



**HAL**  
open science

## Bayesian nonparametric spatial prior for traffic crash risk mapping: a case study of Victoria, Australia

Jean-Baptiste Durand, Florence Forbes, Cong Duc Phan, Long Truong, Hien D Nguyen, Fatoumata Dama

► **To cite this version:**

Jean-Baptiste Durand, Florence Forbes, Cong Duc Phan, Long Truong, Hien D Nguyen, et al.. Bayesian nonparametric spatial prior for traffic crash risk mapping: a case study of Victoria, Australia. Australian and New Zealand Journal of Statistics, 2022, Special Issue: Geoff McLachlan Festschrift, 64 (2), pp.171-204. 10.1111/anzs.12369 . hal-03138803v2

**HAL Id: hal-03138803**

**<https://inria.hal.science/hal-03138803v2>**

Submitted on 7 Sep 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Bayesian nonparametric spatial prior for traffic crash risk mapping: a case study of Victoria, Australia

J.-B. Durand<sup>1\*</sup>, F. Forbes<sup>1</sup>, C.D. Phan<sup>2</sup>, L. Truong<sup>2</sup>, H.D. Nguyen<sup>2,3</sup>, and F. Dama<sup>1</sup>  
*Inria, LJK, Statify Team, La Trobe University, University of Queensland*

## Summary

We investigate the use of Bayesian nonparametric (BNP) models coupled with Markov random fields (MRF) in a risk mapping context, to build partitions of the risk into homogeneous spatial regions. In contrast to most existing methods, the proposed approach does not require an arbitrary commitment to a specified number of risk classes and determines their risk levels automatically. We consider settings in which the relevant information are counts and propose a so-called BNP Hidden MRF (BNP-HMRF) model that is able to handle such data. The model inference is carried out using a variational Bayes Expectation–Maximisation algorithm and the approach is illustrated on traffic crash data in the state of Victoria, Australia. The obtained results corroborate well with the traffic safety literature. More generally, the model presented here for risk mapping offers an effective, convenient and fast way to conduct partition of spatially localised count data.

*Key words:* Road safety; Traffic crashes; Risk mapping; Bayesian nonparametrics; Markov random field; Variational Bayes Expectation–Maximisation algorithm

## 1. Introduction

Traffic-related injuries and deaths are major problems in contemporary societies. Social economic losses from traffic crashes, in particular from motor vehicle crashes, can be enormous. This makes road and traffic safety a major concern, worldwide. The nondecreasing relationship between crash casualties and population suggests that safety improvements could be gained from a better prediction of crash occurrences. Traffic crashes are complex events involving the interactions of various factors. In particular, since road transport involves distances by nature, most studies call for spatial analysis to account for geographical locations

---

\*Author to whom correspondence should be addressed.

<sup>1</sup> Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Inria Grenoble Rhone-Alpes, 655 av. de l'Europe, 38335 Montbonnot, France

<sup>2</sup> School of Engineering and Mathematical Sciences, La Trobe University, Bundoora, Australia

<sup>3</sup> School of Mathematics and Physics, University of Queensland, St. Lucia, Australia

Email: Jean-Baptiste.Durand@inria.fr

*Acknowledgment.* The authors are partly supported by the Inria project Lander.

Opinions and attitudes expressed in this document, which are not explicitly designated as Journal policy, are those of the author and are *not* necessarily endorsed by the Journal, its editorial board, its publisher Wiley or by the Australian Statistical Publishing Association Inc.

15 and environments in which crashes occur. The goal is often to accurately predict the risks at  
16 different locations (Lord & Mannering 2010) and to link these risk values to other variables  
17 for interpretability, or to assess the impact of several risk factors (Theofilatos & Yannis 2014;  
18 Papadimitriou et al. 2019) and road safety measures (Elvik et al. 2009). The hope is to  
19 identify the potential causal sources of crashes, and to apply appropriate control procedures  
20 and protection measures; see e.g., Truong & Currie (2019).

21 In this work, our goal is not to predict the number of crashes or the related risk, as  
22 per se, although predictions are an output of our model. We primarily aim at highlighting  
23 areas with different risk levels, with respect to various covariates, such as population density,  
24 traffic density, signalisation density, etc. The interest of such partitioning is to highlight spatial  
25 heterogeneity, to locate high risk areas (so-called risk hot spots), and to determine whether the  
26 data exhibit some structure in space that could be analysed or directly interpreted. Moreover,  
27 aggregation of connected regions with low numbers of crashes and the same risk level can  
28 also be used to increase the effective sample size dedicated to estimating that risk level.

29 Such partitions can be obtained by applying risk mapping models. Standard risk  
30 mapping models usually produce a continuous estimation of the risk that requires a post-  
31 processing classification step to obtain clearly delimited risk zones, usually based on an  
32 arbitrary choice of risk levels or of risk intervals. Examples of ways to set these thresholds are  
33 illustrated in the Crash Risk Mapping Technical Specifications report of the European Road  
34 Assessment Programme: <https://eurorap.org/crash-rate-mapping/>.

35 Most statistical methods for risk mapping of aggregated data (e.g., Mollié 1999;  
36 Richardson et al. 1995; Pascutto et al. 2000; Lawson et al. 2000) are based on a Poisson log-  
37 linear mixed model and follow the so-called BYM model of Besag, York & Mollié (1991),  
38 and extended by Clayton & Bernardinelli (1992), which is called the convolution model by  
39 Mollié (1996). This model is based on a Hidden Markov Random Field (HMRF), where the  
40 latent intrinsic risk field is modeled by a Markov field with continuous state space, namely a  
41 Gaussian Conditionally Auto-Regressive (CAR) model. Other developments in this context  
42 concern spatio-temporal mapping (Knorr-Held & Richardson 2003; Robertson et al. 2010;  
43 Lawson & Song 2010) and multivariate risk mapping (Knorr-Held, Rasser & Becker 2002;  
44 MacNab 2010).

45 For all of these procedures, the model inference results in a real-valued estimation of the  
46 risk at each location and one of the main reported limitations is that local discontinuities  
47 in the risk field are not modelled (see e.g., Green & Richardson 2002), leading to  
48 potentially oversmoothed risk maps. Also, in some cases, coarser representations where  
49 areas with similar risk values are grouped are desirable (e.g., Abrial et al. 2005). Grouped  
50 representations have the advantage of providing clearly delimited areas in which more  
51 focused studies could be conducted to better understand the crashes determinants. These areas

52 at risk can be viewed as clusters as in Knorr-Held & Rasser (2000), but we prefer to interpret  
53 them as risk classes, as per the seminal work of Schlattmann & Böhning (1993) and Böhning,  
54 Dietz & Schlattmann (2000), and with additional spatial constraints, by Green & Richardson  
55 (2002) and Alfo, Nieddu & Vicari (2009).

56 Indeed, geographically separated areas representing different clusters can have similar  
57 risks and be grouped in the same class. Consequently, the classes can be less numerous than  
58 the number of clusters, and their interpretation is made easier for decision-makers. Using the  
59 BYM model, it is possible to derive such a grouping from the output, using either fixed risk  
60 ranges (usually difficult to choose in practice), or more automated clustering techniques (see  
61 e.g., Fraley & Raftery 2007). In any case, this post-processing step is likely to be sub-optimal.  
62 For this reason, there have been several attempts to design procedures that can directly model  
63 such a risk classification.

64 Green & Richardson (2002) propose replacing the continuous risk field by a partition  
65 model, involving the introduction of a finite number of risk levels and allocation variables to  
66 assign each area under study to one of these levels. Spatial dependencies are then taken into  
67 account by modeling the allocation variables as a latent discrete state-space Markov field,  
68 such as the Ising (two classes or states) or Potts (more than two classes) model (Chandler  
69 1987; Stoehr 2017). In the same spirit, in the work by Fernandez & Green (2002), the spatial  
70 dependence is pushed one level higher, resulting in a more flexible model, but with a more  
71 difficult parameter estimation problem. These various attempts are based on discrete HMRF  
72 modelling and all use MCMC techniques for inference, which can seriously limit and even  
73 prevent their application to large data sets in a reasonable time.

74 An additional problem that arises in the application of HMRF is that of choosing the  
75 number of states, or classes. When data are independent, the typical approach is to use  
76 a penalised likelihood or information theoretic method, such via the Akaike information  
77 criterion, Bayesian information criterion, integrated likelihood, or other similar approaches.  
78 We note that some of these approaches have been extended to the HMRF framework (see, e.g.,  
79 Forbes & Peyrard 2003). A drawback of such approaches is that they are computationally  
80 wasteful. That is, one is required to estimate models corresponding to different numbers  
81 of classes, and only the best of these models is then used for further inference. In order  
82 to alleviate this shortcoming, fully Bayesian methods such as the reversible jump Markov  
83 chain Monte Carlo (Green 1995; Miller & Harrison 2018) may be considered, although such  
84 procedures may actually require a greater amount of overall computational effort, regardless  
85 of wastefulness.

86 In this work, to handle discontinuities in the spatial structure of the risk without having  
87 to arbitrarily choose their number of classes, we propose to operate in the framework of  
88 Bayesian nonparametric (BNP) methods (Ghosal & Van der Vaart 2017). More specifically,



89 we build on methods recently proposed for the modelling of continuous observations by Lü,  
 90 Arbel & Forbes (2020). We extend the approach, referred to as BNP-HMRF, to the modelling  
 91 of count data. We derive the corresponding variational Bayes Expectation–Maximisation  
 92 (VBEM) algorithm for the model estimation. The approach is then illustrated on traffic crash  
 93 data in the state of Victoria, Australia. The analysis provides risk zones and risk levels that  
 94 are globally coherent with other findings in the literature.

95 The proposed BNP-HMRF model is explained in Section 2. The model implementation  
 96 using variational approximation is detailed in Section 3. The application to crash risk  
 97 mapping of Victoria is detailed in Section 4. An assessment of the Markov random field  
 98 hyperparameter estimation based on simulated data is proposed in Section 5. A conclusion  
 99 and perspectives are provided in Section 6.

## 100 2. BNP-HMRF model for count data

101 The study aims at providing some risk mapping of traffic crashes, based on data  
 102 regarding geographical zones. Since, on average, the number of traffic crashes increases  
 103 with respect to other variables characterising the traffic importance (e.g., population size,  
 104 traffic intensity, length of road network), the numbers of crashes have to be normalised with  
 105 respect to at least one of these variables. The obtained ratios provide quantities that we may  
 106 interpret as risks. One objective is then to account for some spatial heterogeneity regarding  
 107 the observed risks. The model described in the following lines aims at clustering regions with  
 108 similar risks to provide a labelled map, where each label is associated with some risk level.

Consider  $J$  regions, where we let  $y_j$  represent the number of crashes occurring in region  
 $j \in \{1, \dots, J\}$ , characterised by a normalisation variable  $N_j$ , e.g., the population size of  
 region  $j$ . For the sake of clarity, we first assume that there is a finite set of  $K$  risk levels  
 $\mathbf{\Lambda} = \{\lambda_0, \dots, \lambda_{K-1}\}$ , that are ordered so that  $\lambda_k$  is the  $(k + 1)$ th smallest level. Since the  
 risk level associated with region  $j$  is not known in advance, a variable  $z_j \in \{0, \dots, K - 1\}$   
 is introduced to indicate the assigned risk level, i.e.  $z_j = k$ , when region  $j$  is at risk level  
 $\lambda_k$ . When region  $j$  is at risk level  $\lambda_k$ , the number of traffic crashes  $y_j$  is then assumed to be  
 Poisson distributed with mean  $\lambda_k N_j$ . That is,  $y_j$  conditioned on  $z_j = k$  has probability mass  
 function

$$p(y_j | z_j = k; \mathbf{\Lambda}, N_j) = \mathcal{P}(y_j; \lambda_k N_j), \quad (1)$$

109 where  $\mathcal{P}(\cdot; \lambda_k N_j)$  denotes the probability mass function of the Poisson distribution with  
 110 mean parameter  $\lambda_k N_j$ . Generically,  $p(y_j | z_j = k; \mathbf{\Lambda}, N_j)$  is referred to as an emission

111 distribution. As a consequence, the mean number of crashes is a linear function of  $N_j$ :  
 112  $E[y_j|z_j = k; \mathbf{\Lambda}, N_j] = \lambda_k N_j$ .

113 The goal is then to estimate the risk levels  $\mathbf{\Lambda} = \{\lambda_0, \dots, \lambda_{K-1}\}$  and the most likely  
 114 risk mapping of each region, through the most likely values of the variables  $z_j$ s. In practice,  
 115 risk levels are likely to vary smoothly across regions. It is more likely that neighbouring  
 116 regions have the same risk level with possible abrupt changes from one region to another  
 117 if they have contrasting characteristics. Thus, for a better estimation of risk levels, a Markov  
 118 random field (MRF) model is used for the set of labels  $\mathbf{z} = \{z_j, j \in J\}$ , to account for spatial  
 119 dependencies between connected regions.

120 Formally, the regions are seen as the vertices of a graph  $G$ . They are connected by an  
 121 edge in the graph whenever they share a boundary, although other types of connections could  
 122 be considered (e.g., they either share a boundary or have a common neighbour, etc.). The  
 123 probability for neighbouring regions having either a similar or different label is controlled  
 124 by some scalar positive parameter denoted by  $\beta$ . The higher the value of  $\beta$ , the more likely  
 125 neighbouring regions are at the same risk level.

126 The number  $K$  of risk levels is not usually known in advance and has to be chosen  
 127 adaptively by users. To avoid this commitment to a fixed number  $K$ , we propose an extension  
 128 of the model that does not restrict the levels to a finite number  $K$ . This extension is based  
 129 on so called Dirichlet Process Mixtures (Lü, Arbel & Forbes 2020) and is referred to as a  
 130 Bayesian Non-Parametric Hidden Markov Random Field (or more concisely, BNP-HMRF).

