Statistical Methods in Medical Research 2005; **14**: 35–59

A comparison of Bayesian spatial models for disease mapping

Nicky Best, Sylvia Richardson and Andrew Thomson Small Area Health Statistics Unit, Department of Epidemiology and Public Health, Imperial College Faculty of Medicine, Norfolk Place, London W2 1PG, UK

With the advent of routine health data indexed at a fine geographical resolution, small area disease mapping studies have become an established technique in geographical epidemiology. The specific issues posed by the sparseness of the data and possibility for local spatial dependence belong to a generic class of statistical problems involving an underlying (latent) spatial process of interest corrupted by observational noise. These are naturally formulated within the framework of hierarchical models, and over the past decade, a variety of spatial models have been proposed for the latent level(s) of the hierarchy. In this article, we provide a comprehensive review of the main classes of such models that have been used for disease mapping within a Bayesian estimation paradigm, and report a performance comparison between representative models in these classes, using a set of simulated data to help illustrate their respective properties. We also consider recent extensions to model the joint spatial distribution of multiple disease or health indicators. The aim is to help the reader choose an appropriate structural prior for the second level of the hierarchical model and to discuss issues of sensitivity to this choice.

1 Introduction

Disease mapping studies aim to summarize spatial variation in disease risk, in order to assess and quantify the amount of true spatial heterogeneity and the associated patterns, to highlight areas of elevated or lowered risk and to obtain clues as to the disease aetiology. For example, Jarup *et al.*[1] used disease mapping methods to investigate evidence of a link between environmental exposures to endocrine disrupting chemicals and prostate cancer incidence in UK. They argued that exposure to environmental carcinogens is unlikely to be evenly distributed geographically, which may give rise to variation in disease occurrence that is detectable in a spatial analysis. As no clear evidence of localized geographical clustering was found, they concluded that geographically varying environmental factors were unlikely to operate strongly in the aetiology of prostate cancer in UK. Buntinx *et al.*[2] used disease mapping methods to proactively ‘screen’ Belgian cancer registry data for evidence of elevated cancer rates in one or a group of municipalities, as part of a surveillance program aimed at addressing

Address for correspondence: Nicky Best, Small Area Health Statistics Unit, Department of Epidemiology and Public Health, Imperial College Faculty of Medicine, St. Marys Campus, Norfolk Place, London W2 1PG, UK. E-mail: n.best@ic.ac.uk

10.1191/0962280205sm388oa

public concerns about disease clusters. Any clusters thus identified were then further analysed and adjusted for area level background factors, and if the increased risk remained, a protocol was established to carry out separate epidemiological studies to investigate possible causes.

Closely related to disease mapping is the spatial analysis of health performance indicators, such as the application by MacNab[3] to study regional variations in incidence rates of intraventricular haemorrhage in neonates in Canadian Neonatal Intensive Care Units (NICU). The goal of the latter analysis was to help facilitate exploration and implementation of quality improvement programs to reduce unnecessary variations in adverse NICU outcomes.

All these applications share similar statistical problems but, of course, the implication and interpretation of the spatial patterns found are different. In such studies, the level of analysis and level of inference are both at the group level, where the groups are typically composed of people living in defined geographical areas (or in the case of health performance indicators, institutions or health service providers located within geographical areas).

In epidemiology, geographical variations of disease have been a subject of interest for a long time, as exemplified in the monograph by Doll.[4] The advent of routine health data indexed at a small geographical resolution, and thus sparse by nature, posed specific problems of modelling and analysis that were recognized in the 1990s as a generic class of statistical problems involving an underlying (latent) spatial process of interest corrupted by observational noise. Such structures also arise in image analysis or agricultural field trials, and are naturally formulated within the framework of hierarchical models. Subsequently, the field of disease mapping has flourished with a variety of estimation methods and spatial models being proposed for the latent level(s) of the hierarchy. Hierarchical formulations are easily dealt with within a Bayesian paradigm, whether empirical or fully Bayesian, and an overwhelming majority of the work being carried out in this domain has thus been placed in this context. Indeed, as the main goal of disease mapping is the estimation of the overall spatial pattern of risk, borrowing of strength across the whole of the study region and reducing the variance through the use of shrinkage estimators is of great importance.

It is well established that Bayes procedures offer a trade-off between bias and variance reduction of the estimates and, particularly in cases where the sample size is small, have been shown to produce a set of point estimates that have good properties in terms of minimizing squared error loss.[5] This variance reduction is attained through the borrowing of information resulting from the adopted hierarchical structure. This leads to Bayes point estimates that are shrunk towards a value that is related to the distribution of all the units included in the hierarchical structure. The effect of shrinkage is thus dependent on the prior structure that has been assumed and conditional on the latter, being close to the 'true model' in some sense. In consequence, it is to be expected that different spatial priors will lead to different shrinkage and the desirable properties of the estimates thus obtained will depend on the ultimate goal of the estimation exercise. If the aim is to estimate histograms or ranks of the area relative risks, loss functions other than squared error loss should be considered. This has been well illustrated in case studies by

Conlon and Louis[6] and Stern and Cressie.[7] Our review of spatial models addresses the primary context of producing and interpreting sets of point estimates that give a good indication of the overall spatial pattern of risk, both in terms of indicating the presence of heterogeneity in the relative risks and also highlighting whether this heterogeneity is linked to spatial aggregation of areas of similar high or low risk, or to individual areas with unusual risk.

Previous reviews of methods for spatial analysis of disease risk include Marshall,[8] who considers empirical Bayes and some early fully Bayesian methods for disease mapping; Bithell,[9] who discusses nonparametric and model based approaches for both areal and point data; the short introductory chapter on disease mapping by Lawson *et al.*[10] and the recent chapter by Richardson[11] that reviews the use of spatial models in epidemiological applications with an emphasis on linking point level and area level formulations. Marshall's review also covers cluster detection methods, which we do not consider here because their goal is rather different from that of producing a set of area specific risk estimates. Various simulation exercises to study different aspects of the performance of disease mapping models have also been reported. For example, Lawson *et al.*[12] compared a number of disease mapping models using various goodness of fit criteria, whereas Richardson *et al.*[13] report a comprehensive simulation study designed to highlight the amount of smoothing of the risks that is actually performed by various Bayesian disease mapping models and to assess their performance for detecting 'true' raised risk areas in a variety of set-ups.

In this article, we will first give a comprehensive review of the main classes of spatial priors that have been proposed in the disease mapping context (Section 2). Section 3 reports a performance comparison between representative models in these classes, using a set of simulated data to help illustrate their respective properties. The aim is to help the reader choose an appropriate prior and to discuss issues of sensitivity to this choice. We concentrate on studying the *structural sensitivity* to the specification of the model at the second level of the hierarchy, but we do not investigate issues of sensitivity to hyperprior specification at the top level, as this has already been addressed by a number of authors in the context of specific applications. Section 4 extends the set-up to include models for joint mapping of multiple disease or health indicators. This illustrates how the additional information processed in a joint analysis improves the estimation of the spatial patterns for each disease besides giving some additional information on their common structure. The article concludes with a discussion and suggestions for further work.

Note that we have limited our review to discussing disease mapping studies and do not discuss the closely related ecological regression studies where the endeavour is to understand some of the causes of the spatial patterns by relating the spatial variation of disease risk to that of risk factors. These studies pose different problems of interpretation, captured under the commonly used terminology of 'ecological bias', as one of their goals is to make inference at the individual rather than group level. Neither do we consider methods based on point level data, as these require individual geographic locations for both the cases and the population or control data, whereas it is more common for data to be routinely available in aggregated form at the small area level.

2 Spatial prior distributions for disease mapping models

The following generic three level hierarchical model has been discussed by a number of authors[11,14,15] as a natural model for disease mapping based on aggregation of the underlying individual level risks:

$$\begin{aligned} Y_i &\sim \text{Poisson}(e^{S_i} E_i), \quad i = 1, \ldots, n & (1)\\ S_i &\sim p(\cdot|\theta) & (2)\\ \theta &\sim \pi() \end{aligned}$$

where Y_i and E_i are the observed and expected number of cases of disease in area i, respectively, S_i is the log relative risk in area i, $p(\cdot|\theta)$ is an appropriate second stage prior distribution for the $\{S_i\}$ – specification of which forms the focus of this article – and θ are hyperparameters of this second stage model with hyperprior distributions $\pi()$. The expected counts are calculated as $E_i = \sum_j N_{ij} r_j$ where r_j is the disease rate for age–sex strata j in the reference population and N_{ij} is the population at risk in area i, strata j. This model is appropriate if the disease is rare, and if the following assumptions hold: 1) the individual level risks vary randomly *within* areas (i.e., are not spatially clustered), 2) the risk associated with living in area i acts *proportionally* on the baseline risks for each strata (i.e., the strata specific area risks r_{ij} simplify to $S_i \times r_j$). For nonrare diseases, a binomial model for the first stage distribution is more appropriate.[16] In this case, it is not possible to aggregate over age–sex groups as in the Poisson model, so separate models must be specified for each strata; the assumption of random within area variation in risk for each strata must still hold.

