STATISTICS IN MEDICINE
Statist. Med. 2003; **22**:3213–3228 (DOI: 10.1002/sim.1570)

A hierarchical model for spatially clustered disease rates

Ronald E. Gangnon[1,*,†] and Murray K. Clayton[2]

[1]*Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, 610 N. Walnut Street, Madison, Wisconsin 53726, U.S.A.*
[2]*Department of Statistics, University of Wisconsin-Madison, 1210 W. Dayton Street, Madison, Wisconsin 53706, U.S.A.*

SUMMARY

Maps of regional disease rates are potentially useful tools in examining spatial patterns of disease and for identifying clusters. Bayes and empirical Bayes approaches to this problem have proven useful in smoothing crude maps of disease rates. In recent years, models including both spatially correlated random effects and spatially unstructured random effects have been very popular. The spatially correlated random effects have been proposed in an attempt to capture a general clustering in the data. As an alternative, we propose replacing the spatially structured random effect with fixed clustering effects associated with particular areas. A reversible jump Markov chain Monte Carlo (RJMCMC) algorithm for posterior inference is described. We illustrate the model using the well-known New York leukaemia data. Copyright © 2003 John Wiley & Sons, Ltd.

KEY WORDS: Bayesian inference; clustering; disease mapping; leukaemia; Poisson model; reversible jump MCMC

1. INTRODUCTION

Statistical methods for analysing spatial patterns of disease incidence or mortality have been of great interest over the past decade. To a large extent, the statistical approaches taken fall into two broad classes: cluster detection or disease mapping. In cluster detection, one typically adopts the hypothesis testing framework, testing the null hypothesis of a common disease rate across the study region against a 'clustering' alternative [1, 2]. In disease mapping, one typically uses Bayes or empirical Bayes methods to produce smoothed estimates of the cell-specific disease rates suitable for mapping [3, 4]. In this paper we develop a model for spatially clustered disease rates that addresses both of these inferential goals.

As an example, consider the well-known data set consisting of data on leukaemia incidence for a five-year period in an eight-county region of upstate New York. The observed leukaemia

*Correspondence to: Ronald E. Gangnon, Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, 610 N. Walnut Street, Madison, Wisconsin 53726, U.S.A.
†E-mail: ronald@biostat.wisc.edu

Received June 2002
Accepted April 2003

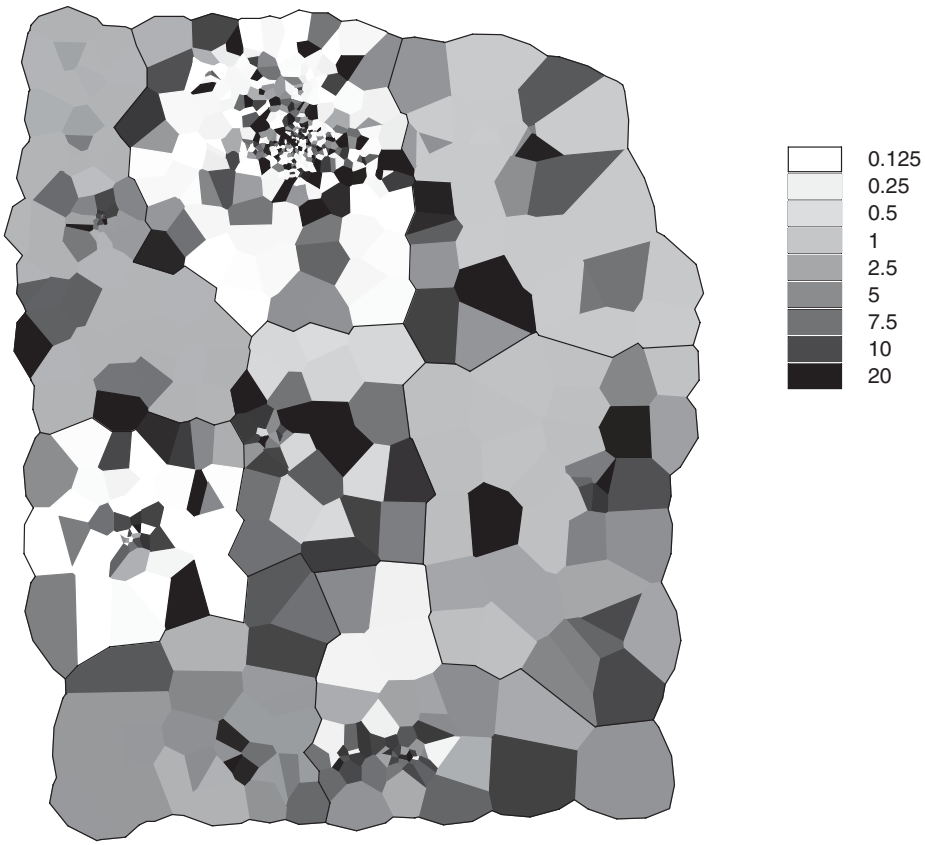

Figure 1. Observed cell-specific five year leukaemia incidence rates for the New York data. Region associated with each cell based on Dirichlet tessellation of cell centroids.

rates for census blocks (in seven counties) or tracts (in Broome county) are displayed in Figure 1. Waller *et al.* [5] provide additional background information about the New York leukaemia data as well as analyses of these data using a number of cluster detection methods, including their own method and the methods of Whittemore *et al.* [6], Openshaw *et al.* [7], and Turnbull *et al.* [8].

Many Bayesian approaches to analysing spatial disease patterns focus on mapping (spatially) smoothed disease rates [3, 4, 9]. Mapping methods produce stable estimates for the cell-specific disease rates by shrinkage to the overall disease rate or by averaging over neighbouring cells. These approaches are most useful for capturing gradual, regional changes in disease rates, and may be less useful in detecting abrupt, localized changes indicative of hot spot clustering. The model proposed by Besag *et al.* [4] incorporates both spatially structured and unstructured random effects in a single model. Ghosh *et al.* [10] have used models of this type to analyse the New York leukaemia data. Waller *et al.* [9] extended this model to incorporate temporal and spatio-temporal effects. Besag *et al.* [11] and Best *et al.* [12] have suggested a prior

specification for the spatially structured random effects more suitable for detection of spatial clusters. None the less, all of these models necessarily assume some form of stationarity of the covariance structure across the study region. As noted by Ferreira *et al.* [13], spatial clusters are, by their nature, regions which are not representative of the entire study region and it therefore seems inappropriate to assume a stationary covariance structure over the entire study region.

A few Bayesian approaches more directly address the disease clustering problem [14–19]. Lawson [14] proposes a point process model for detection of cluster locations when exact case (and control) locations are known. Lawson [18] extends this model to incorporate both localized clustering and general spatial heterogeneity of disease rates. Lawson and Clark [15] describe the application of a point process clustering model to case count data through data augmentation. To apply their model, one imputes locations for each member of the population at risk to produce a point process. One then proposes a clustering model for the point process.