131 The BNP-HMRF model is defined as follows. The set of  $J$  regions under consideration is  
 132 associated to a graph structure  $G = (J, E)$ , where each  $j \in J$  corresponds to a node of  $G$  and  
 133 the set of edges  $E$  represents all pairs of regions with a common boundary. The likelihood  
 134 part of the model is given by (1). The observations are counts  $\mathbf{y} = \{y_j, j \in J\}$  distributed  
 135 independently given  $\mathbf{\Lambda}$  and  $\mathbf{z}$  with for every  $j$ ,

$$p(y_j|z_j = k; \mathbf{\Lambda}, N_j) = \mathcal{P}(y_j; \lambda_k N_j).$$

The risk class labels  $\mathbf{z} = \{z_j, j \in J\}$  are assumed to be distributed as a Markov random field  
 on  $G$  with the following distribution:

$$\begin{aligned} p(\mathbf{z}|\beta, \boldsymbol{\pi}) &\propto \exp\left(\sum_{j=1}^J \ln \pi_{z_j} + \beta \sum_{\{i,j\} \in E} \mathbf{1}_{(z_i=z_j)}\right) \\ &= \left(\prod_{j=1}^J \pi_{z_j}\right) \exp\left(\beta \sum_{\{i,j\} \in E} \mathbf{1}_{(z_i=z_j)}\right), \end{aligned} \quad (2)$$

136 where  $\mathbf{1}_{(z_i=z_j)}$  is the indicator function equal to 1 when  $z_i = z_j$  and 0 otherwise,  $\{i, j\} \in E$   
 137 indicates that  $\{i, j\}$  is an edge in  $G$ ,  $\beta$  is some unknown scalar parameter, and the  $\pi_k$ s are  
 138 weights defined for every  $k \geq 0$  as

$$\pi_k(\boldsymbol{\tau}) = \tau_k \prod_{l < k} (1 - \tau_l),$$

where  $\boldsymbol{\tau}^\top = (\tau_0, \tau_1, \dots)$  is a sequence of independent, identically distributed (i.i.d.) random variables with distribution  $\text{Beta}(1, \alpha)$ . The parameter  $\alpha$  is a hyperparameter, which follows a gamma distribution:

$$\alpha | s_1, s_2 \sim \mathcal{G}(s_1, s_2),$$

while each parameter  $\lambda_k$  in (1) is also distributed according to a gamma distribution:

$$\lambda_k | a_k, b_k \sim \mathcal{G}(a_k, b_k).$$

The characterisation of the  $\pi_k$ s corresponds to a stick-breaking construction (see Lemma 3.4 in Ghosal & Van der Vaart 2017 and Sethuraman 1994, for details) and guarantees that  $\sum_{k=0}^{\infty} \pi_k = 1$ , which in turn ensures that the distribution in (2) is a valid Markov field (see Lü, Arbel & Forbes 2020, for details). The complete hierarchical model can thus be stated, for  $\mathbf{z} = \{z_1, \dots, z_J\}$  and  $k = 0, 1, \dots$ , as follows:

$$\begin{aligned} \lambda_k | a_k, b_k &\sim \mathcal{G}(a_k, b_k), \\ \alpha | s_1, s_2 &\sim \mathcal{G}(s_1, s_2), \\ \tau_k | \alpha &\sim \mathcal{B}(1, \alpha), \\ \pi_k(\boldsymbol{\tau}) &= \tau_k \prod_{l=1}^{k-1} (1 - \tau_l), \\ p(\mathbf{z} | \boldsymbol{\tau}, \beta) &\propto \prod_{j \in J} \pi_{z_j}(\boldsymbol{\tau}) \exp \left( \beta \sum_{\{i,j\} \in E} \mathbf{1}_{(z_i=z_j)} \right), \\ y_j | z_j; \boldsymbol{\Lambda}, N_j &\sim \mathcal{P}(y_j; \lambda_{z_j} N_j), \quad \text{for each } j \in J. \end{aligned}$$

139

### 3. Inference using variational approximation

140 In what follows, we propose an adaptation of the VBEM procedure from Lü, Arbel &  
 141 Forbes (2020) to Poisson emission probabilities.

142 The observed counts are denoted by  $\mathbf{y} = \{y_j, j \in J\}$  and the normalising variables  
 143 are denoted by  $\mathbf{N}_J = \{N_j, j \in J\}$ . The set of parameters to estimate divides into

144 two subsets,  $\Phi^\top = (\beta, s_1, s_2, \mathbf{a}^\top)$  with  $\mathbf{a} = \{a_k, b_k, k = 1, \dots\}$ , which are unknown but  
 145 fixed parameters; and  $\Theta^\top = (\alpha^\top, \Lambda^\top)$ , which are random parameters. The hierarchical  
 146 representation of the model above induces additional latent variables  $(\mathbf{z}^\top, \boldsymbol{\tau}^\top)$ . There is  
 147 no formal difference in the treatment of random parameters and latent variables, but it is  
 148 standard to distinguish between them. The term latent variables usually refers to variables  
 149 whose number increases with the number of observations, while parameters are usually of  
 150 fixed dimension. Parameters  $\Phi$  are estimated by an empirical Bayes principle:

$$\hat{\Phi} = \arg \max_{\Phi} p(\mathbf{y}|\Phi) = \arg \max_{\Phi} \int p(\mathbf{y}, \mathbf{z}, \boldsymbol{\tau}, \Theta|\Phi) d\mathbf{z} d\boldsymbol{\tau} d\Theta.$$

151 The random parameters and latent variables are handled via their posterior density  
 152  $p(\mathbf{z}, \boldsymbol{\tau}, \Theta|\mathbf{y}, N_J, \hat{\Phi})$ . This posterior is not available in close-form due to an intractable  
 153 normalising constant. In this work, we use a variational approximation principle to provide  
 154 an approximation of the true posterior. More specifically, the posterior is approximated using  
 155 a member of a family of distributions  $q$  that factorise and that are truncated appropriately to  
 156 exhibit only a finite number of terms. The infinite state space for each  $z_i$  is dealt with by  
 157 choosing a truncation of the state space to a maximum label  $K$  (Blei & Jordan 2006).

158 In practice, this consists of assuming that the variational distribution  $q(\mathbf{z}) = \prod_{j \in J} q(z_j)$   
 159 and that the  $q_{z_j}$ s satisfy  $q_{z_j}(k) = 0$ , for  $k \geq K$ , and that the variational distribution on  $\boldsymbol{\tau}$  also  
 160 factorises as  $q_{\boldsymbol{\tau}}(\boldsymbol{\tau}) = \prod_{k=0}^{K-2} q_{\tau_k}(\tau_k)$ , with the additional condition that  $\tau_{K-1} = 1$ . We thus  
 161 have the variation approximation, below:

$$q(\mathbf{z}, \boldsymbol{\tau}, \Theta) = q_{\alpha}(\alpha) \prod_{j=1}^J q_{z_j}(z_j) \prod_{k=0}^{K-2} q_{\tau_k}(\tau_k) \prod_{k=0}^{K-1} q_{\lambda_k}(\lambda_k). \quad (3)$$

162 In (3),  $K$  does not represent the number of classes that is actually assumed to exist in the  
 163 data but an upper bound on the number. In practice, the exact value of  $K$  is not critical;  $K$   
 164 has only to be fixed to a value large enough so as to be higher than the maximum expected  
 165 number of classes.

166 The variational approximation procedure is justified by the now standard statement (see  
 167 e.g., Lü, Arbel & Forbes 2020) that for every function  $q(\mathbf{z}, \boldsymbol{\tau}, \Theta)$ , the marginal likelihood  
 168 is lower-bounded, where the lower bound is the Kullback-Leibler divergence between  
 169  $p(\mathbf{y}, \mathbf{z}, \boldsymbol{\tau}, \Theta|N_J, \Phi)$  and  $q(\mathbf{z}, \boldsymbol{\tau}, \Theta)$ ,

$$\ln p(\mathbf{y}|\Phi) \geq \mathbb{E}_{q(\mathbf{z}, \boldsymbol{\tau}, \Theta)} \left[ \ln \frac{p(\mathbf{y}, \mathbf{z}, \boldsymbol{\tau}, \Theta|N_J, \Phi)}{q(\mathbf{z}, \boldsymbol{\tau}, \Theta)} \right]. \quad (4)$$

170 The derived variational algorithm is then an alternate maximisation of this bound with  
 171 respect to each factor in  $q(\mathbf{z}, \boldsymbol{\tau}, \Theta)$  and  $\Phi$ . Each maximisation step has some explicit

172 functional expression although from the computational point of view, some steps may require  
 173 further approximations. Their descriptions at a coarser level in terms of blocks of parameters  
 174 is as follows. The iteration index is denoted by  $(r)$  in the successive update formulas:

175 • VE- $\mathbf{z}$ :

$$q_{\mathbf{z}}^{(r)}(\mathbf{z}) \propto \exp \left( \mathbb{E}_{q_{\theta, \tau}^{(r-1)}} [\ln p(\mathbf{y}, \mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\Theta} | \mathcal{N}_J, \boldsymbol{\Phi}^{(r-1)})] \right); \quad (5)$$

176 • VE- $\boldsymbol{\Theta}, \boldsymbol{\tau}$ :

$$q_{\theta, \tau}^{(r)}(\boldsymbol{\Theta}, \boldsymbol{\tau}) \propto \exp \left( \mathbb{E}_{q_{\mathbf{z}}^{(r)}} [\ln p(\mathbf{y}, \mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\Theta} | \mathcal{N}_J, \boldsymbol{\Phi}^{(r-1)})] \right); \quad (6)$$

177 • VM- $\boldsymbol{\phi}$ :

$$\boldsymbol{\Phi}^{(r)} = \arg \max_{\boldsymbol{\phi}} \mathbb{E}_{q_{\mathbf{z}}^{(r)} q_{\theta, \tau}^{(r)}} [\ln p(\mathbf{y}, \mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\Theta} | \mathcal{N}_J, \boldsymbol{\Phi})]. \quad (7)$$

178 Some of these steps are not specific to the Poisson emissions model and have similar  
 179 forms as the ones derived by Lü, Arbel & Forbes (2020), for Gaussian emissions. These steps  
 180 are briefly recalled below, omitting the iteration index for simplicity.

### VE- $\alpha$ step

We have

$$q_{\alpha}(\alpha) \propto p(\alpha | s_1, s_2) \exp \left( \sum_{k=0}^{K-2} \mathbb{E}_{q_{\tau_k}} [\ln p(\tau_k | \alpha)] \right) = \mathcal{G}(\alpha; \hat{s}_1, \hat{s}_2),$$

with

$$\begin{aligned} \hat{s}_1 &= s_1 + K - 1, \\ \hat{s}_2 &= s_2 - \sum_{k=0}^{K-2} \mathbb{E}_{q_{\tau_k}} [\ln(1 - \tau_k)], \end{aligned}$$

181 and

$$\mathbb{E}_{q_{\tau_k}} [\ln(1 - \tau_k)] = \psi(\gamma_{k,2}) - \psi(\gamma_{k,1} + \gamma_{k,2}), \quad (8)$$

182 where  $\psi(\cdot)$  is the digamma function and  $(\gamma_{k,1}, \gamma_{k,2})$  are the parameters defining  $q_{\tau_k}$  (see  
 183 below).

**VE- $\tau_k$  step**

For every  $0 \leq k < K$ ,

$$q_{\tau_k}(\tau_k) = \mathcal{B}(\tau_k; \hat{\gamma}_{k,1}, \hat{\gamma}_{k,2})$$

184 with

$$\hat{\gamma}_{k,1} = 1 + \sum_{j=1}^J q_{z_j}(k) = 1 + n_k,$$

185 and

$$\hat{\gamma}_{k,2} = \frac{\hat{s}_1}{\hat{s}_2} + \sum_{l=k}^{K-1} n_l,$$

186 where we have used the notation  $n_k = \sum_{j=1}^J q_{z_j}(k)$ . Note also that  $\sum_{k=0}^{K-1} n_k = J$ .

187 **VM- $\beta$  step**

188 The handling of parameter  $\beta$  is particularly nuanced. Standard variational approximations  
 189 focus on solving the posterior intractability resulting from the combination of prior  
 190 distributions and likelihoods, both usually tractable. Unfortunately, when dealing with  
 191 Markov random field priors, the issue does not only arise from the prior/likelihood  
 192 combination but already from the intractable normalising constant of the Markov prior. This  
 193 constant cannot be discarded as it depends on  $\beta$ . Thus, the VM- $\beta$  step has no explicit solution  
 194 and requires further approximation. The additional approximation we propose consists of  
 195 approximating the gradient, with respect to  $\beta$ , of the lower bound (4):

$$\sum_{k=0}^{K-1} \sum_{i \sim j} q_{z_j}^{(r)}(k) q_{z_i}^{(r)}(k) - \sum_{k=0}^{K-1} \sum_{i \sim j} \tilde{q}_{z_j}^{(r)}(k|\beta) \tilde{q}_{z_i}^{(r)}(k|\beta), \quad (9)$$

196 with  $\tilde{q}_{z_j}^{(r)}(z_j|k|\beta)$ , defined by:

$$\tilde{q}_{z_j}^{(r)}(z_j = k|\beta) = \frac{\exp(\ln \pi_k(\tilde{\tau}) + \beta \sum_{i \in N(j)} q_{z_i}^{(r)}(k))}{\sum_{l=0}^{\infty} \exp(\ln \pi_l(\tilde{\tau}) + \beta \sum_{i \in N(j)} q_{z_i}^{(r)}(l))}, \quad (10)$$

197 for all  $k = 0, \dots, K-1$ , where  $N(j)$  is the set of neighbours of  $j$  in  $G$  and  $\tilde{\tau} =$   
 198  $\mathbb{E}_{q_{\tilde{\tau}}}[\tau]$ . Note that in the BNP-HMRF setting, the use of  $\tilde{\tau}$  is also an additional necessary  
 199 approximation; details can be found in Lü, Arbel & Forbes (2020). Regarding  $\beta$ , the  
 200 approximation can be interpreted as the transfer to the Potts prior of the variational  
 201 approximation used for the Markov posterior.

202 The remaining steps are specific to Poisson emission distributions.

**VE- $\lambda_k$  step**

For every  $0 \leq k < K$ , from the general result in (6) it follows that

$$\begin{aligned}
 q_{\lambda_k}(\lambda_k) &\propto \exp\left(\mathbb{E}_{q_z} \left[ \ln \left( p(\lambda_k | a_k, b_k) \prod_{j=1}^J p(y_j | z_j, \lambda_{z_j}, N_j) \right) \right]\right) \\
 &\propto p(\lambda_k | a_k, b_k) \exp\left(\sum_{j=1}^J q_{z_j}(k) \mathbb{E}_{q_{z_j}} \left[ \ln p(y_j | \lambda_k, N_j) \right]\right) \\
 &\propto p(\lambda_k | a_k, b_k) \exp\left(\sum_{j=1}^J q_{z_j}(k) \left( -N_j \lambda_k + y_j \ln(N_j \lambda_k) - \ln(y_j!) \right)\right) \\
 &\propto p(\lambda_k | a_k, b_k) \exp\left(\sum_{j=1}^J q_{z_j}(k) \left( -N_j \lambda_k + y_j \ln(\lambda_k) \right)\right).
 \end{aligned}$$

203 Thus  $q_{\lambda_k}$  is a gamma density  $\mathcal{G}(\lambda_k | \hat{a}_k, \hat{b}_k)$ , where

$$\hat{a}_k = a_k + \sum_{j=1}^J y_j q_{z_j}(k) \quad \text{and} \quad \hat{b}_k = b_k + \sum_{j=1}^J N_j q_{z_j}(k).$$

204 This is a consequence of the conjugacy property of gamma priors for Poisson likelihood  
 205 functions.

206 **VE- $Z_j$  step**

207 For  $j \in J$ , the general result in (5) leads to

$$q_{z_j}(z_j) \propto \exp\left(\mathbb{E}_{q_{\lambda_{z_j}}} \left[ \ln p(y_j | \lambda_{z_j}, N_j) \right] + \mathbb{E}_{q_\tau} \left[ \ln \pi_{z_j}(\tau) \right] + \beta \sum_{i \in N(j)} q_{z_i}(z_j) \right), \quad (11)$$

where  $N(j)$  is the set of neighbours of  $j$  in  $G$ , and for  $z_j = k$ ,

$$\mathbb{E}_{q_\tau} \left[ \ln \pi_k(\tau) \right] = \mathbb{E}_{q_{\tau_k}} \left[ \ln \tau_k \right] + \sum_{l=0}^{k-1} \mathbb{E}_{q_{\tau_l}} \left[ \ln(1 - \tau_l) \right],$$

with

$$\mathbb{E}_{q_{\tau_k}} \left[ \ln(\tau_k) \right] = \psi(\gamma_{k,1}) - \psi(\gamma_{k,1} + \gamma_{k,2}),$$

208 where  $\mathbb{E}_{q_{\tau_l}} \left[ \ln(1 - \tau_l) \right]$  is given by (8).