Other formulations for the underlying individual level (and hence aggregated) risk model are possible, but have been less widely used in epidemiology. For example, in some situations, an *additive risk* model may be more reasonable. Best *et al.*[17] propose a model that allows for a combination of additive and multiplicative effects, equivalent to assuming that the average rate of disease for individuals in age–sex stratum j living in area i is $\lambda_{ij} = \beta_j(r_0 + R_i)$. Here, r_0 is the baseline rate of disease in the study region and $R_i(>0)$ represents the average effect of additive adverse risk factors associated with area i. Age and sex are assumed to affect susceptibility to disease, and hence to act multiplicatively via the strata specific coefficient β_j, which represents the relative risk in strata j compared to the population weighted average risk in either the study region or some reference population. β_j may be estimated or treated as a known standardizing factor. In the latter case, $\beta_j = r_j / (\sum_j r_j N_j^* / \sum_j N_j^*)$ where r_j and N_j^* are, respectively, the disease rate and number of individuals in strata j in the reference population. Under the rare disease assumption, this leads to the following area level hierarchical model:

$$\begin{aligned} Y_i &\sim \text{Poisson}\big((r_0 + R_i) \times M_i\big), \quad i = 1, \ldots, n & (3)\\ R_i &\sim p(\cdot|\theta); \quad R_i > 0 & (4)\\ r_0, \theta &\sim \pi() \end{aligned}$$

where $M_i = \sum_j \beta_j N_{ij}$ and can be thought of as the 'standardized' population in area i (i.e., the actual population weighted by the strata specific relative risks).

The remainder of this section discusses the main classes of spatial prior distributions that have been proposed for the area specific risks $\{S_i\}$ or $\{R_i\}$. The majority of these priors have been developed and applied in the context of the multiplicative models (1) and (2), although in principle, most could also be used to model the excess risks in Equations (3) and (4) following appropriate transformation of the $\{R_i\}$.

2.1 Correlated normal priors

2.1.1 *Jointly specified models*

The multivariate normal distribution is one of the most flexible distributions for representing correlated random variables. Let $\mathbf{S} = \{S_1, \ldots, S_n\}$ denote the vector of area specific spatial random effects in Equation (2). We may specify the dependence structure in terms of an $n \times n$ covariance matrix $\mathbf{\Sigma}$, leading to a second stage prior

$$\mathbf{S} \sim \mathrm{MVN}(\boldsymbol{\mu}, \mathbf{\Sigma})$$

where $\Sigma = \sigma^2 \mathbf{\Omega}$ and Ω_{ij} is the correlation between S_i and S_j. Note that if $\Omega_{ij} = 0$ this implies that S_i and S_j are marginally independent. For reasons of parsimony, it is usual to specify the elements of the correlation matrix as a parametric function of the distance, d_{ij}, between the centroids of each pair of areas, that is, $\Omega_{ij} = f(d_{ij}; \phi)$. However, care is needed to ensure that the chosen function results in a positive definite covariance matrix $\mathbf{\Sigma}$, and there are surprisingly few parametric forms for which this is guaranteed.[18] One of the most widely used choices for $f(\cdot)$ is the exponential decay function $f(d_{ij}; \phi) = \exp(-\phi d_{ij})$ where $\phi > 0$ controls the rate of decrease of correlation with distance, with large values representing rapid decay. Another alternative that exhibits a fairly linear decrease with increasing distance is the disc model, where the correlation between two points is defined as proportional to the intersection area of two discs of common radius ϕ centered on those points.[19] Different shapes of decrease with distance can be modelled using two parameter functions, such as the power exponential family or the Matern class, and anisotropic forms are also available. However, in a disease mapping context, there is rarely much substantive knowledge to guide the choice of functional form, and often quite weak information in the data to estimate the parameters of the correlation function, particularly for more complex forms. Exploratory analysis using variograms can help guide the functional choice, and restricting consideration to the family of one parameter models is recommended unless there are strong substantive reasons for doing otherwise. As high long range correlation of the risks is difficult to distinguish from the effect of the overall mean, it is also important to ensure that the chosen correlation function (and associated hyperpriors) gives near zero correlation at distances within the extent of the study region, to avoid nonidentifiability of the mean and correlation parameters. There are few published examples of the application of such multivariate Gaussian models to disease mapping, although Cook and Pocock[20] and Richardson *et al.*[21] use a non-Bayesian formulation of this model for the area specific errors in ecological regression models and Wakefield *et al.*[14] discuss an application to mapping of cancer risk in the UK. Diggle *et al.*[22] discuss an application to mapping of environmental count data, albeit measured at point rather than areal locations. One important practical limitation of these models for disease mapping applications with even moderately sized study regions (i.e., comprising several hundred areas) is that implementation via Markov chain Monte Carlo (MCMC) algorithms is extremely computationally expensive due to inversion of the $n \times n$ covariance matrix at each iteration.

Kelsall and Wakefield[23] propose a related approach based on specifying a Gaussian random field (GRF) for the underlying *continuous* log relative risk surface at the second level of the hierarchical model. Integrating the GRF over areas, and making use of various approximations, they obtain a multivariate normal distribution for the area specific log relative risks with mean equal to the mean of the underlying GRF and correlation between areas i and j equal to the average correlation between two points randomly chosen from those areas. The approach is attractive in that correlation between areas is induced via a point level correlation function and it is possible to reconstruct the posterior distribution of the underlying continuous risk surface. However, the data will contain little information about spatial dependence at distances less than the size of the smallest areas, so features of the posterior risk surface at a small scale should not be over interpreted.

2.1.2 Conditionally specified (Markov random field) models

Gaussian Markov random fields (MRF) are the most commonly used second stage model in disease mapping, following pioneering papers by Clayton and colleagues[24,25] and Besag *et al.*[26] These models are usually specified by a series of conditional distributions which, in their most general form, may be written as

$$E(S_i|\mathbf{s}_{(-i)}) = \mu_i + \sum_j a_{ij}(s_j - \mu_j), \quad a_{ij} \equiv 0, \; i = 1, \ldots, n \tag{5}$$

$$\text{Var}(S_i|\mathbf{s}_{(-i)}) = \kappa_i > 0, \quad i = 1, \ldots, n \tag{6}$$

where $\mathbf{s}_{(-i)}$ denotes the values of the random effects in all areas except the ith area.[27,28] Such models are also known as Gaussian conditional autoregressions. The μ_i parameters represent 'large scale' spatial trend or gradient at location i, and are usually either assumed constant across locations or specified as a function of covariates. Without loss of generality, we assume $\mu_i = 0 \; \forall i$ here. The a_{ij} are coefficients reflecting local spatial dependence between units i and j. Subject to certain constraints on a_{ij} and κ_i (see below), it can be shown (e.g., Section 2.1 in Besag and Kooperberg[29]) that Equations (5) and (6) define a joint distribution $\mathbf{S} \sim \text{MVN}(\boldsymbol{\mu}, \mathbf{P}^{-1})$ where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)(=\mathbf{0})$, $P_{ii} = 1/\kappa_i$ and $P_{ij} = -a_{ij}/\kappa_i$. As $\mathbf{P}$ is the inverse covariance matrix it must be symmetric, leading to the constraint $a_{ij}\kappa_j = a_{ji}\kappa_i$. A widely used formulation that satisfies this symmetry condition (refer Equations R5 and R6 in Besag *et al.*[26]) is:

$$a_{ij} = \frac{\gamma w_{ij}}{w_{i+}}; \quad w_{i+} = \sum_j w_{ij} \tag{7}$$

$$\kappa_i = \frac{\sigma^2}{w_{i+}}. \tag{8}$$

Note that κ_i must be positive so care is needed if any w_{ij} are negative. w_{ij} is the ijth element of a symmetric $n \times n$ 'weight' matrix $\mathbf{W}$ with diagonal elements $w_{ii} \equiv 0$. A common choice is to set $w_{ij} = 1$ if locations i and j are neighbours and $w_{ij} = 0$ otherwise. γ can be thought of as an autocorrelation parameter that reflects the overall strength of spatial dependence between locations with nonzero weights. To define a proper joint distribution, $\mathbf{P}$ must also be positive definite, which requires that γ be constrained to lie in the range $(1/\lambda_{\text{min}}, 1/\lambda_{\text{max}})$, where λ_{min} and λ_{max} are the minimum and maximum eigenvalues of diag$(1/w_{i+})\mathbf{W}$ (refer p. 471 in Cressie[28]). This turns out to give $\gamma \in (-1, 1)$ irrespective of the neighbourhood structure when parameterization models (7) and (8) are used. This parameterization also yields an intuitive interpretation for the conditional mean $E(S_i|\mathbf{s}_{(-i)})$ as a weighted average of the random effects, s_j, for all locations with non zero w_{ij}, but forces the conditional variance $\text{Var}(S_i|\mathbf{s}_{(-i)})$ to be nonconstant across areas. Stern and Cressie[7] use an alternative parameterization that leads to an invariant conditional variance, but replaces the conditional mean by a weighted *sum*, which seems inappropriate when areas have different numbers of neighbours. The values of λ_{min} and λ_{max} also depend on the neighbourhood structure under Stern and Cressie's parameterization so γ will be defined on a different range for different models.

