Hidden Markov models and disease mapping

Peter J. Green* — University of Bristol, UK.

Sylvia Richardson† — Imperial College, London.

26 March 2002

Abstract

We present new methodology to extend Hidden Markov models to the spatial domain, and use this class of models to analyse spatial heterogeneity of count data on a rare phenomenon. This situation occurs commonly in many domains of application, particularly in disease mapping.

We assume that the counts follow a Poisson model at the lowest level of the hierarchy, and introduce a finite mixture model for the Poisson rates at the next level. The novelty lies in the model for allocation to the mixture components, which follows a spatially correlated process, the Potts model, and in treating the number of components of the spatial mixture as unknown. Inference is performed in a Bayesian framework using reversible jump MCMC. The model introduced can be viewed as a Bayesian semiparametric approach to specifying flexible spatial distribution in hierarchical models.

Performance of the model and comparison with an alternative well-known Markov random field specification for the Poisson rates are demonstrated on synthetic data sets. We show that our allocation model avoids the problem of oversmoothing in cases where the underlying rates exhibit discontinuities, while giving equally good results in cases of smooth gradient-like or highly autocorrelated rates. The methodology is illustrated on an epidemiological application to data on a rare cancer in France.

Some key words: Allocation, Bayesian hierarchical model, Disease mapping, Finite mixture distributions, Heterogeneity, Hidden Markov models, Markov chain Monte Carlo, Poisson mixtures, Potts model, Reversible jump algorithms, Semiparametric model, Spatial mixtures, Split/merge moves.

*Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK.
Email: `P.J.Green@bristol.ac.uk`.

†Department of Epidemiology and Public Health, Imperial College School of Medicine,
Norfolk Place, London, W2 1PG.
Email: `sylvia.richardson@ic.ac.uk`

1 Introduction

1.1 Hidden Markov random fields

Hidden Markov models (HMMs) assume in general terms that the observations form a noisy realisation of an underlying process that has a simple structure with Markovian dependence. The most studied class of HMMs has been that of temporal observations linked to an underlying Markov chain. This formulation originated in engineering, and has since been used in many domains ranging from finance to molecular biology. A comprehensive review of mathematical properties, statistical treatment and applications of hidden Markov chains can found in the recent chapter of Künsch (2001). Robert, et al, (2000) is a recent study that is particularly relevant to the work presented here, because of the variable number of states in the hidden chain, the Bayesian treatment, and the use of reversible jump algorithms in the implementation.

When the data are spatially structured, a natural extension is to hidden Markov random fields, that is, Markov random fields degraded by (conditionally) independent noise. One context in which such models have been much used is image analysis, going back to Besag (1986) and beyond; another is that of disease mapping in epidemiology. By disease mapping, we refer to studies aiming to uncover a potential spatial structure in disease risk when analysing small numbers of observed health events in a pre-defined set of areas. In this case, the noise is related to the rarity of the health event and the size of the population at risk, leading to the low disease counts per area (less than 10) typically found in many studies (Elliott et al, 2000).

In this paper, we consider a class of hidden discrete-state Markov random field models related to an underlying finite mixture model that allows spatial dependence and does not predefine the number of "states" or components of the mixture. Our motivation is dual: we are interested in generalising some of the useful features of hidden Markov chains with an unknown number of discrete states, as well as proposing a flexible alternative to the current Markov random field models commonly used in disease mapping. This motivating context has both stimulated the model development and driven some of the choices made in the implementation.

Although the extension from a hidden Markov chain to a hidden random field is an obvious one in modelling terms, it introduces disproportionate difficulties in implementation. In the case of a linear chain, there are fast methods such as the 'forward-backwards' algorithm for computing likelihoods, and if a full Bayesian approach via MCMC is taken, the normalising constants of the joint prior distribution of the hidden chain are explicitly available. Neither of these is true of the general spatial case, and this has severely limited application of these methods; see Rydén and

Titterington (1998) for a full discussion. One of the aims of the present paper is to demonstrate that implementation is perfectly practicable, at least for a moderate number of hidden variables, and using a model with just two parameters (the number of hidden states, and the strength of interaction).

1.2 Models for spatially correlated count data

We consider the modelling of spatial heterogeneity for count data on a rare phenomenon, observed in a pre-defined set of areas. Throughout, our terminology refers to disease mapping, but we stress that our model is easily translatable to other contexts where spatial heterogeneity is of interest, for example in ecology or agricultural science.

There are many reasons for suspecting heterogeneity in an underlying disease event rate and wanting to characterise it. For example, the discovery of either local discontinuities or smooth gradients can be exploited for further study or action. Indeed, the suspicion of a local excess in disease occurrence or the highlighting of geographical inequalities in health are important public health concerns that can be addressed by an analysis of the spatial heterogeneity. Of course, the analysis must take into account all relevant risk factors that can be assessed at the area level. But it is hardly plausible that all the factors acting on the underlying disease risk can be identified or measured at the required geographical level. Thus, there often remains residual heterogeneity in the disease event rate, which is moreover likely to have a spatial structure inherited from some of the unmeasured or undiscovered risk factors for the disease. Note that epidemiological studies are observational by nature and there is little or no control over the sources of variability. Furthermore, the delicate issue of ecological bias (Greenland and Robins, 1994) has to be born in mind when interpreting sources of variability for disease outcomes analysed at an aggregated level.

Modelling spatial heterogeneity of rare counts has usually been addressed in a hierarchical framework. We shall do the same in the development of our Bayesian approach, and specifically consider a Poisson model at the lowest level of the hierarchy:

$$y_i \sim \text{Poisson}(\lambda_i E_i) \quad \text{independently for } i = 1, 2, \ldots, n \tag{1}$$

where y_i denotes the observed count of disease incidences or deaths in area i, E_i is the expected count based on population size, adjusted for age and sex, etc., and λ_i is an area-specific relative risk variable, the main object of our inference. We use the simple term "risk" to refer to the $\{\lambda_i\}$ in future.

This model may be extended straightforwardly to accommodate dependence on covariates $\{x_{ij}\}$

measured in each area i, so that (1) is replaced, for example, by

$$y_i \sim \text{Poisson}(\lambda_i e^{\sum_j x_{ij}\gamma_j} E_i) \quad \text{independently for } i = 1, 2, \ldots, n. \tag{2}$$

Illustrations of the use of the model both with and without covariates will be found later in the paper.

We now consider the choice of structure for the joint distribution of the $\{\lambda_i, i = 1, 2, \ldots, n\}$ at the next level of the hierarchy, a choice which can be influential on the effective smoothing of the Poisson noise. In the seminal work of Besag, York and Mollié (1991) and Clayton and Bernardinelli (1992), and subsequent papers, loglinear Gaussian models for the $\{\lambda_i\}$ were postulated using a conditional formulation that included a spatial autoregressive component based on contiguity in an undirected graph as well as a term modelling unstructured variability. This approach has been commonly adopted in recent work in disease mapping, and has helped to highlight many interesting features of the geographical distribution of some rare diseases. Alternative formulations of a multivariate Gaussian model for the $\{\log \lambda_i\}$ that directly specify a spatially parametrised covariance matrix have also been discussed (Best, et al, 1999, Wakefield and Morris, 1999). In both cases, the parameters characterising the spatial dependence are constant across the entire study region, although models where the strength of spatial interaction is allowed to vary spatially have also been proposed in other contexts (Clifford, 1986, Aykroyd and Zimeras, 1999). When using these parametric models, there is the potential risk of over-smoothing and masking of local discontinuities, due to the global effect of the parameters. Concern about this is borne out by empirical studies, including that reported in Section 4 below.

There have been several attempts to address this difficulty, which have in common the replacement of a continuously varying random field for $\{\lambda_i\}$ by an allocation or partition model of the form

$$\lambda_i = \lambda_{z_i} \tag{3}$$

where $\{\lambda_j, j = 1, 2, \ldots, k\}$ characterise k different components, and $\{z_i, i = 1, 2, \ldots, n\}$ are *allocation variables* taking values in $\{1, 2, \ldots, k\}$. Moving the spatial dependence one level higher in the hierarchy to the discrete-valued process $\{z_i\}$ has the potential to provide a greater degree of spatial adaptivity, again seen empirically. Note that discreteness in the prior is not imposed on posterior inference, in the sense that, marginalising over the allocations, the posterior mean risk surface from any partition model can provide a smooth estimate of the risk surface. Models that can be described in this framework include the clustering or segmentation models of Knorr-Held and Raßer (2000) and Denison and Holmes (2001), which propose different spatial models for $\{z_i\}$.

In the model investigated in this paper we propose to use a Potts model for $\{z_i\}$, with the number of states and strength of interaction unknown. In contrast to the partition models cited, we retain a Markovian structure for the $\{z_i\}$.

Yet other models in this class are the spatial mixture models introduced by Fernández and Green (2000), in which the spatial dependence is pushed yet one level higher: the $\{z_i\}$ are conditionally independent given weights $w_{ij} = P(z_i = j)$ constructed from Gaussian random fields.

1.3 Mixtures and other related models

Mixture models arise naturally whenever the existence of unknown subpopulations corresponding to different models for the quantity of interest can be hypothesised. They have found applications in many contexts, some of which are illustrated, along with comprehensive accounts of the theory, in the monographs by Titterington et al (1985) and McLachlan and Peel (2000). It is not always possible or advisable to interpret the subgroups of areas identified by such models directly, so an important second perspective on the hidden Markov model adopted here is to view it as a semiparametric approach, following recent developments in both Bayesian and frequentist settings. Indeed, the question of the specification or the potential misspecification of the distribution of latent variables has been the subject of much attention (see for example Roeder, et al, 1996, and Carroll, et al, 1999, for a discussion of such issues in the measurement error context) and mixtures of distributions have been proposed as an alternative. Our proposed model is given extra flexibility by the treatment of the number of allocation classes and the strength of spatial interaction as unknowns, to be estimated together with the Poisson parameters.

Image analysis is another context in which hidden Markov random field models have been much used. Tjelmeland and Besag (1998) provide a systematic study of Markov random fields with higher order interactions, with the aim of producing well-calibrated posterior inference that goes beyond simple restoration. Johnson (1994) allows a variable number of labels as we do, but generates rich geometrical structures by constructing specific non-local potential functions. We stress that these extensions aim to recover high level features of the image, an aim different from ours, of flexible analysis of spatial heterogeneity.

The paper is structured as follows. In Section 2, we present the spatially correlated allocation model. Our Markov chain Monte Carlo (MCMC) implementation, which requires variable dimension moves, is described in Section 3. In Section 4, the performance of the model is analysed on a collection of synthetic data sets designed to test different features of the model in the context of disease mapping data. The results of a simulation study aimed at comparing some aspects of its performance with that of the Markov random field formulation of Besag, York and Mollié (1991) are also presented. An epidemiological application to French cancer data is discussed in Section 5. We conclude with a discussion of extensions and further work.

2 Potts models with Poisson noise

The main novelty of our approach lies in the modelling of the allocation variables $\{z_i\}$ in (3). First, the number of components k is treated as unknown, with prior distribution $p(k)$, typically either truncated Poisson in form, or uniform on some range $\{1, 2, \ldots, k_{\max}\}$. Then, given k, $\{z_i\}$ follows a spatially-correlated process.

The formulation for this process is built on a prescribed undirected graph, which plays the role of the prior conditional independence graph of the hidden random field $\{z_i\}$. Two areas i and i' are said to be neighbours, written $i \sim i'$, $i \in \partial i'$ or $i' \in \partial i$, if they are adjacent with respect to this graph. Typically, areas are taken to be neighbours in this sense if and only if they are spatially contiguous. More sophisticated spatial relationships can be modelled with little difficulty and we give an example of this in Section 4.6. Apart from this specific development, we always use spatial contiguity as our graph structure.

In this paper, we concentrate on the Potts model, an allocation model often used in image processing applications and originating in statistical physics. In contrast to the hierarchical mixture model defined in Richardson and Green (1997), and indeed most mixture models, this formulation does not make use of explicit weights on components.

In the Potts model formulation, the z_i are modelled jointly:

$$p(z|\psi) = e^{\psi U(z) - \theta_k(\psi)} \tag{4}$$

where

$$U(z) = \sum_{i \sim i'} I[z_i = z_{i'}] \quad \text{and} \quad \theta_k(\psi) = \log\left(\sum_{z \in \{1,2,\ldots,k\}^n} e^{\psi U(z)}\right) \tag{5}$$

are the *number of like-labelled neighbour pairs* in the configuration z, and an additive normalising constant. The interaction parameter ψ is non-negative: $\psi = 0$ corresponds to independent allocations, uniformly on the labels $\{1, 2, \ldots, k\}$. The degree of spatial dependence increases with ψ, while allocations remain *marginally* uniform on $\{1, 2, \ldots, k\}$. It is clear that for positive ψ, $p(z|\psi)$ favours probabilistically those allocation patterns where like-labelled locations are neighbours.

To complete the model specification, we must define our prior models for λ, ψ and k.