Gangnon and Clayton [16], Knorr-Held and Raßer [17] and Denison and Holmes [19] each consider a relatively non-parametric Bayesian framework for spatial modelling in which the cells are grouped into 'clusters'. Gangnon and Clayton [16] propose a model for clustering using cell count data in which the study region is divided into several components: a large background area and a relatively small number of clusters. A common rate (or covariate-adjusted risk) within each component is assumed. Knorr-Held and Raßer [17] and Denison and Holmes [19] consider a non-parametric Bayesian framework for modelling cell count data. Although superficially similar to the Gangnon and Clayton [20] model in that cells are grouped into components of constant risk, the models of Knorr-Held and Raßer [17] and Denison and Holmes [19] serve a very different goal. In these models, the components (or clusters of cells) primarily serve as a tool for estimating the underlying risk surface, not as parameters of direct interest. In the model of Gangnon and Clayton [16], the location and composition of the cluster of cells is of primary interest. None of these models includes a spatial heterogeneity component.

In this paper we develop a Bayesian approach to inference about the parameters of a hierarchical model for spatial clustering. This model includes both a fixed effects clustering component for cluster detection and risk estimation along with a spatially unstructured random effects component to capture any extra-Poisson variation as in disease mapping. The clustering component of the model requires the specification of a set of potential clusters and a prior distribution on that set of potential clusters. Our approach allows for multiple clusters and produces posterior estimates of cell-specific and cluster-specific relative risks as well as cell-specific probabilities of cluster membership. In addition, posterior inference about the number of clusters in the data is also possible, and estimates are available both conditional on a fixed number of clusters and unconditionally.

In Section 2 we describe our hierarchical model for spatially clustered disease rates. In Section 3 we describe our implementation of a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm [21] for drawing inferences about the model. In Section 4 we analyse the aforementioned data on leukaemia incidence in upstate New York using the proposed model. Finally, in Section 5, we close with a discussion of alternative model specifications and possible extensions.

2. STATISTICAL MODEL

We begin by defining some notation and a basic statistical model. We consider situations in which the study region is divided into N subregions, or cells. For each cell i, we observe O_i, the number of cases of disease, and n_i, the population at risk in cell i. We assume a Poisson model for the data, that is, $O_i \sim \text{Poisson}(\rho_i n_i)$, where ρ_i is the disease rate in cell i.

We model the cell-specific disease rates using a log-linear model $\log(\rho_i) = \mu + \gamma_i + \varepsilon_i$. There are three basic components in this model: a non-spatial fixed effects component (μ); a spatial clustering component (γ_i), and a spatially unstructured random effects (or extra-Poisson variation) component (ε_i). Our primary interest lies in a prior specification for the spatial clustering component of the model; standard priors are available for the other two components.

In our development here, the non-spatial fixed effects component of the model consists of a single parameter μ. This parameter is related to the overall rate across the study region and is well-identified by the data. We propose using a flat prior for this parameter (a normal prior with large variance serves equally well). In other settings, the non-spatial component of the model could also incorporate the cell-level effects of covariates such as age and sex.

For the spatial heterogeneity effects (ε_i), we follow Waller *et al.* [9] in proposing an exchangeable normal prior for the ε_i's; that is, $\varepsilon_i \sim \text{N}(0, 1/\tau)$. For the parameter τ, we use a proper, but relatively weak, gamma prior distribution for τ. In Section 4 we follow Waller *et al.* [9] in using a gamma distribution with mean 100 and standard deviation 100. Thus, *a priori*, $1/\tau$ falls between 0.003 and 0.40 with 95 per cent probability. To place these values in perspective, a variance of 0.40 implies a roughly 12-fold difference in risk of leukaemia between a cell at the 2.5th percentile of risk and a cell at the 97.5th percentile of risk; a variance of 0.003 implies only a 1.2-fold difference in risk.

The spatial clustering component of the model is $\gamma_i = \sum_{j=1}^{k} \theta_j I_{\{i \in \mathbf{c}_j\}}$, where k is the number of clusters, $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k$ are the sets of cells belonging to the k clusters and $\theta_1, \theta_2, \ldots, \theta_k$ are the log relative risks associated with each cluster (relative to the background risk defined by μ). We develop a prior for the spatial clustering component of the model by successively conditioning on parameters. Given $k, \mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k$ (that is, the number of clusters and their locations), we assign a normal prior for $\theta_1, \theta_2, \ldots, \theta_k$; that is, θ_j iid $\text{N}(0, \sigma_\theta^2)$. In the example, we take σ_θ^2 to be 0.355 so that, *a priori*, the relative risk associated with a cluster falls between 0.25 and 4.00 with 99 per cent probability. The variance σ_θ^2 is not a parameter of interest in this model. In other models with many partitions and without a background component, for example, the models of Knorr-Held and Raßer [17] and Denison and Holmes [19], σ_θ^2 could be a parameter of interest, and a gamma hyperprior could be appropriate. The algorithm described in Section 3 could be easily adapted to such models.

Next, given k (that is, the number of clusters), we select $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k$ independently from a prior distribution on the space of possible clusters; denote this distribution by $p(\mathbf{c})$. Note that, because of the prior independence, two or more clusters may overlap or, in the extreme, even duplicate each other. This potential duplication of clusters could raise some concerns regarding identifiability. However, the prior (and posterior) probability of such duplicate clusters is very small, and sampling with replacement has many computational advantages over sampling without replacement in this setting.

To make our discussion more concrete, we consider a specific set of potential clusters and develop a prior distribution for it. A similar development in a hypothesis testing framework is described in Gangnon and Clayton [20]. We consider circular clusters centred at the cell centroids as potential clusters. We centre clusters at the centroids to avoid empty clusters. The radii of the circles varies continuously from zero up to a fixed maximum radius, $r_{\max}$. If the centroid of a cell falls within the circle, then the whole cell is included in the cluster. Since there are only a finite number of cells, there will only be a finite number of clusters about each cell centroid. To identify these clusters, let $0 = r_{i,(1)} < r_{i,(2)} < \cdots < r_{i,(m_i)} \leqslant r_{\max}$ be the ordered distances from the centroid of cell i to the centroids of all cells, truncated at $r_{\max}$. (If two or more centroids are equidistant from the centroid i, the common distance is only listed once.) Then, the distinct potential clusters about cell i are circles of radii $r_{i,(1)}, r_{i,(2)}, \ldots, r_{i,(m_i)}$. We refer to the cluster centred at the centroid of cell i of radius $r_{i,(j)}$ as cluster i,j for $j = 1, 2, \ldots, m_i$ and $i = 1, 2, \ldots, N$.