The term  $E_{q_{\lambda_k}} \left[ \ln p(y_j | \lambda_k, N_j) \right]$  is obtained as follows:

$$\begin{aligned} E_{q_{\lambda_k}} \left[ \ln p(y_j | \lambda_k, N_j) \right] &= E_{q_{\lambda_k}} \left[ -N_j \lambda_k + y_j \ln(N_j \lambda_k) - \ln(y_j!) \right] \\ &= -N_j E_{q_{\lambda_k}} \left[ \lambda_k \right] + y_j \ln(N_j) + y_j E_{q_{\lambda_k}} \left[ \ln(\lambda_k) \right] - \ln(y_j!), \end{aligned}$$

209 where due to the fact that  $q_{\lambda_k}$  is a gamma density  $\mathcal{G}(\lambda_k | \hat{a}_k, \hat{b}_k)$  (see above),  $E_{q_{\lambda_k}} \left[ \lambda_k \right] =$   
 210  $\hat{a}_k / \hat{b}_k$  and  $E_{q_{\lambda_k}} \left[ \ln(\lambda_k) \right] = \psi(\hat{a}_k) - \ln(\hat{b}_k)$ .

### 211 **VM- $(a_k, b_k)$ step**

212 For every  $0 \leq k < K$ , the VM- $(a_k, b_k)$  step is conducted by maximising

$$E_{q_{\lambda_k}} \left[ \ln \frac{p(\lambda_k | a_k, b_k)}{q_{\lambda_k}(\lambda_k)} \right],$$

213 which is equivalent to minimising the Kullback-Leibler divergence between  $p(\lambda_k | a_k, b_k)$  and  
 214  $q_{\lambda_k}$ . Since both densities are gamma, the minimum is obtained whenever  $p(\lambda_k | a_k, b_k) = q_{\lambda_k}$   
 215 and thus  $(a_k, b_k) = (\hat{a}_k, \hat{b}_k)$ .

### 216 **VM- $(s_1, s_2)$ step**

217 The VM- $(s_1, s_2)$  step is conducted by maximising

$$E_{q_\alpha} \left[ \ln \frac{p(\alpha | s_1, s_2)}{q_\alpha(\alpha)} \right].$$

218 By a similar argument as for the VM- $(a_k, b_k)$  step,  $(s_1, s_2) = (\hat{s}_1, \hat{s}_2)$ , since both densities  
 219 are in the gamma family.

220 Finally, to start the iterative procedure, initial values of the parameters have to be  
 221 provided and the choice generally has an effect on the quality of the final estimates. Here  
 222 we resort to several runs of the  $k$ -Means algorithm with random initial labels and using the  
 223 ratio  $Y_j / N_j$  (referred to as the ‘case ratio’) as input data. For each run, the final clustering  
 224 yields a value  $\hat{z}_j$  for each  $z_j$ , and  $q_{z_j}(k) = \mathbf{1}_{(k=\hat{z}_j)}$ . Parameters  $\hat{\Lambda}_k$  are initialised as the  
 225 sample means of observations within initial class  $k$ . Initial values of parameters  $a_k$  and  $b_k$  are  
 226 defined so that  $E_{q_{\lambda_k}} \left[ \lambda_k \right] = \hat{\Lambda}_k$  and  $\text{var}_{q_{\lambda_k}} \lambda_k = \min_{\ell, \hat{\Lambda}_\ell > 0} \hat{\Lambda}_\ell$ . For each of these possible  
 227 initial variational parameters from the multiple runs, we keep the value that yields the highest  
 228 free energy (4). The computation of the free energy is detailed in Appendix I. In this step,  
 229 initial values are also required for  $(s_1, s_2, \beta)$ , which are arbitrarily set to  $(1.4, 1, 0)$ .



230

#### 4. Application to traffic crash risk mapping

231

232

233

234

235

236

237

238

239

The results presented here comprise of data from Victoria, Australia. Crash data between 2014 and 2018 were obtained from Victoria’s open data directory (see, Truong & Currie 2019). The number of traffic crashes were aggregated at statistical areas level 2 (SA2s), which are medium-sized functional areas within the Australian Statistical Geography Standard (ASGS). The total number of SA2s in Victoria is 458, excluding several SA2s that have no population according to the Australian Bureau of Statistics (ABS) 2016 census. The SA2 scale has the advantage of offering a good compromise between the number of regions and the spatial resolution, thus ensuring both tractability of algorithms and the interpretability of results.

240

241

242

243

244

245

246

247

248

249

Since an absolute validation of the risk mapping is difficult, we resort to covariates to assess whether risk level classes encode relevant and contrasted characteristics between different classes. In practice, the set of covariates is the same as the set of possible normalising variables. The primary set of covariates is given in the upper part of Table 1. Additional covariates are derived from the primary set of covariates, whenever they appear to make sense as normalising variables. These derivatives are listed in the lower part of Table 1. The road density variable (RnDens) roughly corresponds to the squared ratio of the road length on the edge length of a square region. The VtrFAR19 variable represents the traffic density as opposed to the absolute traffic load VktFAR19. The VtpopFAR19 variable represents the traffic load per inhabitant.

250

#### 4.1. Method

251

252

253

254

255

256

We distinguish between two kinds of analyses: exploratory analyses and full runs of the model. Exploratory analyses aim at quickly assessing the potential of several variables as normalising factors for the risk. In this context we resort to reduced numbers of random initialisations and iterations (1,000 and 300, respectively), while full runs may require more initialisations and iterations for some variables (see hereinafter). The methodology regarding exploratory analyses is the following:

257

258

259

260

261

262

263

1. A principal component analysis (PCA) is performed on the variables listed in Table 1. If a group of variables appears to be strongly correlated to the same axis, we keep only one variable in the group and discard all others.
2. For each possible normalising variable  $N_j$ , the ratio  $(y_j/N_j)_{1 \leq j \leq J}$  are computed, quantised into seven bins and represented on a map. Boundaries between bins are defined by sample quantiles. These bins are anticipated to be close to the expected initial partition.

Table 1. Available covariates: primary set (upper part) and additional derived ones (lower part) with their description sometimes truncated.

Code	Description
AREASQ16	Square area in km <sup>2</sup>
pop16	Population in 2016 (resident)
VktFAR19	Vehicle-Kilometres Travelled on Freeways and Arterial roads
RnNoItcNS	Number of non-signalised intersection (excl. roundabout)
RnNoItcS	Number of signalised intersection
RnNoItcR	Number of roundabout intersection
RnNoItc	Total number of intersection (sum of 3 types)
RnLen	Total length of all road network (km)
RnSZ7080	Length of posted speed 70 - 80 km/h road (km)
RnSZ90100	Length of posted speed 90 - 100 km/h road (km)
RnSZo100	Length of posted speed over 100 km/h road (km)
RnDens	road density ( $RnLen^2/AREASQ16$ )
PopDens	population density ( $pop16/AREASQ16$ )
VtrFAR19	VktFAR19/RnLen, Vehicle-Kilometres Travelled on Freeways etc. per km of road
VtpopFAR19	VktFAR19/pop16, Vehicle-Kilometres Travelled on Freeways etc. per inhabitant
PropSign	proportion of signalised intersections over total number of intersections

- 264 3. The numbers of crashes are plotted against the normalising variable. If the model was  
265 adequately characterising the observed data, for each risk level, the average number of  
266 crashes should be a linear function of  $N_j$ , so that the 2D plot should exhibit sets of  
267 points clustered around a number of lines that pass through the origin. In particular,  
268 if there was only one risk level, the whole data set would exhibit a linear structure.  
269 Therefore, as a first evaluation, a linear regression model is estimated and the regression  
270 line is presented. If needed, non-linear transforms of  $N_j$  are considered (exponential,  
271 logarithm, monomials) until the figure exhibits linear structures. This step is purely  
272 exploratory.
- 273 4. The BNP-HMRF is estimated with  $K = 10$  and provides for each region its probability  
274 to be at each risk level. A segmentation or partition of the graph into  $K$  segments is  
275 then obtained by assigning each region/node to the level with the highest probability.  
276 The segmentation is represented as a  $K$ -color map.
- 277 5. The 2D plot described in Step 3 is drawn again, now using different colours, where  
278 one colour is used for each label. To assess label separation, lines passing through the  
279 origin are drawn using slopes  $E_{q\lambda_k} \left[ \lambda_k \right]$ . These are compared with linear regression  
280 models estimated separately for each class by least squares. Label separation is also  
281 quantified by label marginal entropies represented on a map in grey scale.

282 6. If the number of estimated risk classes in the data is less than  $K$ , the model is  
283 considered as potentially relevant.

284 In the latter case, VBEM is run again with more initial values and iterations if necessary (up  
285 to 5,000 and 600, respectively), again with  $K = 10$ . If the relative growth of free energy falls  
286 under  $10^{-5}$  before the specified maximum number of iterations is reached, the algorithm is  
287 considered to have converged and is stopped. Then analyses of variance are performed on  
288 each covariate (except those discarded in step 1) to interpret the risk classes and to examine  
289 how their values segregate within regions. Tests are performed at significance level 0.01 and  
290  $p$ -values are provided, as well as per-label boxplots, for variables that yield  $p$ -values below  
291 0.01. If the number of risk classes is  $K$ , it may indicate that the BNP-HMRF model has  
292 trouble finding well-separated classes and that the number of classes could be arbitrarily high  
293 when the number of regions increases. In that case, a model with continuous risk levels could  
294 be more appropriate.

295 The data, code, python notebook and system environments (conda, docker) required  
296 to reproduce these analyses and results are available at [https://gitlab.inria.fr/  
297 statify\\_public/anzj-crash/-/tree/master/notebooks](https://gitlab.inria.fr/statify_public/anzj-crash/-/tree/master/notebooks).

## 298 4.2. Results

### 299 4.2.1. Normalising variable selection

300 The first plane in the variable space of the PCA is depicted in Appendix II as  
301 Figure 12. The following variables: RnNoItc, RnLen, RnSZ7080, RnDens, and RnSZ90100  
302 were highly correlated with AREASQ16 and thus were discarded. The variables RnSZo100  
303 and VtpopFAR19 had low variability over the regions and were thus not considered  
304 as informative. Among the remaining variables, we illustrate here our procedure by  
305 considering the population size (variable pop16) and the traffic density (variable VtrFAR19)  
306 as normalising variables in turn. Travelled distance-related variables and population size have  
307 been widely used as direct and indirect crash exposure, respectively, in previous crash risk  
308 analyses (Lord & Mannering 2010; Elvik 2014; Truong & Currie 2019). Results with other  
309 normalising variables are provided in Appendix II.

### 310 4.2.2. Risk with respect to the population size

311 We now consider the population size of a region as  $N_j$ , i.e., risks are clustered according  
312 to the impact of population size on the number of crashes. For illustrative purposes, the map  
313 obtained from Step 2 of the exploratory analysis is included in Appendix II as Figure 11. The  
314 model led to seven clusters, among which four have negligible frequencies (see Figures 1

315 and 3). The frequency corresponds to the size of the cluster divided by the total number of  
 316 regions. In the right-hand part of Figure 1, the legend indicates the lower bound of each bin  
 317 on case ratio, except for the last bin where the upper bound is also provided. A close-up view  
 318 on Melbourne is available in Appendix II as Figure 9. Here we do not provide any detailed  
 319 interpretation for 8 regions in four clusters (clusters labelled as 5 to 9) that essentially are  
 320 outliers (the regions correspond to zones with 6 to 184 inhabitants, where 9 to 209 crashes  
 321 occur); see Figures 3 and 4 a). The estimate  $\hat{\beta} \approx 0.34$  indicates rather low spatial aggregation  
 322 of clusters.

323 For a given region  $j$ , the marginal state entropy is defined as

$$H[q_{z_j}] = - \sum_{k=0}^{K-1} q_{z_j}(k) \ln q_{z_j}(k)$$

324 and measures the uncertainty regarding the value of  $Z_j$ , which is between 0 and  $\ln K$ :  
 325  $H[q_{z_j}] = 0$  means that  $q_{z_j}(k_0) = 1$  for some value  $k_0$  while  $H[q_{z_j}] = \ln K$  means that  $q_{z_j}$  is  
 326 a uniform distribution. The cumulative marginal entropy is defined as

$$\sum_{j=1}^J H[q_{z_j}].$$

327 Here the cumulative marginal entropy is 20.7, indicating moderate uncertainty regarding the  
 328 labels in some regions (compared to the other models considered, hereafter), although the  
 329 marginal state entropy is below 0.2 in the majority of the regions (see Figure 2). Note that  
 330 due to the model eliminating some states  $k$  during the estimation process (if and only if  
 331  $\max_j q_{z_j}(k) < 0.5$ ), the remaining labels may not have consecutive indices since they are  
 332 identified with their initial indexing in  $\{0, \dots, K - 1\}$ .

333 We ordered the labels by increasing values of risk levels. It can be seen from Figure 3  
 334 that the fitted linear regression lines are in accordance with slopes induced by expected risk  
 335 levels, indicating well-separated classes. There is however some larger discrepancy between  
 336 the two lines in risk level 4, where the regression line has a larger slope than the expected risk  
 337 level due to possible confusions between levels 4 and 2 in some regions.

338 Risk level 0 is related to peripheral regions that are close to the capital of Melbourne,  
 339 and enclaves, which are often regional towns or rural centres with substantial residential  
 340 developments (see Figure 1; high-resolution versions of all maps are available online). These  
 341 are small regions (Figure 4 b) with high population sizes (Figure 4 a), high absolute traffic  
 342 (Figure 4 c) and high traffic densities (Figure 4 d). Risk level 2 is related to peripheral and  
 343 central regions. These are medium-sized zones with medium population sizes, absolute traffic  
 344 and traffic densities. Risk level 4 is related to far peripheral and hypercentral regions (relative

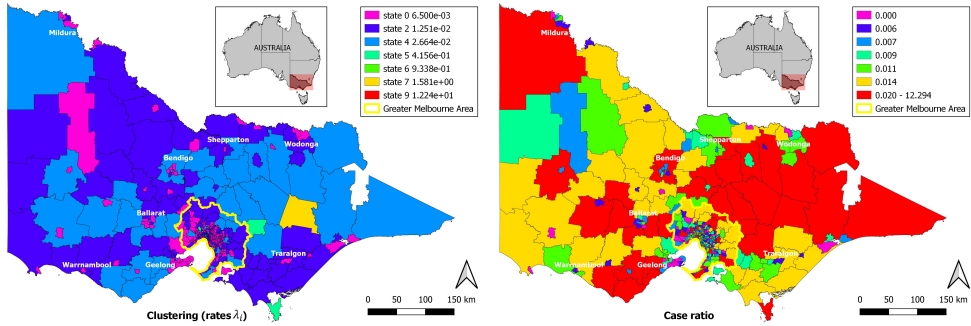


Figure 1. Left-hand part: Risk mapping with respect to population size (variable pop16). Right-hand part: Segmentation using quantiles on ratio. In that part of the figure, the legend indicates the lower bound of each bin on case ratio (sample quantile of order 1/8), except for the last bin where the upper bound is also provided.

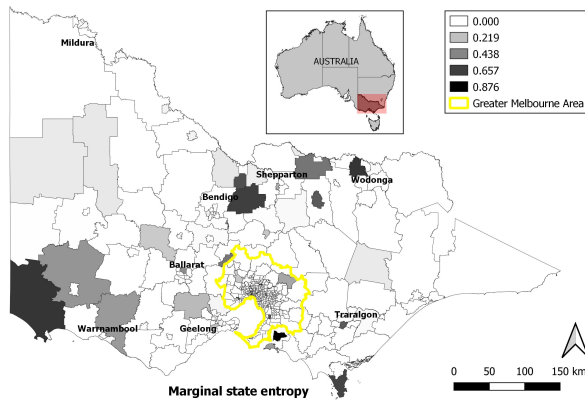


Figure 2. Marginal state entropy for each region regarding risk levels with respect to population size.

345 to the capital). These are sparsely populated, have varying sizes, with high absolute traffic and  
 346 traffic densities. The four variables considered in Figure 4 are well discriminated by the risk  
 347 levels, with ANOVA  $p$ -values between  $10^{-10}$  and  $10^{-15}$ , regarding the effects of the classes.  
 348

349 Since regions with higher population sizes have lower risks, it is possible that population  
 350 has some non-linear effect, which is an avenue for further investigation.

