

Bayesian models for the detection of high risk locations on Portuguese motorways

S.M. Azeredo-Lopes & J.L. Cardoso

Transportation Department, Planning, Traffic and Road Safety Division, National Laboratory for Civil Engineering, Lisboa, Portugal

ABSTRACT: Hierarchical Bayesian regression models, with differing hyper-prior distributions, are considered as accident prediction models to be fitted on data collected over several years on the Portuguese motorway network. A sensitivity analysis is performed by way of simulation to investigate the practical implications of the choice of informative hyper-priors (Gamma, Christiansen and Uniform) and non-informative Gamma, as well as various sample sizes and years of aggregated data, on the results of a road safety analysis, in particular, at detecting high accident risk locations. It was concluded that informative hyper-priors were best at detecting hotspots when small sample sizes were considered. For bigger samples the various hyper-priors produced equivalent outcomes. Furthermore, more accurate results were obtained when more years of data were analyzed.

1 INTRODUCTION

The identification of high accident risk locations, also referred to as hotspots, is the first step of the highway safety management process. There are several methods currently used for hotspot identification: Cheng & Washington (2005, 2008), Schlüter et al. (1997), Geurts et al. (2006), Miaou & Song (2005), Washington & Cheng (2005), Elvik (2008), Miranda-Moreno et al. (2007) employ and compare such various methods. Recently there has been a growing debate on the employment of empirical Bayes methods (EB) and the so-called full Bayes approaches or hierarchical Bayes methods, Miranda-Moreno & Fu (2009), Persaud et al. (2010). The present paper is concerned with obtaining and applying appropriate hierarchical Bayesian models for use in traffic safety studies when employing a ranking criterion for hotspot identification.

Traffic safety studies that have been performed using data collected on Portuguese roads and motorways are described in Azeredo-Lopes & Cardoso (2007a,b, 2009) and have consisted on the employment of classical statistical methods comprising the use of generalized linear models (see details in McCullagh & Nelder 1990) where the number of accidents or casualties are regarded as the response or dependent variable and the model parameters are estimated by maximum likelihood of iteratively least squares methods. However, statistical Bayesian methods have increasingly been applied in traffic safety studies over recent years, mainly resulting from the availability of the

software package WinBUGS (Spiegelhalter et al. 2003, Lunn et al. 2000). Examples of which include Schlüter et al. (1997), Heydecker & Wu (2001), Miaou & Lord (2003), Carriquiry & Pawlovich (2005), Song et al. (2006), Li et al. (2008), Lan et al. (2009), Persaud et al. (2010) and Cafiso et al. (2010). Bayesian methods account in a better way, than the classical methods, for the uncertainty in the data and provide more detailed causal inferences (Carriquiry & Pawlovich 2005) as well as more flexibility in selecting accident count distributions (Lan et al. 2009). The Bayesian calculation combines prior information and current information to derive estimates for the expected safety of the sites that are being evaluated. In the context of traffic accident analysis, the prior information is the expected accident frequency from a group of similar sites and the current information is the site-specific observed accident frequency.

In a hierarchical Bayesian analysis, the parameters of the prior distributions depend in turn on additional parameters with their own priors that are also referred as hyper-priors, see for e.g. Carlin & Louis (2000) and Gelman et al. (2004). When these hyper-priors densities are chosen to be “vague” or “non-informative” they will guarantee to play a minimal role in the posterior distribution. The rationale for using non-informative prior distributions is often said to be to “let the data speak for themselves”, so that inferences are unaffected by information external to the current data (Gelman et al. 2004). There is already one study performed on Portuguese motorway data employing Bayesian

hierarchical models where non-informative hyper-priors were used, see Azeredo-Lopes & Cardoso (2010). There has been, however, no investigation on the effect of the hyper-prior choice adopted on model parameters on the Portuguese traffic safety research. As Lord & Miranda-Moreno (2008) have found out, the specification of hyper-priors may not be trivial when modeling accident data characterized by a low sample mean and small sample size. Under these conditions, the use of vague hyper-priors can be problematic leading to inaccurate posterior estimates, and the results sensitive to the distribution choice.

The aim of this paper is to investigate, in the Portuguese context, the performance of alternative hyper-prior distributions for modeling the dispersion parameter in hierarchical Poisson-Gamma models under various sample sizes and time periods. For this purpose a framework to incorporate available evidences from past similar studies developed by Miranda-Moreno et al. (2009) is used to formulate informative hyper-priors for the dispersion parameter. In order to achieve this aim, a simulation study is carried out to compare alternative distribution choices (Gamma, Uniform and Christiansen) and hyper-prior specifications (vague versus informative). The performance of alternative hyper-prior specifications is evaluated according to the model capacity to detect the “true” hotspots and its corresponding Poisson Mean Differences test, Risk, Percentage Deviation Value and the Spearman correlation coefficient. A comparative performance is also made in terms of parameter estimate accuracy and DIC (deviance information criterion) values (Spiegelhalter et al. 2002).