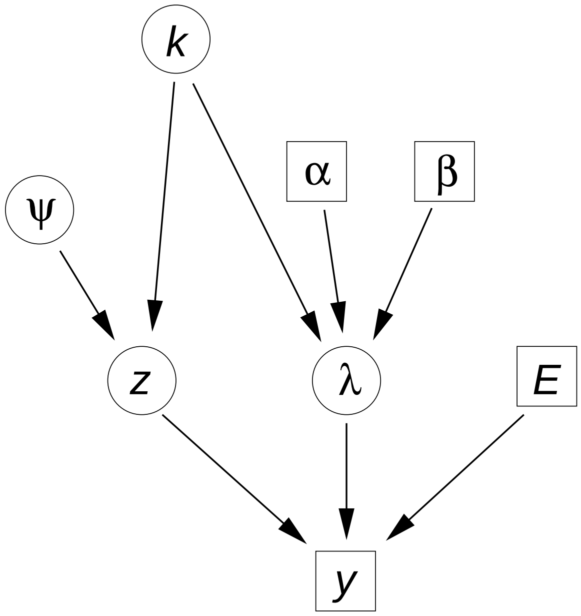

Figure 1: DAG for the Potts spatial mixture model.

We place an independence prior on $\{\lambda_j, j = 1, 2, \ldots, k\}$:

$$\lambda_j \sim \Gamma(\alpha, \beta) \quad \text{independently for } j = 1, 2, \ldots, k.$$

Although we have occasionally considered also a hierarchical model in which $\beta \sim \Gamma(b_1, b_2)$, we usually take the hyperparameters α and β as fixed. Our standard choice is $\alpha = 1$, $\beta = \sum_i E_i / \sum_i y_i$. Usually in epidemiological applications, $\sum_i E_i = \sum_i y_i$, hence $\beta = 1$; in other cases, this choice makes the analysis equivariant to multiplicative misspecification of the $\{E_i\}$. Although it is not strictly necessary, we prefer to ensure identifiability of the labelling of mixture components by indexing the $\{\lambda_j\}$ in numerically increasing order: $\lambda_1 < \lambda_2 < \cdots < \lambda_k$. Thus the joint prior for λ becomes

$$p(\lambda|k, \alpha, \beta) = k! I[\lambda_1 < \lambda_2 < \cdots < \lambda_k] \prod_{j=1}^{k} \frac{\beta^\alpha \lambda_j^{\alpha-1} e^{-\beta\lambda_j}}{\Gamma(\alpha)}$$

The whole issue of labelling and the impact of ordering on MCMC performance is comprehensively discussed in the discussion and rejoinder to Richardson and Green (1997); see also Stephens (2000) for an alternative approach.

We take $p(\psi)$ to be a discrete distribution: Uniform on the values $\{0, 0.1, \ldots, \psi_{\max}\}$. The uniformity is arbitrary – as usual, other forms of distribution could be substituted after sampling, using importance reweighting. The discreteness, which we do not believe has a significant impact on our inference, is for the sake of computational convenience, since the normalising constants $\theta_k(\psi)$ can then be pre-computed offline and stored in a table, with no approximation or interpolation necessary at run-time. Finally, our prior on the number of components k is uniform on the values $\{1, 2, \ldots, k_{\max}\}$.

Prior simulations are useful to inform the choice of $\psi_{\max}$. For the Potts model on the contiguity graph of the French départements, used in our studies in Section 4 below, we found that the value $\psi = 1.0$ is a very high level of interaction: the average prior probability that two neighbouring regions have the same label when $\psi = 1.0$ and $k = 2$ is 0.96, only declining to 0.70 when $k = 8$. We thus choose default values of $\psi = 1.0$ and $k = 10$ that seem sufficient to ensure flexibility for all practical purposes, but these could easily be increased if thought unsuitable for a specific graph. In fact, exceptionally, in one of our examples below, we extend $\psi_{\max}$ up to 1.2.

All the above specifications are expressed somewhat loosely, in the interests of economy of notation. Each of the model ingredients is really a conditional distribution for the stated variable, conditional on both its immediate parameters, and hyperparameters higher up in the directed acyclic graph (DAG) describing the model; see Figure 1. Thus the joint distribution of all variables corresponding to the Potts model formulation is

$$p(k)p(\psi)p(\lambda|k, \alpha, \beta)p(z|k, \psi)p(y|\lambda, z).$$

3 Markov chain Monte Carlo

Naturally, Markov chain Monte Carlo (MCMC) methods are needed to fit these spatial mixture models.

Our sampler for the Potts mixture model uses four different moves, each updating a subset of the variables, under detailed balance with respect to the posterior distribution, which form an irreducible Markov chain when used together in a deterministic scan. In this general structure, our computational method follows other recent work, including that of Richardson and Green (1997). Three of the four are standard fixed-dimension moves, the fourth proposes to change dimension by increasing or decreasing the number of components.

3.1 Fixed dimension moves

The three fixed dimension moves update respectively the allocations z, the spatial interaction parameter ψ and the component parameters $\{\lambda_j\}$.

The allocations z are updated by a Gibbs kernel. The full conditional from which an update for z_i is drawn is

$$p(z_i = j|\cdots) \propto e^{-\lambda_j E_i} \lambda_j^{y_i} e^{\psi n_{ij}},$$

where $n_{ij} = \#\{i' \in \partial i : z_{i'} = j\}$ is the number of neighbours of i currently assigned to component j. Note that in contrast to the simple random sample mixture case of Richardson and Green (1997), here the $\{z_i\}$ are not conditionally independent given all other variables, so that they may not be updated simultaneously.

The interaction parameter ψ has full conditional

$$p(\psi|\cdots) \propto p(\psi)e^{\psi U(z)-\theta_k(\psi)},$$

a discrete distribution on a finite grid of values, like the prior. A random walk Metropolis kernel, proposing perturbations of ± 0.1 with equal probability, is convenient for updating ψ.

Finally, we need to update the parameters $\{\lambda_j\}$. An approach to the simultaneous update of these, maintaining the order restriction, exploits the following simple trick that we have not seen elsewhere. We propose simultaneous independent zero-mean normal increments to each $\log \lambda_j$; the modified values of λ are then placed in increasing order to give $\{\lambda'_j\}$ say. Remarkably, the fact that the proposal density we are using is actually a sum of $k!$ rather complicated terms, due to the re-ordering, does not matter: the terms that appear in the sums in the numerator and denominator of the Metropolis-Hastings ratio are in constant proportion, and so cancel out. The acceptance probability for the complete set of updates, formed from prior ratio, likelihood ratio and a Jacobian for the log transformation, reduces to

$$\min\left\{1, \prod_{j=1}^{k}\left[\left(\frac{\lambda'_j}{\lambda_j}\right)^{\alpha+\sum_{i:z_i=j} y_i} \exp\{-(\lambda'_j - \lambda_j)(\beta + \sum_{i:z_i=j} E_i)\}\right]\right\}.$$

3.2 Variable dimension move

Changing the number of components under detailed balance with respect to the posterior requires a reversible jump move (Green, 1995). We follow the general idea of a random choice between splitting an existing component into two, and merging two existing components into one, as used by Richardson and Green; the probabilities of these alternatives are b_k and d_k when there are currently k components. As well as incrementing or decrementing k, the move entails modifying the vector λ accordingly, and reallocating observations into the new components(s) as necessary.

In contrast to Richardson and Green, here the re-allocation part of the proposal is not done independently for each observation, but performed in a way that approximately respects the spatial structure of the Potts model. This is with the usual aim of increasing the probability of the move's acceptance; exact detailed balance is of course assured by correctly calculating the acceptance probability in terms of the model and proposal probabilities.

We will describe the split move in some detail. First a component to be split, j say, is chosen uniformly at random from $\{1, 2, \ldots, k\}$. This is replaced by two components which we label '$-$' and '$+$', with λ values generated by

$$\lambda_- = \lambda_j u^c \quad \text{and} \quad \lambda_+ = \lambda_j u^{-c}$$

where u is generated from U(0,1), and c is a proposal spread parameter that we set at 0.1. If $\lambda_- < \lambda_{j-1}$ or $\lambda_+ > \lambda_{j+1}$ (with appropriate modifications if $j = 1$ or k), the move is rejected, as the mis-ordered vector has zero density under the ordered prior. Those observations i currently allocated to component j are then dynamically re-allocated between $-$ and $+$. We scan over such i, and the probability with which z_i is set to $-$ rather than $+$ is

$$\frac{e^{\psi n_- - \lambda_- E_i} \lambda_-^{y_i}}{e^{\psi n_- - \lambda_- E_i} \lambda_-^{y_i} + e^{\psi n_+ - \lambda_+ E_i} \lambda_+^{y_i}},$$

where n_- and n_+ are the numbers of areas adjacent to i already proposed for allocation to $-$ and $+$ in this scan. This choice has the affect of mimicking the Potts model term in the target distribution to favour proposed allocations with spatial coherence. As the scan proceeds, the probability P_{alloc} of the allocation actually generated is accumulated. Denote the proposed new allocation vector by z'. Following very similar logic to Richardson and Green (1997, equation (11)), the acceptance probability for this complete proposal is $\min\{1, R\}$ where

$$\begin{aligned} R &= \prod_{i:z_i'=-} e^{-(\lambda_- - \lambda_j)E_i} \left(\frac{\lambda_-}{\lambda_j}\right)^{y_i} \prod_{i:z_i'=+} e^{-(\lambda_+ - \lambda_j)E_i} \left(\frac{\lambda_+}{\lambda_j}\right)^{y_i} \\ &\times \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{\lambda_-\lambda_+}{\lambda_j}\right)^{\alpha-1} e^{-\beta(\lambda_- + \lambda_+ - \lambda_j)} (k+1) \frac{p_{k+1}}{p_k} \\ &\times \exp\{\psi(U(z') - U(z)) + \theta_k(\psi) - \theta_{k+1}(\psi)\} \times \frac{d_{k+1}}{b_k P_{\text{alloc}}} \times \frac{2c\lambda_j}{u}. \end{aligned}$$

3.3 Approximating the Potts model partition function

While the MCMC moves for z and λ make no reference to $\theta_k(\psi)$, values of this normalising constant, the logarithm of what is known as the partition function in statistical physics, are needed for the update for ψ and the split/merge move. Since in our model both k and ψ are discrete, we need to evaluate $\theta_k(\psi)$ on a grid of (k, ψ) pairs. These are computed in offline simulations, specific to the assumed neighbourhood graph, and provided in a look-up table for our MCMC sampler to use.

We have found the following simple method for estimating $\theta_k(\psi)$ easy to use and reliable on graphs of the size we have encountered in real disease-mapping applications.

The thermodynamic integration approach

This method has a long history; according to Gelman and Meng (1998) it has been used in statistical physics since the 1970's, and Ogata and Tanemura (1984) are responsible for its first use in spatial stochastic processes.

Consider the Potts model with k labels on a graph with n vertices, defined in (4) and (5). Differentiating $\theta_k(\psi)$, we obtain

$$\frac{\partial}{\partial\psi}\theta_k(\psi) = \frac{\partial}{\partial\psi}\log\sum_{z\in Z} e^{\psi U(z)} = \sum_{z\in Z} U(z)p(z|\psi) = E(U(z)|\psi, k), \tag{6}$$

the expectation of $U(z)$ when z is distributed according to the assumed Potts model. Here $Z = \{1, 2, \ldots, k\}^n$ is the set of possible labellings of the graph. But $\theta_k(0) = \log\sum_{z\in Z} 1 = n\log k$, so

$$\theta_k(\psi) = n\log k + \int_0^\psi E(U|\psi', k)\mathrm{d}\psi'. \tag{7}$$

In particular, note that when $k = 1$, $U(z)$ is the constant n_E, the number of edges in the graph, so that $\theta_1(\psi) = \psi n_E$.

Equation (7) is the basis of a simple method for estimating $\theta_k(\psi)$: the expectation is replaced by a sample average in a MCMC simulation of the Potts model (for specific (k, ψ')). In our implementation, the integral is computed by numerical integration of a cubic spline smooth of the simulated averages with respect to ψ'.

Improving partition function estimates using the Potts model mixture sampler

Suppose we place a prior $p(k, \psi)$ on (k, ψ), and conduct a MCMC simulation of the distribution of (k, ψ, z) assuming an approximate trial value $\widetilde{\theta_k}(\psi)$ for the log partition function. This can be accomplished by the posterior sampler for the Potts model mixture model derived in Sections 3.1 and 3.2, with the data suppressed, and likelihood terms omitted from the model. We are simulating from the joint distribution

$$\propto p(k, \psi)\exp(\psi U(z) - \widetilde{\theta_k}(\psi)) \tag{8}$$

so the marginal distribution for (k, ψ) is

$$\propto p(k, \psi)\exp(\theta_k(\psi) - \widetilde{\theta_k}(\psi)). \tag{9}$$

This observation can be used in two ways. First, comparing the prior $p(k, \psi)$ to the observed frequencies $\widehat{p}(k, \psi)$, say, provides a check on the departure of $\widetilde{\theta_k}(\psi)$ from $\theta_k(\psi)$. Secondly, if we instead equate (9) to $\widehat{p}(k, \psi)$ we can solve to give improved estimates of $\theta_k(\psi)$, namely

$$\theta_k(\psi) = \widetilde{\theta_k}(\psi) + \log(\widehat{p}(k, \psi)/p(k, \psi)) \tag{10}$$

up to an additive constant. This device takes advantage of the often improved mixing offered by variable-dimension samplers. It cannot cope with very poor initial estimates, but works well as a supplement to thermodynamic integration or any other method for deriving the normalising constant.