Our prior distribution on the set of potential clusters is developed as an approximation to the uniform selection of a circle from the study region, slightly modified to account for the discreteness of the clusters. Specifically, we first select a cluster centre and then, conditional on that centre, select a cluster radius. The cluster centre is selected as the centroid of the cell to which a point sampled from a uniform distribution over the study area belongs. The radius of the circle is then selected at random from a uniform distribution on $[0, r_{\max}]$. Thus, the prior probability of selecting cluster i,j is

$$p(i,j) = \frac{a_i}{A} \frac{r_{i,j+1} - r_{i,j}}{r_{\max}}$$

where a_i is the area of cell i, A is the area of the study region, and $r_{i,m_i+1} = r_{\max}$. In Figure 2, we display the probability of a cell belonging to a cluster selected from the prior distribution. We note that this probability is roughly constant for all cells.

Finally, we select a prior distribution for k, the number of clusters. One possibility would be a distribution on the non-negative integers such as the Poisson, geometric or negative binomial distributions. Another possibility, which we generally prefer, is to restrict k to the values $0, 1, 2, \ldots, k_{\max}$ for some positive integer $k_{\max}$. In most problems, selecting a maximum number of clusters $k_{\max}$ should not be too difficult. In the example, we assign $k_{\max} = 10$. On this restricted space, we typically place a flat (discrete uniform) prior distribution.

3. POSTERIOR CALCULATION

If the clusters (both number and location) were fixed, simulation from the posterior using Markov chain Monte Carlo techniques would be quite straightforward. The structure of the problem is that of a hierarchical generalized linear model. Bayesian techniques for analysing GLMs are discussed in Gelman *et al.* [22], and we follow their approach. The normal prior distributions for μ, $\theta_1, \theta_2, \ldots, \theta_k$ and $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ are conjugate to a quadratic (normal) approximation to the Poisson likelihood, and it is thus relatively easily to sample from approximations to the full conditional distributions of the parameters. We do so in an application of a Metropolis–Hastings algorithm [23]. However, we use the approximation as a tool to develop a proposal distribution; the exact Poisson likelihood is used for inference.

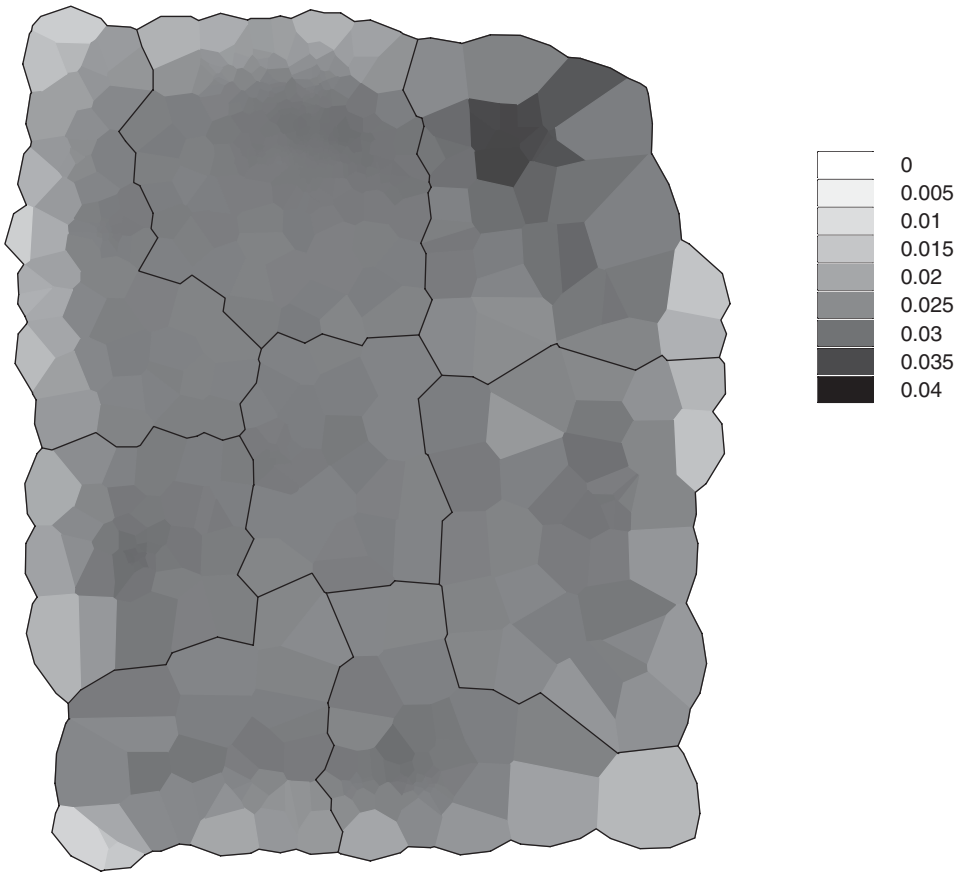

Figure 2. Probability that each cell belongs to a single cluster selected from the prior distribution on the set of potential clusters.

To make this discussion more concrete, we explicitly describe the Metropolis–Hastings steps for updating μ. To propose a new value for μ, we need O_t (the total case count in the study region), E_{t0} (the current value for the parameter $E_t = \sum_{i=1}^{N} \rho_i n_i$, the expected number of cases in the entire study region), μ_0, the current value for the parameter μ, and the prior mean and variance for μ, denoted by η and σ_μ^2, respectively. Based on a local quadratic approximation to the Poisson likelihood, a proposed new value of μ, denoted by μ', is drawn from a normal distribution with mean

$$\eta_p = \frac{E_{t0}}{E_{t0} + 1/\sigma_\mu^2}\,\mu_0 + \frac{1/\sigma_\mu^2}{E_{t0} + 1/\sigma_\mu^2}\,\eta + \frac{O_t - E_{t0}}{E_{t0} + 1/\sigma_\mu^2}$$

and variance

$$\sigma_p^2 = \frac{1}{E_{t0} + 1/\sigma_\mu^2}$$

The new value μ' is accepted with probability

$$\min\left\{1, \frac{\phi(\mu,\eta_p',\sigma_p^{2\prime})}{\phi(\mu',\eta_p,\sigma_p^2)}\frac{\phi(\mu',\eta,\sigma_\mu^2)}{\phi(\mu,\eta,\sigma_\mu^2)}\frac{l(O_t,E_t')}{l(O_t,E_{t0})}\right\}$$

otherwise the current value μ is retained. In this equation, η_t' and E_t' are calculated as η_p and E_{t0} above using μ' in place of μ_0, $\phi(\cdot,\mu,\sigma^2)$ is the density of a normal random variable with mean μ and variance σ^2 and $l(y,\mu)$ is the likelihood of a Poisson random variable with observed count y and mean μ. The definitions of Metropolis–Hastings steps for the other parameters $\theta_1,\theta_2,\ldots,\theta_k$ and $\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_N$ follow the same template.

The gamma prior distribution for τ (the inverse of the prior variance for $\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_N$) is also conjugate, so samples for τ can be obtained using the Gibbs sampler. In particular, if the prior distribution for τ follows a gamma(a,b) distribution (mean a/b and variance a/b^2), the full conditional distribution of τ is gamma$(a+N/2, b+\sum_{i=1}^N \varepsilon_i^2/2)$.