2 HIERARCHICAL BAYESIAN MODELS FOR ACCIDENT DATA

The Bayesian approach treats the mean accident frequency at a site as an unknown quantity and, in the absence of site-specific previous information, the mean accident frequency is described by a prior distribution, the dispersion of which represents its uncertainty. Through the Bayes theorem, the prior distribution is updated with the use of site-specific observations of accident occurrence, resulting in a posterior distribution which represents the knowledge of the mean accident frequency by combining the information from the prior distribution and the observations in an appropriate way (for more details see Gelman et al. 2004 and Congdon 2006).

Several models have been proposed in traffic safety literature for analyzing accident data. These models range from the standard negative binomial (see Lord 2006, Hauer 2002, Zhang et al. 2007, Park & Lord 2008) to more complex models

such as hierarchical Poisson-Gamma and Poisson LogNormal (Lord & Miranda-Moreno 2008) and recently the Conway-Maxwell-Poisson models as in Park & Lord (2009). However, the hierarchical Poisson-Gamma is the most popular. A simple version of this model can be defined as follows (see Lan et al. 2009 and Lord et al. 2005). Considering Y_i , the number of accidents at site i ($i = 1, \dots, n$), over a given period of time T , as Poisson distributed. Then:

$$Y_i | T, \theta_i \sim \text{Poisson}(T \times \theta_i) \quad (1)$$

$$Y_i | T, \theta_i \sim \text{Poisson}(T \times \mu_i e^{\varepsilon_i}) \quad (2)$$

$$e^{\varepsilon_i} \sim \text{Gamma}(\varphi, \varphi) \quad (3)$$

$$\varphi \sim \pi(\cdot) \text{ and } f(\beta) \propto 1 \quad (4)$$

where θ_i is the mean accident frequency per unit time, μ_i is commonly specified as a function of site-specific attributes or covariates (x_i) as $\mu_i = f(x_i; \beta)$ and $\beta = (\beta_0, \dots, \beta_k)$ is a vector of regression parameters to be estimated from the data.

The model error shown in Equation 3 is assumed to follow a Gamma distribution (prior distribution) with equal shape and scale parameters, leading to:

$$E(e^{\varepsilon_i}) = 1 \text{ and } \text{Var}(e^{\varepsilon_i}) = \frac{1}{\varphi} \quad (5)$$

An important element in this model is the hyper-prior distribution assumed on the dispersion parameter φ and denoted by $\pi(\cdot)$ in Equation 4. In the same Equation, $f(\beta)$ denotes the prior density on the regression coefficients β , which is commonly assumed to be flat or non-informative.

The parameter estimates of the model are obtained by posterior inference resulting from Markov Chain Monte Carlo (MCMC) simulation methods such as Gibbs sampling and Metropolis Hasting algorithms (see Gelman et al. 2004) which are implemented in the WinBUGS software. The posterior estimates of the mean accident frequencies are afterwards ranked and used as a criterion for selection of sites for investigation, i.e. high risk sites. The type of hyper-prior distribution will influence the posterior estimate and hence the identification of hotspot sites.

In this study it is investigated the results of using three common hyper-priors on the dispersion parameter φ suggested and employed by Miranda-Moreno et al. (2009):

$$\pi(\varphi; a, b) = \frac{b^a}{\Gamma(a)} \varphi^{a-1} e^{-b\varphi}, \quad \varphi > 0, a > 0, b > 0 \quad (6)$$

$$\pi(\phi; a) = \frac{a}{(a + \phi)^2}, \quad \phi > 0, a > 0 \quad (7)$$

$$\pi(\phi; a) = \frac{1}{b - a}, \quad a < \phi < b \quad (8)$$

Equations 6, 7 and 8 represent the Gamma, Christiansen and Uniform distributions, respectively. The parameters a and b in each equation are called hyper-parameters. The distribution represented by Equation 7 was first suggested by Christiansen & Morris (1997) where a is the hyper-prior guess for the median of the dispersion parameter ϕ .

2.1 Incorporation of prior knowledge

The incorporation of prior knowledge on $\pi(\cdot)$ in Equation 4 was performed according to the methodology followed by Miranda-Moreno et al. (2009). It includes two main steps used for the construction of informative hyper-priors (Schlüter et al. 1997). It requires the use of information about previous studies concerning the mean and plausible range of ϕ . Considering a Gamma distribution for ϕ with mean and variance given, respectively, by:

$$\eta = E(\phi) = a/b \text{ and } \sigma^2 = Var(\phi) = a/b^2 \quad (9)$$

where a and b are the Gamma parameters in Equation 6 assumed now to be fixed. Combining the equations results in:

$$a = \eta \times b \text{ and } b = \eta / \sigma^2 \quad (10)$$

In order to obtain previous information about the parameter ϕ it is necessary to have access to studies concerning accident data from similar roadway elements and jurisdictions. Given that there are not many published studies involving Portuguese motorway data, the values used for the incorporation of prior knowledge were the dispersion parameters obtained by Azeredo-Lopes & Cardoso (2010).

3 SIMULATION STUDY

The reason to use simulated data for investigating whether the proposed models work well for identifying the high accident risk locations was best described by Cheng & Washington (2005) who stated that, with real data, the analyst never knows beforehand which sites are truly hazardous, thus leading to a difficult situation when trying to count false positives and negatives. However, with simulation it is possible to establish sites that are hazardous and assess whether the methods employed can correctly identify them.

