Modelling spatially correlated data via mixtures: a Bayesian approach

Carmen Fernández* Peter J. Green†
University of St Andrews University of Bristol

January 2002

Abstract

This paper develops mixture models for spatially indexed data. We confine attention to the case of finite, typically irregular, patterns of points or regions with prescribed spatial relationships, and to problems where it is only the weights in the mixture that vary from one location to another. Our specific focus is on Poisson distributed data, and applications in disease mapping. We work in a Bayesian framework, with the Poisson parameters drawn from gamma priors, and an unknown number of components. We propose two alternative models for spatially-dependent weights, based on transformations of autoregressive gaussian processes: in one (the Logistic normal model), the mixture component labels are exchangeable, in the other (the Grouped continuous model), they are ordered. Reversible jump Markov chain Monte Carlo algorithms for posterior inference are developed. Finally, the performance of both of these formulations is examined on synthetic data and real data on mortality from rare disease.

Some key words: Disease mapping, Grouped continuous model, Logistic normal, Poisson mixtures, Reversible jump MCMC.

1 Introduction

There are two main motivations for this work. One is methodological – to extend the range of contexts in which mixture-based models play a role to cases of spatially indexed data; the other is

*School of Mathematics and Statistics, University of St Andrews, North Haugh, St Andrews KY16 9SS, U.K. Email: carmen@mcs.st-and.ac.uk.
†Department of Mathematics, University of Bristol, Bristol BS8 1TW, U.K. Email: P.J.Green@bristol.ac.uk.

applied – to develop and investigate new spatially correlated models for geographical epidemiology with the aim of alleviating some possible difficulties with existing disease mapping methods.

Mixture models have a long pedigree, with interest in them going back many decades. These models have been used both in situations where the components of the mixture represent subgroups in a heterogeneous population, and in settings where the mixture formulation is merely a convenient parsimonious form for flexible density estimation. The books by Titterington, Smith and Makov (1985) and Lindsay (1995) provide general background. In recent years, research interest has been re-generated, first by the introduction of the Bayesian approach, with computational implementation by Markov chain Monte Carlo (MCMC) (see, for example, Robert, 1996, for a review) and, secondly, by allowing the number of components to vary in this Bayesian context (as in Nobile, 1994, and Richardson and Green, 1997). These developments have increased the power and flexibility of mixture-based models, by enabling full simultaneous posterior inference about all parameters, by exposing the possibility for 'multiple explanations' in some data sets, by avoiding the difficulties of (non-regular) inference about the number of components in a frequentist setting, and by offering the possibility of fitting data better through model-averaging.

Beyond the independent random sample setting in which basic mixture models are always formulated, this recent activity has stimulated novel statistical models for several situations in which data are more structured. These include regression and autoregression (West *et al.*, 1994), measurement error (Müller and Roeder, 1997; Leblond *et al.*, 2000), modelling random effects in mixed models (Watier *et al.*, 1999), analysis of factorial experiments (Nobile and Green, 2000) and, more generally, hidden switching models (Robert *et al.*, 2000; Frühwirth-Schnatter, 2001). In all of these diverse contexts, the mixture formulation makes a powerful contribution to fitting or adjusting for the effects of heterogeneity in populations, and it is reasonable to believe the opportunities for such contributions are not yet exhausted.

One setting in which mixtures have not been previously exploited is that of spatially indexed data. In fact, there are many problems in this category. Data may be indexed by points, by regions, or by cells of a lattice. Mixtures may be introduced at various levels in such problems. In this paper, we confine attention to the case of finite, typically irregular, patterns of points or regions with prescribed spatial relationships, and to problems where it is only the weights in a mixture model that vary from one location to another. This is still a very wide class of problems. Our spatial mixture formulations apply in broad detail to many of these; one possible application that will not be pursued here is to agricultural field trials (see Fernández and Green, 1999). Our specific focus in this paper is on Poisson distributed data, and applications in disease mapping.

There is an extensive epidemiological statistics literature that deals with issues of extra-Poisson variation and unmeasured covariates in spatially distributed disease data. Typically, the aim is to produce reliable maps of disease (see Mollié, 1996, for an accessible account from a Bayesian viewpoint). The model introduced by Besag *et al.* (1991) has been extensively used in this context. This model contains two random terms, one to represent unstructured heterogeneity and the other one to capture spatial correlation. There are, however, some issues of identifiability and interpretability of the spatial and unstructured heterogeneity parts, and the model does not allow naturally for discontinuities in the risk surface. As Mollié (1996) puts it "It would be useful to develop other forms of prior, to take into account the presence of discontinuities in the spatial structure of the relative risks". We will explore the extent to which our mixture approach can handle this, and use a version of the Besag *et al.* model as a benchmark against which we compare our results. Somewhat related approaches designed to deal with clusters of constant risk and discontinuities were proposed by Schlattmann and Böhning (1993) and Knorr-Held and Rasser (2000).

Our presentation reflects both the general and specific motivations for the work. In Section 2, we review aspects of the basic mixture model that remain relevant in the spatial context, and in Section 3 introduce two new formulations for spatially dependent weights. The disease mapping problem is introduced in Section 4, where the remainder of our model set-up is developed. In Section 5, we briefly discuss the computational implementation of our models via MCMC (further details on this are provided in an Appendix). Section 6 presents results on the performance of the models on both synthetic and real data sets, and we conclude in Section 7 with general discussion and some possibilities for future work.

2 The mixture sampling model

We consider conditionally independent but not identically distributed observations y_i $(i = 1, \ldots, n)$ from a mixture model

$$p(y_i|k, w, \lambda) = \sum_{j=1}^{k} w_{ij} f(y_i|\lambda_j), \tag{2.1}$$

where k is the number of components. The distribution under the j^{th} component $(j = 1, \ldots, k)$ is denoted by $f(\cdot|\lambda_j)$ where f is a density or probability mass function of specified form, the vector $\lambda = (\lambda_1, \ldots, \lambda_k)$ groups the component-specific parameters, and the weights $w = (w_{ij})_{i,j}$ satisfy $w_{ij} > 0$ with $\sum_{j=1}^{k} w_{ij} = 1$ for all i. Note that in contrast to the standard setup, here the weights vary with i. In this paper, we will assume that k, w and λ are unknown and subject to inference.

The well-known representation of mixture models by means of a hidden allocation process will prove useful both for the purpose of interpretation and for the numerical computations. For n independent observations from a mixture of k components, the representation is as follows:

(i) Introduce n independent discrete variables $z_1, \ldots, z_n$, with the multinomial distribution $p(z_i = j|k, w, \lambda) = w_{ij}$, for $j = 1, \ldots, k$.

(ii) Assume that $y_1, \ldots, y_n$ are independent given $z = (z_1, \ldots, z_n)$ with $p(y_i|z, k, w, \lambda) = f(y_i|\lambda_{z_i})$.

If we now integrate out z_i using its multinomial distribution, we obtain the mixture model in (2.1). Thus, z_i can be interpreted as an allocation variable, in the sense that it assigns observation y_i to one of the mixture components. This suggests that mixtures will be appropriate to model a heterogeneous population, where the assignment of observations to groups is unknown. In addition, they are often used as a flexible but reasonably parsimonious way to represent non-standard distributions, for example, as an alternative to the use of nonparametric methods.

3 Modelling spatial dependence through priors on weights

A crucial difference between the mixture model in (2.1) and the case of i.i.d. sampling previously examined in the literature (*e.g.*, Diebolt and Robert, 1994, Richardson and Green, 1997), is that the weights in (2.1) are indexed by i, so they are allowed to vary from observation to observation. In our context, i will be a spatial index which represents a location in the space where the observations take place, and we will introduce spatial dependence through the prior distribution of the weights. In simple terms, the basic behaviour that we try to capture is that observations that correspond to nearby locations are more likely to have similar values of the weights (that is, similar allocation probabilities) than observations from locations that are far apart. Our analysis will be conditional on a given graph with the locations as vertices, and certain designated pairs of nearby locations as edges; this graph will be the conditional independence graph of a component of our models, and thus we are working with Markov random fields indexed by i. Pairs of locations connected by an edge will be called neighbours.

We will propose two different models to allow for spatially correlated weights. In both models, spatial dependence will be introduced by means of a Markov random field with p.d.f.

$$p(x|h) = c(h) \exp\left[-\frac{1}{2}\left\{h \sum_{i\sim i'} (x_i - x_{i'})^2 + \sum_{i=1}^{n} x_i^2\right\}\right], \tag{3.1}$$

where $x = (x_1, \ldots, x_n)$ and $\sum_{i\sim i'}$ denotes the sum over all pairs of neighbours with each pair counted only once. The parameter h is non-negative and $c(h)$ is the appropriate integrating constant:

$$c(h) = (2\pi)^{-n/2} \prod_{i=1}^{n} (1 + hg_i)^{1/2}, \tag{3.2}$$

where $g_1, \ldots, g_n$ denote the eigenvalues of a matrix $A = (a_{ii'})$ coding the adjacencies, with $a_{ii} = \nu_i$ (number of neighbours of location i), and off-diagonal elements $a_{ii'} = -1$ if locations i and i' are neighbours and zero otherwise. Note that, for a given graph, $g_1, \ldots, g_n$ only need to be computed once, and can then be stored for any further analyses using the same graph. From (3.1) it is clear that neighbouring locations are induced to have similar values of the corresponding elements in x, and this effect is more pronounced as the value of h increases. The limiting case $h = 0$ corresponds to independence across locations, whereas as $h \to \infty$ the distribution in (3.1) tends to a degenerate one where neighbouring locations are forced to have exactly the same value of the corresponding elements in x.

The model in (3.1) is very closely related to the Gaussian conditional autoregressive model, extensively used in the disease mapping literature to obtain spatially smoothed estimates of relative risks; see, for example, Clayton and Kaldor (1987), Besag *et al.* (1991), Clayton and Bernardinelli (1992), and Waller *et al.* (1997). The term $\sum_{i=1}^{n} x_i^2$ in the exponent in (3.1) does not appear in the conditional autoregressive model, as in, for example, Besag *et al.* (1991), although it does in the formulation proposed recently by Leroux *et al.* (2000). The inclusion of this term makes (3.1) a proper distribution, and is necessary in order to obtain a proper posterior distribution in our mixtures setting. Despite the similarity of (3.1) to models previously considered, our modelling strategy is quite novel in that (3.1) will be used to obtain a spatially correlated prior for the weights in the mixture model. In the next two subsections we explain how we achieve this.