An appealing feature of the MRF prior (5)–(8) is the possibility to make inference about the overall degree of spatial dependence in the disease risks by estimating γ. However, interpretation of γ is not straightforward (also refer Sun *et al.*[30] for an interpretation of γ as a spatial shrinkage factor), and values close to the maximum $(1/\lambda_{\text{max}})$ are needed to reflect even moderate spatial dependence.[31] On the other hand, if $\gamma = 0$, this indicates independence between areas, but in this case, the MRF prior parameterized using Equations (7) and (8) does not reduce to the usual independent hierarchical normal prior for the log relative risks because the variance is not constant across areas. This led Besag *et al.*[26] to propose the following alternative second stage prior model (henceforth denoted BYM)

$$S_i = V_i + U_i, \quad i = 1, \ldots, n \tag{9}$$

$$V_i \sim \text{Normal}\,(0, \sigma_v^2) \tag{10}$$

$$U_i|\mathbf{u}_{(-i)} \sim \text{Normal}\left(\frac{\sum_j w_{ij}u_j}{w_{i+}}, \frac{\sigma_u^2}{w_{i+}}\right) \tag{11}$$

The $\{U_i\}$ follow an *intrinsic* autoregression, obtained by setting γ to its (upper) limiting value of 1, and can be thought of as representing the spatial component of between area variation in disease risk. Although the univariate conditional prior distributions (11) are well defined, the corresponding joint prior distribution for $\mathbf{U}$ is now improper (undefined mean and infinite variance). A proper posterior distribution will nevertheless be obtained, subject to the usual requirement for proper hyperprior distributions on variance components in hierarchical formulations.[32] The $\{V_i\}$ represent the geographically unstructured component of heterogeneity in disease risk. Posterior inference about the degree of spatial dependence is then based on the proportion of the total (marginal) variation in the $\{S_i\}$ captured by each component (where the marginal variance of the U_i's is estimated empirically at each MCMC iteration[25]).

As explained earlier, the MRF models (5) and (6) are equivalent to specifying a multivariate normal model for the joint distribution of the area specific random effects, but parameterizing the dependence structure in terms of the *precision* matrix, $\mathbf{P}$ (representing conditional independence assumptions), rather than the covariance matrix $\Sigma = \mathbf{P}^{-1}$ (representing marginal independence assumptions). Some care is needed when choosing and interpreting $\mathbf{P}$, as conditional independence $(P_{ij} = 0) \not\Rightarrow$ marginal independence $(\Sigma_{ij} = 0)$ and vice versa. However, there are important computational advantages to modelling the precision matrix, as the conditional independence assumptions are readily exploited by MCMC algorithms so that MRF models can be efficiently implemented without the need for matrix inversion. MacNab[3] proposed a related multivariate normal model that is also parameterized in terms of the precision matrix. MacNab assumes that $\mathbf{S} \sim \text{MVN}(\mathbf{0}, \sigma^2\mathbf{D}^{-1})$ with $\mathbf{D} = \rho\mathbf{P} + (1-\rho)\mathbf{I}$, where $\mathbf{P}$ is defined as for the MRF model, $\mathbf{I}$ is the $n \times n$ identity matrix and $\rho \in [0, 1]$ can be interpreted as a measure of spatial dependence in the sense that if $\rho = 0$ or 1, the model reduces to the Gaussian independence prior (10) or the intrinsic autoregression (11), respectively. This model may thus be seen as an alternative to the BYM model, but with the advantage that it avoids the potential identifiability problem encountered with the BYM prior, whereby only the sum of $U_i + V_i$ is well identified by the data.[33] MacNab reports broadly similar results in a comparison of the two formulations when applied to modelling small area health service outcome and utilization rates, although her model was less sensitive to the hyperpriors and led to a slightly better fit in her particular application.

Another extension of the BYM model has been suggested by Lawson and Clarke.[34] Their model includes a mixture of Gaussian and non-Gaussian (median based) conditional autoregressive components, with the latter designed to pick up discrete jumps in the relative risk structure in an attempt to avoid over smoothing of the risk surface.

2.2 Semi-parametric spatial models

With any of the parametric specifications described earlier, the amount of smoothing performed (e.g., controlled by the parameters σ_u^2 and σ_v^2 in the BYM model) is affected globally by all the areas and is not *adaptive*. Concerns that such parametric models could oversmooth the relative risk surface have led several authors to develop semi-parametric spatial models, which replace the continuously varying spatial distribution for $\{S_i\}$ by discrete allocation or partition models with each cluster or component having a constant unknown relative risk. The number of clusters or components is unknown and treated as a variable to be estimated. The common features of such models are that they allow discontinuities in the risk surface and make fewer distributional assumptions, while effecting nevertheless a necessary amount of smoothing and borrowing of strength by allowing areas to be allocated to the same cluster. Estimates of $\{S_i\}$ are obtained by averaging over a large number of cluster configurations weighted by the corresponding posterior probabilities. The models differ in the way they allocate areas to clusters or mixture components.

This class of models has, on the face of it, the potential to over fit the data by creating a limitless number of unnecessary clusters or components and one might think it necessary to include explicit penalization of the model dimension. In fact, as argued by Denison *et al.*,[35] the Bayesian framework with sensible choice of hyperpriors for the model parameters will automatically favour a simpler model over a more complex one if the data is supported by the former, because the prior mass will be more concentrated around the observed data for the lower dimensional prior and hence, the simpler model will have higher posterior probability. This automatic parsimony is sometimes referred to as Occam's razor.

2.2.1 *Mixture model with spatially dependent allocations*

Green and Richardson[36] propose a mixture model for the S_i values where the allocation of each area to a risk category follows a spatially correlated process. This model can be seen as an extension of hidden Markov models to hidden discrete state MRF, that is, MRF degraded by (conditionally) independent noise. The chosen allocation model is the Potts model, frequently used in image processing, which involves an interaction parameter that controls the degree of spatial dependence. The number of 'states' or components of the mixture is not predefined and is estimated as part of the model. The model is specified as follows:

$$\begin{aligned}
\exp(S_i) &= \eta_{z_i}, \quad i = 1, \ldots, n \\
z_i \text{ (allocation variable)} &\in \{1, \ldots, k\} \quad \text{(chosen according to the Potts model)} \\
p(z|\psi, k) &= \mathrm{e}^{\psi U(\mathbf{z}) - \delta_k(\psi)} \quad \text{(Potts model)} \\
\eta_j &\sim \text{Gamma}(\alpha, \beta), \quad j = 1, \ldots, k \\
k \text{ (number of components)} &\sim \text{Unif}(1, c_{\max})
\end{aligned}$$

where $\psi > 0$ is the interaction parameter to be estimated and $U(\mathbf{z}) = \sum_{i \sim i'} I[z_i = z_{i'}]$ is the number of like labelled pairs of neighbouring areas. Thus the allocation of an area i to a component j will be favoured by having neighbouring areas in the same component j, the more so the larger the value of ψ. This allocation mechanism not only captures prior beliefs about spatial similarity of risks in nearby areas, but also allows noncontiguous areas to belong to the same component. The last term in the Potts model, $\delta_k(\psi) = \log\left(\sum_{\mathbf{z} \in \{1,2,\ldots,k\}^n} \mathrm{e}^{\psi U(\mathbf{z})}\right)$, is the normalizing constant, where the sum is overall possible configurations of the allocation for the n areas.

The number of components in the mixture, k, is uncertain and so implementation of this model uses a reversible jump MCMC (RJMCMC) algorithm.[37] The normalizing constant needs to be evaluated for each particular neighbourhood structure and this is done off-line using a grid of values for ψ. Note that the model is parsimonious in that the parameters η_j are associated with each component and not with each distinct geographic cluster as is the case for the partition models discussed in Section 2.2.2. Nevertheless, clusters of contiguous areas sharing the same value of η_j are created by this model. The choice of value, $c_{\max}$, for the upper bound of the prior on k should, therefore, reflect the anticipated variability in risk – typically, up to about 10 components should be sufficient to describe the different levels of risk across the study region, although if the disease risk is particularly variable, a slightly higher value for $c_{\max}$ may be required.

2.2.2 Spatial partition models

A class of closely related semi-parametric models are the spatial partition models introduced by Knorr-Held and Raßer[38] and Denison and Holmes.[39] As in any partition model, it is assumed that there is a set of k nonoverlapping clusters of areas, each with constant relative risk, and k is treated as unknown. These models differ technically in the way the clusters are defined and in their hyperprior specification but their spirit is very close. We have chosen to use Knorr-Held and Raßer's (henceforth KHR) model in the comparisons reported in Section 3 and thus shall give fuller details of its specification here.

In KHR's model, k of the areas are chosen at random as so-called 'cluster centres' $g_j, 1 \leq j \leq k$. Conditional on these, the remaining areas are allocated to cluster j if that cluster centre is closer than any other in terms of the minimal number of area boundaries that have to be crossed to reach it. Conditional on k, each configuration of centres is assumed equally likely a priori. This leads to the following model:

$$\begin{aligned}
S_i &= \eta_{z_i}, \quad i = 1, \ldots, n \\
z_i &\in \{1, \ldots, k\} \quad \text{(chosen according to allocation mechanism} \\
&\qquad\qquad\qquad\quad \text{described earlier)} \\
\log \eta_j &\sim \text{Normal}(\mu, \sigma^2), \quad j = 1, \ldots, k \\
k \text{ (number of clusters)} &\sim \text{Unif}(1, c_{\max}) \text{ or geometric}
\end{aligned}$$

Note that all areas in a cluster are contiguous, whereas in the spatial mixture model described earlier, areas in the same component are not necessarily contiguous. Hence, the number of clusters, k, in KHR's model will tend to be much higher than the number of components in the mixture model. In general, one might want to set $c_{\max}$ equal to the number of areas to allow for the possibility that each area forms its own cluster (corresponding to independent risks in each area). Again, as k is uncertain, an RJMCMC algorithm is required to implement the KHR model.