#### 351 4.2.3. Risk with respect to the traffic density: VtrFAR19 variable

352 We now set  $N_j$  to be the ratio VtrFAR19 of vehicle-kilometers travelled on freeways  
 353 and arterial roads divided by the total road length in the region, i.e., risks are clustered  
 354 according to the impact of VtrFAR19 on the number of crashes.  $N_j$  can be null in some  
 355 regions, which leads to a degenerate model, so we set the minimal value to one in VktFAR19

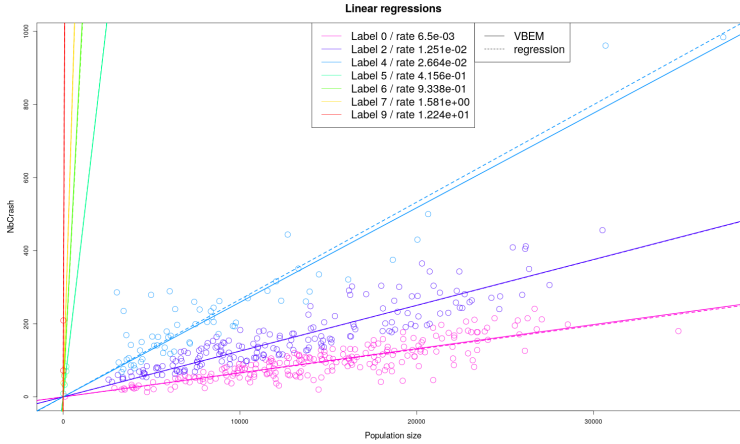


Figure 3. Cluster-wise regressions of crash numbers on population size.

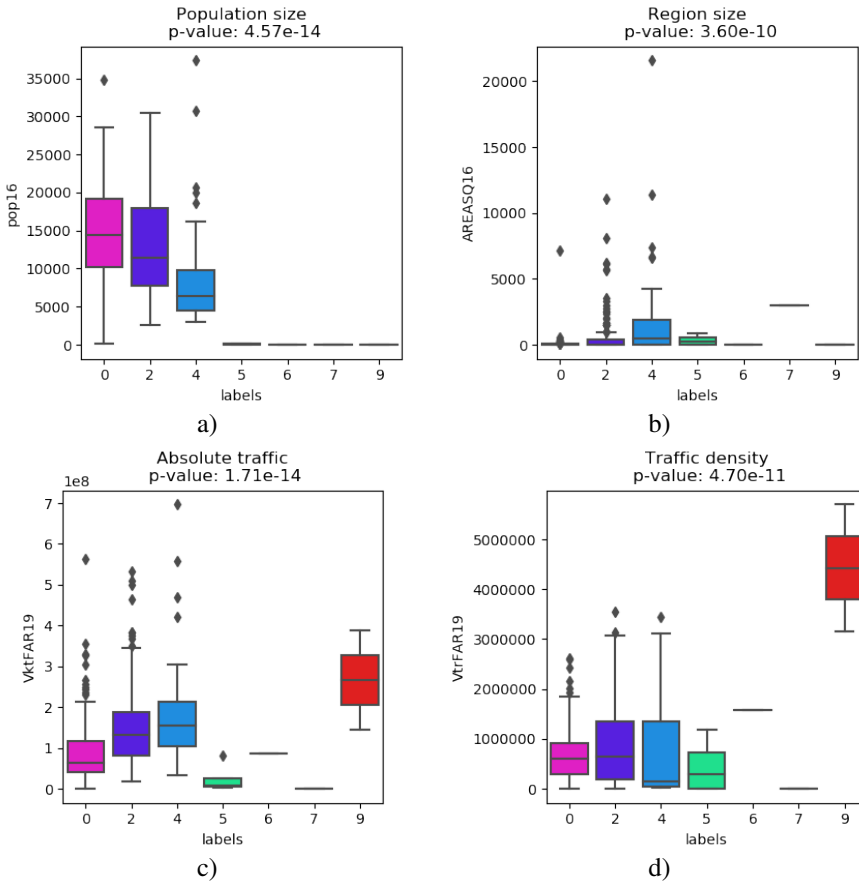


Figure 4. ANOVA of: a) pop16 (population size), b) AREASQ16 (region size), c) VktFAR19 (absolute traffic) and d) VtrFAR19 (traffic density) on risk levels with respect to population size.

356 when computing the ratio (this happens in the zone of French Island, only). We obtain nine  
357 clusters, among which two have negligible frequencies (see Figures 5 and 7). A close-up view  
358 on Melbourne is available in Appendix II as Figure 10. The estimate  $\hat{\beta} \approx 0.54$  indicates good  
359 spatial aggregation of regions. The cumulative marginal entropy is 15.8, indicating mostly  
360 low uncertainty regarding risk levels. From Figure 6, the entropy is very close to 0 in a large  
361 majority of regions and close to 0.693 (i.e.,  $\ln 2$ ) in a small number of regions, which is a  
362 value that represents equiprobability between two risk levels. It can be seen from Figure 7  
363 that linear regression lines are in accordance with slopes induced by expected risk levels,  
364 indicating well-separated classes. There is however some larger discrepancies between the  
365 two values in risk levels 0 and 5, where the regression lines have a larger slopes than the  
366 expected risk levels, due to possible confusions between levels 0 and 2, and levels 5 and 6, in  
367 some regions.

368 Risk level 0 is related to regions in the close periphery of the centre (see Figure 5).  
369 These are small regions (see Figure 8 b) with high traffic densities, medium population sizes  
370 and population densities (see Figures 8 c, a, and d, respectively). Risk level 2 is related to  
371 regions in the close periphery of the capital city centre as well as hypercentral regions, and  
372 central regions and enclaves. These are small regions with intermediate traffic densities, high  
373 population sizes and medium to high population densities. Risk level 3 is related to regions in  
374 the close periphery of the centre as well as hypercentral regions, central regions and enclaves.  
375 These are small regions with low traffic densities, high population sizes and low population  
376 densities. Risk level 5 is related to peripheral regions. These are medium-sized regions with  
377 very low traffic densities, population sizes and population densities. Risk level 6 is related to  
378 far peripheral regions. These are large regions with very low traffic densities, population sizes  
379 and population densities. The four variables considered in Figure 8 are well discriminated by  
380 the risk levels, with ANOVA  $p$ -values between  $10^{-21}$  and  $10^{-69}$  regarding the effects of  
381 classes.

382 Since regions with higher VtrFAR19 have lower risks, it is possible that VtrFAR19 has  
383 some non-linear effect, which we again leave to further investigations.

#### 384 **4.2.4. Combining classes issued from two variables**

385 The analyses performed in Subsections 4.2.2 and 4.2.3 yield different mappings. It  
386 would be possible to define new classes as pairs of classes issued from each model. However  
387 this would lead to more complex interpretations. Indeed, there are strong dependencies  
388 between both classes. A  $\chi^2$  independence test rejects the assumption of independence with  
389  $p$ -value of  $10^{-14}$ . To visualise associations between classes, a correspondence analysis (CA)  
390 was performed. We ignored outlier classes and also class VtrFar19\_7 (meaning class 7 in the

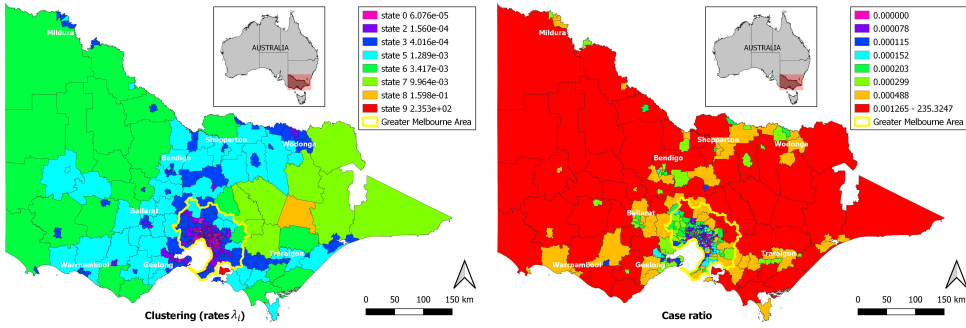


Figure 5. Risk mapping with respect to the traffic density (variable VtrFAR19). Left-hand part: traffic per region. Right-hand part: segmentation using quantiles on ratio.

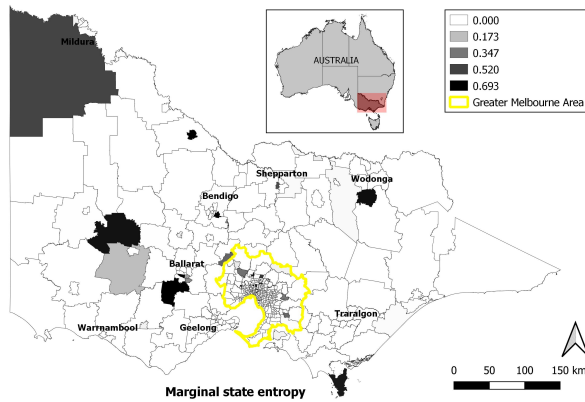


Figure 6. Marginal state entropy for each region regarding risk levels with respect to traffic density VtrFAR19.

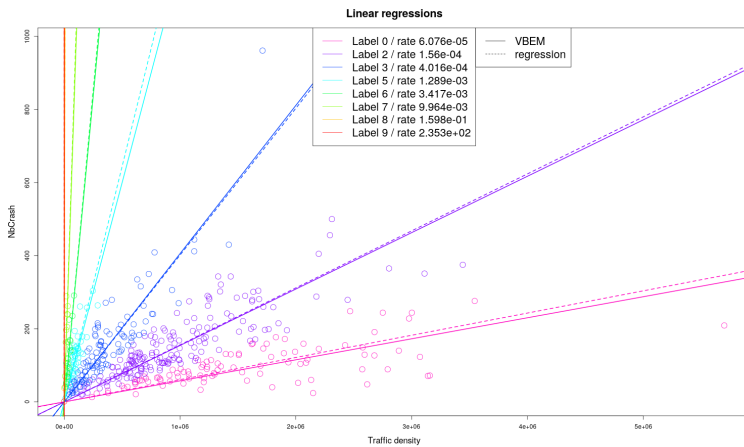


Figure 7. Cluster-wise regressions of crash numbers on traffic density VtrFAR19



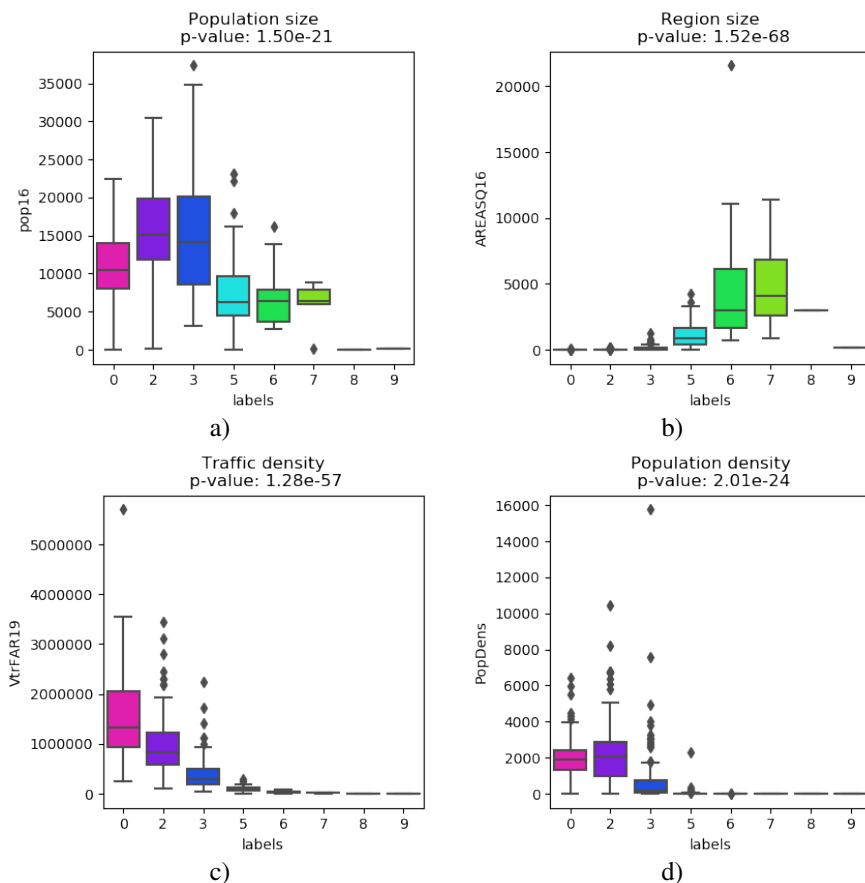


Figure 8. ANOVA of: a) pop16 (population size), b) AREASQ16 (region size), c) VtrFAR19 (traffic density), d) population density on risk levels with respect to traffic density VtrFAR19.

391 model obtained with normalising risk VtrFar19), which perfectly corresponds to Pop16.4  
 392 (higher risk levels in both models). The first CA plane is represented in Appendix II as  
 393 Figure 13. This shows that labels VtrFar19\_2 and VtrFar19\_3 are mainly associated into  
 394 Pop16.0, while Pop16.4 and VtrFar19.5 are strongly associated, VtrFar19.6 is associated  
 395 with Pop16.2 and VtrFar19.0 is split between Pop16.0 and Pop16.2. As a conclusion, the  
 396 orders or risk levels in each model are preserved, except for level 6 in VtrFar19.

#### 397 4.2.5. Risk with respect to other variables

398 Among the different choices of  $N_j$  considered for risk normalisation, the following  
 399 ones yielded some reduced numbers of classes: VktFAR19, PropSign and PopDens. The  
 400 associated results are presented in Appendix II. In contrast, RnNoItcNS, RnLen and RnDens

401 had increasing numbers of classes whenever  $K$  increased further beyond 10, indicating  
 402 some failure in modelling meaningful classes of risks when applying our approach to those  
 403 variables.

404 **5. Assessment of  $\beta$  estimates yielded by VBEM**

405 The behaviour of VBEM regarding estimation of the  $\beta$  hyperparameter is assessed using  
 406 simulated data. The data sets were simulated using 3 states and the values of  $\lambda_k$  obtained for  
 407 the three main clusters in the true data from Subsection 4.2.2:  $\lambda_1 = 0.0065$ ,  $\lambda_2 = 0.013$  and  
 408  $\lambda_3 = 0.027$ . We used the same graph and population values as Subsection 4.2.2. We compare  
 409 two different settings associated with either  $\beta = 0.3$  (estimated value on true data) or  $\beta = 0$   
 410 (case of spatially independent states). For each set of parameters, 50 simulated data sets are  
 411 generated, yielding an estimate  $\hat{\beta}$  for each of them. The Markov random field was simulated  
 412 using the SpaceEM<sup>3</sup> software (Vignes et al. 2011). Simulation relies on 100 Gibbs sampling  
 413 iterations.

414 **5.1. Results when  $\beta = 0.3$**

415 The statistics regarding estimation of hyperparameter  $\beta$  and parameters  $\lambda_k$  (estimated  
 416 as  $E_{q_{\lambda_k}}[\lambda_k]$ ) are provided in Table 2.

417 Regarding the number of clusters: the estimated number of clusters was 3 in 31 samples,  
 418 4 in 11 samples and between 5 in 8 samples. Whenever the estimated number of clusters was  
 419 3, we computed the label discrepancy, defined as the percentage of values in segmentation  
 420 that match the true labels.