The novelty in the current problem is the unknown number (and locations) of the clusters. A number of additional transitions must be proposed to account for the varying number of clusters. A general approach to accounting for a varying numbers of parameters is the reversible jump Markov chain Monte Carlo (RJMCMC) algorithm [21].

In addition to the steps for fixed clusters described above, we use the following three steps:

1. ADD. Propose a new cluster $\mathbf{c}_{k+1}$ and its associated parameter θ_{k+1} for the model.
2. DROP: Propose a cluster to remove from the model.
3. CHANGE. Propose a new cluster location for a cluster currently in the model (maintaining the same value for the associated θ).

The ADD and DROP steps are counterparts of each other, while the CHANGE step is its own counterpart. In each iteration of the algorithm, one of these three steps is proposed with probability $p_a(k)$, $p_d(k)$ and $p_c(k)$, respectively. Note that these probabilities depend on the current value of the parameter k. In the subsequent example, we take $p_a(k)=p_d(k)=p_c(k)=1/3$ for $0<k<k_{\max}$. For $k=0$, $p_a(k)=1$. For $k=k_{\max}$, $p_d(k)=1/3$ and $p_c(k)=2/3$.

The ADD step consists of two parts. First, we propose the new cluster $\mathbf{c}_{k+1}$. Although we could use a random selection from the prior distribution, such a choice would likely be quite inefficient. Instead, we attempt to better utilize information from the data. To do this, for each potential cluster, we first find the posterior mode (conditional on all the current parameter values) for its associated log relative risk. The posterior mode is $\hat{\theta}_{\mathbf{c}}=(O_{\mathbf{c}}-E_{\mathbf{c}})/(E_{\mathbf{c}}+1/\sigma_\theta^2)$, where $O_{\mathbf{c}}$ is the number of cases in the cluster, $E_{\mathbf{c}}$ is the current value for the expected number of cases in the cluster and σ_θ^2 is the prior variance for the θ's (the prior mean is assumed to be 0). We then select the proposed new cluster with probability proportional to the posterior density. In particular, we propose cluster $\mathbf{c}$ with probability

$$p_{\text{select}}(\mathbf{c})=\frac{p(\mathbf{c})\phi(\hat{\theta}_{\mathbf{c}},0,\sigma_\theta^2)l(O_{\mathbf{c}},e^{\hat{\theta}_{\mathbf{c}}}E_{\mathbf{c}})}{\sum_{\mathbf{c}} p(\mathbf{c})\phi(\hat{\theta}_{\mathbf{c}},0,\sigma_\theta^2)l(O_{\mathbf{c}},e^{\hat{\theta}_{\mathbf{c}}}E_{\mathbf{c}})}$$

After the cluster $\mathbf{c}_{k+1}$ is selected, a value for its associated log relative risk, θ_{k+1}, is proposed using the normal approximation described earlier, that is, sampled from a normal distribution with mean $\hat{\theta}_{\mathbf{c}}$ and variance $1/(E_{\mathbf{c}}+1/\sigma_\theta^2)$.

The reversing DROP step is quite simple. One of the k current clusters is selected at random (with probability $1/k$) to be dropped from the model. The acceptance probabilities for the ADD and DROP steps then take the following forms.

For the ADD step (letting $\mathbf{c} = \mathbf{c}_{k+1}$)

$$\min\left\{1, \frac{p_{\mathrm{d}}(k+1)}{p_{\mathrm{a}}(k)} \frac{p(k+1)}{p(k)} \frac{1}{k+1} \frac{p(\mathbf{c})}{p_{\mathrm{select}}(\mathbf{c})} \frac{\phi(\theta_{k+1}, 0, \sigma_\theta^2)}{\phi(\theta_{k+1}, \hat{\theta}_{\mathbf{c}}, 1/(E_{\mathbf{c}} + 1/\sigma_\theta^2))} \frac{l(O_{\mathbf{c}}, \mathrm{e}^{\theta_{k+1}} E_{\mathbf{c}})}{l(O_{\mathbf{c}}, E_{\mathbf{c}})}\right\}$$

For the DROP step (letting $\mathbf{c} = \mathbf{c}_k$)

$$\min\left\{1, \frac{p_{\mathrm{a}}(k-1)}{p_{\mathrm{d}}(k)} \frac{p(k-1)}{p(k)} k \frac{p_{\mathrm{select}}(\mathbf{c})}{p(\mathbf{c})} \frac{\phi(\theta_k, \hat{\theta}_{\mathbf{c}}, 1/(E_{\mathbf{c}} + 1/\sigma_\theta^2))}{\phi(\theta_k, 0, \sigma_\theta^2)} \frac{l(O_{\mathbf{c}}, \mathrm{e}^{-\theta_k} E_{\mathbf{c}})}{l(O_{\mathbf{c}}, E_{\mathbf{c}})}\right\}$$

Note that, without loss of generality, we can assume the kth cluster is chosen to be dropped. If it is not, simply relabel the clusters so that it is.

The CHANGE step is simple as well. We select one of the k clusters at random and fix the associated parameter θ. Again, without loss of generality, we may assume the kth cluster is chosen. We then drop the cluster from the model and select a new cluster with probability proportional to the posterior density (based on the fixed θ_k). The probability that cluster $\mathbf{c}$ is selected as the new kth cluster is then given by

$$p_{\mathrm{select}}(\mathbf{c}) = \frac{p(\mathbf{c})\phi(\theta_k, 0, \sigma_\theta^2) l(O_{\mathbf{c}}, \mathrm{e}^{\theta_k} E_{\mathbf{c}})}{\sum_{\mathbf{c}} p(\mathbf{c})\phi(\theta_k, 0, \sigma_\theta^2) l(O_{\mathbf{c}}, \mathrm{e}^{\theta_k} E_{\mathbf{c}})}$$

The acceptance ratio for this step is identical to one, so it is always accepted.

4. EXAMPLE: NEW YORK LEUKAEMIA DATA

We now present an example of the application of our methodology. The New York leukaemia data set consists of data on leukaemia incidence between 1978 and 1982 in eight counties in upstate New York: Broome; Cayuga; Chenango; Cortland; Madison; Onondaga; Tioga, and Tompkins. The two largest cities in the study region are Syracuse in Onondaga county and Binghamton in Broome county.

The eight-county region is divided into 790 cells. In seven of the counties the cells are census block groups; in Broome county, the cells are larger census tracts. For each cell, the population at risk, count of leukaemia cases and geographic centroid are available. A few cases could not be assigned to a single cell due to incomplete location data. These cases are fractionally assigned to the possible cells in proportion to the cell populations. Additional background information on the New York leukaemia data is available elsewhere [5]. The observed leukaemia rate for each cell is displayed in Figure 1 using the Dirichlet tessellation of the cell centroids. Note that this tessellation provides only an approximation to the true cells. No obvious clusters are evident in this figure.