To allocate areas to a cluster, Denison and Holmes make use of the idea of Voronoi tessellations as a clustering device. Each area is treated as a point location and represented using the co-ordinates of its centroid. The Voronoi tessellation generating points are assumed to be located at any point in the region of interest. Conditional on a set of k generating points (analogous to the g_j in KHR's model), the jth partition element is then composed of areas with centroids closer (in Euclidean distance) to the jth generating point than to any other generating point.

Note that the cluster construction adopted by KHR can be seen as a modification of Voronoi tesselations model for discrete irregular space, and hence is arguably more appropriate for small area data. Another difference between the two partition models is that Denison and Holmes use gamma priors for the η_j instead of lognormal distributions to exploit the Poisson-gamma conjugacy for an efficient implementation of their model. Finally, we remark that these partition models do not retain a Markovian structure for the $\{z_i\}$ in contrast to the spatial mixture model.

A related spatial partition model is that developed by Gangnon and Clayton.[40] This model also groups areas into clusters, but its primary focus is not the flexible modelling of the risk surface but that of making inference about the number, location and composition of clusters. These are modelled as being superimposed on a heterogeneous background risk surface using a specific class of priors for their shape.

2.3 Spatial moving average models

Spatial moving average models are a flexible class of models that have been used to describe continuous spatial processes, particularly in geostatistical applications [e.g., refer the special issue of *Statistical Modelling* (2002) on this topic]. Such models are constructed by integrating a simple two dimensional random noise process (e.g., a grid of iid Gaussian random variables) with a smoothing kernel that is a function of distance and, possibly, location. The kernel can be thought of as a device to 'smear out' the random noise process in two dimensional space to give a smooth surface. Note that any stationary Gaussian process can be expressed as a convolution of iid Gaussian noise and an appropriate kernel. The advantage of the spatial moving average formulation over direct modelling of the covariance function (Section 2.1.1) is that a rich class of kernel functions can be considered for modelling specific features (e.g., nonstationarity, edge effects) of the spatial dependence structure,[41] while still preserving the propriety of the underlying covariance function. Computation is also more efficient as inversion of large matrices is not required. Furthermore, non-Gaussian processes can be specified using spatial moving averages via appropriate distributional assumptions for the noise process.

Spatial moving average models have been developed primarily for continuous processes, and so have not been widely used in a disease mapping context. However, Best *et al.*[17,42] proposed a discrete version of a gamma moving average process to model geographical variations in childhood respiratory illnesses. Their model is based on the additive risk models (3) and (4), where the area specific random effects R_i represent unmeasured spatially varying *excess* risks. The second stage model for each R_i is constructed by specifying an arbitrary grid of latent iid gamma random variables γ_j ($j = 1, \ldots, m$ where m is the total number of grid cells defining the latent process) covering the study region. These are then convolved with a kernel matrix whose elements, k_{ij}, represent the relative contribution of the latent variable in grid cell j to the random effect (area specific excess risk) in area i. Best and colleagues assume an isotropic, stationary Gaussian kernel function, although other kernel forms are easily accommodated. Formally, the second stage model is as follows:

$$R_i = \sigma \sum_{j=1}^{m} k_{ij}\gamma_j, \quad i = 1, \ldots, n \tag{12}$$

$$\gamma_j \sim \text{Gamma}(\alpha_j, \tau), \quad j = 1, \ldots, m$$

$$k_{ij} = \frac{1}{2\pi\phi^2} \mathrm{e}^{-d_{ij}^2/2\phi^2}$$

where σ can be thought of as a scale factor for the spatial random effects, d_{ij} is the distance between the centroid of area i and the centroid of latent grid cell j and ϕ is the spatial range parameter governing how rapidly the influence of the latent gamma random variables on the area specific excess risk declines with distance. An appealing interpretation of this model is to view the gamma random variables as representing the location and magnitude of unmeasured risk factors, and the area specific random effects as representing the cumulative effect of these risk factors in each area, weighted by their distance from the area according to the kernel 'weights' k_{ij}. However, this model does necessitate more 'fine tuning' than, say, the semi-parametric or BYM models, in that the number and size of the latent grid cells must be specified in advance by the user, and it is not always clear how best to choose these (the continuous space version of this model[43] avoids this problem). Note that if the latent gamma random variables are defined on the same areal partition as the disease outcome data, and the kernel is chosen to be the indicator function ($k_{ij} = 1$ if $i = j$ and 0 otherwise) then this model reduces to assuming independent conjugate gamma priors for the relative risk in each area. Estimation is via MCMC using a data augmentation scheme to exploit the Poisson-gamma conjugacy of the full conditionals for the γ_j parameters.

Hyperprior specification is discussed in detail by Best *et al.*[17] Note, however, that the prior shape (α_j) and precision (τ) parameters of the latent gamma variables should be chosen such that γ_j has prior mean proportional to the area of the *j*th latent grid cell (hence the subscript for α_j to accommodate nonregular grids). This makes the model spatially extensible in the sense that any partition of the latent gamma random variables will lead to identical probability distributions for the kernel weighted sums in Equation (12). Because the unobserved risk factors are typically defined on a smaller geographic partition than the disease outcome data, maps of the posterior risk surface can be constructed for other geographical partition, simply by evaluating the kernel sum (12) at these required locations. A potential limitation, however, is that the model relies on the assumption of additive risks which may not always hold in practice, and it does not guarantee that the Poisson rate in each area, ($r_0 + R_i$), will be estimated to be less than 1 (although this is unlikely to be a problem provided the disease is rare).

3 Comparison of model performance

We now present an illustrative comparison of some of the spatial priors discussed in Section 2 using a small simulation study. Data were simulated for a region in northern England covering the town of Huddersfield and surrounding rural areas to the south and west. The region was divided into 170 census enumeration districts (ED) with 1991 census population counts ranging from 129 to 746 (median 475). The true risk surface was based on assuming that the disease rate was positively associated with exposure to ambient levels of NO_2 (estimated concentrations of which were obtained from a study by Briggs *et al.*,[44] and tended to mimic the pattern of the major road network in the region) and with exposure to pollution from two hypothetical point source locations (one located in an area with high NO_2 concentrations, and one in a rural area with low NO_2 concentrations). The resulting risk surface was thus a mixture of areas with background rates, and linear and circular clusters of areas with elevated rates [Figure 1(a)], with true average relative risk per ED ranging from 0.8 to 3.0 (median = 0.9, 90th percentile = 1.3). Five replicate sets of disease counts were simulated by generating cases from binomial distributions with the appropriate disease rate located at each postcode within the study region and then summing over postcodes within each ED. Five different hierarchical models were then fitted to the simulated datasets, each assuming a Poisson likelihood for the first level and a different spatial prior distribution for the second level (Table 1). These spatial prior models were chosen on the grounds that they provide a representative selection of the distributions in each class discussed in Section 2 and because software is available to fit them. Models 1, 2 and 5 can be implemented in WinBUGS,[45] whereas user friendly Fortran and C code, respectively, is available for fitting models 3 (available from the authors on request) and 4 (available from www.stat.uni-muenchen.de/~rasser/bdcd). The datasets and the code for the models fitted in WinBUGS are available from the WinBUGS website (www.mrc-bsu.cam.ac.uk/bugs/examples.shtml).

A single (RJ)MCMC chain was run for each model and dataset for a total of 50 000–70 000 iterations, the first 20 000 of which were discarded as burn-in in each case. In models 2–5, each took only a few minutes to run on a 2.4 GHz pentium 4 PC (or equivalent), although precalculation of the normalizing constant for the MIX model involves MCMC integration for each value of ψ and can take some time (in our case, several hours) if a fine grid of values is chosen. The multivariate normal (EXP) model took ~3.5 h to run.

3.1 Model complexity and fit

The complexity and fit of the various models were compared using the deviance information criterion (DIC).[46] The DIC for each model is the sum of the posterior mean deviance, $\overline{D}$ (defined as minus twice the log likelihood, to reflect model fit), and an estimate of the 'effective' number of parameters, p_D. The latter can be thought of as a penalty term reflecting the model complexity or degrees of freedom. In classical nonhierarchical models, this concept is well defined, but in Bayesian hierarchical models, the shrinkage properties of the prior effectively restrict the freedom of the

Table 1 Spatial priors used for the area specific random effects in the simulation study

Spatial prior	Label	Priors on hyperparameters
1. Multivariate normal with exponential correlation	EXP	$\sigma^{-2} \sim$ Gamma(0.001, 0.001) $\phi \sim$ Unif(0.00005, 0.05) (distance measured in m)
2. Besag, York and Mollié's MRF model	BYM	$\sigma_u^{-2} \sim$ Gamma(0.05, 0.0005) $\sigma_v^{-2} \sim$ Gamma(0.05, 0.0005)
3. Green and Richardson's spatial mixture model	MIX	$\alpha = 1$; $\beta = \sum_i Y_i / \sum_i E_i$ $k \sim$ Unif(1, 10)
4. Knorr-Held and Raßer's partition model	KHR	$\mu = 0$; $\sigma^2 \sim$ Inv Gamma(1,0.01) $k \sim$ Unif(1, 170)
5. Gamma moving average model	GMA	$\phi = 2.5$ km; r_0, $\sigma \sim$ Gamma($0.575, 0.575 \times 2/\overline{Y}$) where $Y =$ overall disease rate (=0.012) $\gamma_j \sim$ Gamma($0.01 \times$ AREA$_j$, 0.01) where AREA$_j = 1.1$ km^2 = area of j^{th} latent cell, $j = 1, \ldots, 280$

Note: Gamma priors are parameterized in terms of the shape and inverse scale parameters.

model parameters. The p_D term proposed by Spiegelhalter *et al.* aims to capture the amount of shrinkage achieved by the hierarchical prior and (at least for approximately normal likelihoods) can be shown to be equal to the ratio of the information in the likelihood to the total information in the posterior (i.e., likelihood + prior). Thus a value of p_D that is small relative to the number of data points indicates that the prior structure provides a lot of information about the parameters and hence leads to considerable borrowing of strength, whereas a p_D close to the number of data points indicates that the hierarchical prior does not contribute much information and hence that there is little borrowing of strength across units. It is therefore of interest to compare the estimated p_D for each model, to gain some insight into the amount of relevant structural information provided by the second stage prior in each case.