Table 2

parameter	$\beta$	$\lambda_1$	$\lambda_2$	$\lambda_3$	label discrepancy
mean	0.55	$6.4 \times 10^{-3}$	$1.2 \times 10^{-2}$	$2.4 \times 10^{-2}$	1.1%
median	0.57	$6.4 \times 10^{-3}$	$1.3 \times 10^{-2}$	$2.7 \times 10^{-2}$	1.0%
standard deviation	0.09	$4.2 \times 10^{-3}$	$2.4 \times 10^{-3}$	$5.5 \times 10^{-3}$	$3.7 \times 10^{-3}$
minimum	0.31	$5.6 \times 10^{-3}$	$6.9 \times 10^{-3}$	$1.2 \times 10^{-2}$	0.2%
maximum	0.74	$7.1 \times 10^{-3}$	$1.4 \times 10^{-2}$	$2.8 \times 10^{-2}$	1.7%

421 Although beta was overestimated in our experiments, we observed that this bias did not  
 422 affect critically the estimation of the others variables and parameters, which remained well  
 423 estimated.

## 424 5.2. Results when $\beta = 0$

425 For one sample, estimation of  $\beta$  failed since the gradient function defined by (9) was very  
 426 flat and from a numerical point of view, there was a large set of values of  $\beta$  cancelling the  
 427 approximated gradient. On the 49 remaining samples, the mean estimated  $\beta$  and the median  
 428 were 0.03, the standard deviation was 0.04, the minimum was -0.05 and the maximum was  
 429 0.14. The test of the null hypothesis  $E[\hat{\beta}] = 0$  against the alternative  $E[\hat{\beta}] \neq 0$  suggested that  
 430  $\beta$  is overestimated (p-value:  $10^{-5}$ ).

431 Regarding estimation of the number of clusters: overestimation in Dirichlet Process  
 432 independent mixture models is well documented and a post-processing treatment is required  
 433 to recover consistency properties (Guha, Ho & Nguyen 2021). This behaviour is illustrated  
 434 in our results: the estimated number of clusters was 3 in 4 samples, between 4 and 5 in 35  
 435 samples and between 6 and 9 in 11 samples.

## 436 5.3. Conclusion on estimation

437 The results regarding  $\beta$  estimates on simulated data suggest that this parameter is  
 438 overestimated, with larger bias when  $\beta$  increases. The number of states is also overestimated,  
 439 possibly with larger bias when  $\beta$  tends to zero.

## 440 6. Conclusion and perspectives

441 The BNP-HMRF model presented here for risk mapping offers a convenient and fast  
 442 approach for conducting segmentation of count data regressions indexed by graphs. For  
 443 example running the 300 iterations required to obtain the results from Subsection 4.2.2 takes  
 444 between 35 seconds and 40 seconds on a Laptop with an Intel Core 8th i7 8665U CPU  
 445 with 4 cores, HT, 1.9Ghz, 4.8Ghz Turbo, 8Mo/UHD 620, using hyperthreading. This running  
 446 time includes selection of the number of classes, which is performed implicitly with state  
 447 probabilities vanishing when states are not relevant.

448 The workload required to add new models to our software is rather low, since the specific  
 449 modifications due to the Poisson assumption can all be embedded into a new class, which is  
 450 added to the package and by using class inheritance principles. This holds if the added model  
 451 remains in the exponential family, using conjugate priors.

452 This models would also offer the possibility to handle missing  $y_j$ s by introducing new  
 453 variational factors  $q_{y_j}$  in our VBEM approximation. Besides handling incomplete data sets,  
 454 this would also allow modellers to perform model selection by cross-validation for example.

455 The proposed model was effective in identifying clusters with distinct risk levels,  
 456 without the requirement of preselecting the number of clusters  $K$ . Furthermore, this is

457 achieved via a computationally efficient VBEM estimation framework. The proposed BNP  
458 model is a significant advancement compared to existing traffic crash mapping approaches,  
459 such as the European Road Assessment Program, the BYM model (Besag, York & Mollié  
460 1991), or other spatial CAR models that appear in the road safety literature (Aguero-Valverde  
461 & Jovanis 2008; Wang & Kockelman 2013; Truong & Currie 2019), which require arbitrary  
462 selection of thresholds for the determination of risk level classes. Like the aforementioned  
463 methods, our approach can also be applied via an MCMC implementation, which may permit  
464 the use of more structured priors. However, such an implementation incurs an expense of a  
465 more computationally intensive estimation procedure, which may not be feasible for large  
466 data sets, where a variational approach may be more appropriate. MCMC implementations  
467 are available for related models (Orbanz & Buhmann 2008) but none is currently available  
468 for our model; thus, we leave the exploration of such procedures to future work.

469 Regarding the estimation of  $\beta$  using simulated data, we note that the proposed approach  
470 is very closed to that in (Forbes & Peyrard 2003). This later mean field approximation  
471 provides good results as reported in (Forbes & Peyrard 2003) (and reproduced using the  
472 SpacEM3 software). Our BNP setting differs in that it involves a problematic expectation  
473 over  $\tau$ , which can be handled in various ways. In the paper, we propose one such solution but  
474 other variants should be investigated. However, note that as in traditional image segmentation,  
475 the Potts model is used as a spatial regularization term and is not *per se* a prior model for an  
476 image or a partition to be recovered. A typical simulation of a Potts model can be indeed  
477 quite far from partitions underlying real data. For this reason, estimating the exact value of  $\beta$   
478 is not the primary target. In practice, although beta is overestimated in our experiments, we  
479 observed that it does not affect critically the estimation of the other parameters, which remain  
480 well estimated.

481 This relative insensitivity to the exact  $\beta$  value is not surprising and has been observed  
482 and discussed in other context involving HMRF modelling (Pereyra et al. 2013; Forbes &  
483 Raftery 1999). Nevertheless, we can propose several ways to prevent this overestimation.  
484 A straightforward possibility is to add an exponential prior on  $\beta$  (Chaari et al. 2012). More  
485 refined attempts include studying in more details variants of the approximation made over  $\tau$  to  
486 recover performance similar as in traditional parametric HMRF models. These investigations  
487 are left for future work.

488 The number of states also tends to be overestimated, possibly with larger bias when  $\beta$   
489 is close to zero. This point could be addressed in the future by running some post-processing  
490 algorithm adapted from (Guha, Ho & Nguyen 2021) on the cluster values.

491 The proposed model will support the Safe System approach, being adopted in many  
492 countries, by identifying crash hot zones for prioritised treatments. Detailed analyses of  
493 these clusters showed that regions with higher traffic densities tend to have lower traffic

494 density-based crash risk levels, while regions with higher population sizes tend to have lower  
 495 population-based crash risk levels. These findings corroborate well with the traffic safety  
 496 literature. It is well-established that crash risks tend to decrease with increasing exposure,  
 497 such as population or the number of road users (safety-in-number effect, Elvik 2014).

498 Regarding further application to traffic crashes, the model could be extended to consider  
 499 multiple crash exposure variables. However, the possibility to use several risk-normalising  
 500 variables leads to multiplying classes and makes their interpretation more difficult. To solve  
 501 this issue, we could build meta-labels by considering for each region a risk signature. This  
 502 signature would be a vector of risk levels, each component corresponding to a specific  
 503 variable  $N_j$ . However, this would greatly increase the dimension of the state space to  
 504 be considered. We could in addition consider either spatial clustering on those signatures  
 505 (possibly equipped with some metrics) or continuous latent variables obtained with multiple  
 506 correspondence analysis, introducing continuous latent variables in our model. Moreover,  
 507 our approach could be compared with clustering results obtained from CAR models, which  
 508 would require some ranges for grouping CAR random effects as discussed in Section 1.

509 Ideally, all possibly relevant variables  $N_{i,j}$  should be included into a unique multiple  
 510 regression model such as

$$p(y_j|\lambda_k, N_{1,j}, \dots, N_{I,j}) = \mathcal{P}(y_j|\lambda_0 + \sum_{i=1}^I \lambda_{k,i} N_{i,j})$$

511 or

$$p(y_j|\lambda_k, N_{1,j}, \dots, N_{I,j}) = \mathcal{P}\left(y_j \mid \exp\left[\lambda_0 + \sum_{i=1}^I \lambda_{k,i} N_{i,j}\right]\right), \quad (12)$$

512 and

$$p(y_j|\lambda_k, N_{1,j}, \dots, N_{I,j}) = \mathcal{P}\left(y_j \mid \exp\left[u_j + \lambda_0 + \sum_{i=1}^I \lambda_{k,i} N_{i,j}\right]\right), \quad u_j \sim \mathcal{N}(0, \tau^2), \quad (13)$$

513 where (12) and (13) have the respective advantages of enabling non-positive linear predictors  
 514 and modelling over-dispersion (see also Waller & Carlin 2010). However, such models do  
 515 not yield explicit VE-steps in the VBEM algorithm. Thus, further approximations would be  
 516 required, MCMC estimation still being an alternative to overcome this additional difficulty.  
 517 Although zero-inflated count data did not seem to invalidate our analyses and results, they  
 518 may have to be considered in other applications, either at the level of Dirichlet Processes (as  
 519 discussed in Canale et al. 2017) or via alternative emission densities.

520 To model centrifugal/centripetal effects through a radius  $r$  or more generally, the effect  
 521 of continuous latent variables  $z_j$ , we could define the risk  $\Lambda$  as a stochastic, monotonic

522 function of  $z_j$ . Another possibility would be the extension of a scalar  $\beta$  by a function  
 523 parameterized by the difference in  $z_j$  values between two regions.

524 BNP-HMRF models could also be extended to handle multivariate count data,  
 525 particularly to address modelling problems in ecology where counts correspond to the  
 526 number of observed species. The emission densities could thus be replaced by Joint Species  
 527 Distribution Models (JSDMs), in which spatial dependencies and heterogeneity sources are  
 528 usually modelled with univariate CARs (see, e.g., Saas & Gosselin 2014) but are ignored in  
 529 the multivariate count data.

### 530 **Acknowledgment**

531 We thank the anonymous reviewers for providing us with insightful comments and  
 532 suggestions.

## 533 **Appendix I**

### 534 **Free energy computation**

Here, we present how to compute the free energy (4). This is mainly used to define a  
 stopping criterion in VBEM, which is why we ignore terms that do not depend on hyper-  
 parameters or variational parameters. The free energy can be decomposed into two terms:

$$\begin{aligned} \mathbb{E}_{q(\mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\Theta})} \left[ \ln \frac{p(\mathbf{y}, \mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\Theta} | N_J, \boldsymbol{\Phi})}{q(\mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\Theta})} \right] &= \mathbb{E}_{q(\mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\Theta})} [\ln p(\mathbf{y}, \mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\Theta} | N_J, \boldsymbol{\Phi})] \\ &\quad - \mathbb{E}_{q_{\mathbf{z}}} [\ln q_{\mathbf{z}}] - \mathbb{E}_{q_{\boldsymbol{\tau}}} [\ln q_{\boldsymbol{\tau}}] - \mathbb{E}_{q_{\boldsymbol{\Theta}}} [\ln q_{\boldsymbol{\Theta}}], \end{aligned}$$

535 where the last three terms are entropies and the first term has already been already calculated  
 536 in the E-step.

### 537 **Entropy terms**

538 In this section, we provide the expressions for  $H[q_{\mathbf{z}}] = -\mathbb{E}_{q_{\mathbf{z}}} [\ln q_{\mathbf{z}}(\mathbf{z})]$ ,  $H[q_{\boldsymbol{\tau}}] =$   
 539  $-\mathbb{E}_{q_{\boldsymbol{\tau}}} [\ln q_{\boldsymbol{\tau}}]$  and  $H[q_{\boldsymbol{\Theta}}] = -\mathbb{E}_{q_{\boldsymbol{\Theta}}} [\ln q_{\boldsymbol{\Theta}}(\boldsymbol{\Theta})]$ . Firstly,

$$H[q_{\mathbf{z}}] = \sum_{j=1}^J H[q_{z_j}] = - \sum_{j=1}^J \sum_{k=0}^{K-1} q_{z_j}(k) \ln q_{z_j}(k),$$

540 where  $q_{z_j}(k)$  is given in (11). Next,

$$H[q_\theta] = H[q_\alpha] + \sum_{k=0}^{K-1} H[q_{\lambda_k}]. \quad (14)$$

541 The first term above can be derived directly from the known entropy expression for a gamma  
542 density:

$$H[q_\alpha] = \hat{s}_1 - \ln \hat{s}_2 + \ln \Gamma(\hat{s}_1) + (1 - \hat{s}_1)\psi(\hat{s}_1).$$

543 The other terms are obtained as follows:

$$-H[q_{\lambda_k}] = E_{q_{\lambda_k}} \left[ \ln(\lambda_k) | \hat{a}_k, \hat{b}_k \right] = \psi(\hat{a}_k) - \ln(\hat{b}_k). \quad (15)$$

544 Moreover,

$$H[q_\tau] = \sum_{k=0}^{K-1} H[q_{\tau_k}],$$

545 where for every  $k$ ,

$$\begin{aligned} H[q_{\tau_k}] &= \ln B(\gamma_{k,1}, \gamma_{k,2}) - (\gamma_{k,1} - 1)[\psi(\gamma_{k,1}) - \psi(\gamma_{k,1} + \gamma_{k,2})] \\ &\quad - (\gamma_{k,2} - 1)[\psi(\gamma_{k,2}) - \psi(\gamma_{k,1} + \gamma_{k,2})] \end{aligned}$$

546 and  $B(\gamma_{k1}, \gamma_{k2}) = \Gamma(\gamma_{k1})\Gamma(\gamma_{k2})/\Gamma(\gamma_{k1} + \gamma_{k2})$ .

## 547 E-step terms

The  $E_{q_{(z, \tau, \Theta)}} [\ln p(\mathbf{y}, \mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\Theta} | \mathbf{N}_J, \boldsymbol{\Phi})]$  term decomposes into 5 terms:

$$\begin{aligned} E_{q_{(z, \tau, \Theta)}} [\ln p(\mathbf{y}, \mathbf{z}, \boldsymbol{\tau}, \boldsymbol{\Theta} | \mathbf{N}_J, \boldsymbol{\Phi})] &= \sum_{j=1}^J E_{q_{z_j} q_\Lambda} [\ln p(y_j | z_j, \boldsymbol{\Lambda})] + E_{q_\alpha} [\ln p(\alpha | s_1, s_2)] \\ &+ \sum_{k=0}^{K-1} E_{q_{\tau_k} q_\alpha} [\ln p(\tau_k | \alpha)] + \sum_{k=0}^{K-1} E_{q_{\lambda_k}} [\ln p(\lambda_k | a_k, b_k)] + E_{q_z q_\tau} [\ln p(\mathbf{z} | \boldsymbol{\tau}, \beta)]. \end{aligned}$$

548 If the free energy is computed at the end of each VBEM iteration, as per Section 3 (VM-  
549  $(s_1, s_2)$  and VM- $(a_k, b_k)$  steps), we have  $(s_1, s_2) = (\hat{s}_1, \hat{s}_2)$  and  $(a_k, b_k) = (\hat{a}_k, \hat{b}_k)$ . Thus, in  
550 the free energy  $E_{q_\alpha} [\ln p(\alpha | s_1, s_2)]$  cancels out with  $H[q_\alpha]$ . Similarly,  $E_{q_{\lambda_k}} [\ln p(\lambda_k | a_k, b_k)]$   
551 cancels out with  $H[q_{\lambda_k}]$  in (14).