In the numerical experiments that follow, we use thermodynamic integration, based on runs of length 50 000 for each (k, ψ) combination, followed by the improvement just described.

A full analysis of the problem of estimating normalising constants is given by Gelman and Meng (1998), who discuss several methods that are more sophisticated, but also more cumbersome; these will handle more challenging problems, and might be needed in adapting our methods to bigger graphs or a wider range of (k, ψ) values.

3.4 Implementation of the BYM model

As mentioned in the introduction, we make use of the method of Besag, York and Mollié (1991) as a standard for comparison in our experiments. This is based on a hierarchical model; given the values of the variance parameters, the logarithms of the risks have a certain multivariate normal distribution, *a priori*. In contrast, in our model, conditional on the allocations and hyperparameters, the risks are gamma distributed. To eliminate the impact on our comparisons of this basic difference, the model we implement, and refer to below as the BYM model, is a minor re-formulation of the model of Besag, York and Mollié.

We suppose that the risk in area i is $\lambda_0 e^{u_i+v_i}$ where, conditional on α, β, τ_u and τ_v, the terms λ_0, $u = (u_i)$ and $v = (v_i)$ are independent, with

$$\begin{aligned}
\lambda_0 &\sim \Gamma(\alpha, \beta) \\
p(u|\tau_u) &\propto \exp(-\tau_u \sum_{i\sim i'} (u_i - u_{i'})^2/2) \\
p(v|\tau_v) &\propto \exp(-\tau_v \sum_{i=1}^{n} v_i^2/2)
\end{aligned}$$

where both $\{u_i\}$ and $\{v_i\}$ are constrained to sum to 0. (In the case of a disconnected graph, we would apply these constraints separately in each connected component, but in fact we do not need this in our examples below.) In this formulation, if τ_u and τ_v go to ∞, we obtain the standard non-spatial conjugate model, which also arises if k is fixed at 1 in our mixture model. Allowing $k > 1$ or $\tau_u, \tau_v < \infty$ are alternative approaches to fitting spatial heterogeneity. We assume $\tau_u \sim \Gamma(\alpha_u, \beta_u)$ and $\tau_v \sim \Gamma(\alpha_v, \beta_v)$ *a priori*, with $\alpha_u = \beta_u = \alpha_v = \beta_v = 0.1$.

Constructing a sampler for the BYM model requires some care both because of the strong spatial interaction among the u variables, and the high correlation *a posteriori* between u and τ_u and between v and τ_v. In addition, the sum-to-zero constraints we impose in our form of the model pose extra difficulties. Recent work by Knorr-Held and Rue (2002) has focussed on the first of these problems. Our sampler, which seems to perform quite adequately, explicitly addresses the second and third. The variables in question are updated in two blocks: (u, τ_u) and (v, τ_v), using moves of a similar design. Taking the (u, τ_u) block as an example, we use a Metropolis-Hastings proposal on the whole vector u, making simultaneous independent Gaussian perturbations, constrained to sum to 0. The proposal is accepted or rejected by reference to the target distribution in the usual way, but with τ_u integrated out. That is, the acceptance ratio for the update from u to u' is

$$\left(\frac{\beta_u + (1/2)\sum_{i\sim j}(u'_i - u'_j)^2}{\beta_u + (1/2)\sum_{i\sim j}(u_i - u_j)^2}\right)^{\alpha_u + n/2} \exp\left\{\sum_{i=1}^{n} y_i(u'_i - u_i) - \lambda_0(e^{u'_i + v_i} - e^{u_i + v_i})\right\}.$$

Following this, whether or not the update for u is accepted, τ_u is updated by a Gibbs move, drawing a new value from its full conditional $\Gamma(\alpha_u + n/2, \beta_u + (1/2)\sum_{i\sim j}(u_i - u_j)^2)$. We adjust the spread of the Gaussian perturbations on u and v in pilot runs to achieve reasonable acceptance rates. On larger graphs, it might prove necessary to apply such perturbations to subsets of the areas instead of the whole graph. Finally, λ_0 is updated by a Gibbs kernel.

4 Model performance and comparison

4.1 The simulated data sets

In this section, the distinguishing features of the model and its performance will be investigated on simulated data sets. Throughout we use the spatial layout of the 94 mainland French départements. To test different characteristics of the model, we generated three data sets corresponding to contrasting geographical features of the underlying simulated risks. Specifically, the "Block4" case refers to a situation of a background value of λ equal to 0.7 with 4 groups of areas having higher values of λ equal to 1.5. These 4 groups consist either of a single well-populated area, or of a group of 5 rural départements or are on the border (see Figure 2). "North-South" simulates a simple North/South divide with λ equal to 0.8 in the North and 1.2 in the South (see Figure 3), while "Gradient NS" corresponds to risks smoothly (linearly) decreasing from North to South (see Figure 4). For each data set, the observed number of events were simulated as

$$y_i \sim \text{Poisson}(\lambda_i E_i) \quad \text{independently for } i = 1, 2, \ldots, n$$

where the expected numbers of events were chosen on the basis of the French population structure and correspond to real data on two types of cancer.

For "Block4" and "North-South", they correspond to the expected number of deaths for larynx cancer in females for the period 1986–1993 and range from 2 to 58, while for "Gradient NS", they are about 3 times larger, and correspond to those of the gall bladder data set analysed by Mollié (1996), where risks were found to have a gradient-like structure. The right hand panel of each figure displays the maximum likelihood estimate of λ_i, y_i/E_i, commonly referred to as the Standardised Mortality Ratio (SMR) in the epidemiological literature.

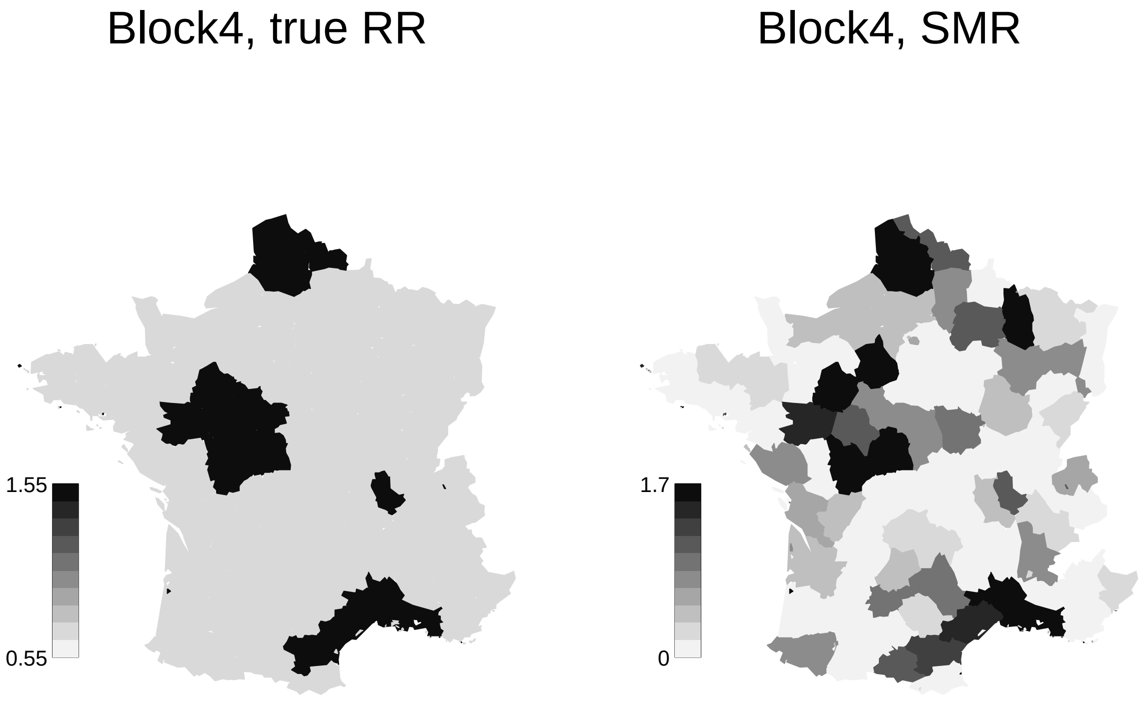

Figure 2: True risks and observed SMRs for the "Block4" data set.

4.2 Output analysis and criteria

All the results displayed correspond to runs of 500 000 sweeps of the algorithm presented in section after a burn in of 20 000 sweeps. The mixing performance of the split and merge moves was satisfactory, with acceptance rates generally around 10%, except in cases where the data supports a low number of components, as in the "North-South" example where the acceptance rate drops to 5%. From the samples, different summaries of the posterior distribution can be computed. Our main interest is in the spatial variation of the $\{\lambda_{z_i}\}$ and associated posterior probabilities.

The section will conclude by reporting some simulation results that aim to compare some aspects of the performance of our spatial mixture model to that of the BYM model. Our criteria

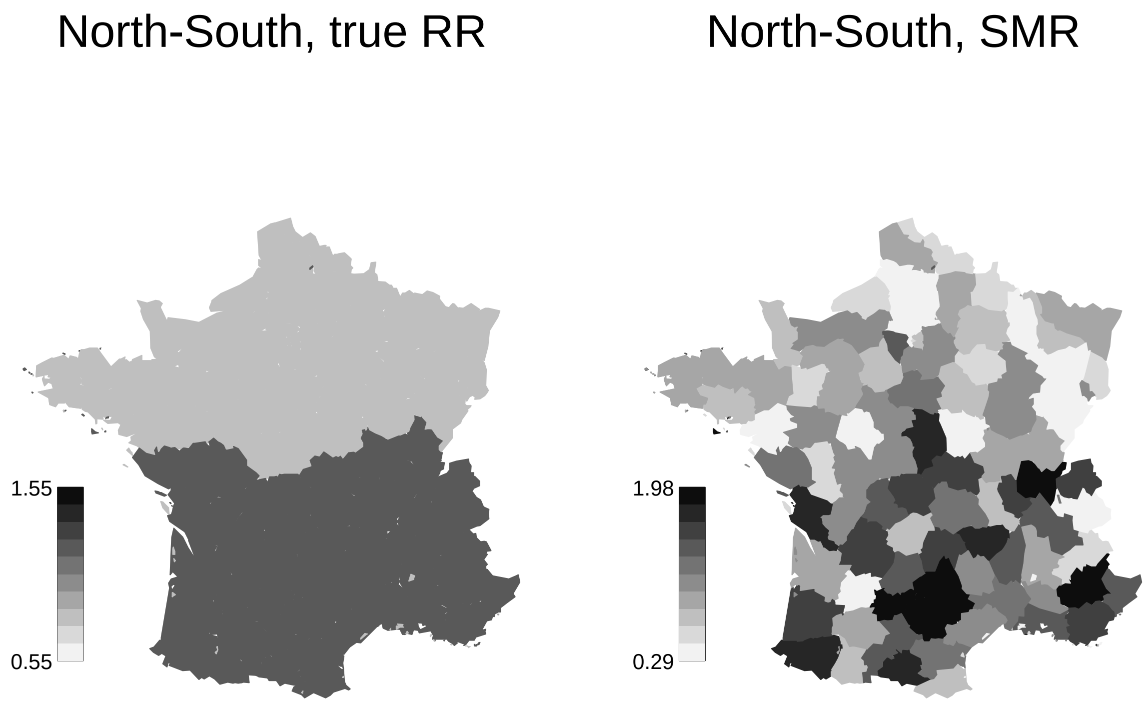

Figure 3: True risks and observed SMRs for the “North-South” data set.