For our analysis of the New York leukaemia data, we utilized the prior described in Section 2 with a maximum cluster radius ($r_{\max}$) of 20 km. Following Gelman and Rubin [24], we ran five independent Markov chains. Each chain used a run-in of 100000 iterations, and 1 million further iterations to obtain the sample of models, keeping every 100th in the generated

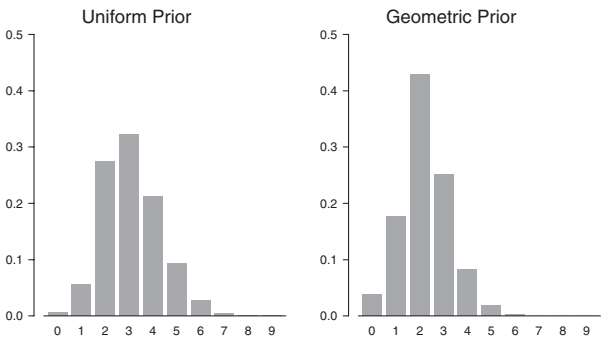

Figure 3. Posterior distribution of the number of clusters k based on 50000 Markov chain Monte Carlo simulations. The first histogram is the posterior distribution based on a uniform prior for k. The second histogram is the posterior distribution using a geometric distribution with success probability $1/2$ as a prior for k.

sample. A subset of parameters were graphically monitored across the five chains. Each of the chains appeared to converge by that point, and there were no substantial differences in the samples across the chains.

In Figure 3, we present the posterior distribution of the number of clusters k included in the model. Based on this distribution, there appears to be strong evidence for clustering; the posterior probability of the no-cluster model is 0.006. The posterior mode for k is 3 (posterior probability 0.322), but there is substantial posterior probability associated with k values of 2 (0.275) and 4 (0.211). Thus, we have fairly strong evidence of clustering in the data, but somewhat equivocal evidence for the correct number of clusters (2, 3 or 4).

In truth, a uniform prior on the number of clusters may be unrealistic. A more defensible prior would likely place higher weight *a priori* on models with few clusters than on models with many clusters. To illustrate the effects of such a prior choice on inference, we consider the impact of assigning a geometric prior (with failure probability $1/2$) to k. The posterior samples based on the uniform prior provide an importance sample for the posterior based on the geometric prior; the importance sampling weight for a model with k clusters is proportional to 0.5^k. The posterior for k based on this second prior is provided in Figure 3. Compared with the posterior based on a flat prior, this distribution is shifted substantially towards models with $k=2$. With this revised prior, there is relatively little support for a model with $k=4$ (posterior probability 0.082).

In Figure 4 we display the posterior means for the cluster risks associated with each cell $E[\exp(\sum_{j=1}^{k}\theta_j I_{\{i\in\mathbf{c}_j\}})]$ for $k=3$, the modal number of clusters. In Figure 5 we display the posterior probability that a cell belongs to a cluster $\Pr(\sum_{j=1}^{k} I_{\{i\in\mathbf{c}_j\}}>0)$ for $k=3$, the modal number of clusters. These figures show convincing evidence for two areas of clustering in the New York leukaemia data and suggestive evidence for a third area of clustering. The term 'areas of clustering' is used instead of 'clusters' to indicate that the data support many different specific clusters in a particular area. The first area of clustering is located in Broome county in the southern portion of the study region and is associated with an increased incidence of leukaemia. This area includes the city of Binghamton. The second area of clustering is located

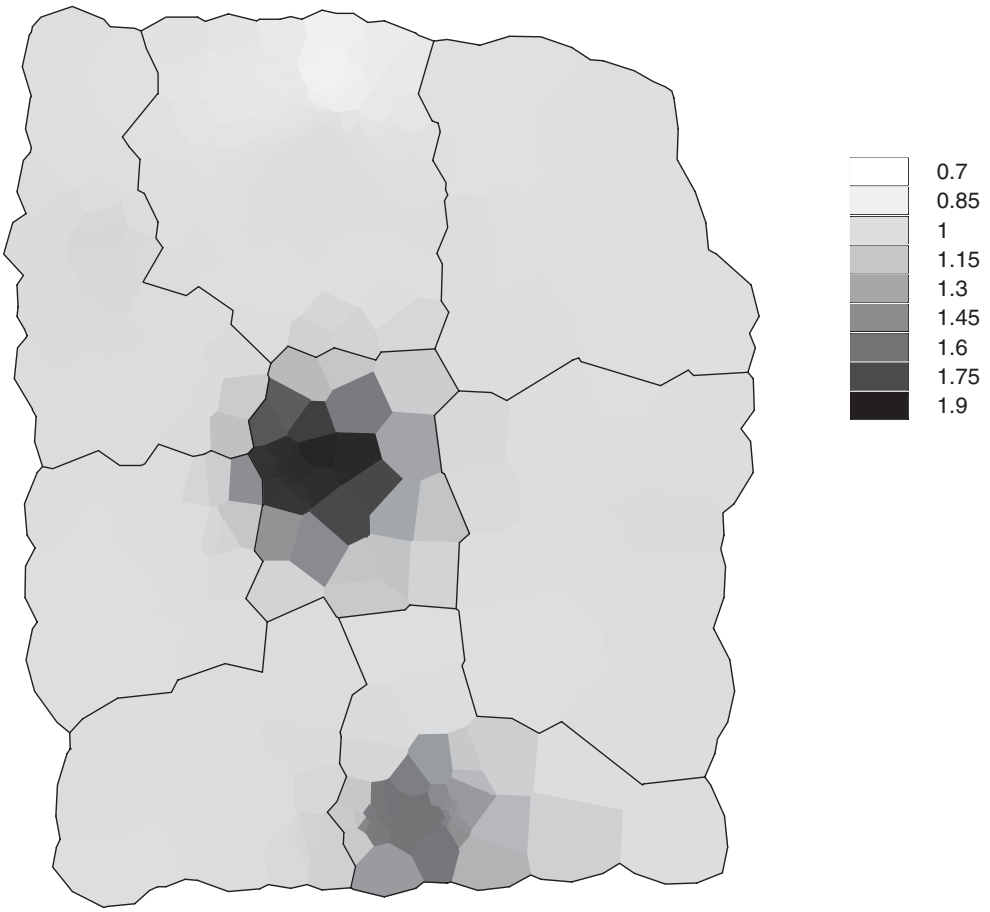

Figure 4. Posterior mean of the cluster risk associated with each cell, $E[\exp(\sum_{j=1}^{k}\theta_j I_{\{i\in\mathbf{c}_j\}})]$, for $k=3$ clusters.

in Cortland county in the centre of the study region and is also associated with an increased incidence of leukaemia. The third area of clustering is located in Onondaga county, north of Syracuse, and is associated with a decreased incidence of leukaemia.