We note that in models with negligible prior information, p_D will be approximately equal to the actual number of parameters and DIC will be approximately equivalent to Akaike's information criterion. Like AIC, DIC can be justified as an approximation to cross validation, and Spiegelhalter *et al.* also show that it can be interpreted approximately as the expected posterior loss in prediction when adopting a particular model. Use of this criterion to compare the performance of disease mapping models thus seems appropriate, as DIC can be thought of as selecting the model that leads to the best prediction of the risk surface (out of the models considered) in the study region of interest.

For each dataset, the DIC for the 'best' model (i.e., with the smallest DIC) is given in Table 2, together with the difference between this and the DIC for every other model.

Table 2 Posterior mean deviance ($\overline{D}$), effective number of parameters (p_D) and model comparison criterion (DIC) for each model and dataset

Datasets	Single disease models					Joint disease models	
	EXP	BYM	MIX	KHR	GMA	MVBYM	SHARE
				$\overline{D}$			
1	205.8	178.1	173.4	174.6	200.6	170.4	177.1
2	180.7	173.4	160.6	194.2	211.9	164.7	171.6
3	174.8	158.7	157.6	174.0	222.0	155.0	158.5
4	193.6	153.3	152.0	152.1	199.0	147.4	150.6
5	185.1	162.4	139.9	173.7	186.2	152.3	161.3
				p_D			
1	18.5	27.4	34.4	27.8	9.4	37.0	24.6
2	43.5	33.3	44.9	19.8	11.7	42.5	29.2
3	51.8	37.3	35.3	20.7	10.9	43.7	31.3
4	14.0	29.6	34.9	21.6	8.0	36.7	28.6
5	6.1	18.5	28.9	12.9	4.8	29.8	17.3
				DIC[a]			
1	+21.9	+3.1	+5.4	**202.4**	+7.6	+5.0	−**0.7**
2	+18.7	+1.2	**205.5**	+8.5	+18.1	+1.7	−**4.7**
3	+33.7	+3.1	**192.9**	+1.8	+40.0	+5.8	−**3.1**
4	+33.9	+9.2	+13.2	**173.7**	+33.3	+10.4	+5.5
5	+22.4	+12.1	**168.8**	+17.8	+22.2	+13.3	+9.8

[a] DIC for the 'best' single disease model for each dataset is shown in bold, with all other DIC values shown as differences from this value; where the DIC for a joint disease model is smaller than that for the 'best' single disease model, this difference is also highlighted in bold.

Spiegelhalter *et al.* suggest that models with DIC values within 1 or 2 of the 'best' model are also strongly supported, those with values between 3 and 7 of the 'best' are only weakly supported and models with a DIC more than 7 higher than the 'best' are substantially inferior. Table 2 also gives the separate contributions of fit ($\overline{D}$) and complexity (p_D) for each model and dataset. We first note that all models lead to considerable borrowing of strength, with between about 10 and 40 'effective' parameters needed to fit the 170 data points. The GMA model consistently has the smallest p_D, indicating that this prior contains a lot of structural information. However, DIC does not support the model due to the overall lack of fit: for Poisson likelihoods, $\overline{D}$ should have approximate sampling expectation $E(\chi^2_n) = n = 170$ if the model is true, whereas $\overline{D}$ is much higher than 170 for the GMA model. This is hardly surprising as the truth is a multiplicative risk model, whereas the GMA model (unlike the others considered here) assumes additive risks. The KHR model has the next smallest values of p_D, again indicating strong structural information in the prior. In this case, the overall model fit is also reasonable and DIC indicates that this is the best or close to best model out of those considered for three of the five datasets. However, for the other two datasets, KHR appears to have overshrunk as p_D is considerably lower than for most of the other models, and DIC indicates that it is substantially inferior to the best supported model. The MIX model behaves somewhat different, generally having the largest value of p_D (and hence the least borrowing of strength). However, this greater apparent complexity is compensated for by a good fit to the data ($\overline{D}$ is consistently smaller than for the other models, and indeed, shows a tendency for the MIX model to overfit the data in some cases, although it should be emphasised that the χ^2_n approximation is not always reliable if the data for each area are sparse). DIC shows that MIX is the best supported model for three datasets, is weakly supported for one dataset and is substantially inferior to KHR for the other. The BYM model is intermediate between MIX and KHR in terms of both fit and complexity, and receives some support according to DIC for most of the datasets, although is never the best supported model. The multivariate normal model (EXP) is substantially inferior to all other models except GMA. There is no consistent pattern to the borrowing of strength achieved by this model, but the lack of support according to DIC suggests that this prior does not provide appropriate information about the spatial structure.

3.2 Variance and bias

Figure 1 shows maps of the true risk surface and the average (across datasets) posterior mean relative risk for each model (where the relative risk for the GMA model is calculated as the ratio of the estimated rate for each area to the average rate for the study region). The BYM and KHR maps look quite similar, although the latter, in particular, is somewhat too smooth. The MIX map is the least spatially smooth (reflecting the higher p_D for this model), and is better at capturing some of the areas with elevated rate in the north and western central part of the study region, but it also tends to overpartition the background areas. These features are partly due to the MIX model's ability to borrow strength across noncontiguous areas. Both the GMA and EXP maps oversmooth the high risk areas in the north, suggesting that the prior spatial dependence structure imposed by these models (which is based on an isotropic

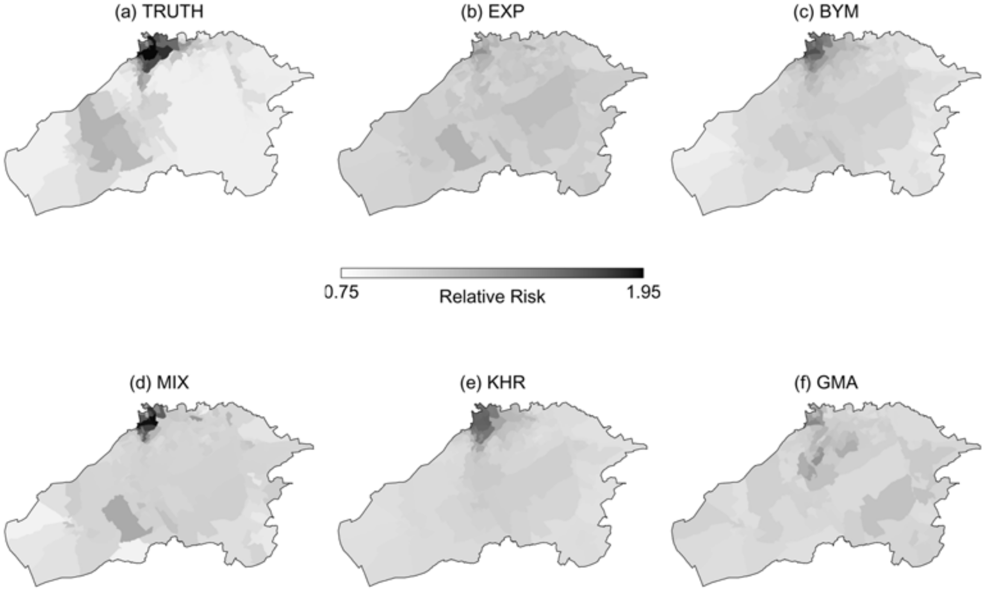

Figure 1 Maps of the true relative risk and the average (across datasets) posterior mean relative risk estimated by each model. Note that three areas in map (a) have true relative risk > 1.95, but are shaded as if they had relative risk $= 1.95$ to allow differences between the remaining areas to be more clearly distinguished by the shading scheme.

stationary correlation function with exponential decay in both cases) is inappropriate for these data.

Figure 2 shows the average percentage bias (difference between posterior mean relative risk or rate and the true value, as a percentage of the truth) across the five datasets for selected EDs under each model. Graph (a) shows the five EDs with the highest true relative risks (1.8–3.0), graph (b) shows a random subset of five EDs with moderate true relative risks (1.0–1.3) and graph (c) shows a random subset of five EDs with background true relative risks (all 0.8). For each model, the median (5 and 95% quantiles) percentage bias for all EDs in the relevant category is also given at the bottom of the graphs. All models yield similar amounts of bias for the background areas (except EXP, which has higher bias), although there is a wider spread of biases for the GMA model. For moderate risk areas, BYM and KHR produce the least biased estimates, indicating that the mechanism by which these models borrow strength performs well for such areas. MIX yields the highest bias here, reflecting the tendency for this model to produce multimodal posteriors when there is some uncertainty about whether an area should be allocated to a background risk component or a high risk component. In contrast, MIX produces the least biased estimates for the most extreme high risk areas, because in this case the allocation mechanism is able to clearly identify that these areas belong to a high risk component. The EXP and GMA models tend to yield the most biased estimates of high risk areas – yet further indication that these priors tend to over-shrink the extreme risks.