3.1 Posterior analysis and accident risk estimator

Miranda-Moreno & Fu (2009) state that the most popular ranking criteria, or safety measures, for hotspot identification are the posterior expected number of accidents, the posterior probability of excess, the posterior expectation of ranks and the posterior probability that one site is the worst. The posterior expectation of θ_i , perhaps the most popular in the safety literature, is denoted and defined as:

$$\tilde{\theta}_i = E[\theta_i | y_i] = \int_0^{\infty} \theta_i p(\theta_i | y) d\theta_i \quad (11)$$

where $p(\theta_i | y)$ is the posterior distribution of θ_i . This criterion is a point estimate of the underlying mean number of accidents on the long term. To select a list of sites for safety analyses the n sites under analysis are sorted based on their posterior mean number of accidents and then select the top r locations ($r < n$) as hotspots. The r is the number of locations that exceed a certain cutoff or threshold value usually obtained by finding the 90% probability quantile, or other suitable quantiles taking into account available budgets. In this study the 50% quantile was chosen. This procedure is the most appropriate when the aim is to identify a list of sites exceeding a certain critical value, denoted here by c . This strategy is represented by the following rule. If:

$$\tilde{\theta}_i > c \quad (12)$$

then site i belongs to the high risk locations list. Thus, the “true” hotspots are first identified from the true accident frequency distribution, which are then compared with the “detected” hotspots determined with alternative hyper-prior distributions. The performance of alternative distributions on ϕ is evaluated by comparing the “detected” with the “true” hotspots.

3.2 Data

The data consisted of measurements recorded on the Portuguese motorway network for periods of one year ranging from 1999 to 2007, subdivided in bidirectional segments with 500 meters of length. The traffic volume and the number of accidents were recorded for each segment (the x_i specific attributes). The data corresponding to each year included around 4600 motorway sections (segments that contained missing values were excluded). Two new sets of data were created by aggregating years 2003 to 2007 ($T = 5$ years) and 2005 to 2007 ($T = 3$ years).

3.3 Simulation design

The simulation methodology considers sets of sites with known safety status. This means that the model parameters and probability distribution of mean accident frequency at each site are known and are used to create the “true” high risk locations and to generate accident data. These generated accident data is afterwards used for identifying the high risk locations after model fitting, i.e. the “detected” hotspots.

The simulation experiment consists of the following specific steps.

Step 1. Randomly selecting n sites from real data.

From each set of data ($T = 3$ and $T = 5$) further sets of n sites ($n = 50, 100$ and 400) were randomly selected.

Step 2. Generating the “true” mean frequency for each site i .

Assume that the accidents at the given locations follow a Poisson-Gamma model with known fixed parameters (these parameters were defined previously from the data corresponding to $T = 3$ and $T = 5$):

$$\hat{\phi}, \hat{\beta} \quad (13)$$

The “true” mean accident frequency at each site i is generated as follows:

$$e^{\varepsilon_i} | \hat{\phi} \sim \text{Gamma}(\hat{\phi}, \hat{\phi}) \quad (14)$$

$$\theta_i^{rue} = \mu_i e^{\varepsilon_i} \quad (15)$$

$$\mu_i = f(x_i; \hat{\beta}) \quad (16)$$

Step 3. Obtaining the “true” list of hotspots.

A critical value c is specified depending on the data set under analysis. This c -value is the 50% quantile of the mean accident frequencies (obtained previously for each data set $T = 3$ and $T = 5$). Then the selection criterion in Equation 12 is applied. That is, if each value obtained by Equation 15 (the “true” θ_i) for each i is greater than c then site i is defined as a “true” hotspot.

Step 4. Generating random Poisson samples.

For each site, simulate accident data based on the “true” θ_i , obtained in Equation 15:

$$Y_i^{sim} | \theta_i^{rue} \sim \text{Poisson}(\theta_i^{rue}) \quad (17)$$

Step 5. Obtaining the posterior expectation and the detected hotspots.

Given the accident data simulated in Equation 17 and the site-specific attributes x_i , the various models with hyper-priors of types given by

Equations 6, 7 and 8, as well as a non-informative Gamma are fitted.

The estimates from the model’s parameters (β and ϕ) and the posterior expectation of θ_i (given by Equation 11) are obtained with the use of Markov Chain Monte Carlo (MCMC) simulation methods which are already implemented in the WinBUGS software package. Once the posterior expectation of θ_i is obtained for each site and for each model considered (with particular hyper-prior specifications), the “detected” hotspots are identified if Equation 12 holds, else the site is a “non-detected” hotspot.

Step 6. Computation of the performance evaluation criteria.

For each model, the “true” hotspot list obtained in Step 3 is compared with the “detected” hotspot list obtained in Step 5 by employment of the performance evaluation criteria.

In order to obtain statistically reliable estimates, each scenario is replicated 25 times repeating steps 1 to 6. The average of the 25 replications is reported in chapter 5.