In Figure 6 we display the posterior probability that a cell belongs to a cluster $\Pr(\sum_{j=1}^{k} I_{\{i\in\mathbf{c}_j\}} > 0)$ averaged across all values of k using the discrete uniform prior for k as an index of the absolute strength of the evidence for clusters in particular locations. The overall evidence for the areas of clustering in Broome and Cortland counties is quite strong (posterior probabilities of 0.86 and 0.80, respectively); the evidence for the area of clustering in Onondaga county is more modest (posterior probability of 0.33).

To this point, we have only explored inferences about the clustering component of the model. We now examine the spatial heterogeneity component of the model (ε_i). For $k=0$ (no clusters), the posterior median for the variance $1/\tau$ is 0.048 with a central 95 per cent

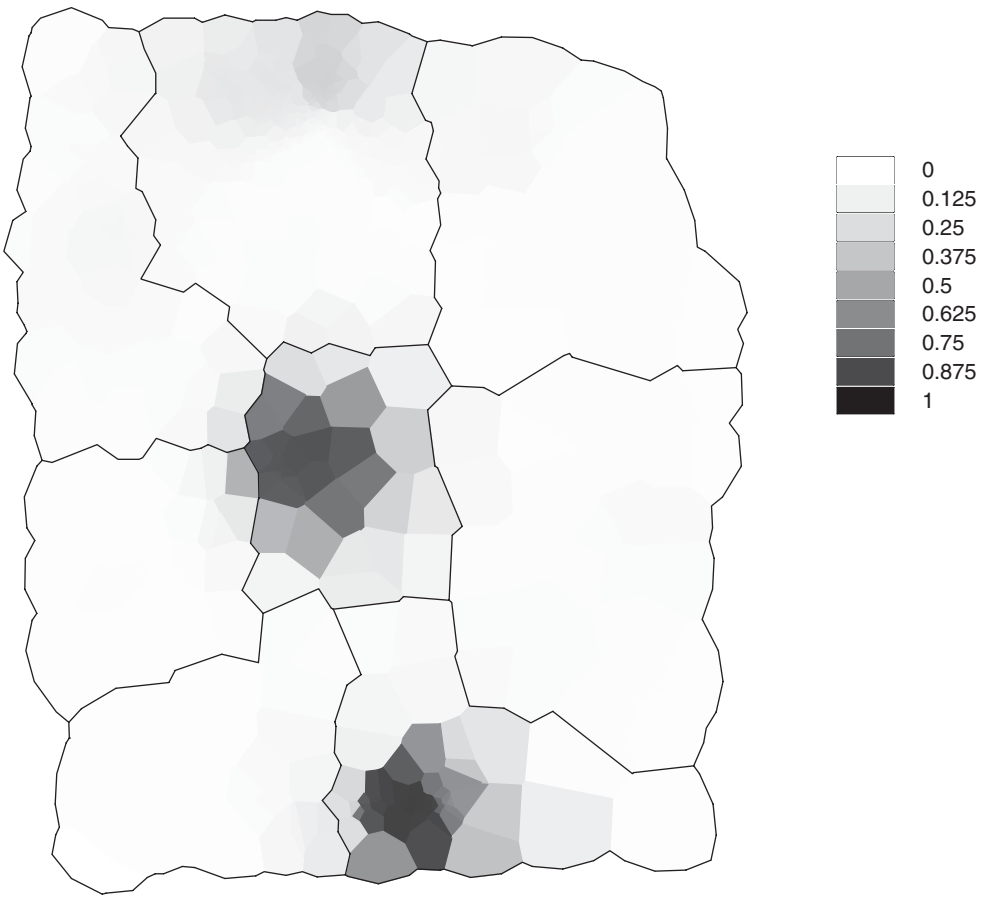

Figure 5. Posterior probability that each cell belongs to a cluster, $\Pr(\sum_{j=1}^{k} I_{\{i \in \mathbf{c}_j\}} > 0)$, for $k = 3$ clusters.

posterior credible interval of $(0.005, 0.18)$. For $k = 3$ (the modal number of clusters), the posterior median is 0.015 with a central 95 per cent posterior credible interval of $(0.003, 0.10)$. These observations suggest that the clustering component of the model may be explaining much of the apparent heterogeneity in leukaemia rates.

In Figure 7 we present the posterior means for the disease rate in each cell for $k = 0$ (no clusters). In Figure 8 we present the posterior means for the disease rate in each cell for $k = 3$. The second map shows clear evidence of the three areas of clustering that we have previously discussed, but it also shows evidence of variations in risk within the areas of clustering. We note that the model with no clusters shows little consistent evidence of elevated or lowered risks in the potential areas of clustering.

Many analyses of the New York leukaemia data have been published. The majority of the previous analyses have been based on hypothesis testing methods solely or primarily aimed at detecting a single cluster with an elevated risk of disease. These methods have generally detected clustering in either Broome county or Cortland county [5]. Some methods [20, 25]

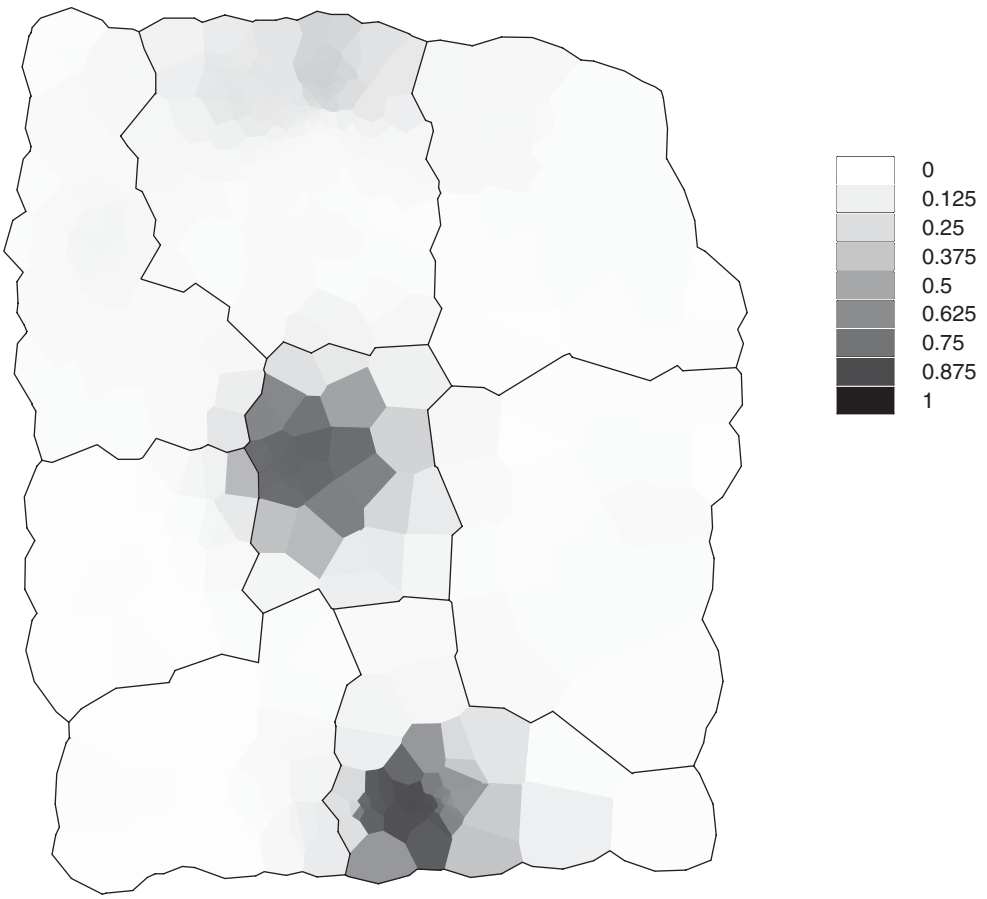

Figure 6. Posterior probability that each cell belongs to a cluster, $\Pr(\sum_{j=1}^{k} I_{\{i \in \mathbf{c}_j\}} > 0)$, based on the discrete uniform prior for k.