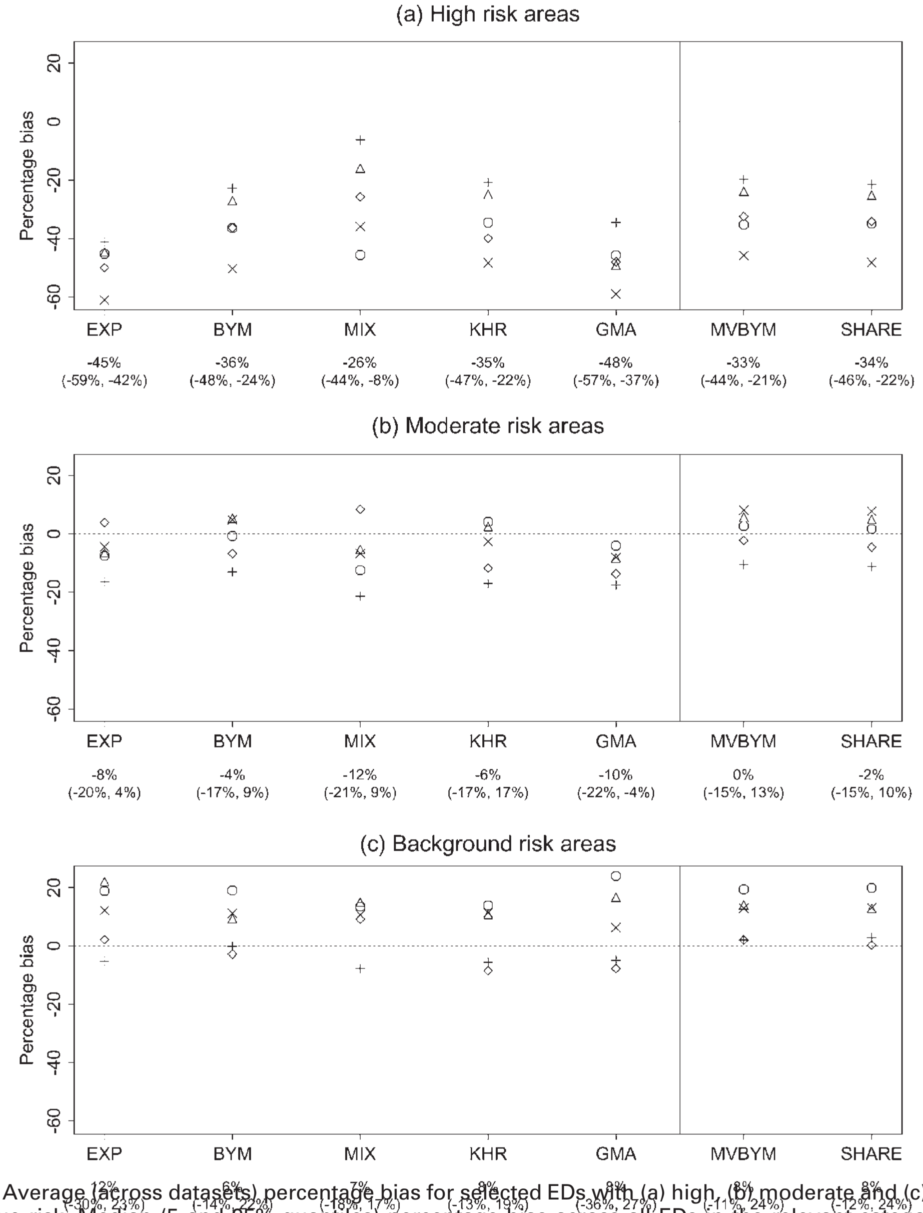

Figure 2 Average (across datasets) percentage bias for selected EDs with (a) high, (b) moderate and (c) background true risk. Median (5 and 95% quantiles) percentage bias across all EDs in the relevant category are shown at the bottom of each plot.

As noted earlier, shrinkage is a trade of between bias and variance reduction. The GMA model yielded posterior risk estimates with a coefficient of variation (CV) typically less than half that of the other models. However, we have already seen that the point estimates for this model have high bias, suggesting that in this case the bias–variance trade-off is too heavy in favour of variance reduction. This observation is consistent with the small p_D values for this model, indicating considerable borrowing of strength (hence the increased precision). BYM and EXP have similar CVs, both slightly lower than KHR (except for the high risk areas where KHR is more precise). The MIX model risk estimates have the highest posterior CV (around twice that of BYM). Again, this may partly reflect the skewness or slight multimodality of the posterior distributions for some background and moderate risk areas due to uncertainty about which mixture component the area should be allocated to.

3.3 Exceedence probabilities

Several authors recommend mapping or reporting the posterior probability that the relative risk or rate in each area exceeds a specified threshold of interest.[25] Richardson *et al.*[13] extend this idea and consider various decision rules for classifying whether area i has an increased risk based on how much of the posterior distribution of the relative risk parameter $\exp(S_i)$ exceeds a reference threshold. Figure 3 shows the proportion of EDs classified as having increased risk according to the decision rule that at least 75% of the posterior distribution of $\exp(S_i) > 1$, plotted against the true relative risk. (Note that for the GMA models, the decision rule was modified to require that at least 75% of the posterior distribution of the area specific rate $(r_0 + R_i)$ exceeds the average rate for the study region.) Results for BYM and KHR are very similar: both models perform reasonably well, classifying over 50% (80%) of areas with true relative risk >1.4 (>2.0) as being at increased risk. In contrast, EXP fails to pick up the majority of areas with true increased risk, again reflecting the tendency for this model to overshrink the

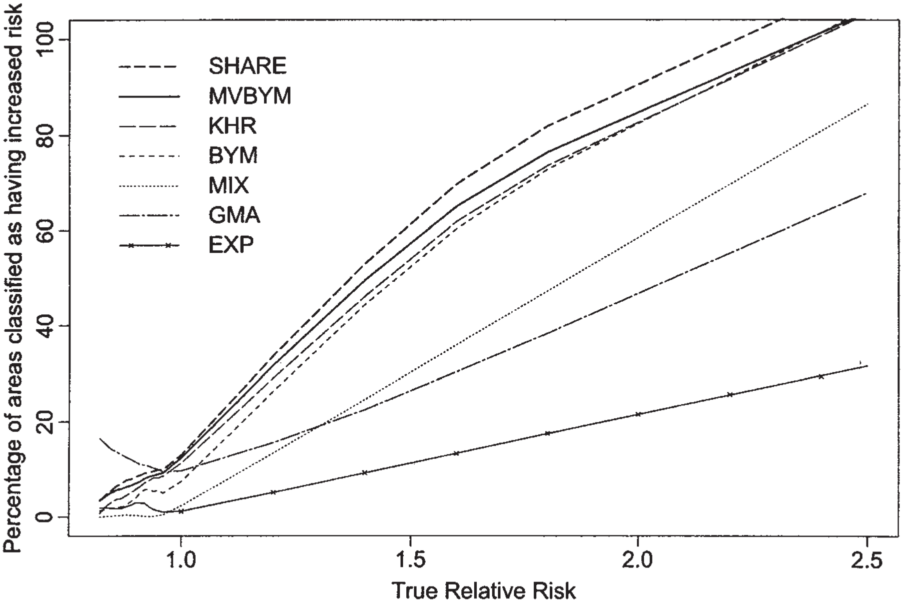

Figure 3 Proportion of EDs in all five datasets classified as having increased risk (i.e., at least 75% of the posterior distribution for the relative risk exceeds 1). Lines have been smoothed by grouping areas with similar true risks and applying a scatterplot smoother.

extreme risks. MIX and GMA perform intermediately: neither are very good at detecting areas with moderately raised risk, but the MIX model, in particular, does reasonably well at detecting the most extreme risk areas. However, Richardson *et al.*[13] showed that, while the decision rule used in Figure 3 was near optimal for the BYM model, a different decision rule such as at least 5% of the posterior distribution of $\exp(S_i) > 1.5$ was better able to discriminate between background and increased risk areas for the MIX model, and so the comparison presented in Figure 3 is not entirely fair. Other decision rules may also be more appropriate for some of the other models, but it is beyond the scope of the present article to explore this issue in detail.