3.4 Computational details

The simulated data were generated with the software R (R Development Core Team 2010). The posterior quantities of interest were estimated using WinBUGS which allows Bayesian analysis to be performed using MCMC methods. A connection between the two softwares was made with the R package R2 WinBUGS developed by Sturtz et al. (2005). The Negative Binomial model was fitted to the observed data using the R MASS package developed by Venables & Ripley (2002).

The simulations were replicated 25 times (Geedipally & Lord, 2010) and afterwards the averages of the various performance criteria were computed.

The MCMC simulations had the following specifications: 3 Markov chains, 11000 iterations per chain, 9000 burn-in iterations, 5 iterations thinned, resulting in a total of 1200 iterations used for parameter estimation. These ensured that the MCMC converged to the target distributions. Verifying and certifying that convergence occurs is essential when the aim is to produce reliable results for the posterior distribution. In order to certify that the 25 simulations converged, the statistic *Rhat* of the Gelman-Rubin diagnostic was used, for more details see Gelman et al. (2004), Congdon (2006) and Ntzoufras (2009).

3.5 Model specifications

The functional form of Equation 16, also used by Persaud et al. (2004) and Cafiso et al. (2010) was the following:

$$\mu_i = \exp(\text{offset}(\text{Length}_i) + \beta_0 + \beta_1 \times \log(\text{AADT}_i)) \quad (18)$$

where AADT_i is the average annual daily traffic and Length_i is the length of the segment (the x_i specific attributes) and β_0 and β_1 are the regression coefficient parameters.

The “true” regression estimates and the “true” dispersion parameter (see Equation 13) obtained before the simulation process, are displayed in Table 7 (in the *Real values* row) for $T = 3$ and $T = 5$ years.

The informative hyper-prior parameters were built based on past evidences reported in Azeredo-Lopes & Cardoso (2010) according to the method described in section 2.1.

The prior distributions for the β_j parameters were assumed non-informative $N(0, 10^3)$ to reflect the lack of precise knowledge of the value of these coefficients.

The informative and non-informative hyper-priors for φ considered in the study were:

1. Non-informative Gamma with $a = b = 0.001$
2. Informative Christiansen with $a = 1.952$
3. Informative Uniform with $a = 1.5$ and $b = 3.0$
4. Informative Gamma with $a = 14.284$ and $b = 6.705$.

4 PERFORMANCE EVALUATION CRITERIA

The methods used to evaluate the performance of the four types of hyper-prior distributions were those employed by Miranda-Moreno et al. (2009), Miranda-Moreno & Fu (2009) and Cheng & Washington (2008) being the latter two adapted so that two lists of accident frequencies (the “true” and the “simulated”) could be compared.

4.1 Accuracy measures

The accuracy measures will assess the model’s capacity to detect the “true” hotspots. Table 1 shows the classification of outcomes (Type I and Type II errors, of detected and non-detected sites as hotspots) according to any hotspot identification method.

The measures of accuracy defined using the notations displayed in Table 1 are: $\text{FDR} = V/D$, $\text{FNR} = R/(n-D)$, $\text{SENS} = S/n_1$ and $\text{SPEC} = U/n_0$.

Small values of FDR (false discovery rate) and FNR (false negative rate) are expected when a method or a model perform well. SENS is the sensitivity or proportion of sites that have been correctly detected as hotspots, it can also be interpreted as the method capacity to detect a

Table 1. Possible outcomes when n sites are classified by a given hotspot identification method.

	Number of sites “detected” as non hotspots	Number of sites “detected” as hotspots	Total
Number of “true” non-hotspots	U	V	n_0
Number of “true” hotspots	R	S	n_1
	n-D	D	n

“true” hotspot in a group under analysis. Miranda-Moreno et al. (2007) argue that values of SENS greater than 0.80 are difficult to obtain. SPEC, the specificity, represents the proportion of non-hotspots that have been correctly classified as “non-hotspots”. It is expected to obtain SPEC values close to 1.

Note that the lower the FDR and FNR are the better the results are; conversely, the higher the SENS and SPEC are, the better the outcome is.

4.2 Poisson mean differences test

The Poisson means differences test (PMD) was developed by Cheng & Washington (2008) with the aim to differentiate false identifications and quantify the adverse consequences resulting from erroneous identification of unequal importance. This test statistic is the sum of the absolute difference of the “true” θ_i (Equation 15) associated with the falsely identified sites and critical value of the “true” θ_i . The latter was considered to be the 90% quantile of all the “true” θ_i . The model with the larger PMD is the less desirable (Geedipally & Lord 2010).

4.3 Risk

Risk is the proportion of the total number of errors (Type I and Type II) and the number of sites under analysis. Risk values close to zero are expected when a method or a model performs well (Geedipally & Lord 2010).

4.4 Percentage deviation value

The percentage deviation value described by Miranda-Moreno & Fu (2009) was adapted to quantify the effects of changing the hyper-prior specifications. It is calculated by dividing the number of accident locations that are common in the “true” and “simulated” lists of hotspots by the total number of hotspots selected from

the top of a list obtained by sorting the list of the mean accident frequencies for a specific model according to the list of the “true” accident frequencies.