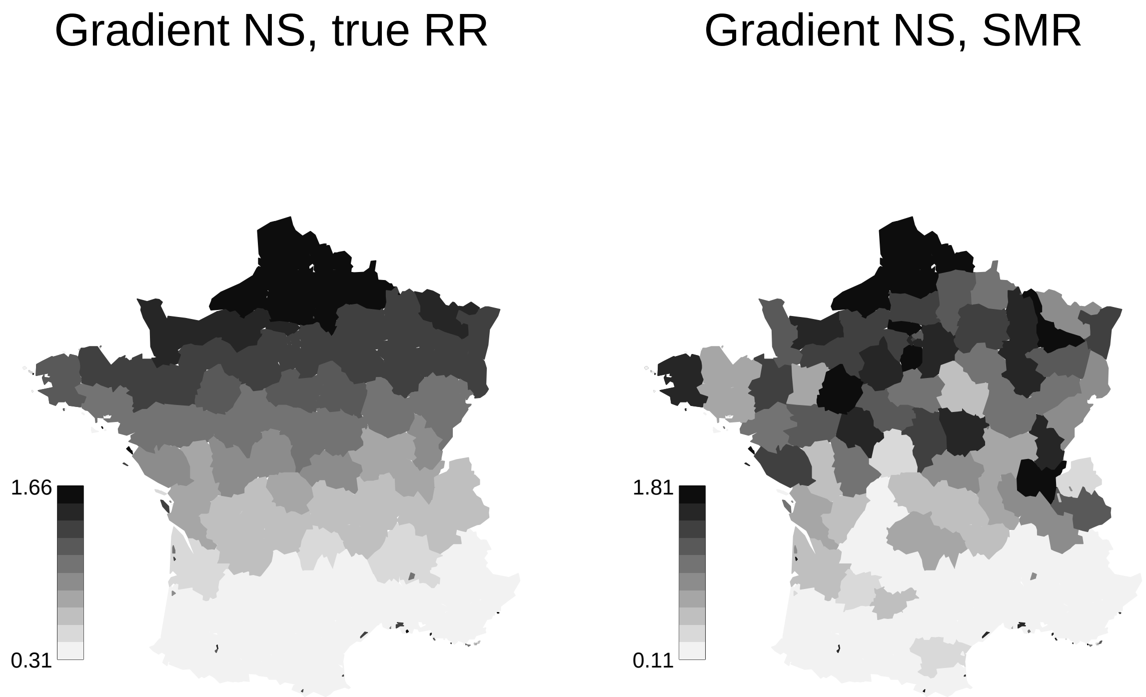

Figure 4: True risks and observed SMRs for the “Gradient NS” data set.

for comparison are fairly straightforward. For simulated data, we know the "true" underlying risks, and these will be denoted by $\{\lambda_i^t\}$. We then calculate for each area i, $\text{mse}_i = E((\lambda_i^t - \lambda_i)^2|y)$, the (posterior) mean square error, where λ_i corresponds to λ_{z_i} for the mixture model and to $\lambda_0 \exp(u_i + v_i)$ for the BYM model. To summarise the performance over the whole map, we compute RAMSE$=(\sum_i \text{mse}_i/n)^{1/2}$, the root averaged mean square error. Since this criterion has the disadvantage that it penalises multiplicative overestimation of a risk more than underestimation, we also compute a corresponding criterion on the log scale, i.e. replacing above λ_i^t and λ_i by $\log \lambda_i^t$ and $\log \lambda_i$ respectively. The corresponding summary over the whole map, denoted by RAMSEL, now treats symmetrically a risk which is, say, doubled or halved.

Turning to a measure applicable to real data, where the true risk map is not available, and which aims to balance fit and complexity, we have computed the Deviance Information Criterion (DIC) proposed by Spiegelhalter et al (1998). It is the sum of two terms: $E(D|y)$, the posterior expected deviance and p_{D}, a penalty term. $E(D|y)$ is evaluated from the MCMC output in a standard way; at each sweep values of the parameters are produced, from which can be calculated the Poisson deviance, $D = 2\sum_i(y_i \log(y_i/\mu_i) - y_i + \mu_i)$, where $\mu_i = \lambda_{z_i} E_i$. The penalty p_{D} is the difference between the posterior expectation of this deviance and the deviance at the posterior mean of the parameters. Here, we have used the posterior mean of the $\{\lambda_{z_i}\}$ for the mixture model and that of $\lambda_0 \exp(u_i + v_i)$ for the BYM model. For each data set and each model, we report the value of DIC, of the posterior expected deviance, $E(D|y)$, and that of p_{D}, which can be interpreted as a measure of model complexity (see also Best, et al, (1999) for the use of DIC in comparison of spatial models).

4.3 Posterior inference on k and ψ

Figure 5 displays the joint posterior distribution of k and ψ for the 3 data sets. There are clear differences between the "images". There is support for low values of k in the first 2 data sets, while larger k and ψ are necessary to fit the gradient-like structure. The posterior for ψ concentrates around higher values when there exists relatively large "clusters" (that is, neighbouring areas with the same λ) in the true setup. The different types of correlation between k and ψ show the adaptivity of the mixture to different geographical patterns.

For "North-South", there is a clear peak of $p(k|y)$ for $k = 2$, the true number of components, while for "Block4", the mode of $p(k|y)$ is at $k = 3$, with the mixture model preferring to fit more than one component to model the variability of the large number of areas having the background risk.

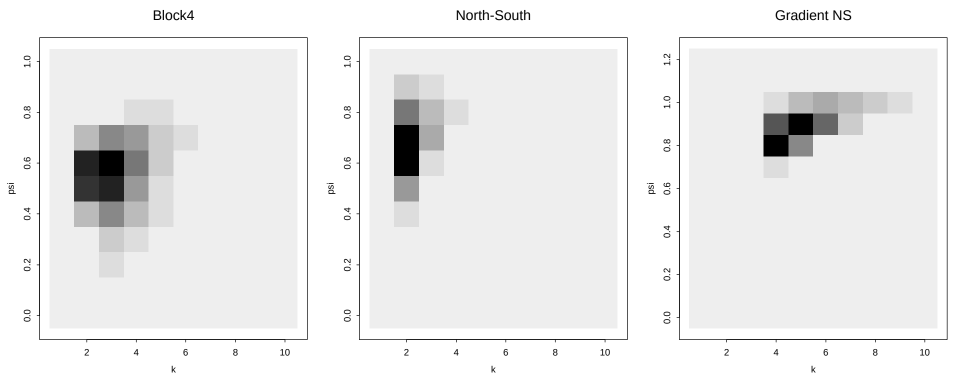

Figure 5: Joint posterior distribution of k and ψ for the 3 simulated data sets.

4.4 Posterior estimates of the $\{\lambda_{z_i}\}$

Figures 6 (left panel) and 7 show the posterior mean of λ_{z_i} for the 3 simulated data sets. Visually comparing those to the true simulated risks shows an excellent match. Due to model averaging, the posterior means of the λ_{z_i} are smoothly varying over the space, and not step-like. The flexibility of the mixture to adapt to very different patterns of risks is apparent. The right panel of Figure 6 displays the map of posterior standard deviations of the λ_{z_i} for "North-South". It is noticeable that variability is a little higher on the border areas between the contrasting zones; of course this is also modulated by the size of the expected counts. As mean estimates can be a misleading summary in cases of high variability or skewness, Figure 8 (left panel) displays for "Block4" a representation for each area of the posterior distribution of the $\{\lambda_{z_i}\}$ as a 5 bin histograms with break points at 0.7, 0.9, 1.1, and 1.3. We see that the histograms corresponding to areas of simulated elevated risk in the 4 blocks are clearly right skewed in comparison to the majority of left skewed histograms corresponding to the background areas. It is also interesting to display maps of posterior probabilities that the risk in each area exceeds certain thresholds. These can easily be computed from the output. In Figure 8 (right panel), we see that by looking at Prob(RR $>$ 1) in the "North-South" example, we have a clear indication that the Southern areas have more elevated risks than the Northern ones.

Another posterior summary that is easily obtainable from our mixture model is the posterior distribution of the allocations z_i between different components, conditional on values of k. Such allocation graphs are simple and visually effective at isolating areas of particularly high or low risk. But their interpretation is conditioned by the separation between the components and in the examples that we looked at, we did not find that such graphs uncover new features not visually apparent in the histogram of the $\{\lambda_{z_i}\}$, One feature that is important to point out is the wide posterior variability of these allocations. Indeed, *a priori*, our model assumes that for each k the

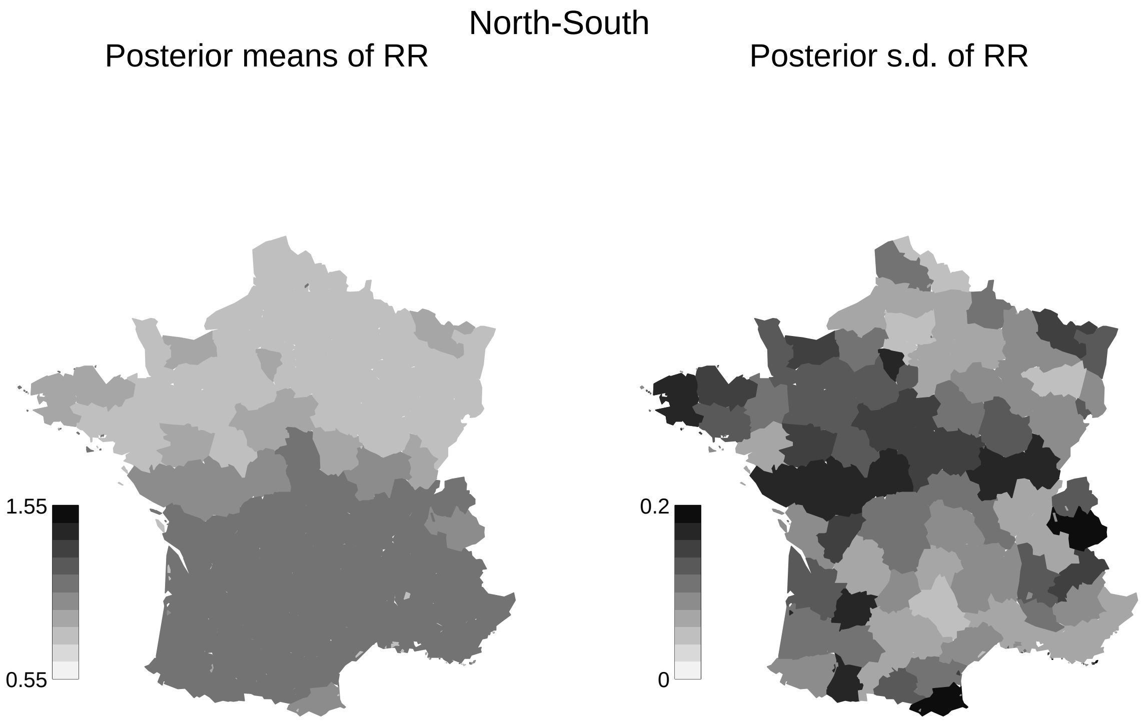

Figure 6: Mixture estimates of the risks for "North-South", posterior means (left panel), posterior standard deviations (right panel).

different labels are equally probable. This prior assumption, akin to that of having uniformly distributed weights which is usually made in mixture models, nevertheless allows the posterior allocations to be far from uniform when there is information in the data. This can be seen in Figure 9 where the box plots of the modal allocation probabilities are displayed for the "Block4" and the "Gradient NS" data sets and a selected range of values of k. To be precise, for each area i and each k, we determine $\max_j\{P(z_i = j|k, y)\}$ and form a box plot of these probabilities over i.

For "Block4", most of the modal allocation probabilities are above 0.90 when $k = 2$, and reflect the real contrast in the data. When $k \geq 3$, the areas with background risk are split between several components with close values of λ, and their allocation probabilities are much closer to their prior mean of $1/k$ (indicated by a dot in Figure 9) as could be expected, since there is little information in the data about these further components. For "Gradient NS", there is substantial structure in the data and the modal allocation probabilities reflect this, being well above $1/k$ for all values of k.

