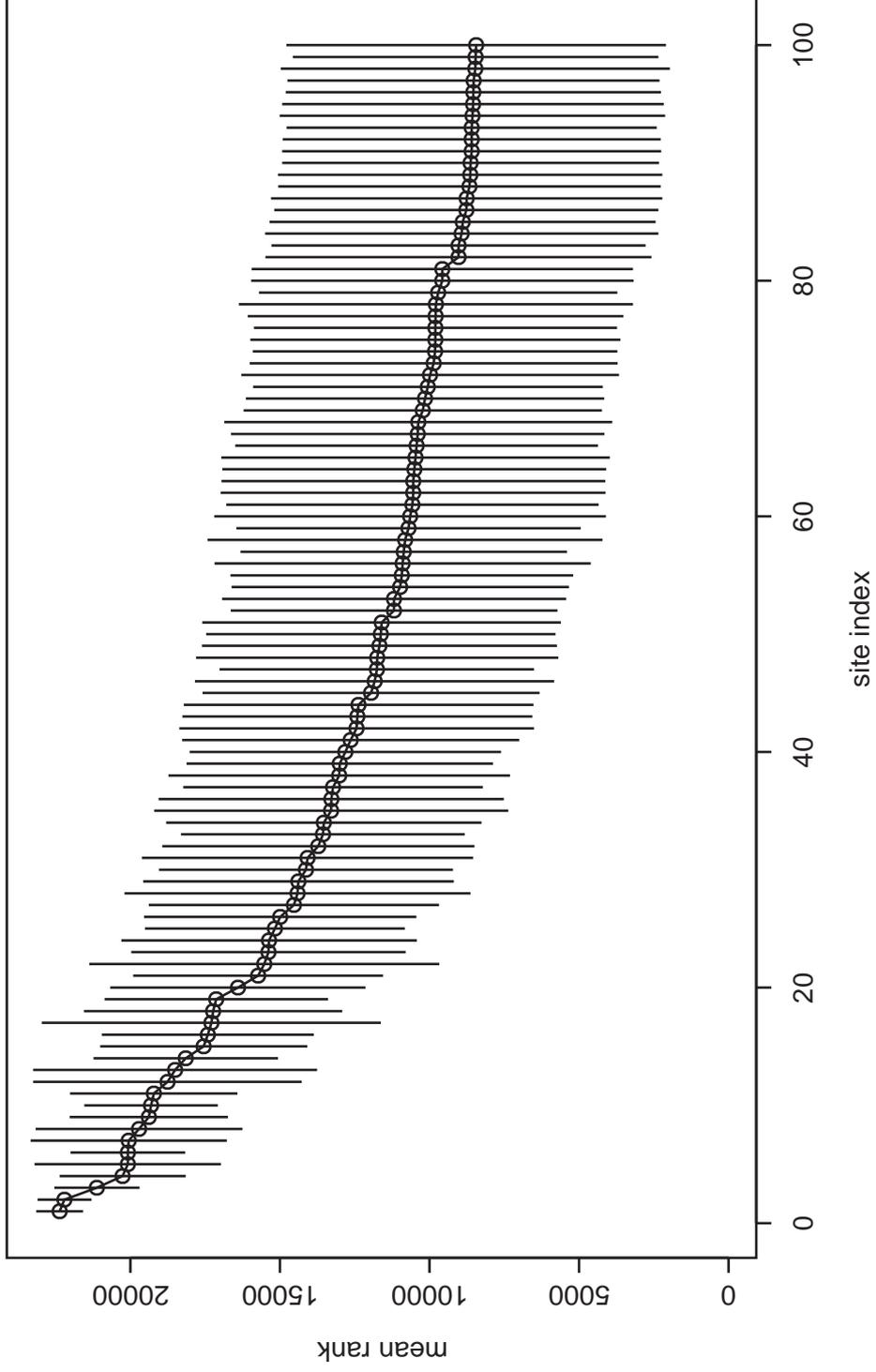


Figure 1: Posterior mean ranks for a random sample of 100 sites. The lines represent an interval \pm standard deviation from the mean rank