The percentage deviation only gives information about the number of hotspot locations that do not appear in both “true” and simulated ranked lists. Therefore, the greater the value the worse the outcome since it suggests that there are more hotspots that do not appear in both lists, that is, the model results in a higher number of hotspots with respect to the “true” list. The total number of hotspots selected from the top of this list was chosen to be 3, 5, 8, 10 and 15 to ensure that various possibilities were taken into account.

4.5 Spearman correlation coefficient

Miranda-Moreno & Fu (2009) employed the Spearman correlation coefficient, ρ , to evaluate the degree of association between the “true” and “simulated” ranks ordered on the basis of any hotspot identification technique. The value of ρ can vary from +1 to -1. A value close to +1 suggests that the “true” versus the “simulated” ranks are positively and linearly related. In the present context ρ evaluates the degree of association between the “true” and the “simulated” ranks for the four types of hyper-prior distributions considered.

5 RESULTS

The simulation results for the various scenarios considered are summarized in Tables 2 to 9 and represent the mean values of the 25 simulations performed.

Table 2. Hotspot identification errors for $n = 50$ sites and $T = 3$ and 5 years.

Hyper-prior	FDR	FNR	SENS	SPEC	PMD	Risk
<i>Time period = 3 years</i>						
(1)	0.392	0.064	0.266	0.980	1.370	0.080
(2)	0.372	0.060	0.338	0.972	1.408	0.083
(3)	0.349	0.065	0.271	0.981	1.345	0.079
(4)	0.372	0.063	0.281	0.980	1.344	0.079
<i>Time period = 5 years</i>						
(1)	0.738	0.017	0.188	0.989	0.696	0.027
(2)	0.567	0.014	0.354	0.988	0.606	0.026
(3)	0.583	0.016	0.208	0.993	0.436	0.022
(4)	0.458	0.015	0.271	0.994	0.403	0.021

(1) Non-info. Gamma, (2) Info. Christiansen, (3) Info. Uniform, (4) Info. Gamma.

Table 3. Hotspot identification errors for $n = 100$ sites and $T = 3$ and 5 years.

Hyper-prior	FDR	FNR	SENS	SPEC	PMD	Risk
<i>Time period = 3 years</i>						
(1)	0.364	0.040	0.388	0.977	1.655	0.060
(2)	0.353	0.036	0.464	0.976	1.571	0.057
(3)	0.312	0.040	0.398	0.980	1.570	0.058
(4)	0.317	0.040	0.395	0.978	1.602	0.060
<i>Time period = 5 years</i>						
(1)	0.311	0.026	0.294	0.992	1.343	0.033
(2)	0.332	0.021	0.469	0.988	1.348	0.031
(3)	0.285	0.024	0.375	0.991	1.305	0.032
(4)	0.328	0.023	0.380	0.994	1.336	0.032

(1) Non-info. Gamma, (2) Info. Christiansen, (3) Info. Uniform, (4) Info. Gamma.

Table 4. Hotspot identification errors for $n = 400$ sites and $T = 3$ and 5 years.

Hyper-prior	FDR	FNR	SENS	SPEC	PMD	Risk
<i>Time period = 3 years</i>						
(1)	0.375	0.054	0.414	0.973	6.736	0.075
(2)	0.371	0.052	0.436	0.972	6.623	0.074
(3)	0.346	0.054	0.411	0.976	6.566	0.072
(4)	0.347	0.054	0.416	0.976	6.570	0.073
<i>Time period = 5 years</i>						
(1)	0.288	0.022	0.396	0.993	4.176	0.028
(2)	0.270	0.021	0.411	0.993	4.079	0.027
(3)	0.286	0.022	0.376	0.994	4.134	0.028
(4)	0.284	0.022	0.387	0.994	4.135	0.028

(1) Non-info. Gamma, (2) Info. Christiansen, (3) Info. Uniform, (4) Info. Gamma.

The best overall results were obtained by the models with informative Gamma and Christiansen hyper-priors (see rows (4) and (2) in Tables 2 and 3). The model performing worse at detecting the “true” hotspots involved the non-informative Gamma (rows (1) in Tables 2 and 3). For $n = 400$ (Table 4) the type of hyper-prior considered becomes less relevant. This also agrees with previous works results, e.g. Lord (2006) and Miranda-Moreno et al. (2009), where the former states that the problem of low mean can also be solved by increasing the number of roadway elements involved in the analysis, i.e. increase the sample size.

Overall the best outcomes shown in Tables 2 to 4 were obtained when the data analyzed

Table 5. Percentage deviation and ranking correlation for $T = 3$ years and for $n = 50, 100$ and 400 sites.