4.5 Clusters

We define a cluster to be a set of like-labelled areas connected by paths from neighbour to neighbour. To be precise, areas i and j are in the same cluster if there is a path $i = l_0 \sim l_1 \sim \cdots \sim l_r = j$ such

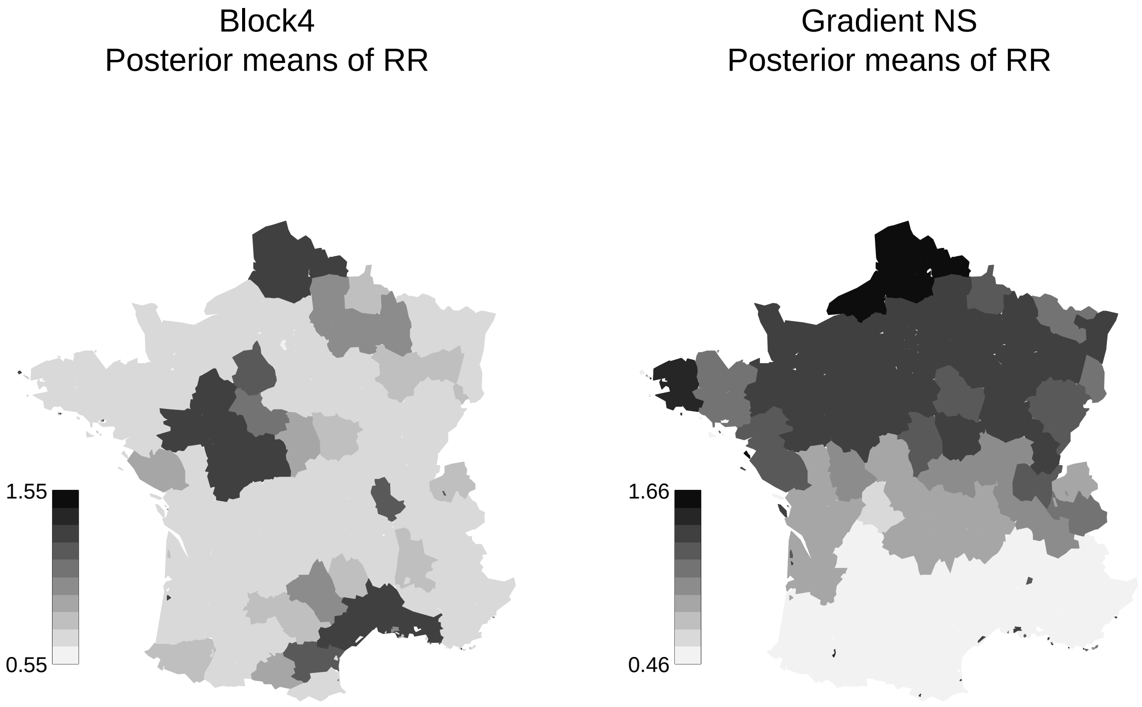

Figure 7: Mixture estimates of the risks for "Block4" (left panel) and "Gradient NS" (right panel): posterior means.

that z_{l_p} is the same for all $0 \leq p \leq r$. Note that in the segmentation approach of Knorr-Held and Raßer (2000), each cluster is differently labelled, while in our model, disconnected areas can have the same label. It is therefore interesting to compare the prior and posterior distributions of the number of clusters m. This can be done conditionally on a fixed value of k, and we will do this in our synthetic examples for the "true" and the modal k, or integrating over k using all the output.

Figure 10 displays the prior distribution of m (integrating over the uniform priors for k and ψ), as well as the posterior distribution of m for the 3 data sets (integrating over k (full line), conditional on the value of k corresponding to the "true k" when it exists (dotted line), conditional on the mode of $p(k|y)$ (dashed line)). We see peaked patterns for the three data sets, contrasting with the flat shape and extended tail of the distribution of m for the prior model. This shows that the spatial pattern in the data resulted in concentrating the posterior distribution of m on smaller values.

A simple structure for $p(m|y)$ is apparent for "North-South" and "Gradient NS" with hardly any shift between the conditional and the overall cluster distribution. On the other hand, for "Block4", the bimodality of $p(m|y)$ indicates hesitation between a simple and a more complex clustering pattern arising when the areas with the background risks are split between more than

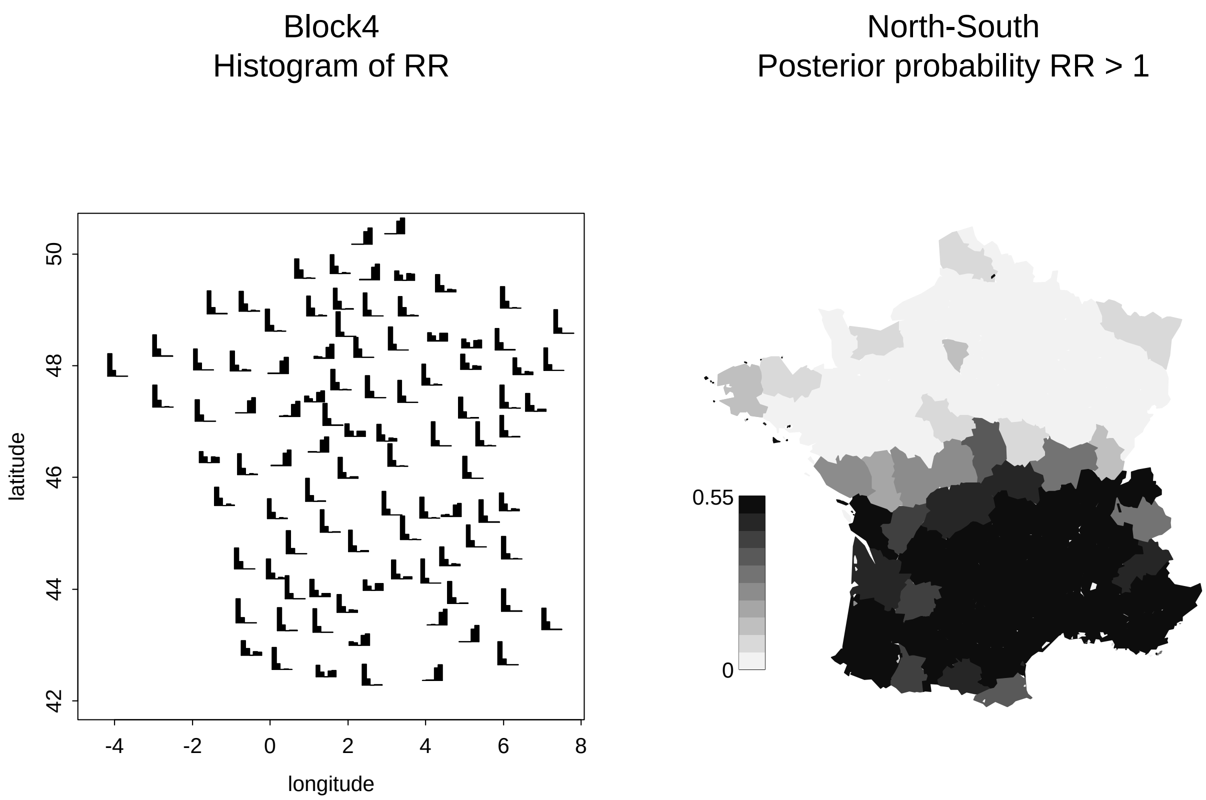

Figure 8: Examples of posterior summaries obtained using the mixture model, histograms of posterior risks for "Block4" (left panel), posterior probability of risk >1 for "North-South" (right panel).

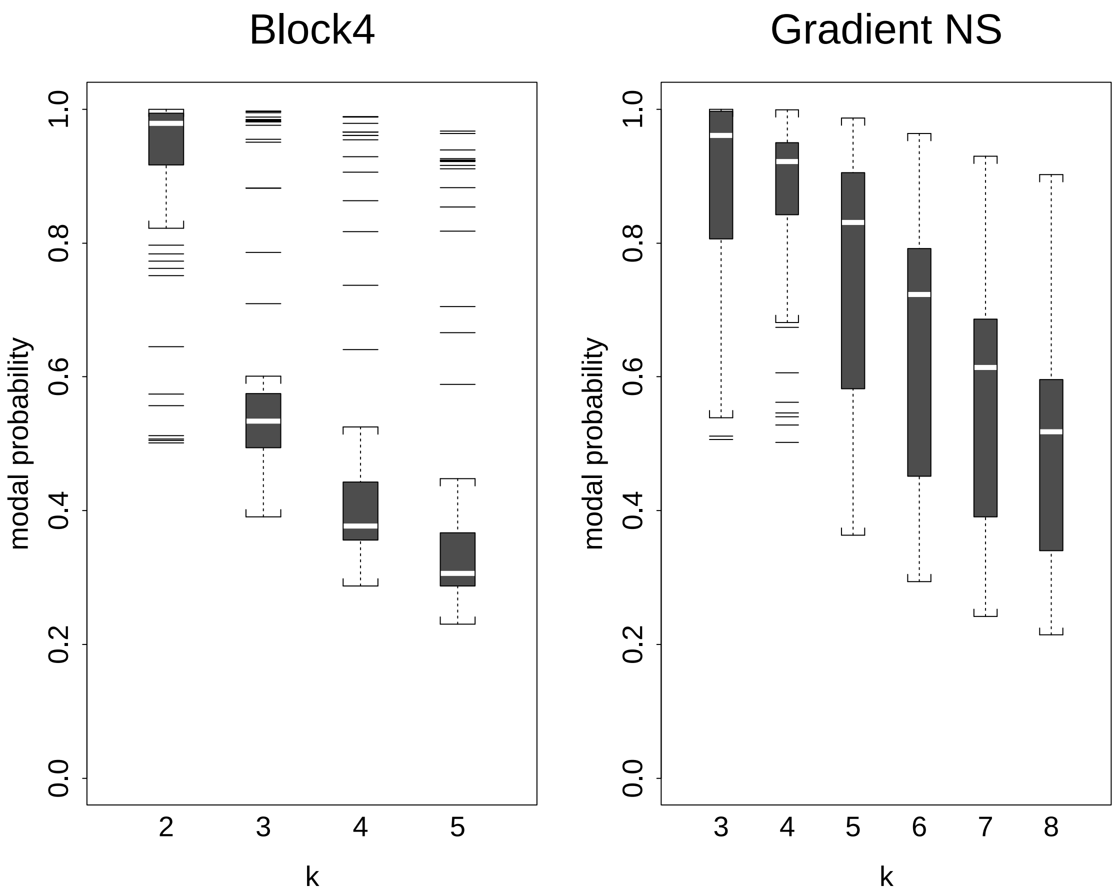

Figure 9: Box plots of the allocation probabilities for Block4 and Gradient NS data sets. The dots represent the prior means of $1/k$.

one component. We see flexibility of our mixture model in generating different cluster patterns. Clearly, the use of the Potts model in conjunction with variable k has not "frozen" the cluster pattern, as might have been anticipated from the experience of Tjelmeland and Besag (1998) using the Potts model with a fixed k in a related context in image analysis.

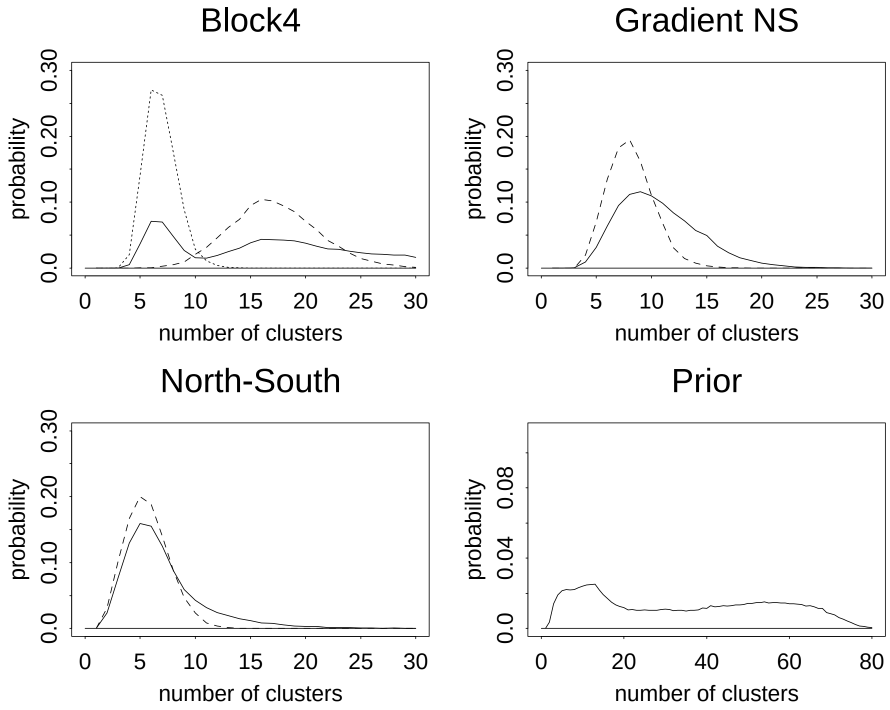

Figure 10: Posterior distributions of the number m of clusters obtained using the mixture model: integrated over k (full line), conditional on "true" k (dotted line), conditional on the posterior mode of k (dashed line). In the "Gradient NS" case, there is of course no true k; in the "North-South" case, the true and posterior modal k coincide. For the "prior" case, we plot only the integrated distribution, conditional on $k > 1$.

4.6 Assumptions and sensitivity

Our hierarchical formulation involves different levels of assumptions: distributional, quantitative (in relation to hyperparameter specification), and structural.

The distributional assumption for the observed counts has to be adapted to the data. For disease mapping, it is standard to use the Poisson assumption, but in other cases where spatial hidden Markov models are employed, for example in ecological applications, this component of the model could be replaced by alternative distributions such as the binomial or negative binomial.

As in any mixture-like problem, one might anticipate that some aspects of the model are sensitive to the choice of prior distribution for the component parameters. We have chosen to use Gamma distributions for the λs with $\alpha = 1$, $\beta = \sum_i E_i / \sum_i y_i$. We conducted a sensitivity study of this choice by letting α take the alternative values of 0.4 or 2.5, with β adjusted correspondingly so that $\alpha/\beta = \sum_i y_i / \sum_i E_i$; we also allowed an additional level in the hierarchy and treated β as random, with $\beta \sim \Gamma(b_1, b_2)$. The posterior distributions of the area-specific risks were highly stable under these various choices. For the three data sets, neither the posterior means nor the posterior standard deviations vary by more then 0.03 from their values under the standard choice $\alpha = 1$, $\beta = \sum_i E_i / \sum_i y_i$. This is a welcome feature of the model. On the other hand, as anticipated from other mixture studies that we have conducted (Richardson and Green, 1997), inference on k and z is less stable, with a tendency for the model to fit more components as the variability of the Gamma distribution is decreased. This sensitivity is the reason to be careful not to over-interpret the posterior on k, and to use the mixture simply for exploring interesting features of the heterogeneity.