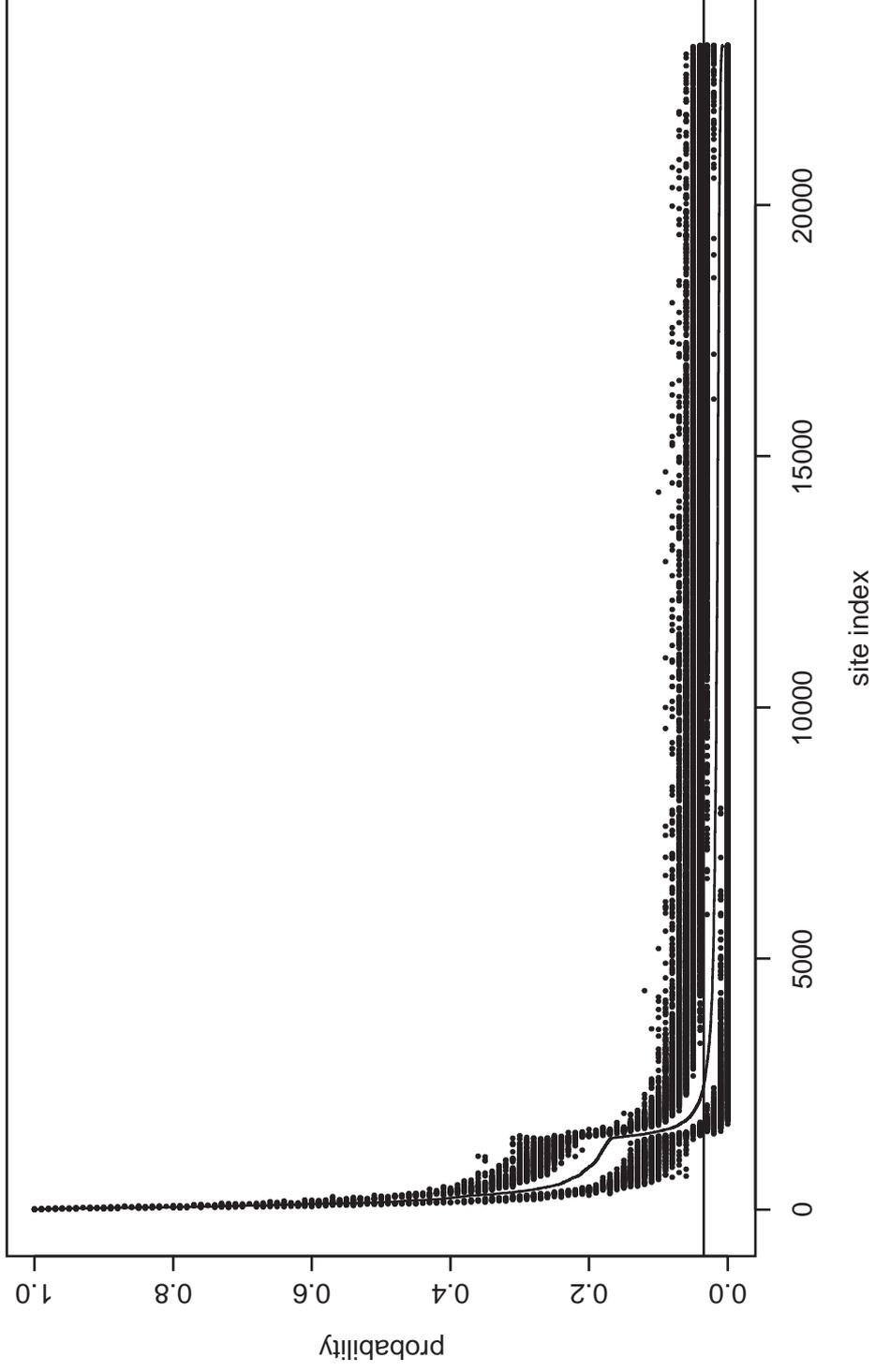


Figure 2: The estimated probability that the site belongs to the 800 worst sites for all accident locations in Flanders, together with a simple confidence interval