Hyper-prior	Percentage deviation, Ranking correlation					ρ
	$r = 3$	$r = 5$	$r = 8$	$r = 10$	$r = 15$	
$n = 50$						
(1)	0.587	0.504	0.515	0.468	0.397	0.648
(2)	0.640	0.536	0.515	0.488	0.411	0.644
(3)	0.600	0.480	0.505	0.468	0.419	0.649
(4)	0.613	0.496	0.510	0.468	0.416	0.629
$n = 100$						
(1)	0.547	0.512	0.475	0.452	0.419	0.678
(2)	0.547	0.520	0.470	0.440	0.435	0.682
(3)	0.507	0.504	0.470	0.432	0.413	0.685
(4)	0.547	0.512	0.475	0.452	0.419	0.678
$n = 400$						
(1)	0.507	0.448	0.455	0.488	0.491	0.711
(2)	0.493	0.440	0.425	0.472	0.467	0.715
(3)	0.507	0.440	0.440	0.472	0.464	0.713
(4)	0.507	0.448	0.435	0.484	0.461	0.715

(1) Non-info. Gamma, (2) Info. Christiansen, (3) Info. Uniform, (4) Info. Gamma.

Table 6. Percentage deviation and ranking correlation for $T = 5$ years and for $n = 50, 100$ and 400 sites.

Hyper-prior	Percentage deviation, Ranking correlation					ρ
	$r = 3$	$r = 5$	$r = 8$	$r = 10$	$r = 15$	
$n = 50$						
(1)	0.520	0.472	0.490	0.468	0.413	0.576
(2)	0.547	0.520	0.495	0.468	0.421	0.598
(3)	0.520	0.512	0.470	0.444	0.395	0.586
(4)	0.533	0.504	0.475	0.472	0.413	0.562
$n = 100$						
(1)	0.560	0.504	0.420	0.432	0.384	0.735
(2)	0.547	0.504	0.445	0.436	0.371	0.743
(3)	0.547	0.480	0.410	0.428	0.368	0.744
(4)	0.547	0.496	0.415	0.428	0.381	0.745
$n = 400$						
(1)	0.387	0.464	0.460	0.452	0.453	0.727
(2)	0.373	0.440	0.470	0.440	0.440	0.730
(3)	0.387	0.456	0.440	0.440	0.456	0.731
(4)	0.387	0.440	0.440	0.444	0.448	0.731

(1) Non-info. Gamma, (2) Info. Christiansen, (3) Info. Uniform, (4) Info. Gamma.

was aggregated over $T = 5$ years as opposed to $T = 3$.

There seems to be no conclusion regarding the sample sizes, n , as far as the Risk is concerned (Tables 2 to 4).

For $n = 50$ the best outcomes of the percentage deviation values are obtained by the informative Uniform and Christiansen hyper-priors (Tables 5 and 6). When the sample sizes increase not many differences are found amongst the percentage

Table 7. True versus posterior estimated values of model parameters ($n = 50$ sites and $T = 3$ and 5 years).

Hyper-prior	β_0 (Rhat)	β_l (Rhat)	φ (Rhat)	DIC
<i>Time period = 5 years</i>				
<i>Real values</i>	-11.284	1.022	1.617	-
(1)	-14.557 (1.319)	1.202 (1.301)	93.276 (1.134)	61.474
(2)	-14.308 (1.558)	1.181 (1.499)	1.673 (1.053)	59.341
(3)	-13.411 (1.333)	1.094 (1.312)	2.229 (1.002)	62.118
(4)	-13.415 (1.481)	1.095 (1.457)	2.095 (1.006)	62.200
<i>Time period = 5 years</i>				
<i>Real values</i>	-11.083	0.963	1.814	-
(1)	-12.510 (1.356)	0.929 (1.358)	103.219 (1.131)	64.862
(2)	-13.231 (1.448)	0.996 (1.424)	1.848 (1.042)	62.321
(3)	-12.886 (1.266)	0.961 (1.266)	2.242 (1.003)	64.350
(4)	-12.957 (1.376)	0.923 (1.374)	2.124 (1.003)	64.410

(1) Non-info. Gamma, (2) Info. Christiansen, (3) Info. Uniform, (4) Info. Gamma.

Table 8. True versus posterior estimated values of model parameters ($n = 100$ sites and $T = 3$ and 5 years).

Hyper-prior	β_0 (Rhat)	β_l (Rhat)	φ (Rhat)	DIC
<i>Time period = 3 years</i>				
<i>Real values</i>	-11.284	1.022	1.617	-
(1)	-14.067 (1.297)	1.170 (1.302)	139.465 (1.395)	109.469
(2)	-14.094 (1.421)	1.173 (1.374)	2.035 (1.050)	105.747
(3)	-14.037 (1.332)	1.167 (1.301)	2.236 (1.003)	108.906
(4)	-13.974 (1.335)	1.160 (1.325)	2.112 (1.004)	108.566
<i>Time period = 5 years</i>				
<i>Real values</i>	-11.083	0.963	1.814	-
(1)	-12.134 (1.313)	0.909 (1.302)	141.562 (1.228)	137.946
(2)	-12.402 (1.299)	0.933 (1.299)	2.769 (1.039)	134.613
(3)	-12.460 (1.256)	0.937 (1.254)	2.235 (1.003)	136.872
(4)	-12.063 (1.247)	0.903 (1.229)	2.118 (1.004)	136.429

(1) Non-info. Gamma, (2) Info. Christiansen, (3) Info. Uniform, (4) Info. Gamma.