The single-parameter Potts model on the nearest-neighbour graph with variable number of labels is a particular formulation of a spatial hidden Markov model that allows computational tractability and, we believe, sufficient flexibility. In some applications, allowing higher-order neighbours might be called for and this can be done without changing the computational strategy. However, it becomes more cumbersome to compute the look-up tables for the log partition function $\theta_k(\psi)$ if there are additional parameters. We have investigated the effect of replacing the 0–1 contiguity coefficients by a piecewise linear function

$$w_{ii'} = 1 \quad (d_{ii'} \leq 60); \quad (120 - d_{ii'})/60 \quad (60 < d_{ii'} \leq 120); \quad 0 \quad (d_{ii'} > 120) \tag{11}$$

of the distance $d_{ii'}$ (in km.) between the administrative centres of each area, and redefining the prior potential function $U(z)$ as $\sum_{i,i'} w_{ii'} I[z_i = z_{i'}]$, giving a modified Potts model with smoother spatial dependence, and effectively larger neighbourhoods.

We found that the posterior means for the area-specific risks are quite robust overall to the modification of the Potts model, as can be seen by comparing Figures 11 and 7. Nonetheless, differences can be seen for some areas and it is interesting to explore these further. For "Block4", the largest difference of estimated risks between the two maps was 0.2, occurring for an area in the middle of the central elevated block. This had a low SMR of 0.96, even though its true risk was 2. With the standard Potts model, the posterior mean risk for this area is raised to 1.12, because of the influence of the contiguous neighbours with elevated risks. On the other hand, with the modified model, the influence of the further-distant areas with a low true risk of 0.7 predominates, and the posterior mean risk for this area is pulled down to 0.92. A similar story can be told for "Gradient NS". The largest difference of risks between the two maps, 0.21, occurs for an area on the Northern border, the estimated risk for that area being pulled down when using the modified model. For both data sets, the DIC was lower and RAMSE was slightly smaller for the standard model. As expected, changing the model influences the posterior distribution of k. For example, mixtures with fewer components were fitted with the modified model in the "Gradient NS" case, reflecting the flexible interaction between the type of spatial structure of the data, the chosen spatial model and the estimated mixture characteristics. This modest investigation has thus indicated that the model shows good adaptivity. A full investigation of the choice of spatial model for the allocations is beyond the scope of this paper.

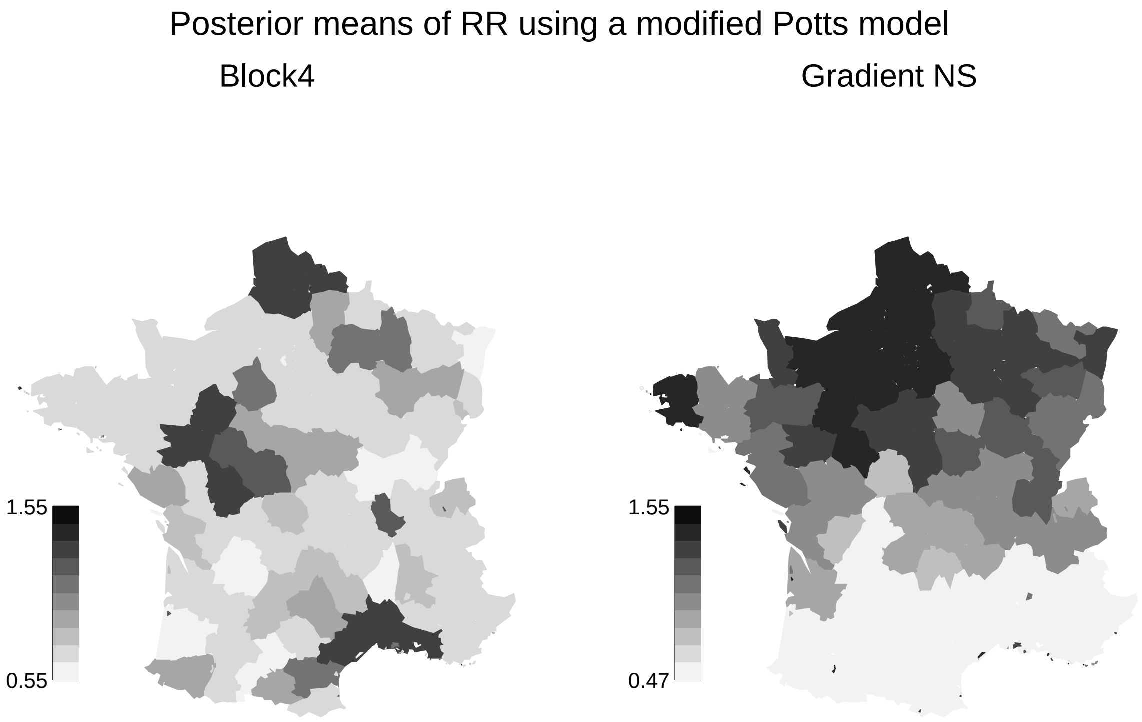

Figure 11: Mixture estimate of the risks for "Block4" (left panel) and "Gradient NS" (right panel) computed under the modified Potts model based on equation (11).

4.7 Comparison with BYM

We first comment briefly on the results of a small simulation study conducted on the basis of the three underlying risk patterns described previously, together with three other set-ups. In the first one, "CAR", the (log) risks follow an intrinsic autoregressive Gaussian model with zero mean based on the contiguity matrix W, where $w_{ij} = 1$ if $i \sim j$. To be precise, the joint distribution of $\{\log(\lambda_i)\}$ is simulated with covariance matrix 0.16 times the generalised inverse of the matrix $\text{diag}(\nu_1, \ldots, \nu_n) - W$, where ν_i is the number of neighbours of area i, and the constant 0.16 is used to scale appropriately the variability of the risk across the map. The corresponding maps of true risk and SMR are displayed in Figure 12. This set-up was chosen to correspond to the spatial model underlying the BYM analysis. The last two examples have no spatial pattern: "Flat" refers to risks displaying no trend or spatial correlation, the "null hypothesis" for an epidemiologist, taken here as drawn at random from a Uniform $(0.9, 1.1)$ distribution and "Over" simulates overdispersed risks using a gamma mixture of Poissons chosen so that $E(y) = 1.5\text{var}(y)$. For those last three data sets, the expected number of deaths are again those of the data set on larynx cancer for women.

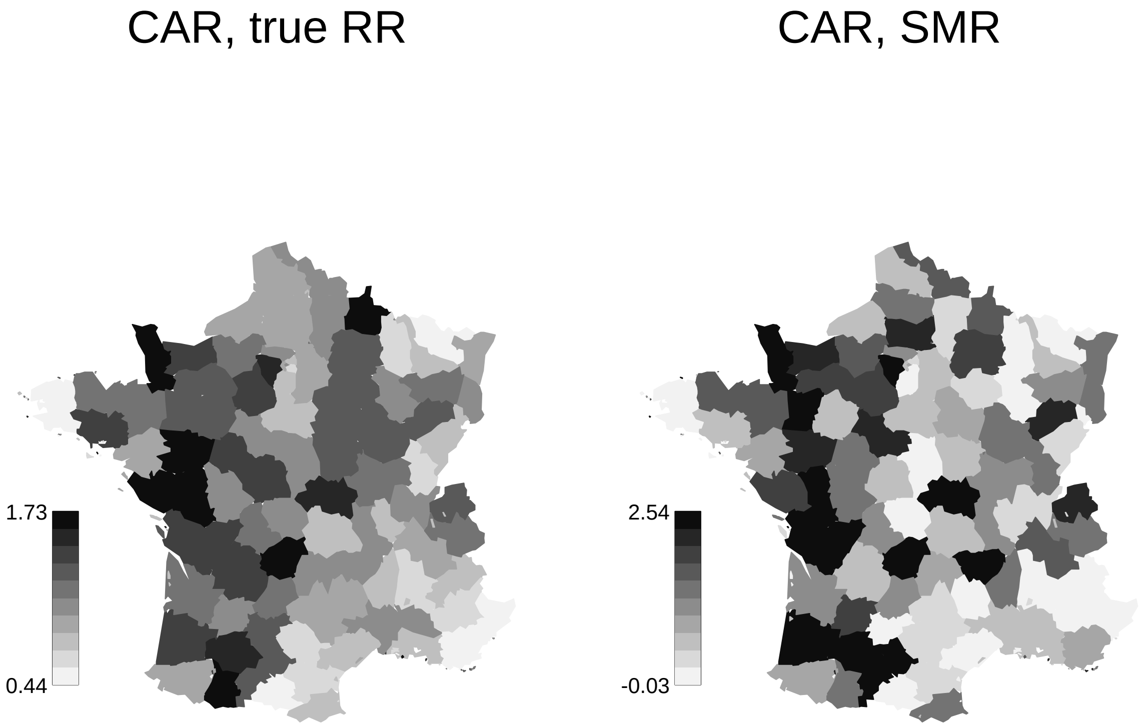

Figure 12: True risks and observed SMRs for the "CAR" data set.

The spatial mixture and the BYM are compared on the basis of RAMSE, RAMSEL, DIC, $E(D)$ and p_{D}. Each line of Table 1 corresponds to the criterion averaged over 5 independent Poisson replications of the data pattern (the first replication for four of the data sets having been displayed previously). Note that out of the six risk patterns considered, only the first two have discontinuities. The comparisons are thus designed to investigate the versatility of the two models in recovering risk patterns for which they are not necessarily well adapted.

Table 1: Simulation results comparing spatial mixture and BYM models.

Data Sets	RAMSE		RAMSEL		DIC		$E(D\|y)$		p_{D}	
	MIX	BYM	MIX	BYM	MIX	BYM	MIX	BYM	MIX	BYM
Block4	0.22	0.27	0.22	0.30	118.2	138.8	89.5	91.1	28.7	47.7
North-South	0.15	0.22	0.15	0.22	116.2	124.1	97.8	87.0	18.4	37.1
Gradient NS	0.22	0.24	0.27	0.26	125.4	129.7	94.6	86.6	30.8	43.1
CAR	0.27	0.28	0.27	0.27	133.8	136.7	102.1	95.1	31.7	41.6
Flat	0.09	0.19	0.09	0.19	93.4	108.6	89.4	77.2	4.0	31.4
Over	0.23	0.26	0.24	0.21	127.5	128.3	112.2	92.1	15.3	36.2

From the table, one can see that the spatial mixture model gives more faithful estimation of the underlying risks, with smaller RAMSE in most cases. The RAMSEL criterion gives a similar picture except for the "Over" data set where it is smaller for BYM. As could be expected, the difference is accentuated when there are contrasting zones, as in "Block4" or in "North-South" or when the pattern is fairly uniform as in "Flat". When the risks are smoothly varying, as in "Gradient NS" or "CAR", the two models give similar results; it is perhaps surprising that the mixture model performs competitively. Moreover, the spatial structure of the $\{\lambda_i\}$ induced by the chosen allocation process leads to a more parsimonious model with consistently lower DIC and p_{D}.

If one accepts the DIC principle at face value, the BYM model overfits the data in 5 of the 6 cases. With the exception of "Block4", the posterior deviance under the BYM model is substantially smaller, but the p_{D} is so much greater that the DIC is at least as large. In the balance between recovering the true underlying scene and fitting the data, the mixture model is clearly less influenced by the noise in the data than the BYM model and able to effect some spatially adaptive smoothing in a variety of situations.

Complementing the results of Table 1, we display the map of posterior means of $\{\lambda_i\}$ estimated by the BYM model for two data sets ("Block4" and "North-South") in Figures 13 and 14 (left panels). Comparing these with those corresponding to our mixture model (Figures 6 and 7), against the maps of the simulated risks (Figures 2 and 3), there is evidence of remaining noise, unsmoothed by the analysis. This illustrates what was quantified by the two mean square error criteria in Table 1, that recovery of the unblurred picture is less effective for the BYM model. It is also interesting to see that the posterior variability of the risks in both data sets is quite different to that of the mixture model, whether one compares the posterior standard deviations for the “North-South” (Figures 6 and 14) or the histograms for “Block4” (Figures 8 and 13). For the BYM model, variability is not increased for areas along discontinuities. Further, the variability of the risks is higher overall, and closely linked to the size of the risks. This phenomenon is also true for smoothly varying risks as in the “Gradient NS” example (results not shown).