showed evidence of clustering in both locations; however, they provided no formal method for evaluating the significance of multiple clusters.

An alternative Bayesian analysis of the New York leukaemia data was described by Gangnon and Clayton [16]. Their method allowed for a much larger class of potential clusters; essentially any connected set of cells was a potential cluster. In contrast, the method described here uses a limited set of potential clusters. The benefits of using a limited set of clusters include a more concrete prior specification (especially for the number of clusters k) and the ability to incorporate extra-Poisson variation into the model.

Gangnon and Clayton [16] found evidence for three clusters associated with an increased risk of leukaemia: areas of clustering in Broome and Cortland counties discussed here and an area of clustering in Onondaga county within the city of Syracuse. The differences in inference likely result from differences in prior specifications and the inclusion of a spatial heterogeneity component in our model. The prior used in Gangnon and Clayton [16] places

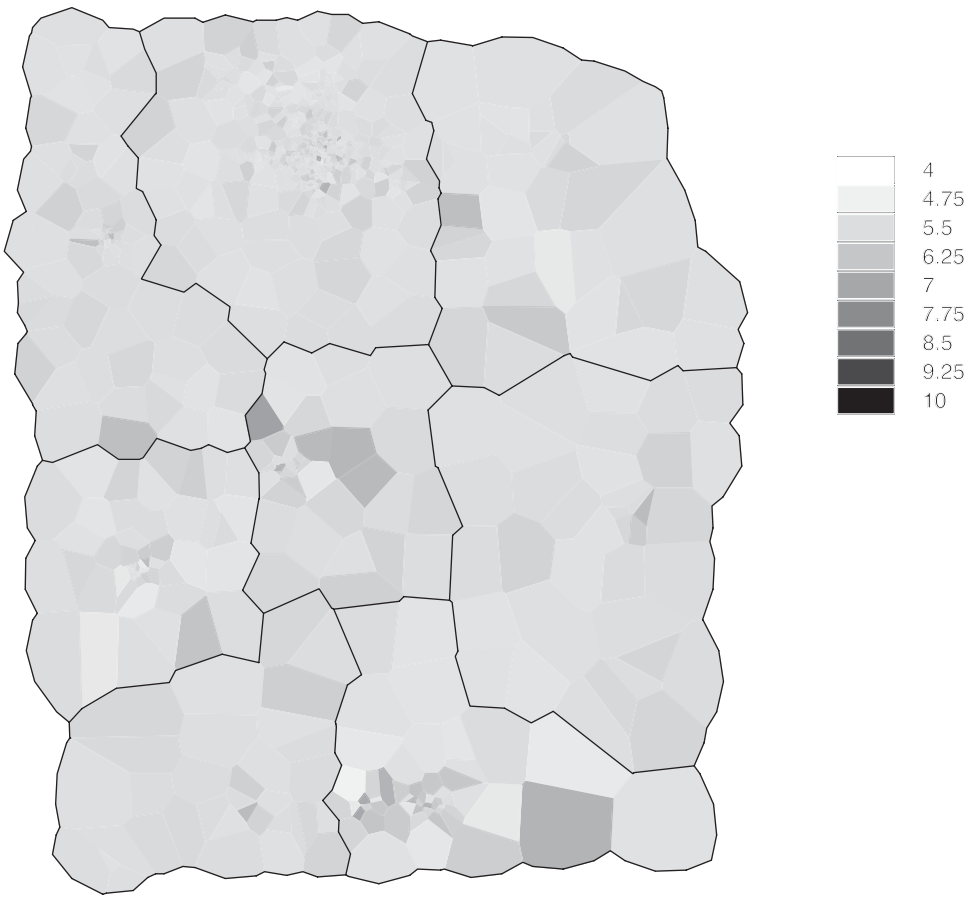

Figure 7. Posterior mean of the leukaemia rate associated with each cell, $E[\exp(\mu + \varepsilon_i)]$, for $k = 0$.

relatively larger weight on the many small overlapping clusters within Syracuse than does the more uniform prior used in our analysis. A recent analysis by Denison and Holmes [19] produced an estimated risk surface that shows apparent evidence for all four features described above. They found compelling evidence for elevated leukaemia risks in Broome and Cortland counties, but did not present formal evaluations of the risks in Onondaga county.

5. DISCUSSION

In this paper we demonstrate the use of a hierarchical model for estimating the locations of spatial clusters and their risks in the presence of extra-Poisson variation using cell count data. The model for clustering effects assumes a discontinuous risk surface with a large background region and a small number of clusters. The model also incorporates spatially unstructured random effects to capture extra-Poisson variation in disease rates. The analysis of the New