552  $\sum_{j=1}^J \mathbb{E}_{q_{z_j, q_\Lambda}} [\ln p(y_j | z_j, \Lambda)]$  **term.**

553 For each  $1 \leq j \leq J$ ,

$$\mathbb{E}_{q_{z_j, q_\Lambda}} [\ln p(y_j | z_j, \Lambda)] = \sum_{k=0}^{K-1} q_{z_j}(k) \mathbb{E}_{q_{\lambda_k^*}} [\ln p(y_j | \lambda_k^*)].$$

In the case of Poisson emission densities,

$$\begin{aligned} \mathbb{E}_{q_{z_j, q_\Lambda}} [\ln p(y_j | z_j, \Lambda)] &= \sum_{k=0}^{K-1} q_{z_j}(k) \left[ y_j \mathbb{E}_{q_{\lambda_k}} [\ln(\lambda_k)] - N_j \frac{\hat{a}_k}{\hat{b}_k} + y_j \ln(N_j) - \ln(y_j!) \right] \\ &= y_j \sum_{k=0}^{K-1} q_{z_j}(k) \mathbb{E}_{q_{\lambda_k}} [\ln(\lambda_k)] - N_j \sum_{k=0}^{K-1} q_{z_j}(k) \frac{\hat{a}_k}{\hat{b}_k} - y_j \ln(N_j) - \ln(y_j!), \end{aligned} \quad (16)$$

554 with  $\mathbb{E}_{q_{\lambda_k}} [\ln(\lambda_k) | \hat{a}_k, \hat{b}_k]$  given in (15). Since the last two terms in (16) depend on the data  
555 only, they do not need to be computed to monitor the convergence of the VBEM algorithm.

556  $\mathbb{E}_{q_{\tau_k, q_\alpha}} [\ln p(\tau_k | \alpha)]$  **term.**

557 Using the expression of a Beta distribution cross-entropy, it follows that

$$\mathbb{E}_{q_{\tau_k, q_\alpha}} [\ln p(\tau_k | \alpha)] = \psi(\hat{s}_1) - \ln \hat{s}_2 + \left( \frac{\hat{s}_1}{\hat{s}_2} - 1 \right) [\psi(\gamma_{k2}) - \psi(\gamma_{k1} + \gamma_{k2})].$$

558  $\mathbb{E}_{q_z, q_\tau} [\ln p(z | \tau, \beta)]$  **term.**

559 This term cannot be computed exactly due to the intractable normalising constant  $\mathcal{K}$  in  
560 the Potts model expression:

$$\ln p(z | \tau, \beta) = \sum_{j=1}^J \ln \pi_{z_j}(\tau) + \beta \sum_{i \sim j} \mathbf{1}_{(z_i = z_j)} - \ln \mathcal{K}(\beta, \tau).$$

561 The first two terms can be computed easily but the last one requires approximation. As for the  
562 estimation of  $\beta$ , we can use a mean-field like approximation and approximate at each iteration  
563 the true value of  $\mathcal{K}(\beta, \tau)$  by  $\tilde{\mathcal{K}}(\beta, \tau)$ , the normalising constant of  $\tilde{q}_z$ , defined in (10):

$$\tilde{\mathcal{K}}(\beta, \tau) = \prod_{j=1}^J \left( \sum_{l=1}^{\infty} \exp(\ln \pi_l(\tau) + \beta \sum_{i \in N(j)} q_{z_i}^{(r)}(l)) \right),$$

564 where as in Section 3,  $\tau$  in the above expression could be replaced by  $\mathbb{E}_{q_\tau}[\tau]$  to avoid the  
565 dependence in a random  $\tau$ . However, this would correspond to a zeroth order approximation



566 and we can do better with a first order approximation as follows:

$$\mathcal{K}(\beta, \boldsymbol{\tau}) \approx \tilde{\mathcal{K}}(\beta, \boldsymbol{\tau}) \exp(\mathbb{E}_{\tilde{q}_z} [V(\mathbf{z}; \boldsymbol{\tau}, \beta) - \tilde{V}(\mathbf{z}; \boldsymbol{\tau}, \beta)]), \quad (17)$$

567 where  $V(\mathbf{z}; \boldsymbol{\pi}, \beta)$  is defined as

$$V(\mathbf{z}; \boldsymbol{\pi}, \beta) = \sum_{j=1}^N \log \pi_{z_j} + \beta \sum_{i \sim j} \mathbf{1}_{(z_i = z_j)}$$

and similarly,

$$\tilde{V}(\mathbf{z}; \boldsymbol{\tau}, \beta) = \sum_{j=1}^N \left( \ln \pi_{z_j}(\boldsymbol{\tau}) + \beta \sum_{i \in N(j)} q_{z_i}^{(r)}(z_j) \right).$$

Note that all terms are tractable:

$$\begin{aligned} \mathbb{E}_{\tilde{q}_z} [V(\mathbf{z}; \boldsymbol{\tau}, \beta) - \tilde{V}(\mathbf{z}; \boldsymbol{\tau}, \beta)] &= \beta \sum_{j=1}^N \sum_{k=0}^{K-1} \tilde{q}_{z_j}(k|\boldsymbol{\tau}, \beta) \left( \frac{\sum_{i \in N(j)} \tilde{q}_{z_i}(k|\boldsymbol{\tau}, \beta)}{2} \right. \\ &\quad \left. - \sum_{i \in N(j)} q_{z_i}^{(r)}(k) \right). \end{aligned}$$

Then,

$$\begin{aligned} \mathbb{E}_{q_z q_\tau} [\ln \mathcal{K}(\beta, \boldsymbol{\tau})] &\approx \mathbb{E}_{q_\tau} [\ln \tilde{\mathcal{K}}(\beta, \boldsymbol{\tau})] + \mathbb{E}_{q_\tau} [\mathbb{E}_{\tilde{q}_z} [V(\mathbf{z}; \boldsymbol{\tau}, \beta) - \tilde{V}(\mathbf{z}; \boldsymbol{\tau}, \beta)]] \\ &\approx \sum_{j=1}^N \mathbb{E}_{q_\tau} \left[ \ln \left( \sum_{l=1}^{\infty} \exp(\ln \pi_l(\boldsymbol{\tau}) + \beta \sum_{i \in N(j)} q_{z_i}^{(r)}(l)) \right) \right] \\ &\quad + \beta \sum_{j=1}^N \sum_{k=0}^{K-1} \mathbb{E}_{q_\tau} \left[ \tilde{q}_{z_j}(k|\boldsymbol{\tau}, \beta) \left( \frac{\sum_{i \in N(j)} \tilde{q}_{z_i}(k|\boldsymbol{\tau}, \beta)}{2} - \sum_{i \in N(j)} q_{z_i}^{(r)}(k) \right) \right]. \end{aligned}$$

568

## Appendix II

569

### Additional figures and results on crash data analysis

570 Here, we present further figures and analyses obtained by using different covariates in  
571 our model.

572 Figure 9 is a close-up on the map in Figure 1 highlighting risk mapping in the regions  
573 of Melbourne. Similarly, Figure 10 is a close-up on the map in Figure 5

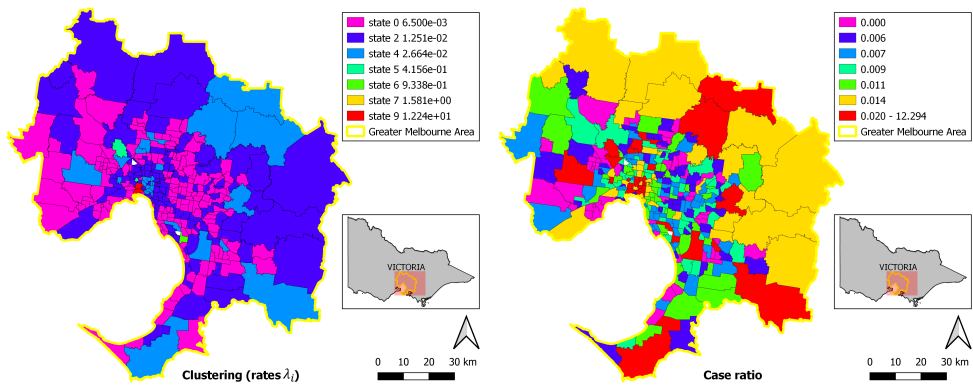


Figure 9. Left-hand part: Risk mapping with respect to population size (variable pop16) in Melbourne. Right-hand part: Segmentation using quantiles on ratio.

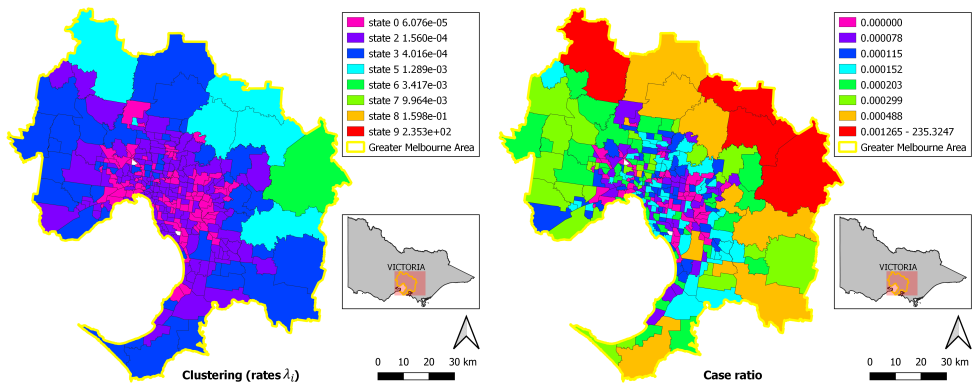


Figure 10. Left-hand part: Risk mapping with respect to the traffic density (variable VtrFAR19) in Melbourne. Right-hand part: Segmentation using quantiles on ratio.

574 Figure 11 depicts the map obtained from Step 2 of the exploratory analysis: population  
 575 size and number of crashes per region together with the segmentation obtained by using 7  
 576 quantiles on the ratio of the ‘number of crashes on population size’.

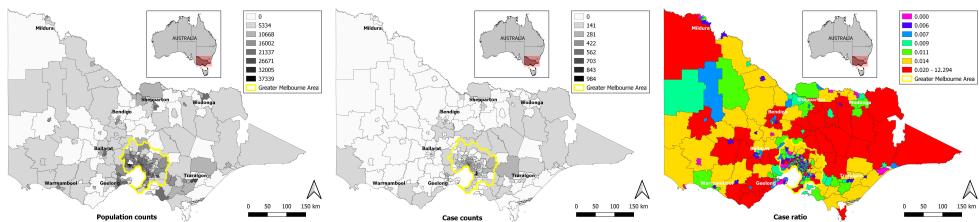


Figure 11. Maps of data and risk ratio used in exploratory using population size (variable pop16). Left: population size per region. Middle: number of crashes per region. Right: segmentation using quantiles on ratio.

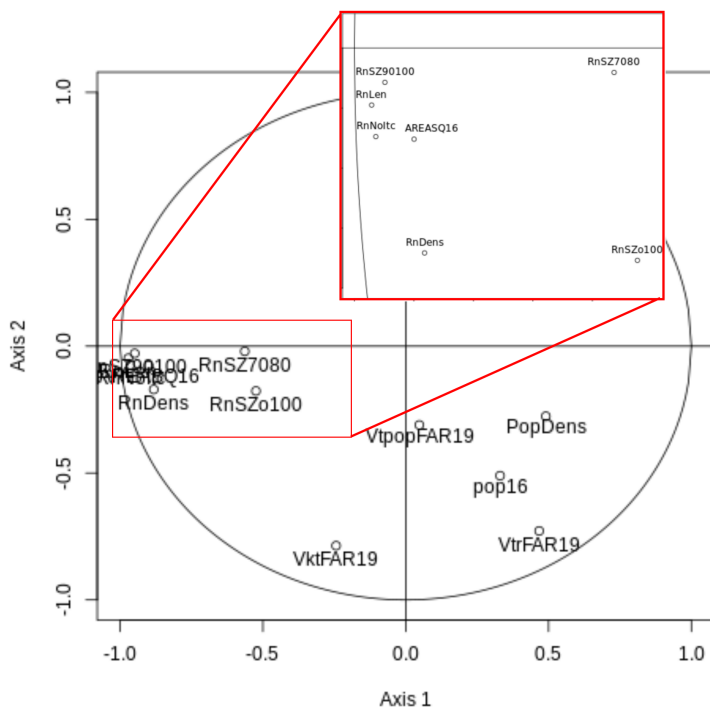


Figure 12. Principal component analysis on covariates: first plane in the variable space. The rectangle framed in red has been zoomed in for the sake of readability.

577 Figure 12 depicts the first plane in the variable space of the PCA performed on  
 578 the following variables: RnNoItc, RnLen, RnSZ7080, RnDens, RnSZ90100, AREASQ16,  
 579 RnSZo100, VtPopFAR19, PopDens, pop16, VktFAR19 and VtrFAR19.

580 Figure 13 depicts the first plane of the correspondence analysis between the labels  
 581 obtained using Pop16 (population size) and VtrFar19 (traffic density) as normalising  
 582 variables.

583 The following subsections reproduce similar analyses as those that appear in  
 584 Subsections 4.2.2 and 4.2.3 but using other variables to normalise risks of crashes.

### 585 Risk with respect to the traffic: VktFAR19 variable

586 We now set  $N_j$  as the vehicle-kilometres travelled on freeways and arterial roads,  
 587 i.e., risks are clustered according to the impact of vehicle-kilometres travelled on the  
 588 number of crashes. The risk appears as somewhat spatially homogeneous and applying the  
 589 model directly does not yield several well-separated clusters. We consider the transform  
 590  $p(y_j|z_j = k; \mathbf{\Lambda}, N_j) = \mathcal{P}(y_j; \lambda_k \sqrt{N_j})$ . Here,  $N_j$  can be null in some regions, which leads

## CA Biplot: Classes for Pop16 and VtrFar19

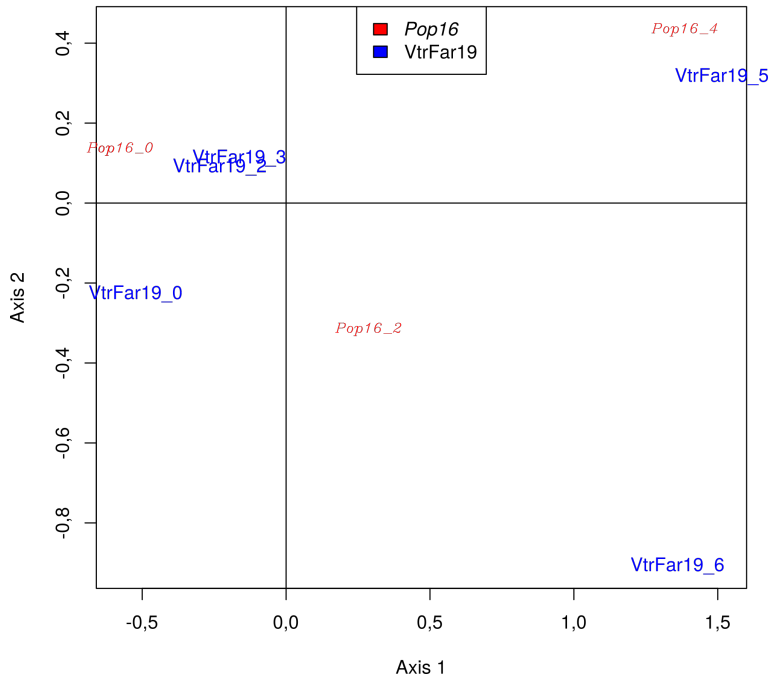


Figure 13. Correspondence analysis between the classes associated with the Pop16 (population size) and VtrFAR19 (traffic density) models: first principal plane.