4 Models for joint mapping of multiple diseases

Many diseases share common risk factors (smoking being an obvious example). Hence there may be advantages to carrying out a joint mapping analysis of two or more related diseases in order to borrow strength *across diseases* as well across nearby areas. Two alternative formulations for the joint spatial analysis of multiple diseases have recently been proposed and are discussed subsequently. For simplicity, we consider the case of two diseases, $d = 1, 2$, so that the first level of the model becomes

$$\begin{aligned} Y_{1i} &\sim \text{Poisson}(E_{1i}\mathrm{e}^{S_{1i}}) \\ Y_{2i} &\sim \text{Poisson}(E_{2i}\mathrm{e}^{S_{2i}}) \end{aligned}$$

4.1 Multivariate MRF models

The MRF models discussed in Section 2.1.2 extend naturally to a multivariate setting[47] by replacing the univariate Gaussian conditional distribution for $S_i|\mathbf{s}_{(-i)}$ with a multivariate conditional distribution for the vector of log relative risks, $\mathbf{S}_i = (S_{1i}, S_{2i})'$, in area i:

$$\mathbf{S}_i|\mathbf{s}_{1(-i)}, \mathbf{s}_{2(-i)} \sim \text{MVN}\left(\sum_j \mathbf{A}_{ij}\mathbf{s}_j, \mathbf{K}_i\right)$$

Here, $\mathbf{A}_{ij} = (a_{ij1}, a_{ij2})' \, \mathbf{I}_{2\times 2}$ (with a_{ijd} defined as in Equation (7) for $d = 1, 2$) and $\mathbf{K}_i = w_{i+}^{-1}\boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma}$ is 2×2 with diagonal elements σ_1^2 and σ_2^2 and off diagonal element ρ representing the (conditional) correlation between the two diseases within an area. As before, we assume that the multivariate equivalents of the μ_i parameters in Equation (5) are zero. Setting $\gamma = 1$ in the specification of the a_{ij}s yields a multivariate version of the intrinsic autoregressive prior (11), which has been investigated by Gamerman *et al.*[48] as a prior for correlated spatially varying regression coefficients. Including an additional vector of unstructured random effects $\mathbf{V}_i = (V_{1i}, V_{2i})' \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_v)$, with $\boldsymbol{\Sigma}_v$ a diagonal matrix, in the second level model for the area specific log relative risks gives the multivariate equivalent of the BYM model.

Gelfand and Vounatsou[49] generalize the proper version of the multivariate MRF model to allow for different autocorrelation parameters, γ_d, for each disease d. This is nontrivial due to the requirement that the elements of $\mathbf{A}_{ij}$ and $\mathbf{K}_i$ satisfy the conditions for symmetry and nonsingularity of the underlying precision matrix. Carlin and Banerjee[50] apply this model to the spatial analysis of survival rates for various types of cancer.

4.2 Shared component model

Knorr-Held and Best[51] proposed a joint formulation that aims to identify shared and disease specific spatially varying patterns of risk by appropriate partitioning of the underlying risk surface for each disease as follows:

$$S_{1i} = \delta\eta_i + \psi_{1i}$$
$$S_{2i} = \frac{\eta_i}{\delta} + \psi_{2i}$$

here η_i captures the effects of unmeasured risk factors that are shared by both diseases (weighted by the scaling parameter δ to allow a different 'risk gradient' to be associated with these shared factors for each disease), whereas ψ_{di} represent unmeasured risk factors specific to one or other of the diseases, $d = 1$ or 2, only. The three components, $\boldsymbol{\eta}$, $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ are assumed to be independent, with each one following a spatial prior distribution. In principle, any suitable spatial distribution could be used: Knorr-Held and Best use the spatial partition model of Knorr-Held and Raßer[38] (Section 2.2.2), whereas in the simulation study described in Section 4.3, we use the BYM prior for each component. Hence the shared component model could be seen as a more flexible model than the multivariate MRF approach, at least for the joint modelling of two diseases. Extension to more than two diseases follows straightforwardly for the multivariate MRF, and is also possible for the shared component model (see Held *et al.*,[52] this volume). Spatial versions of Bayesian factor analysis models may prove to be a useful line of research here.

Another attractive feature of the shared component approach is the ability to separately estimate and map the shared and disease specific components of the risk surface. In contrast, the multivariate MRF approach provides a numeric estimate of the correlation between the overall log relative risks for each disease, but only enables the *overall* area specific relative risk for each disease to be mapped.

4.3 Model comparison

We briefly describe an extension of the simulation study of Section 3 to compare the performance of the multivariate BYM and shared component models with the single disease models already considered. Counts of cases of a second disease were simulated using same procedure as earlier. This disease was assumed to share one of the risk factors for the first disease (NO_2), but with a less strong risk gradient, and was unaffected by the two point source risk factors. Risk of the second disease was also assumed to depend on an additional spatially random risk factor. The correlation between the true ED level relative risks for the two diseases was 0.67. The multivariate

BYM (MVBYM) and shared component (SHARE) models were fitted to the two diseases using the WinBUGS software. Here we focus on results only for the first disease, to enable comparison with the results of the single disease models reported in Section 3.

As might be expected, both the MVBYM and the SHARE models perform quite similarly to their univariate BYM counterpart. For example, maps of the posterior mean relative risks (not shown) look virtually identical to the BYM map [Figure 1(c)], but with some small improvements in terms of reduced bias for the moderate and high risk areas (Figure 2). The SHARE model also has a greater ability to detect areas with a true increased risk (Figure 3). Somewhat surprisingly, the posterior distributions of the risk estimates from both joint models have similar or slightly higher CVs to the single disease BYM model, whereas we might have expected the additional borrowing of strength across the two diseases to result in increased precision. In terms of overall model comparison, the DIC values (shown in Table 2 as differences from the DIC for the best single disease model) indicate that the SHARE model is actually the best supported model for three datasets and has the second smallest DIC for the other two datasets. However, the MVBYM is substantially inferior to the best supported model (SHARE, MIX or KHR) for four datasets and has only weak support for the fifth. The lack of support for the MVBYM is largely due to its greater effective complexity (it has a p_D similar or greater than MIX) which is not compensated for by a better fit. In contrast, the SHARE model, although apparently more complex in terms of the number of random effect parameters (six per area), appears to borrow strength in a more appropriate way and so has fewer effective parameters.

5 Discussion

In this article, we have attempted to offer some technical and practical insights into important features of the main classes of Bayesian hierarchical spatial models available for small area disease mapping. The results of our simulation study suggest that the BYM model and the two semi-parametric models (MIX and KHR), all have good properties for modelling a single disease. A particular feature of the MIX model is its ability to borrow strength across both the local neighbourhood and noncontiguous groups of areas, and this seems to result in less biased point estimates of risk than the other models for the high risk areas. The BYM and KHR models have a tendency to oversmooth the point estimates, but are somewhat better than the MIX model at overall classification of areas into increased or background risk groups (this finding appears to hold even when a more appropriate decision rule is used for the MIX model than that shown in Figure 3[13]).

Representing spatial dependence in disease risk by means of a multivariate normal prior with parametric covariance matrix (e.g., EXP model), while intuitively appealing, does not perform well in practice. Our results show that such models heavily oversmooth the extreme risk estimates and lead to poor inference compared with the other models considered. Results (not shown) of using an alternative correlation structure based on the ‘disc’ function (Section 2.1.1) were very similar to those reported for the EXP model in the simulation study. These models are also computationally slow and cumbersome to implement, and so are not recommended for general use in a disease mapping context.

Performance of the GMA model was also inferior to the BYM, MIX and KHR models for the simulated datasets in Section 3. However, these data were generated under a multiplicative model, whereas the GMA model (unlike all the other models considered) assumes an additive form for the area specific risks. The lack of fit demonstrated by the high values of the posterior mean deviance for this model is therefore not surprising, and we might expect it to perform better in situations where the underlying risks are truly additive. Furthermore, the geographical partition used for the latent gamma random variables and the size of the kernel range parameter were chosen somewhat arbitrarily, and other choices may have yielded better results. In principle, the range parameter could be treated as uncertain and estimated as part of the model (although this is currently very slow to implement using WinBUGS). However, the greater need for fine tuning of the GMA model specification compared with most of the other models considered here count against it as a tool of choice for standard disease mapping. Rather, the particular strength of the GMA model is in the extra flexibility it offers for modelling data on disparate spatial scales, and so it is likely to prove more useful in ecological regression contexts where the latent gamma random variables could be modelled as a function of observed (as well as unobserved) risk factors.

One interesting feature to note is that the mechanism for representing spatial structure in each of the BYM, MIX and KHR models is based on the neighbourhood or adjacency matrix of the study region, which seems to offer more flexibility for capturing appropriate features of the risk surface than the parametric functions of distance used by the multivariate normal and moving average models. Unlike the BYM, MIX and KHR models, the latter two models also assume stationarity of the mean and variance, which may be too restrictive in many disease mapping applications. However, as noted earlier, moving average models are potentially a very flexible class of models, and further work to extend the GMA and other related models to allow nonstationary and anisotropic kernel functions would be a useful area of research.

If data on multiple diseases or health outcomes that share similar geographical patterns are available, then joint modelling of the risk surfaces can lead to improved inference over separate analysis of each outcome. Indeed, even if one outcome is of primary interest, it could still be advantageous to include data on other relevant outcomes in the analysis if these are readily obtainable. In the case of two outcomes, the shared component model appears to perform slightly better than the multivariate BYM model. This may reflect the fact that the MVBYM model forces the correlation between the two outcomes to be constant for all areas, which may be too restrictive, whereas the shared component model allows the shared fraction of risk to vary across the study region. For more than two diseases, the MVBYM (or other versions of the multivariate MRF model) is the obvious choice, although we are currently investigating extensions of the shared component model to handle this situation.

We finish on a practical note by commenting that the importance of using statistical methods for borrowing strength in small area disease mapping studies is now widely recognized in the epidemiological community. However, a (nonexhaustive) search of the major epidemiological journals over the past decade indicates that the BYM model appears to be the only fully Bayesian spatial model to have been used in published applications of disease mapping outside of the statistical literature. The comparisons reported here, and by others elsewhere, suggest that the BYM model remains an appropriate tool for small area disease mapping. However other recently proposed Bayesian spatial models have the potential to offer some improvements in certain applications, and so sensitivity to structural assumptions as well as hyperprior specification should be explored as part of any disease mapping study.