3.1 The Logistic Normal model

For a mixture with k components, our first model requires introducing k independent n-dimensional vectors, $x_j \equiv (x_{1j}, \ldots, x_{nj})$, $j = 1, \ldots, k$, each distributed according to (3.1). Note that although the vectors are independent of each other, each of them incorporates spatial dependence among its n elements. Next, we define the weights using the logistic transform (see, for example, McCullagh and Nelder, 1989, ch. 5), by which the weights for location i take the form

$$w_{ij} = \frac{\exp(x_{ij}/\phi)}{\sum_{l=1}^{k} \exp(x_{il}/\phi)}, \quad j = 1, \ldots, k, \tag{3.3}$$

with $\phi > 0$. Thus, the weights for location i depend on the i^{th} element of each of $x_1, \ldots, x_k$. The dependence structure of the distribution (3.1) induces spatial dependence among the weights.

As the value of h in (3.1) increases, realisations of the x_j processes become smoother, and there is also stronger shrinkage towards zero (the mean of the processes), but the scale parameter ϕ introduced in (3.3) can compensate for this. Note that $w_{ij}/w_{il} = \exp[(x_{ij} - x_{il})/\phi]$ is a monotonic function of ϕ converging to 1 as $\phi \to \infty$ and to either 0 or ∞ when $\phi \to 0$. A smaller value of ϕ can thus alleviate the effects of increasing shrinkage in the x values. The limiting case $\phi = 0$ corresponds to $w_{ij} = 1$ if $x_{ij} = \max\{x_{il} : l = 1, \ldots, k\}$, and zero otherwise, which implies that the allocation variables $z_1, \ldots, z_n$ are deterministic functions of x. The other limiting case ($\phi \to \infty$) leads to $w_{ij} = 1/k$ for all i and j, thus precluding any spatial patterns or randomness in the weights. We will restrict ϕ to be small enough to avoid this undesirable feature.

3.2 The Grouped Continuous model

In this model, we use a single n-dimensional vector $x \equiv (x_1, \ldots, x_n)$, distributed according to (3.1). Motivated by previous work on ordinal data based on grouped continuous models (McCullagh and Nelder, 1989, ch. 5) we define the weights for location i as

$$w_{ij} = \begin{cases} \Psi\left(\tau^{-1}(\delta_1 - x_i)\right) & \text{if } j = 1 \\ \Psi\left(\tau^{-1}(\delta_j - x_i)\right) - \Psi\left(\tau^{-1}(\delta_{j-1} - x_i)\right) & \text{if } j = 2, \ldots, k-1 \\ 1 - \Psi\left(\tau^{-1}(\delta_{k-1} - x_i)\right) & \text{if } j = k \end{cases} \tag{3.4}$$

where $\Psi(\cdot)$ is a specified continuous distribution function, and τ is a scale parameter. We will consider a logistic specification, where $\Psi(r) = [1 + \exp(-r)]^{-1}$, although other choices of $\Psi(\cdot)$ could be accommodated with simple modifications to our computational implementation. The parameters $\delta_1 < \ldots < \delta_{k-1}$ constitute a set of ordered unknown cutpoints, over which we also specify a prior distribution. In our experience, care is required in the choice of this distribution, because the model can easily lead to many empty components in the sense that there is a high prior probability that the values of the allocation variables $(z_1, \ldots, z_n)$ do not span the entire set $\{1, \ldots, k\}$. We consider a prior distribution for the δ_j's with support on $(x_{\min}, x_{\max})$, where $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of $\{x_i : i = 1, \ldots, n\}$. In this way, we minimise the chance of empty components due to several of the δ_j's lying well out of the range of the x_i's. To induce further separation between the values of the δ_j's, we take $\delta_1, \ldots, \delta_{k-1}$ to be the $s^{th}, 2s^{th}, \ldots, (k-1)s^{th}$ order statistics obtained from $ks - 1$ i.i.d. replications, for some $s \geq 1$ (a similar idea appeared in the change-point example of Green, 1995). If we base ourselves on the Uniform distribution, this leads to the following p.d.f. for $\delta = (\delta_1, \ldots, \delta_{k-1})$:

$$p(\delta|k, x) = \frac{(ks-1)!}{[(s-1)!]^k (x_{\max} - x_{\min})^{ks-1}} \prod_{j=1}^{k} (\delta_j - \delta_{j-1})^{s-1} I[x_{\min} < \delta_1 < \ldots < \delta_{k-1} < x_{\max}], \tag{3.5}$$

where we have defined $\delta_0 = x_{\min}$ and $\delta_k = x_{\max}$. Throughout the paper, $I[l]$ will denote a function that takes the value 1 if the logical statement l is fulfilled and zero otherwise. As with $\Psi(\cdot)$, alternative choices to (3.5) only imply minor modifications to our framework.

The spatial dependence incorporated in x implies that neighbouring locations are more likely to have similar values of the weights in (3.4) than locations that are far apart. Similarly to the role of ϕ in the Logistic Normal model, the scale parameter τ in (3.4) also influences the value of the weights and, by making it small we can in part compensate for the shrinkage towards the mean of x that occurs when the parameter h in (3.1) increases. In the limiting situation where $\tau = 0$, the allocation z_i is a deterministic function of x_i and the δ_j's, with $w_{ij} = 1$ if $\delta_{j-1} < x_i < \delta_j$, and zero otherwise. As $\tau \to \infty$, we obtain the limit $w_{i1} = w_{ik} = 0.5$ and $w_{ij} = 0$ for the intermediate components, precluding spatial dependence or randomness in the weights while assigning all the probability to the two extreme components. We shall restrict τ to be small enough to obtain non-negligible prior probabilities for all of the components.

3.3 Priors on spatial interaction parameters

As we have already discussed, the parameter h controls the dependence in the Markov random fields in (3.1). In particular, the limiting case $h = 0$ leads to independent weights across locations. The scale parameters ϕ and τ in (3.3) and (3.4) affect the size and general behaviour of the weights and, as a consequence, also influence the correlation between the allocation variables $z_1, \ldots, z_n$. Although we have provided some intuition as to the role of these parameters, it is not easy fully to understand and separate the roles of h and ϕ in the Logistic Normal (henceforth denoted as LN) model and, similarly, of h and τ in the Grouped Continuous (GC) model. In general terms, we will view them as "smoothing" parameters and, in order to allow extra flexibility and cope with uncertainty about them, we assign prior distributions to them. In our examples, we shall consider h and ϕ (or τ) independent with

$$h \sim \text{ Uniform on } (0, h_{\max}), \quad \phi \sim \text{ Uniform on } (0, \phi_{\max}), \quad \tau \sim \text{ Uniform on } (0, \tau_{\max}), \qquad (3.6)$$

for some positive numbers $h_{\max}$, $\phi_{\max}$ and $\tau_{\max}$. The support intervals will be chosen large enough to allow for interior modes in the posterior distributions, but not so large that the unreasonable features displayed by the weights as ϕ or τ increase could emerge. We have chosen independent Uniform priors so that we can easily assess the impact of the data when looking at the corresponding posterior distributions but, once again, other prior distributions could be easily accommodated in our framework.

4 An application in disease mapping

The objective of disease mapping is to analyse the geographical variation in rates of disease incidence or mortality within a specific area. In recent years there has been considerable scientific and public interest in uncovering spatial patterns in disease, as this can help formulate aetiological hypotheses pointing towards certain causes/deterrents of disease or, for example, suggesting a potential disadvantage of certain groups within the population. The data are usually given in the form of counts (the numbers of observed cases of disease), y_i, for regions that constitute a geographical partition of the area of interest. A preliminary analysis of these and other data, usually incorporating stratification on age and sex, yields an expected count e_i for each region i.

The standardised mortality ratios (SMR's), are defined as y_i/e_i, for $i = 1, \ldots, n$, and correspond to maximum likelihood estimates of relative risk under independent sampling from

$$p(y_i|r_i) = \text{Pois}(y_i|e_i r_i), \tag{4.1}$$

a Poisson distribution with mean $e_i r_i$, where r_i denotes the relative risk in region i. Traditionally, maps of disease display the SMR's using a grey or coloured scale. For rare diseases, where the expected number of cases can be quite small for some of the regions, these maps can be misleading because the most extreme values (which visually dominate the map) typically correspond to low population areas where the variance of the estimator is largest. In addition, this approach fails to account for two features often found in this context: extra-Poisson variation (*i.e.* there is more heterogeneity in the population than would be implied by a Poisson model), and spatial correlation in the relative risks (due, for example, to unmeasured covariates that are spatially correlated).

We now consider using the two spatial mixture formulations of Section 3 in this context of disease mapping. The locations are now the regions in which our data are collected. We take the simplest possible graph structure connecting the locations: regions i and i' are neighbours if and only if they are spatially contiguous. The mixture assumption is to model the relative risk r_i for region i with allocation z_i (see Section 2) as $r_i = \lambda_{z_i}$, giving the mixture representation

$$p(y_i|k, w, \lambda) = \sum_{j=1}^{k} w_{ij} \text{Pois}(y_i|e_i \lambda_j), \tag{4.2}$$

for conditionally independent disease counts $\{y_i\}$. Thus, the relative risk surface is represented as a mixture (due to the random nature of the allocations z_i) of discrete surfaces with levels $\lambda_1, \ldots, \lambda_k$. This should allow for an adaptive type of smoothing, showing discontinuities between regions that require it while applying stronger smoothing in other geographical areas. Modelling of the weights

w_{ij} by either the LN or GC formulations has been discussed, so it only remains to place priors on $\lambda = (\lambda_1, \ldots, \lambda_k)$, the component-specific parameters, and k, the number of components in the mixture. A natural choice for the prior distribution of the λ_j's is a Gamma specification, and we will assume that they are in increasing order. This leads to

$$p(\lambda|k) = k! \prod_{j=1}^{k} f_G(\lambda_j|\alpha, \beta)\, I[\lambda_1 < \ldots < \lambda_k], \tag{4.3}$$

where $f_G(\cdot|\alpha, \beta)$ denotes the p.d.f. of a Gamma distribution with shape parameter α and mean equal to α/β. Whereas ordering the λ_j's merely induces component identifiability in the LN model, it is a structural feature of the GC model, as it implies that neighbouring regions are, a priori, more likely to be allocated to components with consecutive values of the λ_j's than to more "distant" components. For the number of mixture components we take the prior distribution

$$k \sim \text{ Uniform on } \{1, \ldots, k_{\max}\}, \tag{4.4}$$

for a suitably large integer $k_{\max}$. Converting results to those corresponding to another prior, say $p^*(k)$, with the same support is straightforward since $p^*(k|y) \propto p(k|y)p^*(k)$, where $p(k|y)$ and $p^*(k|y)$ respectively denote the posterior for k under the Uniform prior in (4.4) and under the new prior $p^*(k)$. Similar transformations apply to other posterior distributions.