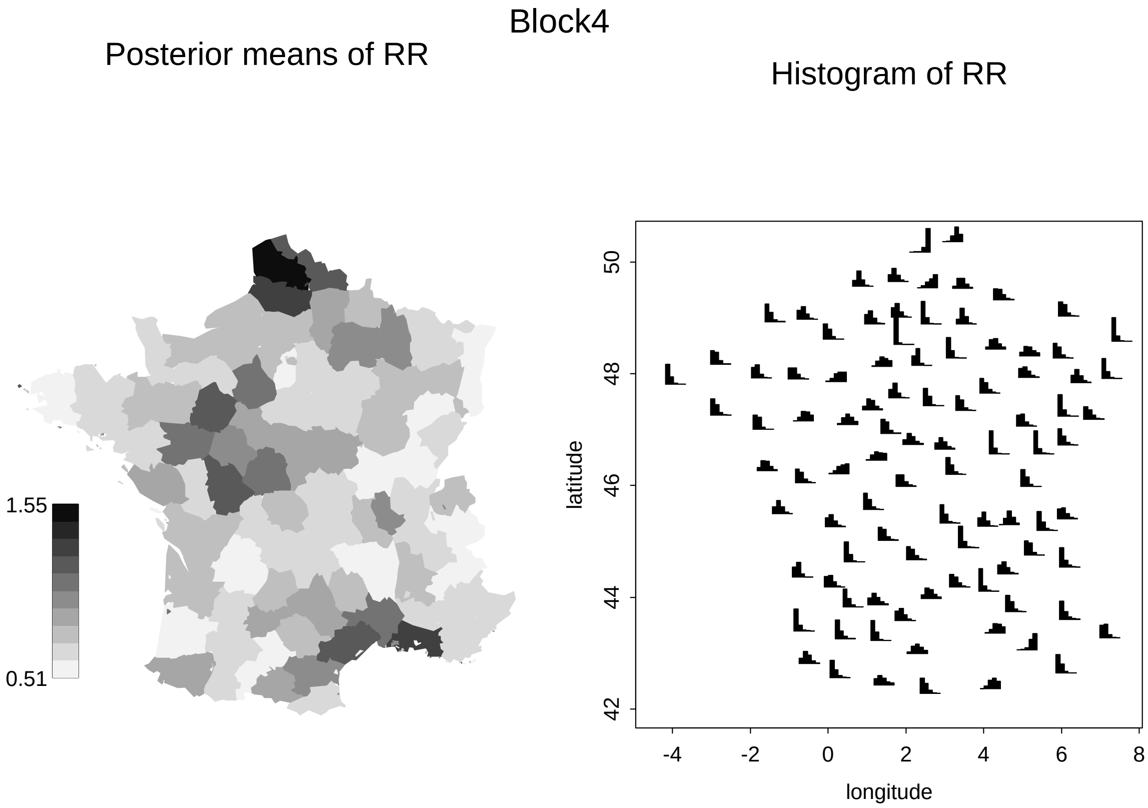

Figure 13: BYM estimates of the risks for “Block4”, posterior means (left panel), histogram of the risks (right panel).

Finally, Figure 15 displays for each area the root mean square error between the simulated and estimated $\{\lambda_i\}$ corresponding to the mixture model (left panel) or the BYM model (right panel) for the “North-South” data set. For the mixture model, the highest errors are along the discontinuity, while for the BYM model, this pattern is less clear and higher errors can be seen over all the Southern areas.

5 Epidemiological application to disease data

The performance of the model is illustrated on data concerning larynx cancer mortality in France at the level of the 94 mainland French départements reported in Rezvani *et al.*, 1997) for the period 1986–1993. For this data set, we also illustrate how the introduction of area level covariates in the

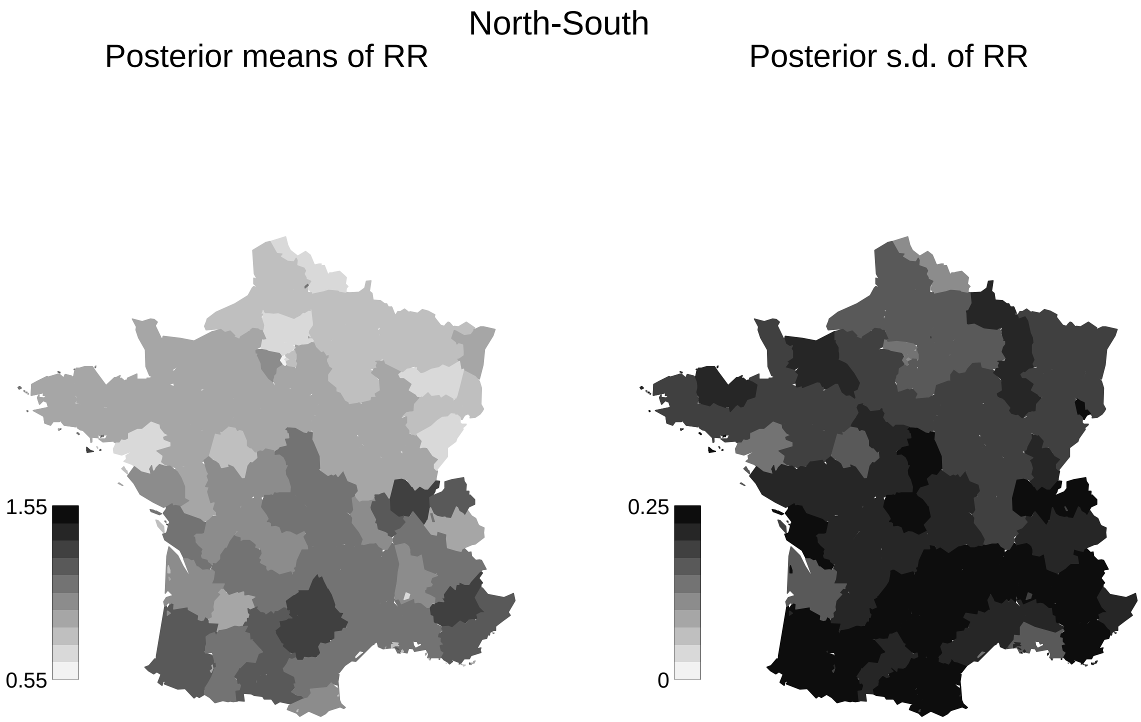

Figure 14: BYM estimates of the risks for "North-South", posterior means (left panel), posterior standard deviations (right panel).

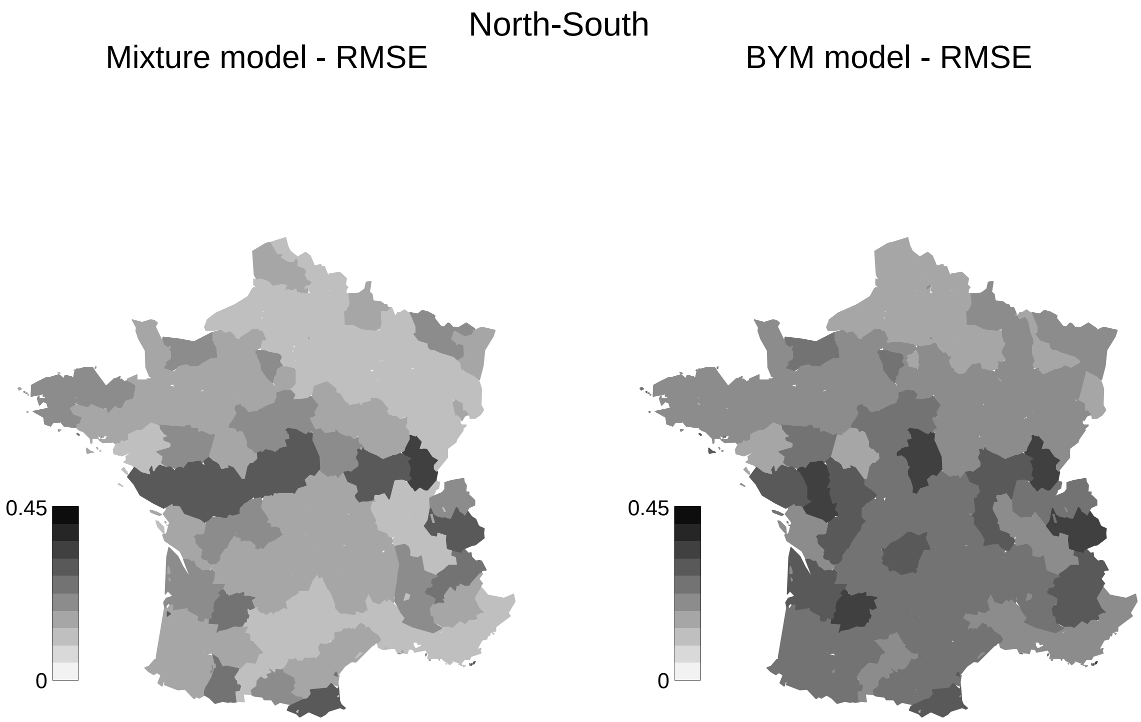

Figure 15: Root mean square errors for the "North-South" data set: comparison between the mixture model (left panel) and the BYM model (right panel).

model, as set out in equation (2), reduces the spatial heterogeneity of the risks.

The update of the regression parameters γ_j in (2) was performed using random walk Metropolis, other updates being as described before, with the necessary adjustment to the likelihood. Acceptance rates were around 49%, using normally distributed perturbations with standard deviations 1.5 times the reciprocal of the range of the corresponding covariate.

Larynx cancer for women is rare, with observed number of deaths per area in this data set ranging from 0 to 148 and SMR ranging from 0 to 2.1. The epidemiology of this cancer site has been mostly studied for men where risk factors such as smoking, alcohol consumption, dietary factors and specific occupational exposure have been brought to light in case-control studies (Austin and Reynolds, 1996). Here, we include two covariates in our model: the per capita sales of cigarettes in 1975, an available proxy for smoking, and an indicator of the urbanisation of the area as recorded in the 1975 census. These variables are both lagged in time to allow for a delay between putative exposure and disease. Under the Potts mixture model adjusted without covariates, the posterior distribution of k peaks at $k = 2$, with $p(k|y) = 0.0, 0.49, 0.23, 0.12$ for $k = 1, 2, 3, 4$ respectively.

Regions of higher risk are apparent in the North and South-East. Even though the posterior means of the risks are fairly similar for the spatial mixture and the BYM model (Figure 16), there is markedly higher variability for the BYM model (Figure 17), a phenomenon that was already described previously.

When the two covariates are introduced in the Potts mixture model, the posterior distribution of k shifts markedly towards the left, with $p(k|y) = 0.31, 0.39, 0.18, 0.07$ for $k = 1, 2, 3, 4$. As expected, the inclusion of the two covariates has explained in part the heterogeneity. In fact, the range of the posterior mean of the residual risks (λs in equation (2)) is substantially reduced to approximately one third of the range of the risks displayed in Figure 16.

This example supports our view that, by combining information from the mixture structure and the display of the posterior distribution of the risk estimates, one is able to characterise the heterogeneity of the risks and investigate how it is affected by the introduction of covariates. The ultimate goal, from a public health point of view, is to uncover sets of covariates that can account for most of the geographical variability of the risks.

6 Concluding remarks

Spatially structured heterogeneity is a common phenomenon which can be tackled in a hierarchical framework using a hidden Markov random field approach, through direct multivariate specification

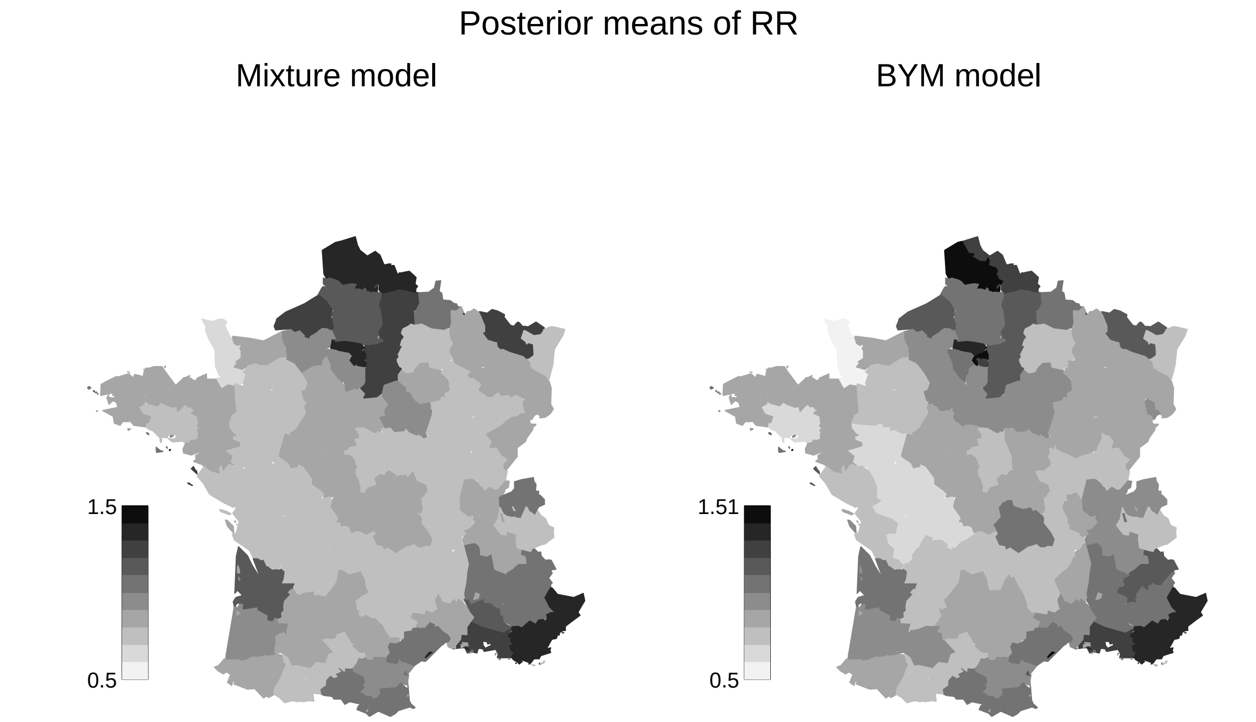

Figure 16: Larynx cancer mortality: comparison of posterior means for the risks obtained using the mixture model (left panel) and the BYM model (right panel).