Table 9. True versus posterior estimated values of model parameters ($n = 400$ sites and $T = 3$ and 5 years).

Hyper-prior	β_0 (Rhat)	β_l (Rhat)	φ (Rhat)	DIC
<i>Time period = 3 years</i>				
<i>Real values</i>	-11.284	1.022	1.617	-
(1)	-12.114 (1.210)	0.996 (1.208)	35.445 (1.264)	491.569
(2)	-12.115 (1.178)	0.996 (1.177)	2.007 (1.063)	484.863
(3)	-12.154 (1.149)	1.000 (1.152)	2.175 (1.007)	490.991
(4)	-12.111 (1.299)	0.996 (1.292)	2.037 (1.011)	489.734
<i>Time period = 5 years</i>				
<i>Real values</i>	-11.083	0.963	1.814	-
(1)	-12.847 (1.308)	0.973 (1.311)	48.531 (1.603)	523.052
(2)	-12.833 (1.373)	0.972 (1.366)	2.394 (1.063)	517.126
(3)	-12.821 (1.210)	0.970 (1.206)	2.181 (1.007)	523.091
(4)	-12.857 (1.315)	0.974 (1.307)	2.047 (1.010)	521.608

(1) Non-info. Gamma, (2) Info. Christiansen, (3) Info. Uniform, (4) Info. Gamma.

deviations. However, the Spearman correlation coefficient, ρ , increases with increasing n , suggesting that the greater the samples the more positively and linearly related are the “true” versus the estimated hotspot ranks. The number of years does not seem to influence any of the measures shown in Tables 5 and 6.

The hyper-priors employed do not seem to influence considerably the posterior estimates of the β parameters (Tables 7, 8 and 9). When the non-informative Gamma is used the posterior estimates of the dispersion parameter ϕ differ substantially from its “true” values. This illustrates that the use of informative hyper-priors can lead to much more accurate parameter estimates than non-informative ones.

In all the scenarios considered the mean $Rhat$ values for all the parameter estimates are close to 1 demonstrating that, on average, the MCMC simulations converged to the sought target distribution.

The models with informative Christiansen hyper-priors produced the smaller DIC values and are thus the best fitted models out of the four considered.

6 OVERALL CONCLUSIONS

This study shows that when developing hierarchical Bayesian accident prediction models with the aim of using its results for hotspot detection in Portuguese motorways it is best to employ informative hyper-priors, namely the Gamma and Christiansen’s for small sample sizes (between 50 and 100). It was also found that more accurate results were obtained when more years of aggregated data were analyzed.

REFERENCES

Azeredo-Lopes, S. & Cardoso, J. 2007a. Accident prediction models for Portuguese motorways. *Informação Científica, LNEC, Portugal*.

Azeredo-Lopes, S. & Cardoso, J. 2007b. Accident prediction models for Portuguese single carriageway roads. *Informação Científica, LNEC, Portugal*.

Azeredo-Lopes, S. & Cardoso, J. 2009. Accident prediction models for bidirectional data on Portuguese motorways. *Informação Científica, LNEC, Portugal*.

Azeredo-Lopes, S. & Cardoso, J. 2010. Modelos de estimativa de frequência de acidentes e vítimas para dados bidireccionais de auto-estradas portuguesas – uma abordagem Bayesiana. *Relatório, LNEC, Portugal*.

Cafiso, S., Graziano, A.D., Silvestro, G. D., Cava, G.L. & Persaud, B. 2010. Development of comprehensive accident models for two-lane rural highways using

exposure, geometry, consistency and context variables. *Accident Analysis and Prevention*. 42: 1072–1079.

Carlin, B. & Louis, T. 2000. Bayes and Empirical Bayes Methods for Data Analysis. Chapman & Hall.

Carriquiry, A. & Pawlovich, M. 2005. From empirical Bayes to full Bayes: methods for analyzing traffic safety data. http://www.dot.state.ia.us/crashanalysis/pdfs/eb_fb_comparison_whitepaper_october2004.pdf (accessed August 16, 2010).

Cheng, W. & Washington, S. 2005. Experimental evaluation of hotspot identification methods. *Accident Analysis and Prevention*. 37: 870–881.

Cheng, W. & Washington, S. 2008. New criteria for evaluating methods of identifying hot spots. *Transportation Research Record: Journal of the Transportation Research Board*. 2083: 76–85.

Christiansen, C. & Morris, C. 1997. Hierarchical Poisson Regression Models. *Journal of the American Statistical Association*. 92: 618–632.

Congdon, P. 2006. *Bayesian Statistical Modeling*. Wiley, Chichester.

Elvik, R. 2008. A survey of operational definitions of hazardous road locations in some European countries. *Accident Analysis and Prevention*. 4: 1830–1835.

Geedipally, S. & Lord, D. 2010. Hot Spot Identification by Modeling Single-Vehicle and Multi-Vehicle Crashes Separately. *Transportation Research Record: Journal of the Transportation Research Board*. 10: 2563.

Gelman, A., Carlin, J., Stern, H. & Rubin, D. 2004. *Bayesian Data Analysis*. Chapman & Hall, London.