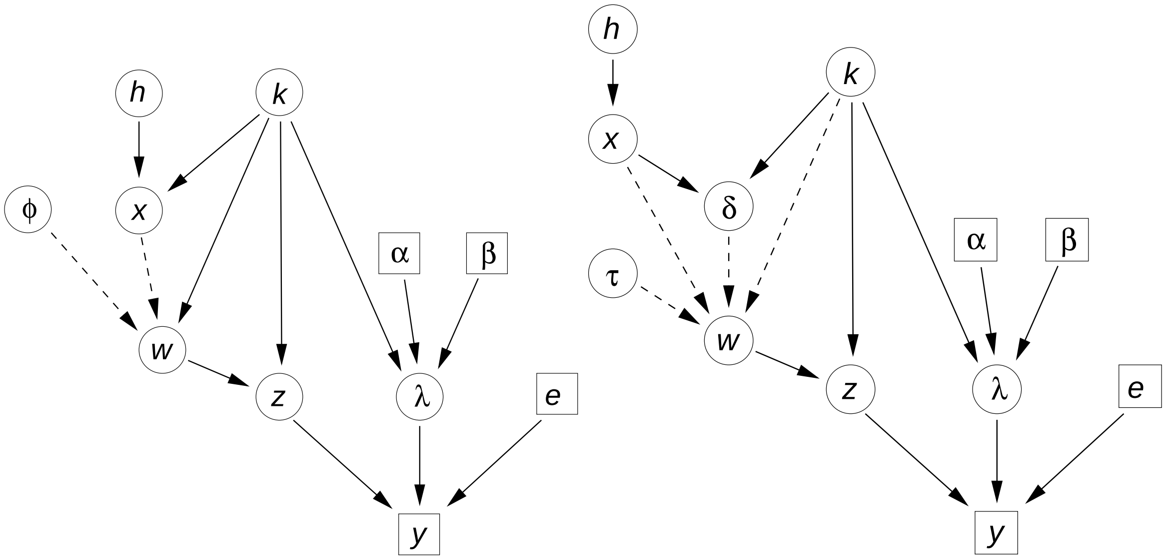

Figure 1: Directed acyclic graphs corresponding to LN and GC spatial mixture models (left and right panels, respectively).

Figure 1 displays direct acyclic graphs for the complete hierarchical Bayesian model under both the LN and GC specifications. Square boxes represent fixed or observed quantities and circles the unknowns. Broken lines denote deterministic (as opposed to stochastic) relationships.

5 Computational implementation

The Bayesian models proposed in this paper are too complex to be amenable to analytical calculations. Hence, we will use MCMC methods (for example, Tierney, 1994; Besag *et al.*, 1995), and in particular reversible jump MCMC (Green, 1995), which can cope with the varying dimensionality of the parameter space induced by the unknown number of mixture components. The parameter space is given by $\Theta = \bigcup_{k=1}^{k_{\max}} \Theta_k$, where Θ_k is the parameter space for a k-component mixture. Richardson and Green (1997) introduced the use of reversible jump MCMC for mixtures with unknown number of components in an i.i.d. setting. Here, we consider the more complex scenario where spatial interaction is present.

5.1 Construction of sampler

To facilitate sampling, we will augment with the hidden allocation variables $(z_1, \ldots, z_n)$ and with the auxiliary variables x distributed as in (3.1). Thus, our aim is to compute the joint posterior distribution of all the variables in each of the graphs in Figure 1. It is worth mentioning a couple of unusual features of our sampler (focussing the discussion on the LN model, although similar comments apply to the GC model). Firstly, since the weights w are a deterministic function of (x, ϕ) (see (3.3)), they cannot be included as an additional random variable in the sampler. Secondly, in most steps of the sampler the allocation variables $(z_1, \ldots, z_n)$ will be integrated out to avoid slow mixing for values of ϕ which are small in relation to the values of x_{ij} (in such cases, the allocations would be virtually deterministic functions of (x, ϕ) since $k-1$ out of the k weights in (3.3) would be negligible). The resulting Markov chain is irreducible and has the required posterior distribution as invariant distribution, so ergodic averages converge to the corresponding posterior expectations. Further details on the sampler can be found in the Appendix. Of course, other samplers would be possible, but this is one that we have found to work well in practice.

5.2 Performance issues

The main difficulty was in finding a proposal generator that led to reasonable acceptance rates in the reversible jump step. This was particularly difficult for the LN model, where increasing the number of components by 1 implies having to propose an n-dimensional auxiliary vector x. After some trial and error, we found that generating the proposed n-dimensional vector from the Markov random field prior in (3.1) led to reasonable acceptance rates (about 7% for the cases we examined, discussed in the following section). To achieve this efficiently, we used Rue's (2001) algorithm for fast simulation of Gaussian Markov random fields. Acceptance rates for the reversible jump move in the GC model were also of the order of 7%. Other parameters drawn through Metropolis-Hastings steps also had reasonable acceptance rates (of course, the actual values depended on the parameter, model and data set considered). In all the data sets analysed in the next section, the sampler was ran for an initial 500,000 draws which were discarded, followed by 1,000,000 draws which were used to present results. This was found to be more than enough for convergence.

We have also carefully monitored serial autocorrelations along the MCMC path, yielding results consistent with these observations. For example, for the GC model applied to the larynx cancer data described in Section 6.3 below, autocorrelations in the h and τ series declined to insignificant levels after about 200 and 80 sweeps, respectively, while the series for individual λ_{z_i} did so after lags varying from 8 to 40. Dependence is greater in posterior samples from the LN model, where the corresponding figures for h, ϕ and λ_{z_i} are 600, 600 and 10 to 50. In both cases, as we make clear, we generate very long runs of 1,000,000 sweeps, after burn-in, so that substantial thinning is possible (in the interests of saving storage space) while still leaving large sub-samples for computing summary statistics. The programs were coded in GNU Fortran 77, and executed on a 750 MHz Pentium PC; running time was approximately 1 hour.

6 Results

We will illustrate the performance of the two spatial mixture models developed earlier in the paper on both real and synthetic data sets, making also some limited comparisons with a standard model.

We will present results for the LN and GC models, described in Section 3. The runs presented here correspond to hyperparameter values $h_{\max} = \phi_{\max} = 10$ and $\tau_{\max} = 0.5$ for the priors in (3.6). These values were chosen so as to allow reasonable flexibility in the degree of prior spatial correlation. For the GC model, the choice $\tau_{\max} = 0.5$ ensures that marginal prior distributions of the allocation variables are approximately Uniform, which always holds under the LN model and seems sensible in the absence of strong prior information. Larger values of $\tau_{\max}$ would favour prior allocations towards the two extreme components, as explained at the end of Subsection 3.2. In addition, we took $\alpha = 1$ and $\beta = 0.69$ in (4.3). When $k = 1$, this corresponds to an Exponential prior distribution for the relative risk with median equal to 1, which seems reasonable. Finally, s was taken equal to 5 in the GC model (see (3.5)) and $k_{\max}$ was assigned the value 10.

Certain quantities in the mixture model, in particular the component-specific parameters ($\lambda_1, \ldots,$ λ_k) and the allocations $(z_1, \ldots, z_n)$, can only be interpreted in the context of a fixed number of components. Thus, when posterior inference on these quantities is reported, it is conditional upon k. On the other hand, region-specific relative risks, defined as $r_i = \lambda_{z_i}$, have an unequivocal interpretation regardless of the value of k, so we present results for them mixing over k. Much of the subsequent discussion will focus on the relative risks $(r_1, \ldots, r_n)$, which are often the quantities of ultimate interest and the only ones that allow for direct comparison with other models.

6.1 A standard model for comparison

The model of Besag *et al.* (1991), henceforth the BYM model, can be written as:

$$\log r_i = \kappa^{1/2} u_i + \eta^{1/2} v_i, \tag{6.1}$$

where

$$p(u_1, \ldots, u_n) \propto \exp\left\{-\sum_{i \sim i'} \frac{(u_i - u_{i'})^2}{2}\right\} \quad \text{and} \quad p(v_1, \ldots, v_n) = \prod_{i=1}^{n} f_N(v_i|0, 1), \tag{6.2}$$

that is, $(u_1, \ldots, u_n)$ jointly follow a Gaussian conditional autoregressive model whereas $v_1, \ldots, v_n$ are independent standard normals. We consider the hyperpriors

$$\kappa^{-1} \sim \text{Gamma}(\kappa_1, \kappa_2) \quad \text{and} \quad \eta^{-1} \sim \text{Gamma}(\eta_1, \eta_2), \tag{6.3}$$

where we make the weakly informative choices $\kappa_1 = \kappa_2 = 0.1$ and $\eta_1 = \eta_2 = 0.001$, as recommended by Best *et al.* (1999). Whenever we refer to the BYM model, we mean the one corresponding to sampling from (4.1) with the prior given through (6.1)–(6.3).

6.2 Synthetic data sets

Here we analyse two synthetic data sets that were designed to test the performance of our models in different situations. Both simulated data sets, as well as the larynx cancer real data analysed in the next subsection, correspond to counts in the 94 mainland French départements. The expected numbers of cases e_i have been always taken equal to those in the larynx cancer data set, which vary from 2.08 for Lozère to 57.77 for Paris, with $\sum e_i = 1339$. The large variation in the expected number of cases for the various regions will make maps based on SMR's unreliable and some form of spatial smoothing necessary.

In constructing the first data set, denoted "blocks", we set most regions to have true relative risk equal to 0.75, but four clusters of regions have relative risk 2. Counts are generated independently from the model in (4.1), using the expected number of cases and relative risks just described. This results in counts y_i that vary from 0 (Lozère) to 91 (Nord) with $\sum y_i = 1397$. The SMR's vary from

0 (Lozère) to 2.93 (Maine-et-Loire). Figure 2 displays the true relative risks and SMR's as well as the medians of the marginal posterior distributions of the relative risks, according to the BYM and GC models. In Table 1, we display for each of the methods the root averaged mean square error of the log relative risk, defined as $\{\sum_i E[(\log(\lambda_{z_i}) - \log(r_i^t))^2]/n\}^{1/2}$, where r_i^t is the true relative risk in region i.