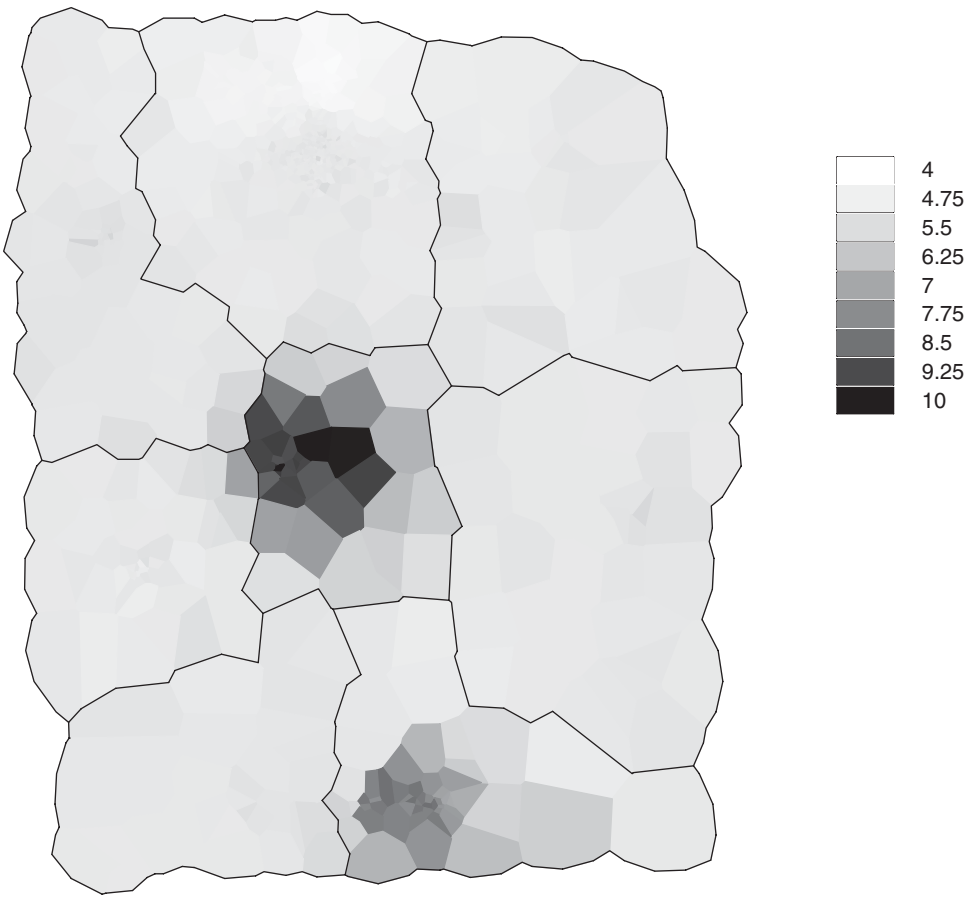

Figure 8. Posterior mean of the leukaemia rate associated with each cell, $E[\exp(\mu + \sum_{j=1}^{k} \theta_j I_{\{i \in \mathbf{c}_j\}} + \varepsilon_i)]$, for $k = 3$ clusters.

York leukaemia data presented here is primarily illustrative. Future analyses will evaluate the impact of different specifications of the clustering model on the resulting inferences, both in the selection of potential clusters and the choice of prior specification. We conclude by briefly commenting on two extensions of this work.

In our presentation we have focused on an approximately uniform prior on the available clusters. This prior is suitable for exploratory studies or routine surveillance in cases where no potential cluster locations have been identified. Likewise, evidence for prespecified clusters can be evaluated in an unbiased fashion using this uniform prior for the clusters. On the other hand, prior knowledge of cluster locations can be incorporated into these models. An informative prior could be postulated for the first cluster. For example, with probability one, the first cluster could be required to overlap a single cell (or a set of cells or one of a set of cells). In such a setting, the first cluster would likely be forced into the model and inference would range over cluster sizes from 1 up to $k_{\max}$. Alternatively, if the presence of the cluster

was less certain, a mixture prior could be formulated for the clusters such that, with some probability, a cluster is drawn from the restricted distribution above, otherwise, a cluster is drawn from the 'uniform' distribution. The extension of these ideas to multiple prespecified clusters is straightforward.

Finally, we note that in many applications it is useful to evaluate the clustering effects after accounting for regional covariates such as demographic composition of the cells or average pollution levels. Since the underlying model is a generalized linear model, the inclusion of such covariates is quite straightforward. One would simply replace the parameter μ with the linear predictor $\mu + \beta^t \mathbf{x}$. One could also easily extend the model to incorporate interactions between the covariate effects and the clusters.

REFERENCES

1. Marshall RJ. A review of methods for the statistical analysis of spatial patterns of disease. *Journal of the Royal Statistical Society*, *Series A* 1991; **154**:421–441.
2. Elliot P, Martuzzi M, Shaddick G. Spatial statistical methods in environmental epidemiology: a critique. *Statistical Methods in Medical Research* 1995; **4**:137–159.
3. Clayton D, Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* 1987; **43**:671–681.
4. Besag J, York J, Mollié A. Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* 1991; **43**:1–59.
5. Waller LA, Turnbull BW, Clark LC, Nasca P. Spatial pattern analyses to detect rare disease clusters. In *Case Studies in Biometry*, Lange N, Ryan L, Billard L (eds). Wiley: New York, 1994; 3–22.
6. Whittemore A, Friend N, Brown BW, Holly EA. A test to detect clusters of disease. *Biometrika* 1987; **74**:31–35.
7. Openshaw S, Craft AW, Charlton M, Birch JM. Investigation of leukaemia clusters by use of a geographical analysis machine. *Lancet* 1988; **1**:272–273.
8. Turnbull BW, Iwano EJ, Burnett WS, Howe HL, Clark LC. Monitoring for clusters of disease: application to leukemia incidence in upstate New York. *American Journal of Epidemiology* 1990; **132**:S136–S143.
9. Waller LA, Carlin BP, Xia H, Gelfand AE. Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association* 1997; **92**:607–617.
10. Ghosh M, Natarajan K, Waller LA, Kim D. Hierarchical Bayes GLMs for the analysis of spatial data: an application to disease mapping. *Journal of Statistical Planning and Inference* 1999; **75**:305–318.
11. Besag J, Green P, Higdon D, Mengersen K. Bayesian computation and stochastic systems (disc: P41-66). *Statistical Science* 1995; **10**:3–41.
12. Best NG, Arnold RA, Thomas A, Waller LA, Conlon EM. Bayesian models for spatially correlated disease and exposure data. In *Bayesian Statistics* 6. 1999; 131–156.
13. Ferreira JTAS, Denison DGT, Holmes CC. Partition modelling. In *Spatial Cluster Modelling*, Lawson AB, Denison DGT (eds). Chapman & Hall: 2002; 125–145.
14. Lawson AB. Markov chain Monte Carlo methods for putative pollution source problems in environmental epidemiology. *Statistics in Medicine* 1995; **14**:2473–2486.
15. Lawson AB, Clark A. Markov chain Monte Carlo methods for putative sources of hazard and general clustering. In *Disease Mapping and Risk Assessment for Public Health*, Lawson AB, Bohning D, Biggeri A, Viel J-F, Bertollini R (eds). Wiley WHO. 1999; chapter 9.
16. Gangnon RE, Clayton MK. Bayesian detection and modeling of spatial disease clustering. *Biometrics* 2000; **56**:922–935.
17. Knorr-Held L, Raßer G. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* 2000; **56**:13–21.
18. Lawson AB. Cluster modelling of disease incidence via RJMCMC methods: a comparative evaluation. *Statistics in Medicine* 2000; **19**:2361–2375.
19. Denison D, Holmes C. Bayesian partitioning for estimating disease risk. *Biometrics* 2001; **57**:143–149.
20. Gangnon RE, Clayton MK. A weighted average likelihood ratio test for spatial clustering of disease. *Statistics in Medicine* 2001; **20**:2977–2987.

21. Green P. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 1995; **82**:711–732.
22. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. Chapman & Hall: 1995.
23. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970; **57**:97–109.
24. Gelman A, Rubin D. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**:457–472.
25. Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. *Statistics in Medicine* 1995; **14**: 799–810.