Acknowledgements

The UK Small Area Health Statistics Unit is funded by the Department of Health, Department of the Environment, Food and Rural Affairs, Environment Agency, Health and Safety Executive, Scottish Executive, National Assembly for Wales and the Northern Ireland Assembly. NGB is partly funded by an MRC Career Development Grant G9803841. The authors would like to thank Leo Knorr-Held and Peter Green for providing computer code for the KHR and MIX models, respectively, and Andrew Thomas and Dave Lunn for their work in implementing spatial models in WinBUGS.

References

1 Jarup L, Best N, Toledano M, Wakefield J, Elliot P. Geographical epidemiology of prostate cancer in Great Britain. *International Journal of Cancer* 2002; **97**: 695–99.

2 Buntinx F, Geys H, Lousbergh D, Broeders G, Cloes E, Dhollander D, Op De Beeck L, Vanden Brande J, Van Waes A, Molenberghs G. Geographical differences in cancer incidence in the belgian province of limburg. *European Journal of Cancer* 2003; **39**: 2058–72.

3 MacNab YC. Hierarchical Bayesian modeling of spatially correlated health service outcome and utilization rates. *Biometrics* 2003; **59**: 305–16.

4 Doll R. The epidemiology of cancer. *Cancer* 1980; **45**: 2475–85.

5 Carlin BP, Louis TA eds. *Bayes and empirical Bayes methods for data analysis*. New York, NY, USA: Chapman & Hall/CRC, 2000.

6 Conlon EM, Louis TA. Addressing multiple goals in evaluating region-specific risk using Bayesian methods. In Lawson A, Biggeri A, Böhning D, Lesaffre E, Viel J-F, Bertollini R eds. *Disease mapping and risk assessment for public health*. Chichester, UK: John Wiley & Sons, 1999: 31–47.

7 Stern H, Cressie N. Inference for extremes in disease mapping. In Lawson A, Biggeri A, Böhning D, Lesaffre E, Viel J-F, Bertollini R eds. *Disease mapping and risk assessment for public health*. Chichester, UK: John Wiley & Sons, 1999: 63–84.

8 Marshall RJ. A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 1991; **154**: 421–41.

9 Bithell JF. A classification of disease mapping methods. *Statistics in Medicine* 2000; **19**: 2203–15.

10 Lawson A, Böhning D, Biggeri A, Lesaffre E, Viel J-F. Disease mapping and its uses. In Lawson A, Biggeri A, Böhning D, Lesaffre E, Viel J-F, Bertollini R eds. *Disease mapping and risk assesment for public health*. Chichester, UK: John Wiley & Sons, 1999: 3–13.

11 Richardson S. Spatial models in epidemiological applications. In Green PJ, Hjort NL, Richardson S eds. *Highly structured stochastic systems*. Oxford: Oxford University Press, 2003: 237–59.

12 Lawson A, Biggeri A, Böhning D, Lesaffre E, Clark A, Schlattmann P, Divino F. Disease mapping models: an empirical evaluation. *Statistics in Medicine* 2000; **19**: 2217–41.

13 Richardson S, Thomson A, Best NG, Elliott P. Interpreting posterior relative risk estimates in disease mapping studies. *Environmental Health Perspectives* 2004; **112**: 1016–25.

14 Wakefield JC, Best NG, Waller LA. Bayesian approaches to disease mapping. In Elliott P, Wakefield JC, Best NG, Briggs DJ eds. *Spatial epidemiology: methods and applications*. Oxford: Oxford University Press, 2000: 104–27.

15 Pascutto C, Wakefield JC, Best NG, Richardson S, Bernardinelli L, Staines A, Elliott P. Statistical issues in the analysis of disease mapping data. *Statistics in Medicine* 2000; **19**: 2493–519.

16 Knorr-Held L, Besag J. Modelling risk from a disease in time and space. *Statistics in Medicine* 1998; **17**: 2045–60.

17 Best NG, Ickstadt K, Wolpert RL, Briggs DJ. Combining models of health and exposure data: the SAVIAH study. In Elliott P, Wakefield JC, Best NG, Briggs DJ eds. *Spatial epidemiology: methods and applications*. Oxford: Oxford University Press, 2000: 393– 414.

18 Ripley BD ed. *Spatial statistics*. New York, NY, USA: John Wiley & Sons, 1981.

19 Richardson S. Statistical methods for geographical correlation studies. In Elliott P, Cuzick J, English D, Stern R eds. *Geographical and environmental epidemiology*. Oxford: Oxford University Press, 1992: 181–204.

20 Cook DG, Pocock SJ. Multiple regression in geographic mortality studies with allowance for spatially correlated errors. *Biometrics* 1983; **39**: 361–71.

21 Richardson S, Guihenneuc C, Lasserre V. Spatial linear models with autocorrelated error structure. *The Statistician* 1992; **41**: 539–57.

22 Diggle PJ, Tawn JA, Moyeed RA. Model-based geostatistics. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 1998; **47**: 299–350 (with discussion).

23 Kelsall J, Wakefield J. Modeling spatial variation in disease risk: A geostatistical approach. *Journal of the American Statistical Association* 2002; **97**: 692–701.

24 Clayton DG, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 1987; **43**: 359–70.

25 Clayton D, Bernardinelli L. Bayesian methods for mapping disease risk. In Elliott P, Cuzick J, English D, Stern R eds. *Geographical and environmental epidemiology: Methods for small-area studies*. Oxford: Oxford University Press, 1992: 205–20.

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

27 Besag J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 1974; **36**: 192–236 (with discussion).

28 Cressie NAC. *Statistics for spatial data*. New York, NY, USA: John Wiley & Sons, 1993.

29 Besag J, Kooperberg C. On conditional and intrinsic autoregressions. *Biometrika* 1995; **82**: 733–46.

30 Sun D, Tsutakawa RK, Kim H, He Z. Spatio-temporal interaction with disease mapping. *Statistics in Medicine* 2000; **19**: 2015–35.

31 Besag J. On a system of two-dimensional recurrence equations. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 1981; **43**: 302–9.

32 Sun D, Tsutakawa RK, Speckman PL. Bayesian inference for CAR (1) models with noninformative priors. *Biometrika* 1999; **86**: 341–50.

33 Eberly LE, Carlin BP. Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. *Statistics in Medicine* 2000; **19**: 2279–94.

34 Lawson A, Clark A. Spatial mixture relative risk models applied to disease mapping. *Statistics in Medicine* 2002; **21**: 359–370.

35 Denison DGT, Holmes CC, Mallick BK, Smith AFM. *Bayesian methods for nonlinear classification and regression*. New York, NY, USA: John Wiley & Sons, 2002.

36 Green P, Richardson S. Hidden markov models and disease mapping. *Journal of the American Statistical Association* 2002; **97**: 1055–70.

37 Green PJ. Reversible jump Markov chain Monte Carlo computation and bayesian model determination. *Biometrika* 1995; **82**: 711–32.

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

39 Denison DGT, Holmes CC. Bayesian partitioning for estimating disease risk. *Biometrics* 2001; **57**: 143–9.

40 Gangnon RE, Clayton MK. Bayesian detection and modeling of spatial disease clustering. *Biometrics* 2000; **56**: 922–35.

41 Higdon DM. Space and space-time modeling using process convolutions. In Anderson CW, Barnett V, Chatwin PC, El-Shaarawi AH eds. *Quantitative methods for current environmental issues*. London, UK: Springer-Verlag, 2002: 37–54.

42 Best NG, Ickstadt K, Wolpert RL, Cockings S, Elliott P, Bennett J, Bottle A, Reed S. Modelling the impact of traffic-related air pollution on childhood respiratory illness. In Gatsonis C, Kass RE, Carlin B, Carriquiry A, Gelman A, Verninelli I, West M eds. *Case studies in Bayesian statistics volume V*. New York, NY, USA: Springer-Verlag, 2002: 183–259 (with discussion).

43 Best NG, Ickstadt K, Wolpert RL. Spatial Poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American Statistical Association* 2000; **95**: 1076–88.

44 Briggs DJ, Collins S, Elliott P, Fischer P, Kingham S, Lebret E, Pryl K, van Reeuwijk H, Smallbone K, van der Veen A. Mapping urban air pollution using GIS: a regression-based approach. *International Journal of Geographic Information Science* 1997; **11**: 699–718.

45 Spiegelhalter DJ, Thomas A, Best N, Lunn D. *WinBUGS User Manual, Version 1.4*, 2002. (On-line user manual, http://www.mrc-bsu.cam.ac.uk/bugs: accessed 20 January 2004).

46 Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B (Statistical Methodology)* 2002; **64**: 583–639.

47 Mardia KV. Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis* 1988; **24**: 265–84.

48 Gamerman D, Moreira ARB, Rue H. Space-varying regression models: specifications and simulation. *Computational Statistics and Data Analysis* 2003; **42**: 513–33.

49 Gelfand A, Vounatsou P. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* 2003; **4**: 11–25.

50 Carlin BP, Banerjee S. Hierarchical multivariate CAR models for spatio-temporally correlated survival data. In Bernardo JM, Berger JO, Dawid AP, Smith AFM eds. *Bayesian Statistics* 7. Oxford: Oxford University Press, 2003: 45–63.

51 Knorr-Held L, Best N. A shared component model for detecting joint and selective clustering of two diseases. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 2001; **164**: 73–86.

52 Held L, Natario I, Fenton S, Rue H, Becker N. Towards joint disease mapping. *Statistical Methods in Medical Research* 2005; **14**: 61–82.