591 to a degenerate model, so we set the minimal value to one. We obtain three clusters, all  
 592 of which are well represented (see Figure 14). The estimate  $\hat{\beta} \approx 0.42$  indicates a moderate  
 593 spatial aggregation of regions. The cumulative marginal entropy is 32.6, indicating some  
 594 rather high uncertainty regarding risk levels in some regions. It can be seen from Figure 15  
 595 that linear regression lines are in accordance with slopes induced by expected risks levels,  
 596 indicating well-separated classes. There is however some larger discrepancies between the  
 597 two quantities in risk level 8, where the regression line has a lower slope than the expected  
 598 risk level, due to possible confusions between levels 8 and 4 in some regions.

599 Risk level 1 is related to peripheral far west and northern regions (see Figure 14).  
 600 These are regions with low traffic densities, population sizes and population densities (see  
 601 Figure 16). Risk level 4 is related to regions in the close periphery of the capital city centre  
 602 and far east regions. These have intermediate traffic densities, population sizes and population  
 603 densities. Risk level 8 is related hypercentral regions and those at the close east periphery.  
 604 These have high traffic densities, population sizes and population densities.

605 We remark that VktFAR19 has no effect on the labels ( $p$ -value 0.69), which seems to  
 606 confirm the linear relationship between VktFAR19 and crash numbers (cluster-wise).

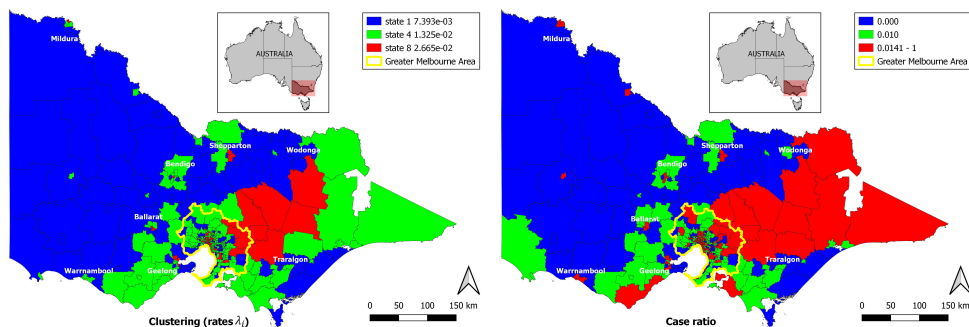


Figure 14. Risk mapping with respect to the traffic (variable VktFAR19). Left-hand part: traffic per region. Right-hand part: segmentation using quantiles on ratio.

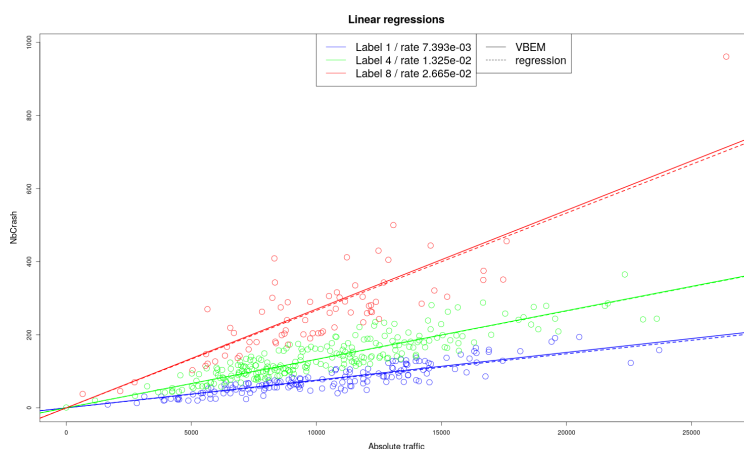


Figure 15. Cluster-wise regressions of crash numbers on VktFAR19.

### 607 **Risk with respect to the proportion of signalised intersections: PropSign variable**

608 The number of signalised intersections (including roundabouts) was generally quite low  
 609 and the number of non-signalised intersections was highly correlated with the number of  
 610 intersections (Pearson correlation: 0.99988). However, the ratio  $r$  of the number of signalised  
 611 intersections on total number of intersections what somehow variable, so we could use it as  
 612 the normalising variable. We subtracted the minimal value (rounded to the closest multiple  
 613 of 0.1, here 0.8) and considered  $N_j = (r - 0.8)^{1.5}$ . In this scenario, we consider that risks  
 614 are clustered according to the impact of relative increase of the proportion of signalised  
 615 intersections with respect to the minimal ratio on the number of crashes. We obtained  
 616 nine clusters, among which four have negligible frequencies (see Figure 7). The estimate  
 617  $\hat{\beta} \approx 0.27$  indicates low spatial aggregation of regions. The cumulative marginal entropy is  
 618 15.8, indicating mostly low uncertainty regarding risk level. Labels are mostly explained

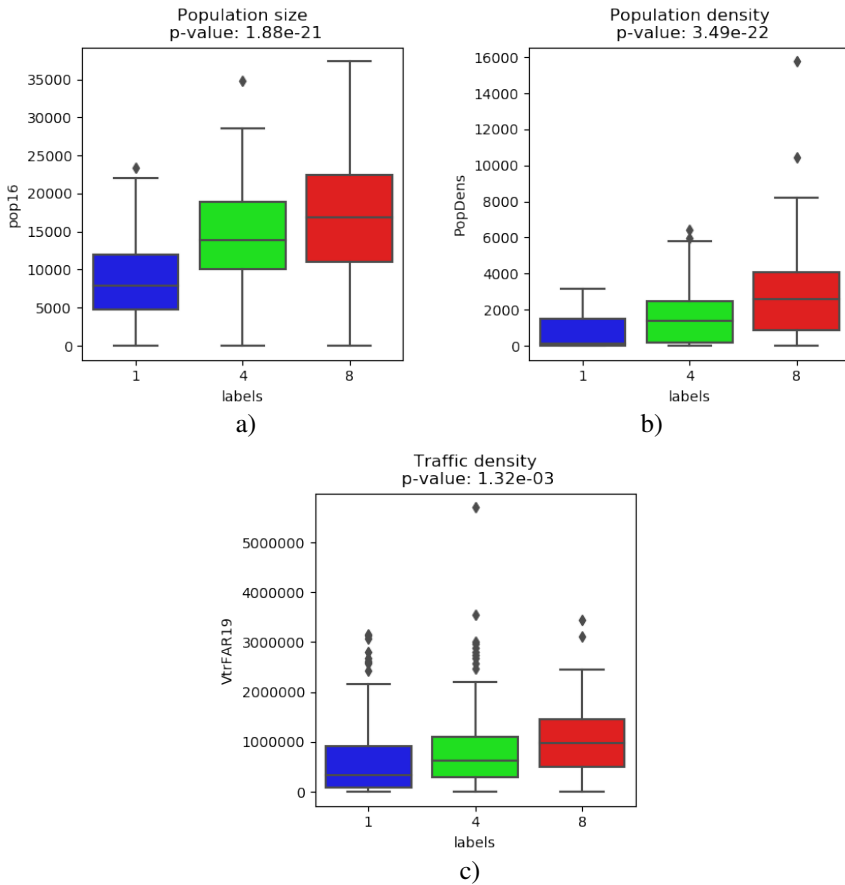


Figure 16. ANOVA of: a) pop16 (population size), b) PopDens (population density) and c) VtrFAR19 (traffic density) on risk levels with respect to absolute traffic intensity VktFAR19.

619 by centripetal gradients. It can be seen from Figure 18 that the accordance between linear  
 620 regression lines and slopes induced by expected risks levels is moderate, indicating possible  
 621 confusions between classes 3 versus 4, 4 versus 5 and 5 versus 6.

622 Risk levels 5, 6, 7 and 9 apply to a very small numbers of regions (with small values of  
 623  $N_j$ , see Figure 17) and can be neglected. Risk level 0, 1 and 2 apply to far peripheral regions  
 624 with low absolute traffics, traffic densities, population sizes and population densities (with  
 625 an increasing risk regarding those four variables). Risk level 3 is less peripheral than 0, 1, 2  
 626 with higher absolute traffics, traffic densities, population sizes and population densities. Risk  
 627 level 4 applies to more central regions than 3, with slightly higher population densities but  
 628 somewhat lower population sizes and traffic.

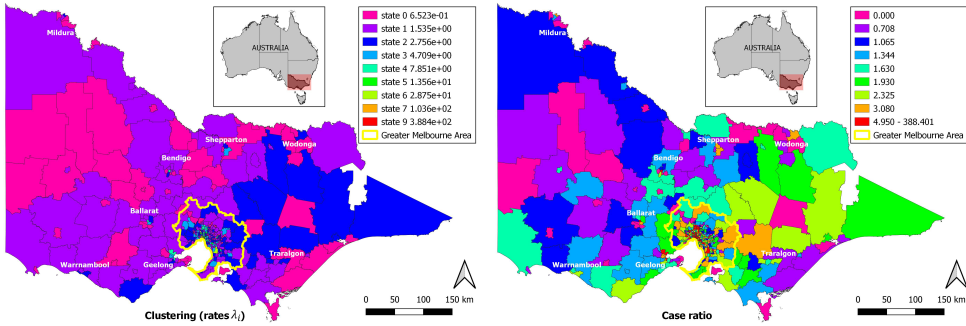


Figure 17. Risk mapping with respect to the ratio of signalised on total number of intersections (variable PropSig). Left-hand part: ratio per region. Right-hand part: segmentation using quantiles on ratio.

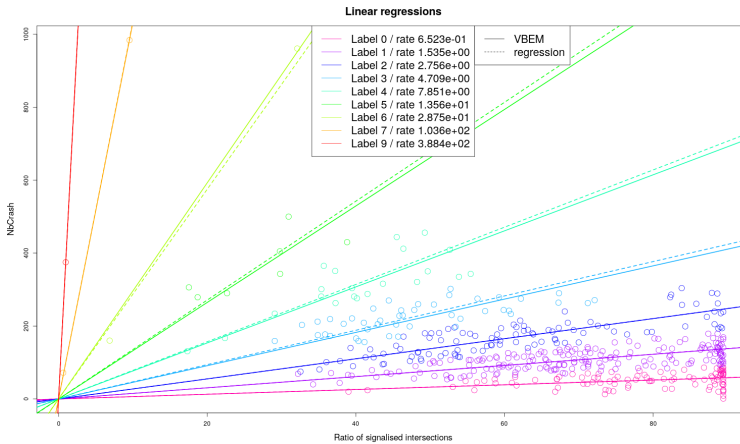


Figure 18. Cluster-wise regressions of crash numbers on the ratio of signalised on total number of intersections.

629 **Risk with respect to the population density: PopDens**

630 We now set  $N_j$  as the population density PopDens, i.e., risks are clustered according  
 631 to the impact of PopDens on the number of crashes. We consider the transform  $p(y_j|z_j =$   
 632  $k; \Lambda, N_j) = \mathcal{P}(y_j; \Lambda, \sqrt{N_j})$ . We obtain nine clusters, among which three have negligible  
 633 frequencies (see Figure 21). The estimate  $\hat{\beta} \approx 0.45$  indicates moderate spatial aggregation of  
 634 regions. The cumulative marginal entropy is 17.7, indicating moderate uncertainty regarding  
 635 risk level. It can be seen from Figure 21 that the accordance between linear regression  
 636 lines and slope induced by expected risks levels is moderate, indicating possible confusions  
 637 between classes 2 versus 3, 3 versus 4 (mainly), 4 versus 5 and 5 versus 6.

638 Risk levels 7 to 9 apply to a very small numbers of regions (hypercentral or  
 639 hyperperipheral, with small  $N_j$ , see Figure 20) and can be neglected. Risk levels 0 to 6 are

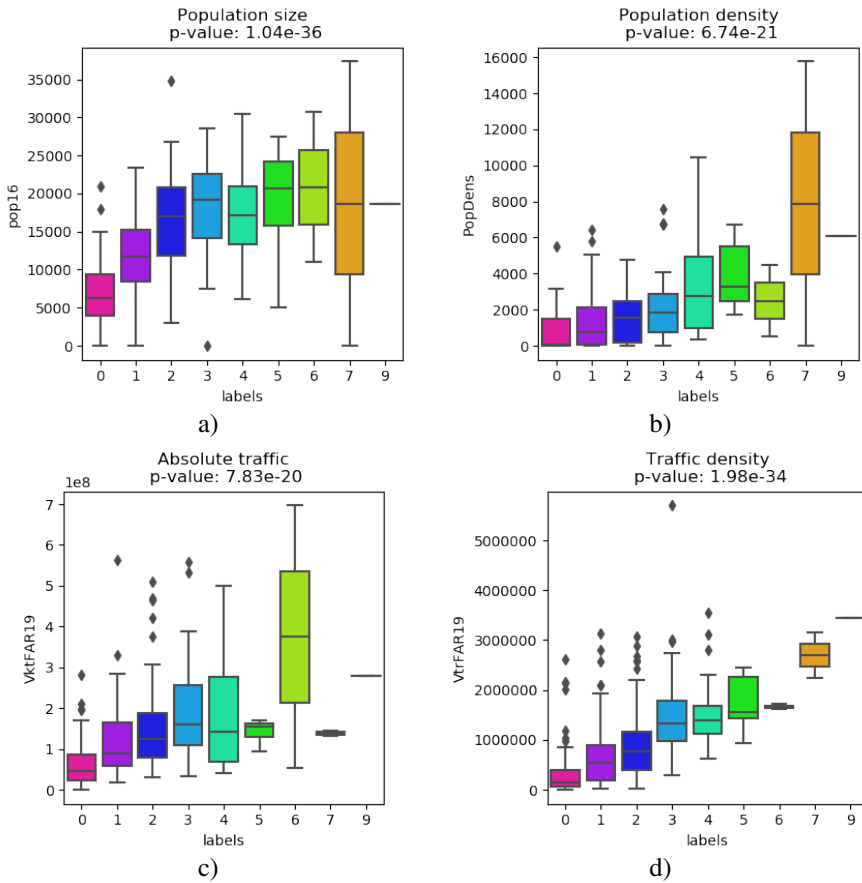


Figure 19. ANOVA of: a) pop16 (population size), b) PopDens (population density), c) VktFAR19 (absolute traffic intensity) and d) VtrFAR19 (traffic density) on risk levels with respect to the proportion of signalised intersections.

640 mostly explained by centrifugal gradients: from small central regions, with high population  
 641 densities, low absolute traffics but high traffic densities to large peripheral regions, with low  
 642 population densities, high absolute traffics but low traffic densities (see Figure 22). Enclaves  
 643 mostly have risk levels 2 or 0, while the risks of their surrounding regions are higher.

644 We remark that if a region has low VktFAR19 and high VtrFAR19, it means that it has  
 645 low absolute traffic but dense traffic, when considering the total road network length.

646 Since regions with higher PopDens have lower risks, it is possible that PopDens has  
 647 some non-linear effect, which is a point to be further investigated.



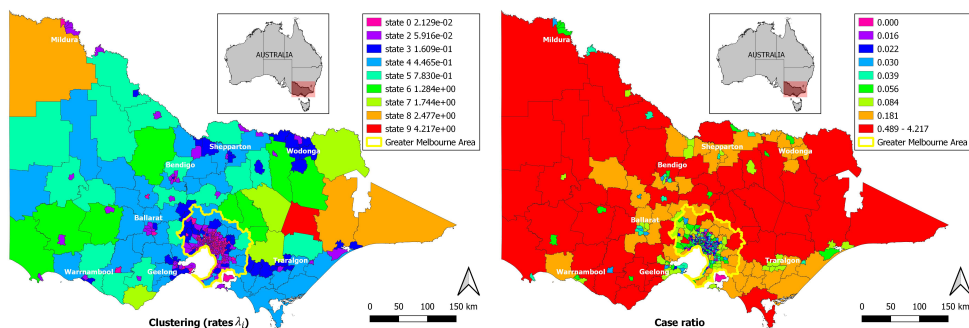


Figure 20. Risk mapping for variable PopDens. Left-hand part: population density per region. Middle part: assigned risk level entropy. Right-hand part: segmentation into a finite number of risk levels.