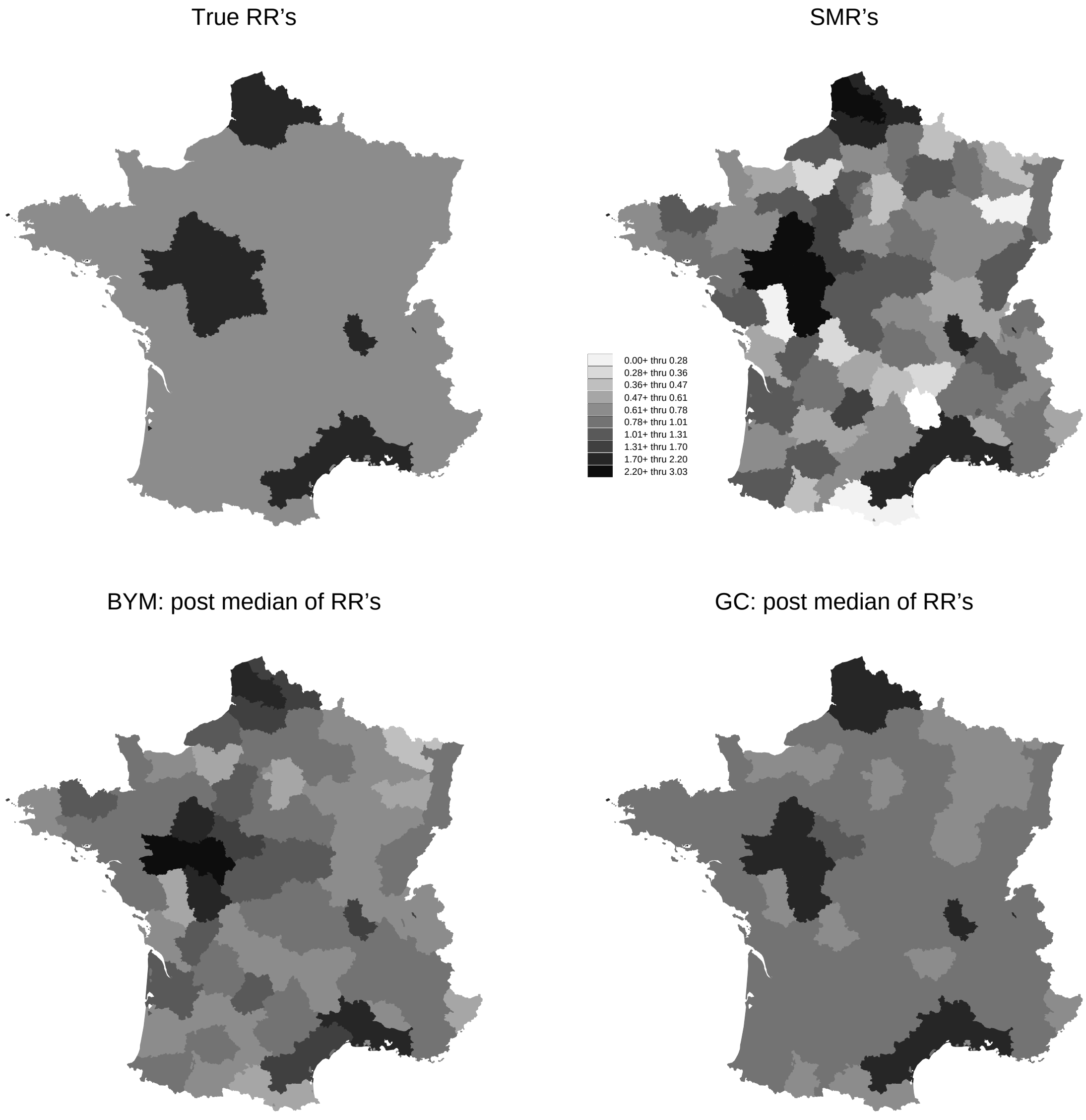

Figure 2: "blocks" synthetic data set: true relative risks, SMR's, and posterior medians of relative risks according to BYM and GC models.

On the basis of mean square error, both mixture models give more accurate marginal inference on the true relative risks than does the BYM approach. The posterior median reconstruction based on the GC model (Figure 2) is more adequate than the one obtained from the BYM model,

Table 1: Simulation results comparing spatial mixture and BYM models; root averaged mean square errors for log relative risk.

data set	LN	GC	BYM
blocks	.265	.229	.332
gradns	.293	.266	.218

which is less successful at detecting boundaries while undersmoothing within clusters of regions with common risk. This is probably due to the fact that, in order to accommodate changes across boundaries, which are not naturally built into the BYM model, the latter needs to undersmooth the entire map. The map obtained using the LN model is quite similar to the one from the GC model, albeit slightly less smooth within the low rate area.

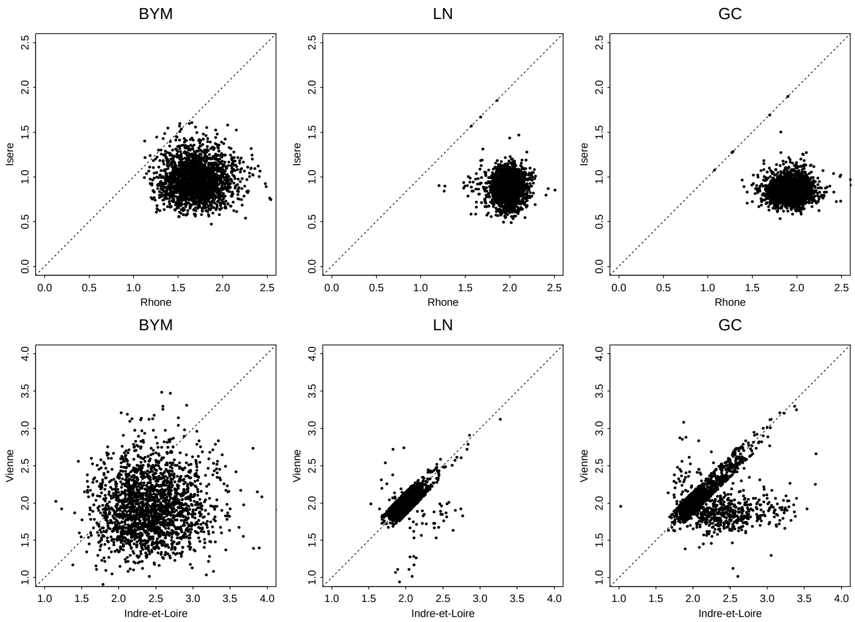

Figure 3: Scatter plots of samples from joint posterior of relative risks for selected pairs of neighbouring regions, for "blocks" data set. Some distortion has been introduced in the diagonal to aid visualisation. Top panels: Rhône and Isère. Bottom panels: Indre-et-Loire and Vienne.

Figure 3 presents scatter-plots corresponding to the joint posterior distribution of the relative risks for selected pairs of neighbouring regions with all three models. The upper panels correspond to Rhône (which has true relative risk 2) and Isère (with relative risk 0.75). Clearly, the mixture models are better at detecting this boundary, in particular, they lead to a smaller spread in the posterior distribution. The lower panels correspond to Indre-et-Loire and Vienne, both with true relative risk 2. The mixture models again perform better in this case. In particular, the BYM model overestimates the relative risk of Indre-et-Loire which has a fairly large SMR (2.79). The LN model does best, the posterior probability on the diagonal is 97%, and clearly located around the true value 2. The GC model puts 73% of the mass on the diagonal.

The second data set, “gradns”, is aimed at testing the performance of our models in a situation where there are no underlying clusters or sharp boundaries between regions but, instead, the true relative risks decrease smoothly from the north to the south of the country. The smallest true relative risk is 0.32 (Pyrénées-Orientales) and the largest is 1.64 (Nord). Counts have been generated according to the model in (4.1) using these relative risks. The counts vary from 0 (Lozére) to 92 (Paris), with $\sum y_i = 1378$. The corresponding SMR's vary from 0 (Lozére) to 2.28 (Meuse). The true relative risks and SMR's are displayed in Figure 4. The other three maps in Figure 4 correspond to the posterior medians of the relative risks, respectively using the BYM, LN and GC models. Here, they all perform quite similarly qualitatively; in terms of posterior mean square error (Table 1) the BYM model is superior, followed by the GC and LN models. We find it encouraging that mixture models are able to come close to the performance of the BYM model in a case where this would seem likely to perform better.

6.3 Larynx cancer data set

These data consist of counts of deaths from larynx cancer among females across the mainland French départements, for the period 1986–93 (Rezvani *et al.*, 1997). The expected number of cases are as explained at the beginning of Subsection 6.2, whereas the actual counts range from 0 (Lozère) to 75 (Paris). For this data we have that $\sum y_i = \sum e_i = 1339$, and the SMR's vary from 0 (Lozère) to 2.12 (Hautes-Alpes). Figure 5 displays the SMR's. First we concentrate on some results specific to the mixture models, and then we present comparisons with the BYM model. To save space, we will only present graphical displays for certain quantities and make comments on others in the text.

Figure 6 plots the posterior distribution of k, the number of components, under the LN model (solid line) and GC model (dotted line). The LN model clearly favours an explanation using just two components, whereas the posterior under the GC model is much more diffuse without strongly favouring any value of k (the modal value in this case is $k = 4$, but the posterior probability attached to it is smaller than 0.2). We have also examined the posterior distribution for (h, ϕ), the spatial interaction parameters in the LN model. The latter is concentrated on large values of h and small values of ϕ, thus favouring the hypothesis of strong spatial dependence. A similar message is