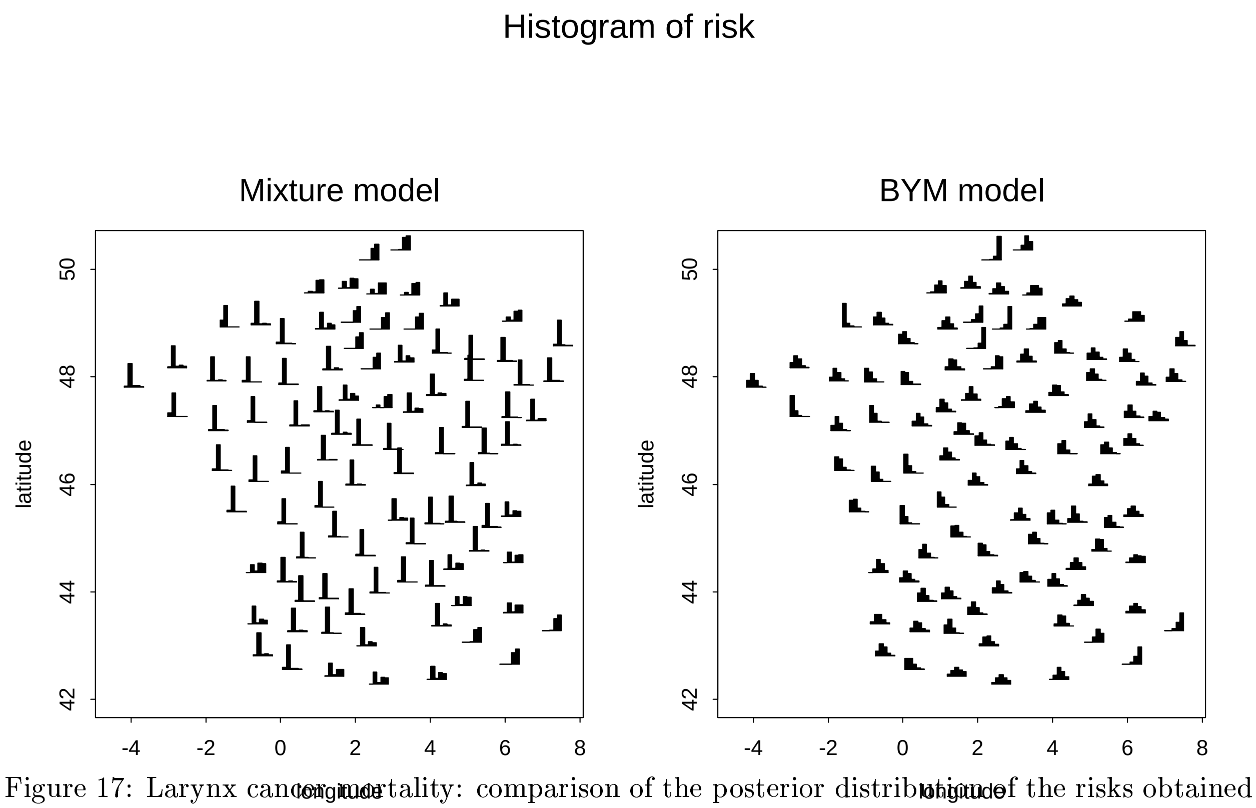

Figure 17: Larynx cancer mortality: comparison of the posterior distributions of the risks obtained using the mixture model (left panel) and the BYM model (right panel).

of the underlying field, or via partition models. In our paper, focussed on rare outcomes and hierarchical Poisson models, we have proposed a new model within the hidden Markov random field framework. Whether the features this model offers are useful depends on the purpose of the modelling exercise. In our study of the performance of the allocation model, we have been strongly influenced by the specific epidemiological context. There, recovery of the underlying Poisson rates is often the prime object of inference, supplemented by quantification of the extent of the underlying variability, and probability statements about areas of high risk.

In terms of recovery of the "hidden state" or "true image", we have shown that the posterior mean risks estimated under the allocation model give a faithful representation of the true risks in a wide variety of situations, encompassing both smooth and discontinuous cases. This flexibility is important, since there is usually little prior information on the underlying risks. Further, we have seen that information from the posterior estimate of the mixture structure can also be helpful in characterising the spatial variability. We have explored some other features of the joint distribution of the risks, such as the number of clusters, but will leave a fuller study to further work.

Although these do not receive much space in the paper, we have conducted numerous checks on the correctness of our MCMC samplers, and on the adequacy of their performance. In particular, we have verified that the prior distribution is recovered if we implement our computations without likelihood or data. We are satisfied that the range chosen for the interaction parameter ψ allows substantial spatial dependence in the allocations, so that higher values of ψ, for which mixing could be slower, are not necessary. Over-long runs have been used on purpose in our examples, the algorithm being quite fast; our sampler, coded in Fortran, makes about 900 sweeps per second on a 300MHz PC.

Further work on model comparison is certainly needed; we regard our comparisons with the BYM model as quite limited, both in terms of the scope and of the criteria chosen for comparison. It would be interesting to extend comparisons to include the models proposed by Knorr-Held and Raßer, and by Fernández and Green, and also to include other non-Gaussian Markov random field approaches. Indeed, when discontinuities are expected it is advisable to replace the quadratic potential, leading to a Gaussian prior for the spatially structured random effects u_i, by an absolute value difference potential. This is discussed in Besag, et al, (1991) and used in Best, et al, (1999), and is related to smoothing using medians instead of means.

Regarding the relevant criteria for comparison and choice, Bayesian model comparison is an active area of research, and there is much debate on the most appropriate approach.

As we have mentioned at the end of Section 2, the model could be extended in several directions. One interesting avenue is the inclusion of covariates with heterogeneous effect, that is, where covariate effects and allocations interact. We have implemented the model outlined in equation (2), where the effect of the covariates is homogeneous. A Poisson mixture model with interaction between allocations and covariates but independent, non-spatial allocations was discussed in Viallefont, Richardson and Green (2002), and there should be no obstacle to extending this to spatially-correlated allocations. Another extension which could be considered is to combine the spatial mixture with a BYM model, in effect replacing λ_0 by λ_{z_i} in Section 3.4. It remains to be seen whether this could usefully combine the best features of both models, or whether there will be gross overparametrisation. A model containing a mixture of Gaussian and non-Gaussian (median based) conditional autoregressive components was recently proposed by Lawson (2001).

Finally, there is great scope for extensions to spatio-temporal modelling, both for epidemiological applications and more generally.

Acknowledgements

We are grateful to Julian Besag, Carmen Fernández, Alan Gelfand, Alex Lewin, Annie Mollié and Christine Monfort for valuable interaction and comments. The work was partially supported by an EPSRC Visiting Research Fellowship, a travel grant from the ESF programme on Highly Structured Stochastic Systems, and the INSERM contract ITM 4TM05F. Most of this work was undertaken while SR was at INSERM, Paris.

References

Austin, D.F. and Reynolds, P. (1996) Laryngeal cancer. In: *Cancer epidemiology and prevention*, eds D. Schottenfeld and J.F. Fraumeni, Oxford University Press, second edition, 619–636.

Aykroyd, R. G. and Zimeras, S. (1999) Inhomogeneous prior models for image reconstruction *Journal of the American Statistical Association*, **94**, 934–946.

Besag, J. (1986) On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society*, B, **48**, 259–302.

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration with applications in spatial statistics (with discussion). *Annals of the Institute of Mathematical Statistics*, **43**, 1–59.

Best, N. G., Arnold, R. A., Thomas, A., Waller, L. A. and Conlon, E. M. (1999) Bayesian models for spatially correlated disease and exposure data In: *Bayesian Statistics 6.* eds Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. Oxford University Press, Oxford, 131–156.

Carroll, R. J., Roeder, K. and Wasserman, L. (1999) Flexible parametric measurement error models. *Biometrics*, **55**, 44–54.

Clayton, D. and Bernardinelli, L. (1992) Bayesian methods for mapping disease risk. In: *Geographical and environment epidemiology: methods for small area studies.* eds Elliott, P., Cuzick, J., English, D. and Stern, R. Oxford University Press, Oxford, 205–220

Clifford, P. (1986) Contribution to the discussion of the paper by Besag (1986), *Journal of the Royal Statistical Society*, B, **48**, 284.

Denison, D. G. T., and Holmes, C. C. (2001) Bayesian partitioning for estimating disease risk. *Biometrics*, **57**, 143–149.

Elliott, P., Wakefield, J.C., Best, N.G. and Briggs, D.J. (2000). *Spatial Epidemiology: Methods and Applications.* Oxford University Press, Oxford.

Fernández, C. and Green, P. J. (2000) *Modelling spatially correlated data via mixtures: a Bayesian approach.* Manuscript, University of Bristol.

Gelman, A., and Meng, X-L. (1998) Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. *Statistical Science*, **13**, 163–185.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Greenland, S. and Robins, J. (1994) Ecological studies: biases, misconceptions, and counterexamples. *American Journal of Epidemiology*, **139**, 747–760.

Johnson, V. E. (1994) A model for segmentation and analysis of noisy images. *Journal of the American Statistical Association*, **89**, 230–241.

Knorr-Held L., and Raßer, G. (2000) Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, **56**, 13–21.

Knorr-Held, L., and Rue, H. (2002) On block updating in Markov random field models for disease mapping. Scandinavian Journal of Statistics, to appear. Available at `http://www.stat.uni-muenchen.de/sfb386/publikation.html`

Künsch, H.R. (2001) State space and hidden Markov models. In: *Complex stochastic systems*, eds Barndorff-Nielsen, O.E., Cox, D.R. and Klüppelberg, C. Chapman and Hall/CRC, London, 109–173.

Lawson, A.B. and Clark, A. (2001) Spatial mixture relative risk models applied to disease mapping. *Statistics in Medicine*, to appear.

McLachlan, G. and Peel, D. (2000) *Finite mixture models.* Wiley, Chichester.

Mollié, A. (1996) Bayesian mapping of disease. In: *Markov Chain Monte Carlo in Practice*, eds Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. Chapman and Hall, London, 359–379.

Ogata, Y. and Tanemura, M. (1984) Likelihood analysis of spatial point patterns. *Journal of the Royal Statistical Society*, B, **46**, 496–518.

Rezvani, A., Mollié, A., Doyon, F. and Sancho-Garnier, H. (1997) *Atlas de la mortalité par cancer en France, période 1986-1993.* Editions INSERM.

Richardson, S. and Green, P. J. (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion) *Journal of the Royal Statistical Society*, B, **59**, 731–792.

Robert, C. P., Rydén, T., and Titterington, D. M. (2000) Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society*, B, **62**, 57–75.

Roeder, K., Carroll, R. J., and Lindsay, B. G. (1996) A semiparametric mixture approach to case-control studies with errors in covariables. *Journal of the American Statistical Association*, **91**, 722–732.

Rydén, T. and Titterington. D. M. (1998) Computational Bayesian analysis of hidden Markov models. *Journal of Computational and Graphical Statistics*, **7**, 194–211.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, B, **64**, to appear.

Tjelmeland, H. and Besag, J. (1998). Markov random fields with higher-order interactions. *Scandinavian Journal of Statistics*, **25**, 415–433.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions.* Wiley, Chichester.

Viallefont, V., Richardson, S. and Green, P. J. (2002) Bayesian analysis of Poisson mixtures. *Journal of Nonparametric Statistics*, to appear.

Wakefield, J. and Morris, S. (1999). Spatial dependence and errors-in-variables in environmental epidemiology, In: *Bayesian Statistics 6.* eds Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. Oxford University Press, Oxford, 657-684.