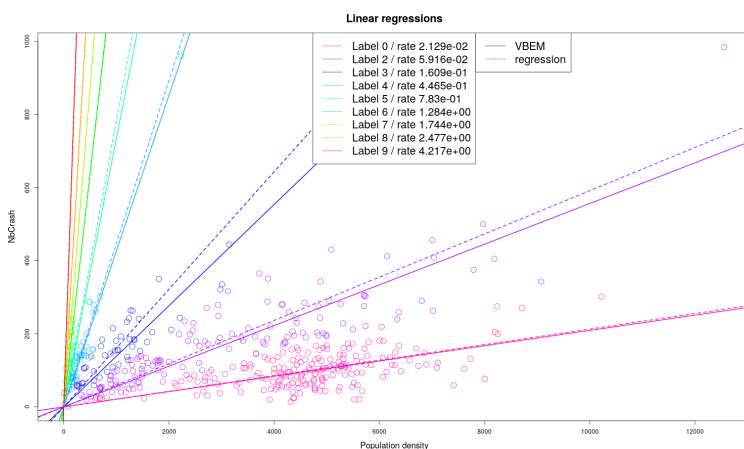


Figure 21. Cluster-wise regressions of crash numbers on PopDens.

648

### References

- 649 ABRIAL, D., CALAVAS, D., JARRIGE, N. & DUCROT, C. (2005). Spatial heterogeneity of the risk of BSE in  
650 France following the ban of meat and bone meal in cattle feed. *Preventive veterinary medicine* **67**, 69–  
651 82. doi:10.1016/j.prevetmed.2004.10.004. URL [http://www.ncbi.nlm.nih.gov/pubmed/](http://www.ncbi.nlm.nih.gov/pubmed/15698909)  
652 15698909.
- 653 AGUERO-VALVERDE, J. & JOVANIS, P.P. (2008). Analysis of road crash frequency with spatial models.  
654 *Transportation Research Record* **2061**, 55–63.
- 655 ALFO, M., NIEDDU, L. & VICARI, D. (2009). Finite mixture models for mapping spatially dependent  
656 disease counts. *Biometrical Journal* **51**, 84–97.
- 657 BESAG, J., YORK, J. & MOLLIE, A. (1991). Bayesian image restoration, with two applications in spatial  
658 statistics. *Annals of the Institute of Statistical Mathematics* **43**, 1–59.
- 659 BLEI, D.M. & JORDAN, M.I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Anal.*  
660 **1**, 121–143.
- 661 BÖHNING, D., DIETZ, E. & SCHLATTMANN, P. (2000). Space-time mixture modelling of public health  
662 data. *Statistics in Medicine* **19**, 2333–2344.

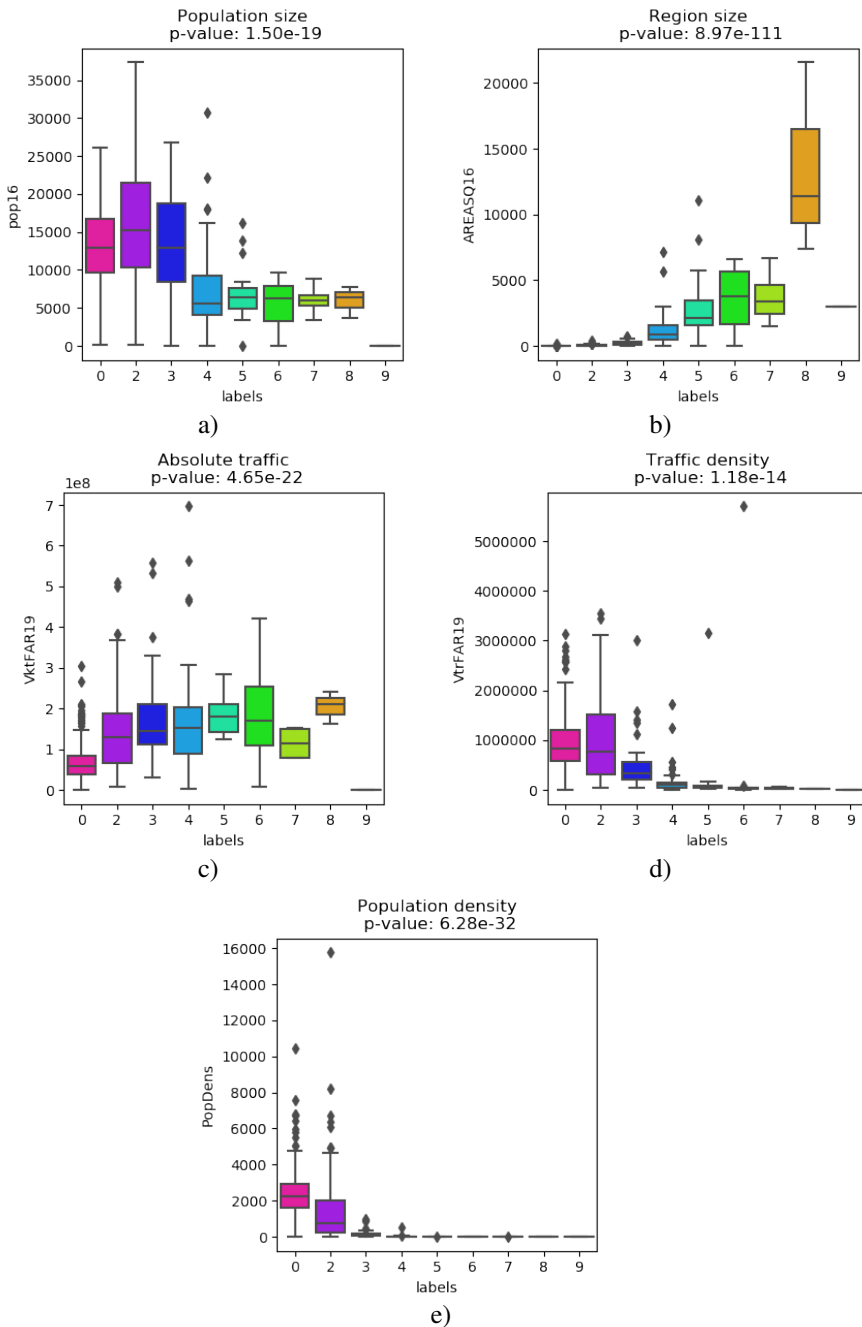


Figure 22. ANOVA of: a) pop16 (population size), b) AREASQ16 (region size), c) VktFAR19 (absolute traffic intensity), d) VtrFAR19 (traffic density) and e) PopDens (population density) on risk levels with respect to population density.

- 663 CANALE, A., LIJOI, A., NIPOTI, B. & PRÜNSTER, I. (2017). On the pitman-yor process with spike and  
664 slab base measure. *Biometrika* **104**, 681–697.
- 665 CHAARI, L., VINCENT, T., FORBES, F., DOJAT, M. & CIUCIU, P. (2012). Fast joint detection-estimation of  
666 evoked brain activity in event-related fMRI using a variational approach. *IEEE transactions on Medical*  
667 *Imaging* **32**, 821–837.
- 668 CHANDLER, D. (1987). *Introduction to modern statistical mechanics*. New York, Oxford: Oxford University  
669 Press. URL <http://opac.inria.fr/record=b1081336>.
- 670 CLAYTON, D. & BERNADINELLI, L. (1992). Bayesian methods for mapping disease risk. *Geographical*  
671 *and Environment Epidemiology: Methods for Small Area Studies*, eds. P.Elliot, J.Cuzik, D.English, and  
672 R.Stern, Oxford, UK:Oxford University Press , 205–220.
- 673 ELVIK, R. (2014). *Towards a general theory of the relationship between exposure and risk*. Institute of  
674 Transport Economics, Oslo, Norway.
- 675 ELVIK, R., VAA, T., HOYE, A. & SORENSEN, M. (2009). *The handbook of road safety measures*. Emerald  
676 Group Publishing.
- 677 FERNANDEZ, C. & GREEN, P. (2002). Modelling spatially correlated data via mixtures: a Bayesian approach.  
678 *Journal of the Royal Statistical Society: Series B (Methodological)* **64**, 805–826.
- 679 FORBES, F. & PEYRARD, N. (2003). Hidden Markov random field model selection criteria based on mean  
680 field-like approximations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**, 1089–  
681 1101.
- 682 FORBES, F. & RAFTERY, A.E. (1999). Bayesian morphology: fast unsupervised Bayesian image analysis.  
683 *Journal of the American Statistical Association* **94**, 555–568.
- 684 FRALEY, C. & RAFTERY, A. (2007). Bayesian regularization for normal mixture estimation and model-based  
685 clustering. *Journal of Classification* **24**, 155–181.
- 686 GHOSAL, S. & VAN DER VAART, A. (2017). *Fundamentals of nonparametric Bayesian inference*, vol. 44.  
687 Cambridge University Press.
- 688 GREEN, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model  
689 determination. *Biometrika* **82**, 711–732.
- 690 GREEN, P.J. & RICHARDSON, S. (2002). Hidden Markov models and disease mapping. *Journal of the*  
691 *American Statistical Association* **97**, 1–16.
- 692 GUHA, A., HO, N. & NGUYEN, X. (2021). On posterior contraction of parameters and interpretability in  
693 Bayesian mixture modeling. *Bernoulli* **27**, 2159–2188.
- 694 KNORR-HELD, L. & RASSER, G. (2000). Bayesian detection of clusters and discontinuities in disease maps.  
695 *Biometrics* **56**, 13–21.
- 696 KNORR-HELD, L., RASSER, G. & BECKER, N. (2002). Disease mapping of stage-specific cancer incidence  
697 data. *Biometrics* **58**, 492–501.
- 698 KNORR-HELD, L. & RICHARDSON, S. (2003). A hierarchical model for space-time surveillance data on  
699 meningococcal disease incidence. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*  
700 **52**, 169–183.
- 701 LAWSON, A., BIGGERI, A., BOEHNING, D., LESAFFRE, E., VIEL, J., CLARK, A., SCHLATTMANN, P.  
702 & DIVINO, F. (2000). Disease mapping models: an empirical evaluation. *Statistics in Medicine* **19**,  
703 2217–2241.
- 704 LAWSON, A. & SONG, H. (2010). Bayesian hierarchical modeling of the dynamics of spatio-temporal  
705 influenza season outbreaks. *Spatial and Spatio-temporal Epidemiology* **1**, 187–195.
- 706 LORD, D. & MANNERING, F. (2010). The statistical analysis of crash-frequency data: A review and  
707 assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice* **44**,  
708 291–305. doi:<https://doi.org/10.1016/j.tra.2010.02.001>. URL <http://www.sciencedirect.com/science/article/pii/S0965856410000376>.
- 709
- 710 LÜ, H., ARBEL, J. & FORBES, F. (2020). Bayesian nonparametric priors for hidden Markov  
711 random fields. *Statistics and Computing* **30**, 1015–1035. URL <https://doi.org/10.1007/>

712 s11222-020-09935-9.

- 713 MACNAB, Y. (2010). On Gaussian Markov random fields and Bayesian disease mapping. *Statistical Methods*  
 714 *in Medical Research* **20**, 49–68.
- 715 MILLER, J.W. & HARRISON, M.T. (2018). Mixture models with a prior on the number of components.  
 716 *Journal of the American Statistical Association* **113**, 340–356.
- 717 MOLLIE, A. (1996). Bayesian mapping of disease. *Markov Chain Monte Carlo in Practice*, eds. W. Gilks, S.  
 718 Richardson, and D. J. Spiegelhalter, London: Chapman and Hall , 359–379.
- 719 MOLLIE, A. (1999). Bayesian and empirical Bayes approaches to disease mapping. In *Disease mapping and*  
 720 *risk assessment for public health*, eds. A. Lawson, A. Biggeri & D. Bohning. Wiley, pp. 15–29.
- 721 ORBANZ, P. & BUHMANN, J. (2008). Nonparametric Bayesian image segmentation. *International Journal*  
 722 *of Computer Vision* **77**, 25–45.
- 723 PAPADIMITRIOU, E., FILTNESS, A., THEOFILATOS, A., ZIAKOPOULOS, A., QUIGLEY, C. & YANNIS, G.  
 724 (2019). Review and ranking of crash risk factors related to the road infrastructure. *Accident Analysis*  
 725 *& Prevention* **125**, 85 – 97.
- 726 PASCUTTO, C., WAKEFIELD, J., BEST, N., RICHARDSON, S., BERNARDINELLI, L., STAINES, A. &  
 727 ELLIOTT, P. (2000). Statistical issues in the analysis of disease mapping data. *Statistics in Medicine*  
 728 **19**, 2493–2519.
- 729 PEREYRA, M., DOBIGEON, N., BATATIA, H. & TOURNERET, J.Y. (2013). Estimating the granularity  
 730 coefficient of a Potts-Markov random field within a Markov chain Monte Carlo algorithm. *IEEE Trans.*  
 731 *Imag. Process.* **22**, 2385–2397.
- 732 RICHARDSON, S., MONFORT, C., GREEN, M., DRAPER, G. & MUIRHEAD, C. (1995). Spatial variation of  
 733 natural radiation and childhood leukaemia incidence in great britain. *Statistics in Medicine* **14**, 2487–  
 734 2501.
- 735 ROBERTSON, C., NELSON, T., MACNAB, Y. & LAWSON, A. (2010). Review of methods for space-time  
 736 disease surveillance. *Spatial and Spatio-temporal Epidemiology* **1**, 105–116.
- 737 SAAS, Y. & GOSSELIN, F. (2014). Comparison of regression methods for spatially-autocorrelated count data  
 738 on regularly- and irregularly-spaced locations. *Ecography* **37**, 476–489.
- 739 SCHLATTMANN, P. & BÖHNING, D. (1993). Mixture models and disease mapping. *Statistics in Medicine*  
 740 **12**, 1943–1950.
- 741 SETHURAMAN, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica* **4**, 639–650.
- 742 STOHR, J. (2017). A review on statistical inference methods for discrete Markov random fields. *arXiv*  
 743 *e-prints* , arXiv:1704.03331v1[04.03331].
- 744 THEOFILATOS, A. & YANNIS, G. (2014). A review of the effect of traffic and weather characteristics on  
 745 road safety. *Accident Analysis & Prevention* **72**, 244 – 256.
- 746 TRUONG, L. & CURRIE, G. (2019). Macroscopic road safety impacts of public transport: A case study of  
 747 Melbourne, Australia. *Accident Analysis & Prevention* **132**, 105270.
- 748 VIGNES, M., BLANCHET, J., LEROUX, D. & FORBES, F. (2011). SpaCEM<sup>3</sup>: a software for biological  
 749 module detection when data is incomplete, high dimensional and dependent. *Bioinformatics* **27**, 881–  
 750 882.
- 751 WALLER, L. & CARLIN, B. (2010). Disease mapping. In *Handbook of spatial statistics*, eds. A. Gelfand,  
 752 P. Diggle, P. Guttorp & M. Fuentes, vol. 2010, chap. 14. Handbook of Modern Statistical Methods. Boca  
 753 Raton: Chapman & Hall, CRC Press, pp. 217–244.
- 754 WANG, Y. & KOCKELMAN, K.M. (2013). A poisson-lognormal conditional-autoregressive model for  
 755 multivariate spatial analysis of pedestrian crash counts across neighborhoods. *Accident Analysis &*  
 756 *Prevention* **60**, 71–84.