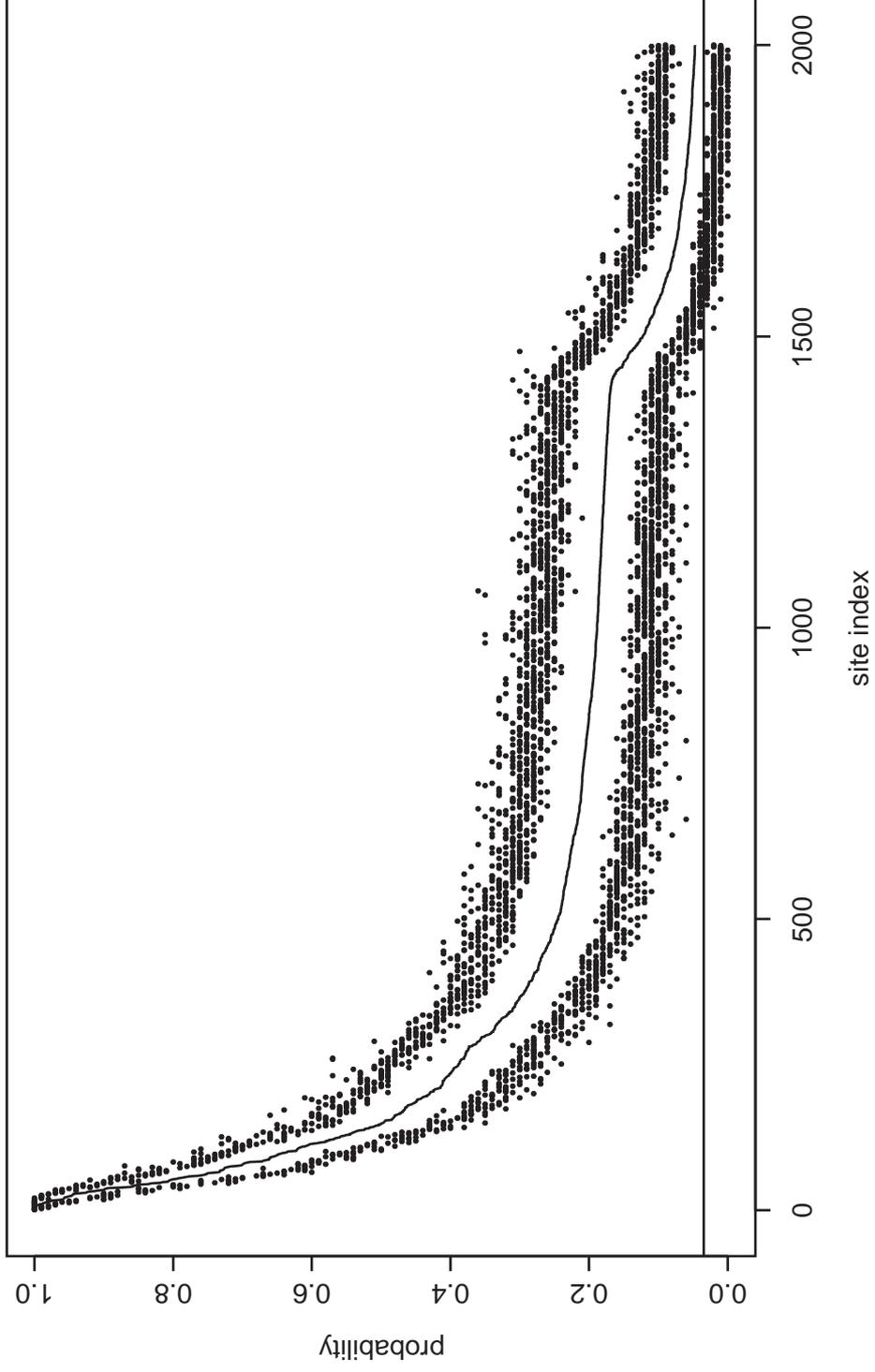


Figure 3: The estimated probability that the site belongs to the 800 worst sites for the 2000 accident locations ranked as worst, including a confidence interval