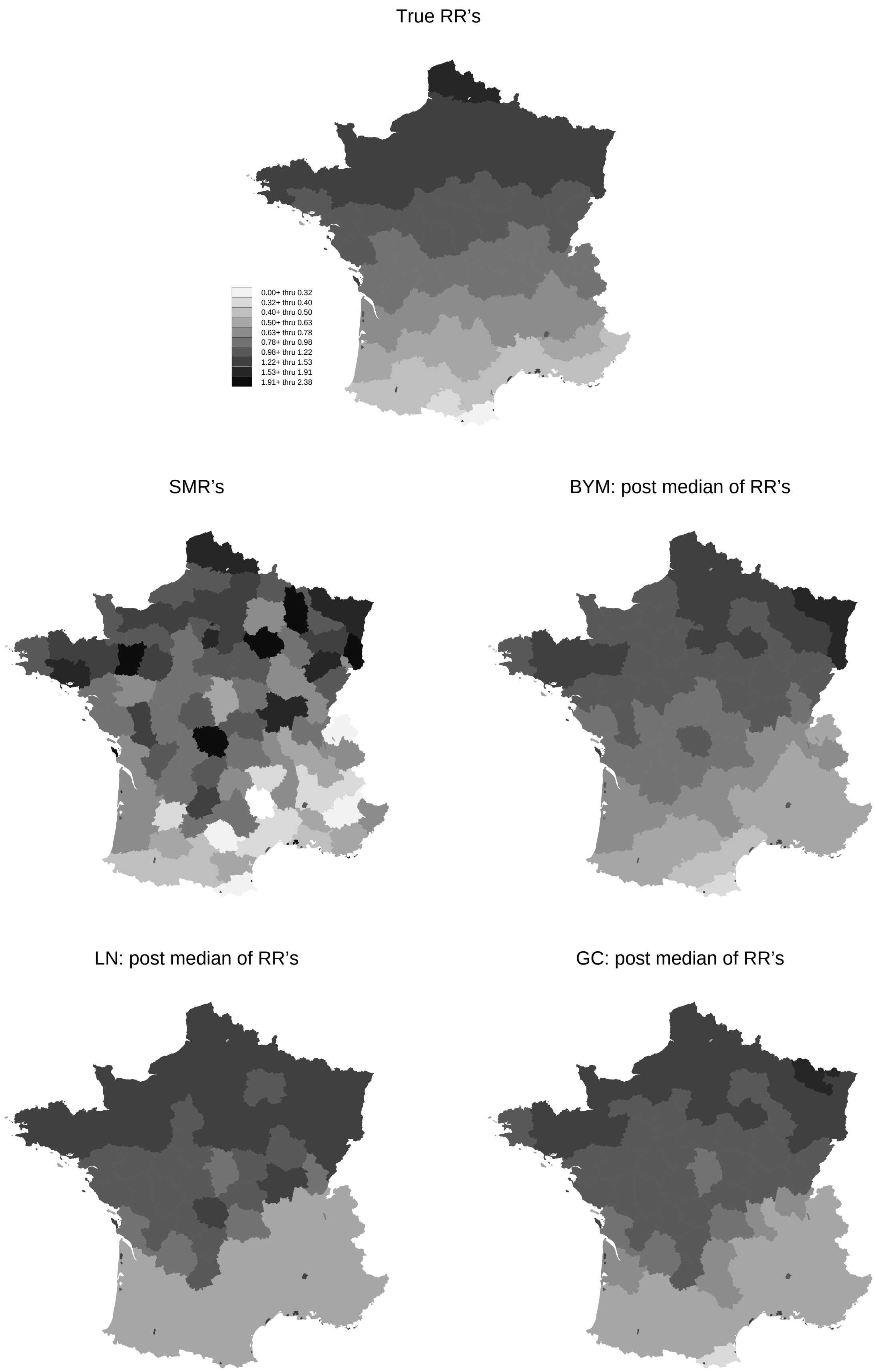

Figure 4: "gradns" synthetic data set: true relative risks, SMR's, and posterior medians of relative risks from BYM, LN and GC models.

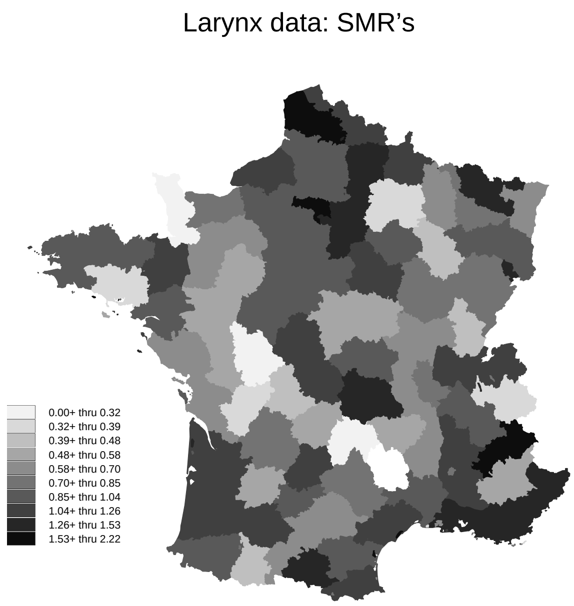

Figure 5: SMR's for the larynx cancer data set.

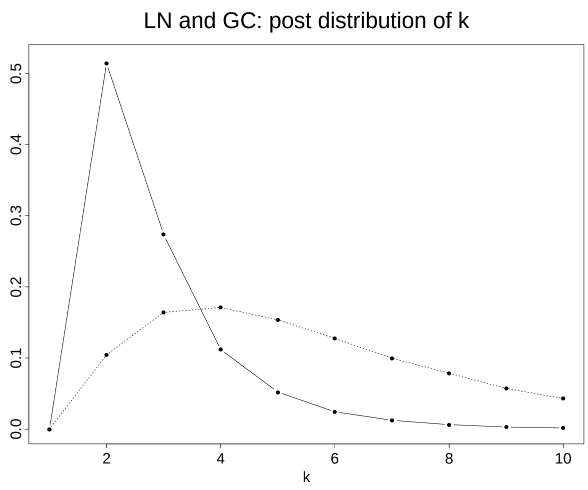

Figure 6: Posterior distributions for the number of components k in spatial mixture models fitted to the larynx cancer data set: LN and GC models, solid and dotted lines respectively.

obtained through the posterior distribution of (h, τ) in the GC model.

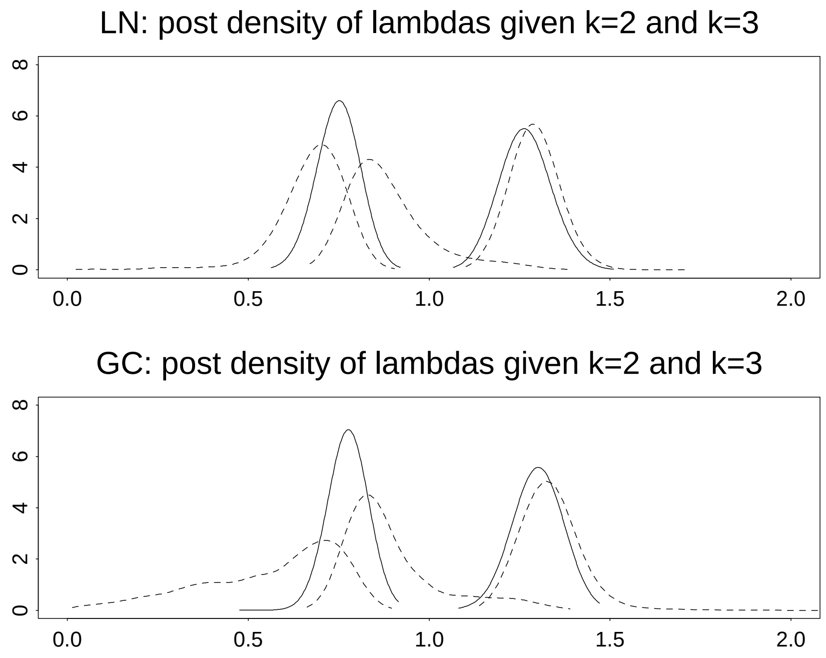

Figure 7: Posterior density estimates of component parameters λ_j for spatial mixture models fitted to larynx cancer data set: LN and GC models in top and bottom panels respectively. Solid lines: $k = 2$, broken lines: $k = 3$.

For the LN model, Figure 7 (upper panel) gives the marginal posterior densities of λ_1 and λ_2 given $k = 2$ (solid lines), and of λ_1, λ_2 and λ_3 given $k = 3$ (broken lines). For these parameters, we present results for $k = 2$ and $k = 3$ because they are the most favoured values under the LN model (see Figure 6). The main effect of going from $k = 2$ to $k = 3$ components is that the first component is split into two. A similar pattern emerges with the GC model (lower panel of Figure 7). We have also examined the marginal posterior distributions of the allocation variables for each of the regions, again conditioning on $k = 2$ and $k = 3$. For $k = 2$, these distributions are virtually identical under the LN and GC models, which makes sense as the same holds for the posterior distribution of the relative risks associated with each of the components (see Figure 7). When $k = 3$, the marginal posterior distributions of the allocations are slightly different under the two models. In general terms, we see that regions that under $k = 2$ were allocated with high posterior probability to the component associated with larger relative risk, tend to remain allocated to the component associated with largest relative risk under $k = 3$. On the other hand, regions that had high probability of being allocated to the component leading to smaller relative risk under $k = 2$, now essentially have that probability split between the components corresponding to smallest and intermediate relative risks. For these regions, the GC model tends to favour the intermediate component over the smallest more often than the LN model, which is consistent with the message from Figure 7 that the smallest relative risk is more extreme under the GC than under the LN model.

The rest of the results presented here relate to the posterior distribution of the relative risks $(r_1, \ldots, r_n)$. As already explained, this distribution is computed by also mixing over k and, of course, these quantities are directly comparable with those in the BYM model. Figure 8 displays histograms of the marginal posterior distributions of the relative risks for each of the regions, respectively under the BYM, LN and GC models. The mixture models lead to less dispersed posterior distributions than the BYM model. Also note that the mixture models lead to bimodal histograms in some regions. This is more apparent with the LN model, probably because it fits fewer (usually two) components, leading to a less smooth posterior distribution for the relative risks. Figure 9 presents the posterior median of the relative risks. The GC model seems to offer an intermediate type of smoothing between those resulting from the BYM and LN models. This is consistent with the fact that the GC model tends to allocate (at least a priori) neighbouring regions to components that are adjacent in terms of their associated relative risks.

In addition to looking at marginal posterior distributions of the relative risks, we have explored joint distributions for some pairs of neighbouring regions. As an illustration, the top panels of Figure 10 present scatter-plots of the joint posteriors of the relative risks for Aisne (corresponding to $i = 2$) and Marne ($i = 50$) under all three models. The SMR for Aisne (1.32) is much higher than for Marne (0.34). In accordance with this, the joint posterior distributions of the relative risks have most of the probability in the region where $r_2 > r_{50}$ and this effect is more pronounced in the LN model. The probability on the diagonal is 8% for the LN model and 12% for the GC model. Note that the scatter-plot for the GC model leads to a situation in between the BYM and LN models, with the mass a bit more spread out and slightly closer to the diagonal than in the LN model. A similar pattern was observed when we looked at other pairs of neighbouring regions. For conciseness, the lower panels of Figure 10 simply present the posterior distribution of the differences in relative risks between regions Puy-de-Dôme and Allier (left panel) and regions Moselle and Meurthe-et-Moselle (right panel). In both cases, the solid line corresponds to the BYM model, whereas the dotted and broken lines respectively represent the LN and GC models. For the mixture models, the spike at zero represents the posterior probability that the relative risks for both regions are identical: this probability is 0.40 for LN model and 0.45 for GC model in the left panel, and 0.22 for both models in the right panel.