Geurts, K., Wets, G., Brijs, T., Vanhoof, K. & Karlis, D. 2006. Ranking and selecting dangerous crash locations: correcting for the number of passengers and Bayesian ranking plots. *Journal of Safety Research*. 37: 83–91.

Hauer, E. 2002. Estimating safety by the empirical Bayes method: a tutorial. *Transportation Research Record*. 1784: 27–31.

Heydecker, B. & Wu, J. 2001. Identification of sites for road accident remedial work by Bayesian statistical methods: an example of uncertain inference. *Advances in Engineering Software*. 32: 859–869.

Lan, B., Persaud, B., Lyon, C. & Bhim, R. 2009. Validation of a Full Bayes methodology for observational before-after road safety studies and application to evaluation of rural signal conversions. *Accident Analysis and Prevention*. 41 (3): 574–580.

Li, W., Carriquiry, A., Pawlovich, M. & Welch, T. 2008. The Choice of Statistical Models in Road Safety Countermeasures Effectiveness Studies in Iowa. *Accident Analysis and Prevention*. 40: 1531–1542.

Lord, D. 2006. Modeling motor vehicle crashes using Poisson-Gamma models: examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis and Prevention*. 38: 751–766.

Lord, D., Washington, S. & Ivan, J. 2005. Poisson, Poisson-Gamma and Zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. *Accident Analysis and Prevention*. 37: 35–46.

Lord, D. & Miranda-Moreno, L. 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-Gamma models for modeling motor vehicle crashes: a Bayesian perspective. *Safety Science*. 46 (5): 751–770.

- Lunn, D., Thomas, A., Best, N. & Spiegelhalter, D. 2000. WinBUGS – a Bayesian modelling framework: concepts, structure and extensibility. *Statistics and Computing*. 10: 325–337.
- Miaou, S.-P. & Song, J. 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. *Accident Analysis and Prevention*. 37: 699–720.
- Miranda-Moreno, L., Labbe, A. & Fu, L. 2007. Bayesian multiple testing procedures for hotspot identification. *Accident Analysis and Prevention*. 39: 1192–1201.
- Miranda-Moreno, L. & Fu, L. 2009. Traffic Study: Empirical Bayes or Hierarchical Bayes? In Sloboda, B.W. (ed.), *Transportation Statistics*: 21–42. J. Ross Publishing, U.S.A.
- Miranda-Moreno, L., Lord, D. & Fu, L. 2009. Bayesian road safety analysis: incorporation of past experiences and effect of hyper-prior choice. *Paper presented at the 87th Annual Meeting of the Transportation Research Board*. 08–1788. Submitted for publication in *Journal of Safety Research*.
- Ntzoufras, I. 2009. *Bayesian Modeling Using WinBUGS*. Wiley, New Jersey.
- Park, B. & Lord, D. 2008. Adjustment of the maximum likelihood estimate of the negative binomial dispersion parameter. *Transportation Research Record*. 2061:9–19.
- Park, B. & Lord, D. 2009. Application of finite mixture models for vehicle crash data analysis. *Accident Analysis and Prevention*. 41 (4): 683–691.
- Persaud, B., Retting, R. & Lyon, C. 2004. Crash reduction following installation of centerline rumble strips on rural two-lane roads. *Accident Analysis and Prevention*. 36: 1073–1079.
- Persaud, B., Lan, B., Lyon, C. & Bhim, R. 2010. Comparison of empirical Bayes and full Bayes approaches for before-after road safety evaluations. *Accident Analysis and Prevention*. 42: 38–43.
- R Development Core Team. 2010. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. ISBN 3–900051–07–0, URL <http://www.R-project.org>. Vienna, Austria.
- Schlüter, P., Deely, J. & Nicholson, A. 1997. Ranking and selecting motor vehicle accident sites by using a hierarchical Bayesian model. *Journal of the Royal Statistical Society, Series D*. 46, No.3: 293–316.
- Song, J., Ghosh, M., Miaou, S. & Mallick, B. 2006. Bayesian Multivariate Spatial Models for Roadway Traffic Crash Mapping. *Journal of Multivariate Analysis*. 97: 246–273.
- Spiegelhalter, D., Best, N., Carlin, B. & van der Linde, A. 2002. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, B*, 64: 583–639.
- Spiegelhalter, D., Thomas, D., Best, N. & Lunn, D. 2003. WinBUGS Version 1.4 User Manual. *MRC Biostatistics Unit*, Cambridge, U.K. <http://www.mrc-cam.ac.uk/bugs>.
- Sturtz, S., Ligges, U. & Gelman, A. 2005. R2 WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software* 12 (3): 1–16.
- Venables, W. & Ripley, B. 2002. *Modern Applied Statistics with S*: Fourth Edition, Springer, New York, ISBN 0–387–95457–0, <http://www.stats.ox.ac.uk/pub/MASS4>.
- Zhang, Y., Ye, Z. & Lord, D. 2007. Estimating the dispersion parameter of the negative binomial distribution for analyzing crash data using a bootstrapped maximum likelihood method. *Transportation Research Record*. 2019: 15–21.