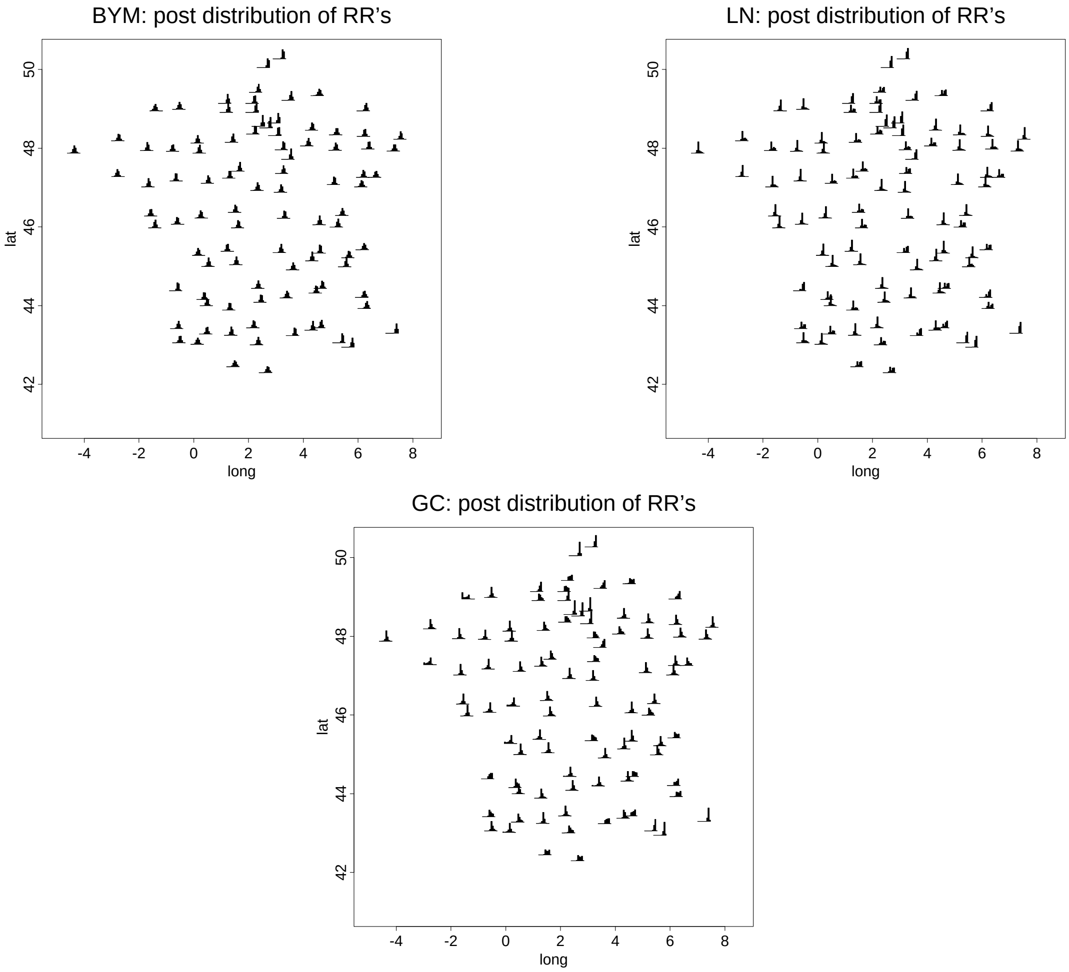

Figure 8: Histograms of the posterior distributions for larynx cancer relative risks in the 94 mainland French départements, each plotted at the approximate centre of the département (with some distortion near Paris for clarity). Ten classes with cutpoints equidistant on logarithmic scale (starting at 0.32 and increasing by a factor 1.21).

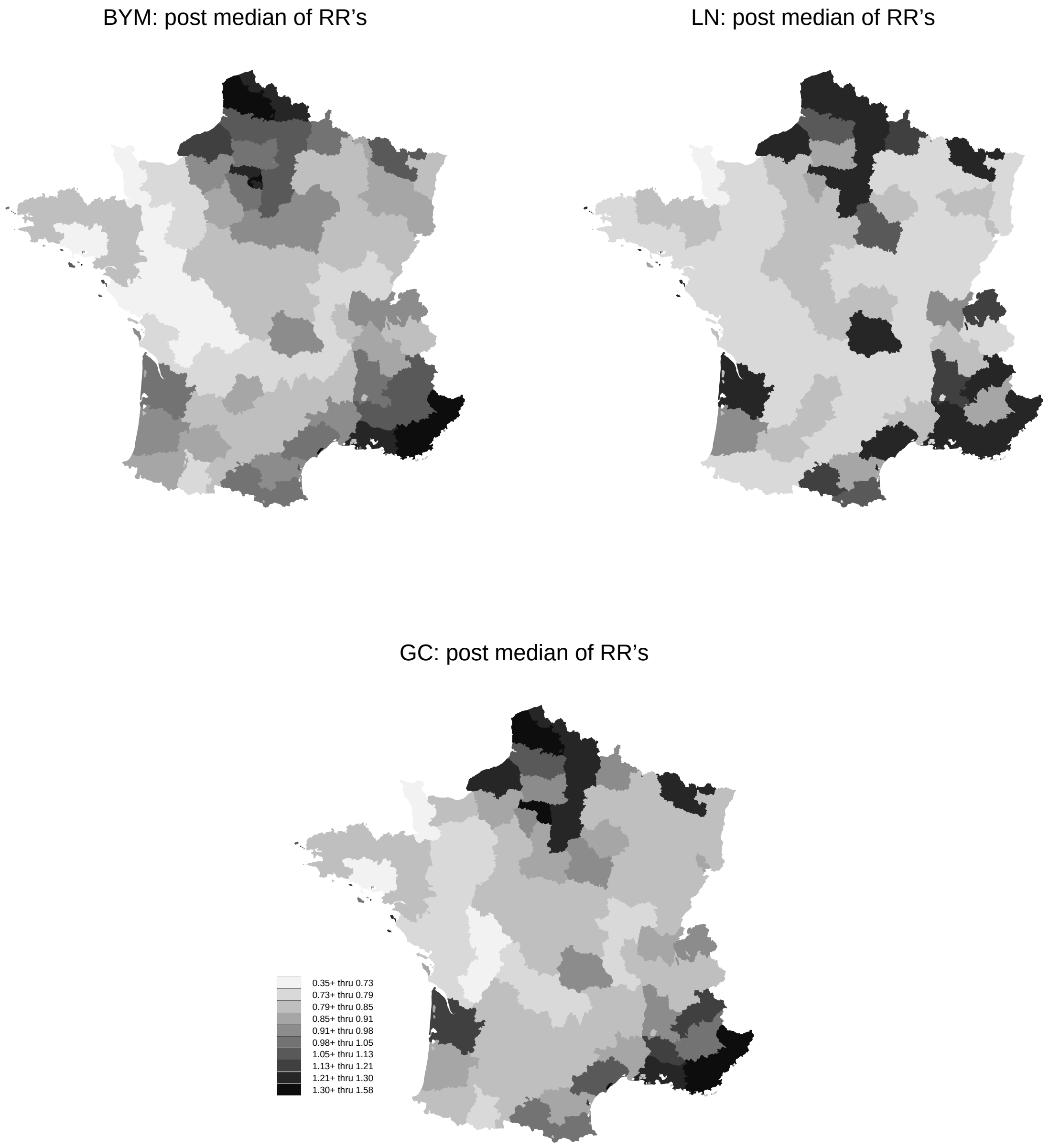

Figure 9: Posterior medians of relative risks for larynx cancer data set, according to BYM, LN and GC models.

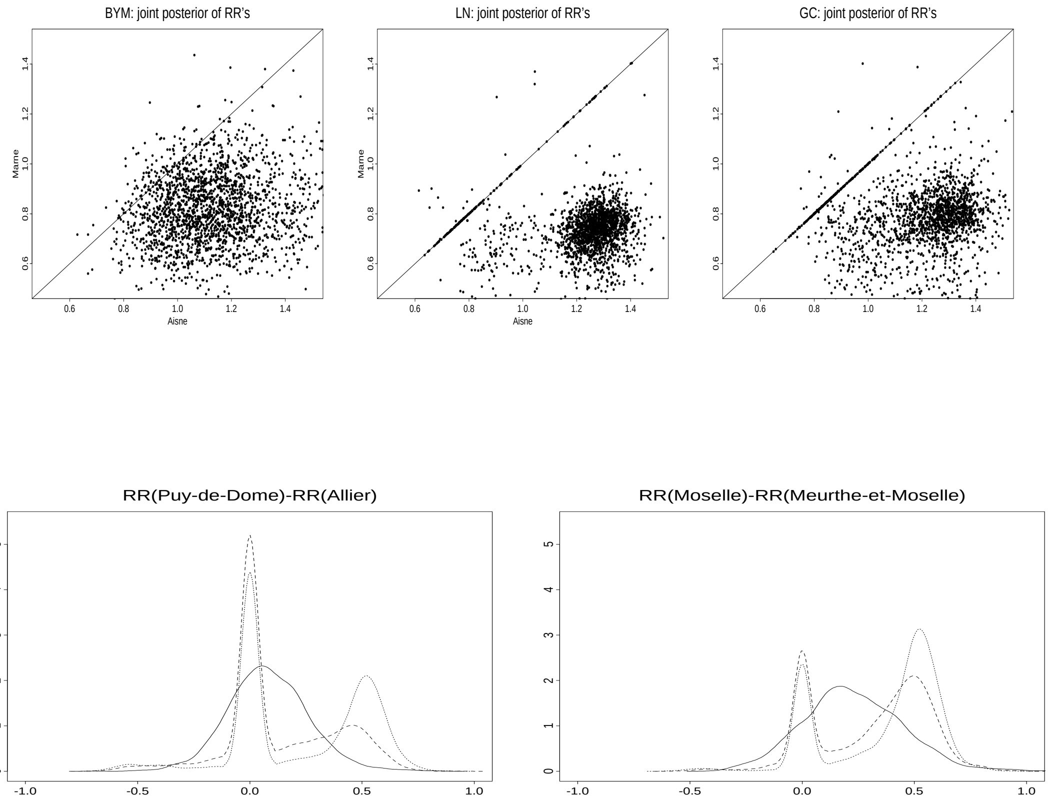

Figure 10: Aspects of the joint posterior distributions of relative risks for pairs of neighbouring regions, for larynx cancer data. Top panels: scatter plots of samples from joint posterior for Aisne and Marne with BYM, LN and GC models. Bottom panels: density estimates of posterior for differences between relative risks for Puy-de-Dôme and Allier (left) and Moselle and Meurthe-et-Moselle (right): solid, dotted and broken lines indicate BYM, LN and GC models.

6.4 Sensitivity to prior assumptions

Here we shall briefly comment on the sensitivity of the results with respect to the prior assumptions. We shall focus the discussion on the larynx data set, although we have also examined robustness for the synthetic data sets and found qualitatively similar results.

First of all, one should acknowledge that in complex models, such as those developed in this paper, it is virtually impossible fully to understand and separate the role of all the parameters that intervene in the model. Furthermore, posterior distributions of some of the parameters will be heavily influenced by prior choices. In our view, one should mostly focus on the quantities of real interest, the relative risks in this case, and examine whether or not inference on such quantities is heavily dependent on prior choices. We have run our programs many times using different prior hyperparameters and even different prior distributions, and found that posterior inference on relative risks is fairly robust to moderate perturbations in the prior.

For the LN model, our basic setting takes $h_{\max} = \phi_{\max} = 10$ in (3.6), and $\alpha = 1$ and $\beta = 0.69$ in (4.3). With this prior choices, the posterior distribution of h (not displayed) has mass fairly evenly spread across the entire interval $(0, 10)$, whereas the posterior distribution of ϕ is mostly concentrated close to 0. As a consequence, we have examined results under a range of higher values of $h_{\max}$. The top left panel of Figure 11 displays histograms of the marginal posterior distributions of the relative risks when $h_{\max} = 100$ (a rather extreme choice) which, as can be seen, are not dissimilar from those presented in Figure 8. The middle left panel corresponds to $p(h) = (1+h)^{-2}$ or, equivalently, a Uniform prior distribution on $(0, 1)$ for the parameter $h/(1 + h)$. This prior does not restrict h to a bounded interval and is motivated by the fact that $h/(1 + h)$ represents the relative weight of each pair of neighbouring locations in (3.1). Again, histograms of posterior distributions of relative risks are quite similar to those in Figure 8. We have also found reasonable robust results to moderate departures from the values $\alpha = 1$ and $\beta = 0.69$ in (4.3). The bottom left panel of Figure 11 corresponds to a random β with a Gamma hyperprior with mean 1 and variance 10. The panels on the right-hand side of Figure 11 correspond to the same departures in terms of h and β from the basic setting $h_{\max} = 10$, $\tau_{\max} = 0.5$, $\alpha = 1$, $\beta = 0.69$ in the GC model. Again, robustness holds in a qualitative sense.

Some sensitivity of the posterior distribution of k, the number of mixture components, to the choice of α and β is to be expected (it is not surprising that assumptions about location and spread of the components affect the number of components fitted by the model). This appears to be more the case in the GC than in the LN model, as Figure 12 illustrates (right panels). The left panels of

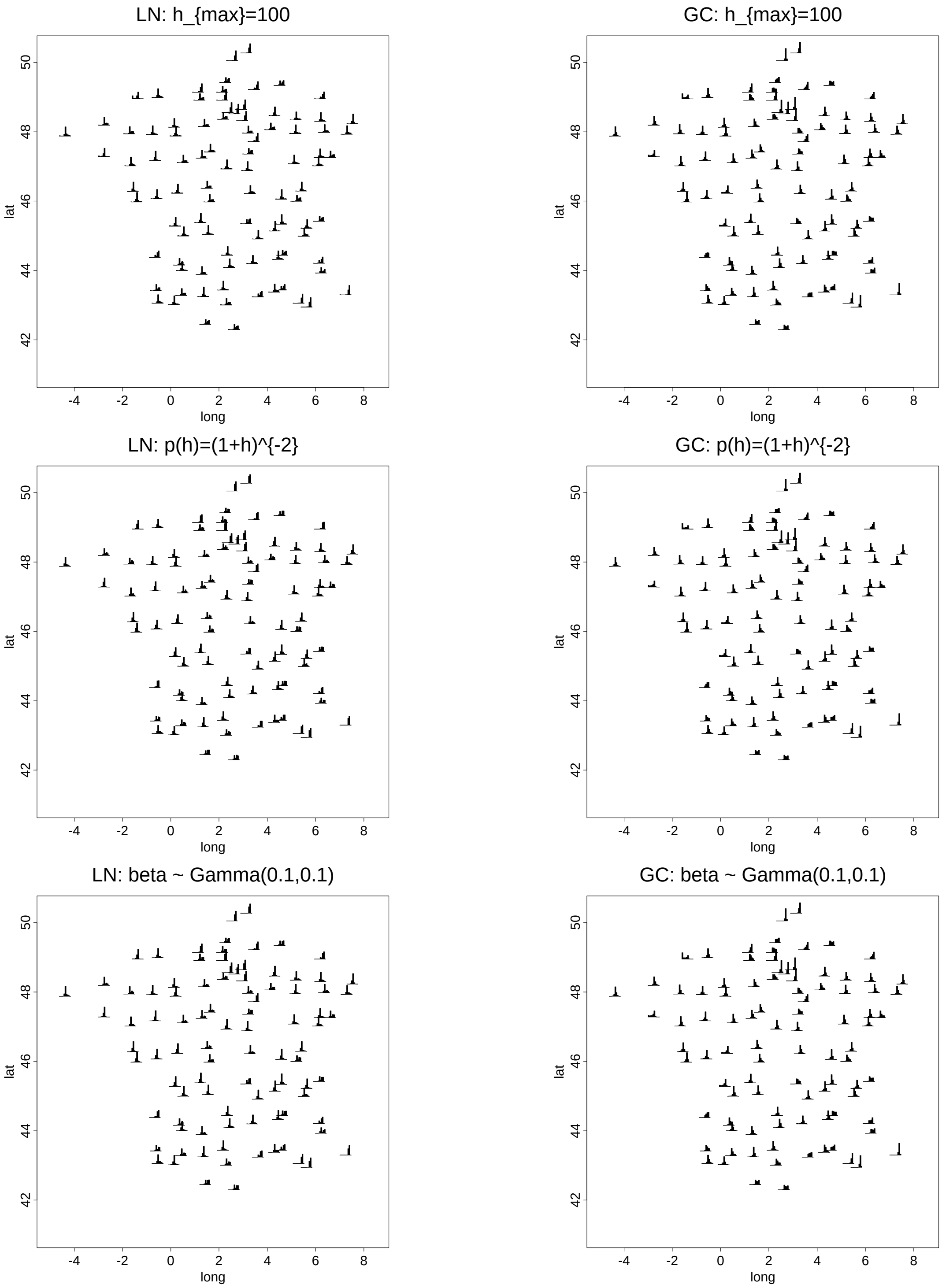

Figure 11: Robustness of posterior distributions for larynx cancer relative risks. Ten class histograms with cutpoints equidistant on logarithmic scale (starting at 0.32 and increasing by a factor 1.21). Titles indicate departures from basic setting in Figure 8, where $h \sim \text{Unif}\,(0, h_{\max} = 10)$ and $\beta = 0.69$.

Figure 12 point towards robustness with respect to the prior distribution of the spatial parameter h.

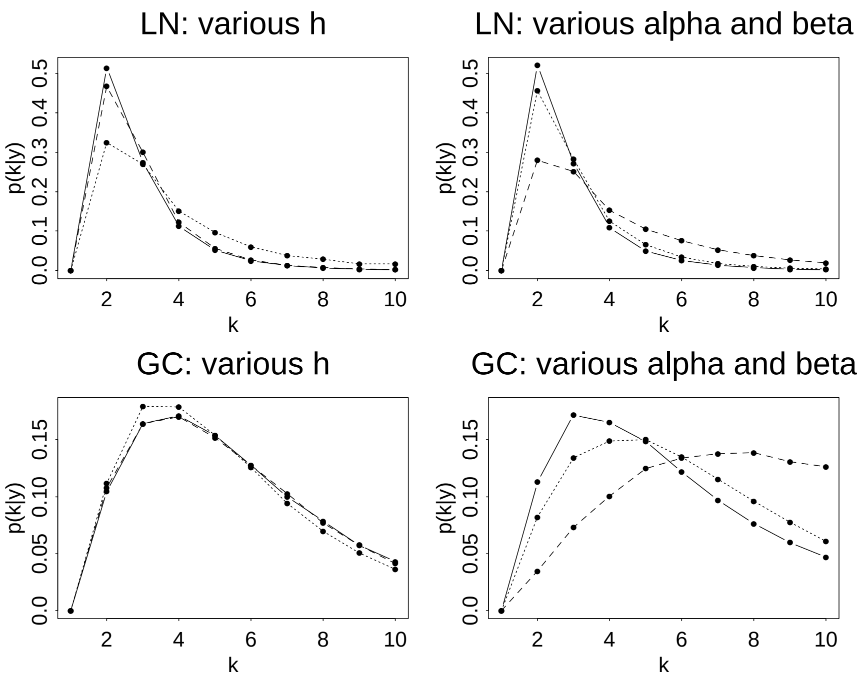

Figure 12: Sensitivity of the posterior distribution of number of components k for larynx cancer data set with LN (top panels) and GC (lower panels) models. Left panels: $\alpha = 1$ and $\beta = 0.69$ throughout, with $h_{\max} = 10$ (solid lines), $h_{\max} = 100$ (dotted lines) and $p(h) = (1+h)^{-2}$ (broken lines). Right panels: $h_{\max} = 10$ throughout, with $\alpha = 1$ and $\beta \sim$ Gamma(0.1,0.1) (solid lines), $\alpha = \beta = 1$ (dotted lines) and $\alpha = \beta = 2$ (broken lines).

7 Conclusion and possible extensions

We believe the spatial mixture models introduced and discussed in this paper provide interesting new tools for those modelling heterogeneity in spatial data. As might have been anticipated, they seem to offer improved performance in the task of estimating the true risk pattern compared to a model based on a continuously-distributed Markov random field, when that true pattern has step-like discontinuities. What is more intriguing is that not much performance is lost in the opposite case of a smooth trend. It is also remarkable that the LN version of the model uncovers such a simple 'explanation' for the larynx data set, as a mixture of just two components.

Further work is needed to draw wider comparisons, which should include other mixture- and partition-based models such as those of Knorr-Held and Rasser (2000) and Green and Richardson

(2000), and also variants of the random field models in which the Gaussian assumption is replaced by use of a non-quadratic pairwise potential function in (6.2), such as the absolute value or log cosh function. Finally, it would be interesting to explore the formulation and performance of related correlated mixture models in other spatial contexts, and also in temporal problems such as signal analysis.

8 Appendix: Details of the sampler

8.1 Implementation for the Logistic Normal model

Denoting by $(k, \lambda, x, h, \phi, z_1, \ldots, z_n)$ the current state of the chain, we follow steps (a)–(f). Note that the allocation variables have been integrated out in steps (a)–(d), whereas they are included in steps (e)–(f). One time step for the chain comprises a complete sweep through the six steps, at the end of which the state is recorded.

(a) Update k, λ and x:

Updating k implies a change in dimensionality for $\lambda = (\lambda_1, \ldots, \lambda_k)$ and $x = (x_1, \ldots, x_k)$, so new values also need to be proposed for them. We use reversible jump to update (k, λ, x) according to $p(k, \lambda, x|y, h, \phi)$.

We propose to update k to k', where $k' = k+1$ (add one component) with probability $b_k \in [0,1]$ and $k' = k-1$ (remove one component) with the remaining probability.

If $k' = k+1$, we propose to update λ and x in the following way. Draw a value λ_* from the prior Gamma(α, β) distribution and an n-dimensional vector x_* from the prior in (3.1). Form $\lambda' = (\lambda'_1, \ldots, \lambda'_{k+1})$ by inserting λ_* in the appropriate position within the ordered vector λ, and x' by inserting x_* in the same place in x. Compute weights w'_{ij}, as in (3.3), using x' in place of x (note that all the weights change).

If $k' = k-1$, we randomly choose a component $*$ using a Uniform distribution on the set $\{1, \ldots, k\}$. Form λ' and x' by removing the value λ_* from λ and the n-dimensional vector x_* from x. Recompute weights w'_{ij} using (3.3) and x' (again all weights change).

According to the reversible jump framework, the acceptance probability for the move that adds one component is

$$\min\left(1, \frac{1-b_{k+1}}{b_k} \prod_{i=1}^{n} \frac{\sum_{j=1}^{k+1} w'_{ij} \mathrm{Pois}(y_i|e_i\lambda'_j)}{\sum_{j=1}^{k} w_{ij} \mathrm{Pois}(y_i|e_i\lambda_j)}\right), \tag{8.1}$$

whereas the acceptance probability for the move that removes one component only requires obvious changes. If the proposal is not accepted the chain keeps the current state.

This step can potentially be slow due to the generation of the n-dimensional vector x_* when $k' = k+1$. This problem is more acute when h is unknown, because then the precision matrix in (3.1) changes at every sweep of the chain and computing a new Choleski decomposition is then required. Thus, instead of a standard algorithm for multivariate normal random generation, we used the algorithm developed by Rue (2001) for fast sampling of Gaussian Markov random fields.

(b) Update x:

We update the n regions sequentially. For region i, the posterior full conditional distribution of $(x_{i1}, \ldots, x_{ik})$ (with allocations integrated out) has p.d.f. proportional to

$$\Big\{\sum_{j=1}^{k} w_{ij} \mathrm{Pois}(y_i | e_i \lambda_j)\Big\} \prod_{j=1}^{k} f_N\Big(x_{ij} \Big| \frac{h \sum_{i' \in \partial i} x_{i'j}}{1 + h\nu_i}, \frac{1}{1+h\nu_i}\Big), \tag{8.2}$$

where $f_N(\cdot | m, v)$ denotes the p.d.f. of a univariate normal distribution with mean m and variance v, and ν_i is the number of neighbours of region i. We use a standard Metropolis-Hastings updating scheme, drawing a candidate value $(x'_{i1}, \ldots, x'_{ik})$ from the k independent normal distributions in (8.2).

(c) Update h:

The posterior full conditional distribution for this parameter has p.d.f. proportional to

$$c(h)^k \exp\Big\{-\frac{h}{2} \sum_{j=1}^{k} \sum_{i \sim i'} (x_{ij} - x_{i'j})^2\Big\}\, I[0 < h < h_{\max}], \tag{8.3}$$

and we use Metropolis-Hastings with candidate generated from prior. Clearly, the ability to evaluate $c(h)$ easily is important in this step.

(d) Update ϕ:

The conditional posterior distribution of this parameter (with allocations integrated out) has p.d.f. proportional to

$$\prod_{i=1}^{n} \Big\{\sum_{j=1}^{k} w_{ij} \mathrm{Pois}(y_i | e_i \lambda_j)\Big\}\, I[0 < \phi < \phi_{\max}]. \tag{8.4}$$

We update ϕ using Metropolis-Hastings with candidate generated from prior.

(e) Update allocations:

In the posterior full conditional, the n allocation variables are mutually independent with

$$p(z_i = j | y, k, \lambda, x, h, \phi) \propto w_{ij} \mathrm{Pois}(y_i | e_i \lambda_j)\, I[j \in \{1, \ldots, k\}], \tag{8.5}$$

from which we can draw directly. Thus, this is just a Gibbs step.

(f) Update λ:

The p.d.f. of the posterior full conditional distribution of λ is proportional to

$$\prod_{j=1}^{k} f_G\Big(\lambda_j \Big| \alpha + \sum_{i:z_i=j} y_i, \beta + \sum_{i:z_i=j} e_i\Big) I[\lambda_1 < \ldots < \lambda_k], \tag{8.6}$$

from which we draw directly (Gibbs).

8.2 Implementation for the Grouped Continuous model

Updating in this model will follow seven steps, with the allocation variables integrated out in the first five. These steps can be summarised as updating: (a) (k, λ, δ), (b) x, (c) h, (d) δ, (e) τ, (f) z, and (g) λ. Most of the steps are quite simple and can be dealt with quite similarly to the LN model. Thus, we will only briefly describe the move we used in step (a).

First, we propose $k' = k + 1$ with probability b_k and $k' = k - 1$ with the remaining probability. If $k' = k + 1$, we draw δ_* from a Uniform distribution on $(x_{\min}, x_{\max})$ and form δ' by inserting δ_* in the corresponding position in the ordered vector δ. Say δ_* occupies position j_* in δ'. Next, we form λ' by substituting λ_{j_*} in λ by the pair $\lambda'_{j_*} < \lambda'_{j_*+1}$, which are generated according to the p.d.f. $2\prod_{l=0}^{1} f_N(\lambda'_{j_*+l}|\lambda_{j_*}, v_s)$ for some positive variance v_s. Compute new weights using (3.4) with δ' in place of δ.

If $k' = k - 1$, we propose to remove one of the elements δ_{j_*} of δ, randomly and uniformly. Next, we substitute the pair $\lambda_{j_*} < \lambda_{j_*+1}$ by λ'_{j_*}, which is generated from a normal distribution with mean $(\lambda_{j_*} + \lambda_{j_*+1})/2$ and variance denoted by v_m. In this way, we form δ' and λ'. Further, we compute the new weights using (3.4).

According to the reversible jump framework, the acceptance probability of the move that adds one component is: zero if the proposed values λ'_{j_*}, λ'_{j_*+1} do not fall in the appropriate range, that is, do not lead to a vector λ' with ordered elements; otherwise, it is $\min(1, Q)$, where

$$Q = \frac{1 - b_{k+1}}{b_k} \left[\prod_{i=1}^{n} \frac{\sum_{j=1}^{k+1} w'_{ij} \mathrm{Pois}(y_i|e_i\lambda'_j)}{\sum_{j=1}^{k} w_{ij} \mathrm{Pois}(y_i|e_i\lambda_j)}\right] \frac{(ks + s - 1)!}{(ks - 1)!(s - 1)!} \frac{k + 1}{k} \left[\frac{(\delta'_{jt} - \delta_{jt-1})(\delta_{jt} - \delta'_{jt})}{(\delta_{jt} - \delta_{jt-1})(x_{\max} - x_{\min})}\right]^{s-1}$$
$$\left[\prod_{l=0}^{1} f_G(\lambda'_{j_*+l}|\alpha, \beta)\right] [f_G(\lambda_{j_*}|\alpha, \beta)]^{-1} f_N\left(\lambda_{j_*}\Big|\frac{\lambda'_{j_*} + \lambda'_{j_*+1}}{2}, v_m\right) \left[2\prod_{l=0}^{1} f_N(\lambda'_{j_*+l}|\lambda_{j_*}, v_s)\right]^{-1}.$$

The acceptance probability for the reverse move is obtained similarly.

Acknowledgements

We are most grateful for numerous stimulating discussions with Sylvia Richardson during the course of this work, and also for the advice of Havard Rue concerning simulation of Gaussian fields. We also thank two anonymous Referees for their constructive comments. The project was partially supported by the EPSRC, and most of the work was conducted while the first author was also at the University of Bristol.

References

Besag, J., Green, P. J., Higdon, D. and Mengersen, K. (1995) Bayesian computation and stochastic systems. *Statistical Science*, **10**, 3–66 (with discussion).

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

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. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith. Oxford: Clarendon Press.

Clayton, D. G. and Bernardinelli, L. (1992) Bayesian methods for mapping disease risk. In *Geographical and Environmental Epidemiology: Methods for Small-Area Studies*, eds. P. Elliot, J. Cuzick, D. English and R. Stern. Oxford: Oxford University Press.

Clayton, D. G. and Kaldor, J. (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, **43**, 671–681.

Diebolt, J. and Robert, C. P. (1994) Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society* B, **56**, 363–375.

Fernández, C. and Green, P. J. (1999) Contribution to discussion of paper by Besag and Higdon. *Journal of the Royal Statistical Society* B, **61**, 722–724.

Frühwirth-Schnatter, S. (2001) Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, **96**, 194–209.

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

Green, P. J. and Richardson, S. (2000) *Hidden Markov models and disease mapping.* Manuscript, University of Bristol.

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

Leblond, L., Richardson, S. and Green, P. J. (2000) *Mixture models in measurement error problems, with reference to epidemiological studies.* Manuscript, INSERM.

Leroux, B. G., Lei, X. and Breslow, N. (2000) Estimation of disease rates in small areas: A new mixed model for spatial dependence In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, eds. M.E. Halloran and D. Berry, pp. 179–191. New York: Springer.

Lindsay, B. G. (1995) *Mixture models: theory, geometry, and applications.* NSF-CBMS Regional Conference Series in Probability and Statistics, Vol. 5, Institute of Mathematical Statistics: Hayward, California.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models, Second Edition.* London: Chapman and Hall.

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

Müller, P. and Roeder, K. (1997) A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika*, **84**, 523–537.

Nobile, A. (1994) *Bayesian analysis of finite mixture distributions.* Ph. D. thesis, Carnegie Mellon University.

Nobile, A. and Green, P. J. (2000) Bayesian analysis of factorial experiments by mixture modelling. *Biometrika*, **87**, 15–35.

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. *Journal of the Royal Statistical Society* B, **59**, 731–792 (with discussion).

Robert, C.P. (1996) Mixtures of distributions: inference and estimation. Chapter 24 (pp. 441–464) of *Practical Markov chain Monte Carlo*, W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds. Chapman and Hall, London.

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.

Rue, H. (2001) Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society* B, **63**, 325–338.

Schlattmann, P. and Böhning, D. (1993) Mixture models and disease mapping. *Statistics in Medicine*, **12**, 1943–1950.

Tierney, L. (1994) Markov chains for exploring posterior distributions. *The Annals of Statistics*, **22**, 1701–1762 (with discussion).

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

Waller, L. A., Carlin, B. P., Xia, H. and Gelfand, A. E. (1997) Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, **92**, 607–617.

Watier, L., Richardson, S. and Green, P. J. (1999) Using Gaussian mixtures with unknown number of components for mixed model estimation. In the *Proceedings of the 14th International Workshop on Statistical Modelling, Graz (Austria)*, July 1999, pp.394–401.

West, M., Müller, P. and Escobar, M. D. (1994) Hierarchical priors and mixture models, with application in regression and density estimation. In *Aspects of Uncertainty: A tribute to Lindley*, A. F. M Smith and P. Freeman, eds. Wiley, New York.