A hierarchical model for spatially clustered disease rates

Ronald E. Gangnon, Ph.D.,
Murray K. Clayton, Ph.D.
A Hierarchical Model for Spatially Clustered Disease Rates

Ronald E. Gangnon

Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison,
610 N. Walnut Street, Madison, Wisconsin 53705, U.S.A.

email: ronald@biostat.wisc.edu

and

Murray K. Clayton

Department of Statistics, University of Wisconsin–Madison,
1210 W. Dayton Street, Madison, Wisconsin 53706, U.S.A.

SUMMARY. Maps of regional disease rates are potential 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, the model of Waller et al. (1997) has proven to be very popular. This model includes both spatial autocorrelation and spatial heterogeneity effects. The spatial autocorrelation effect attempts to capture "clustering" effects in the data, while the spatial heterogeneity effect attempts to capture spatially unstructured variation. In practice, the two sets of effects are generally not separately identified by the data, leading to many challenges in model fitting and interpretation. As an alternative, we propose replacing the spatial autocorrelation effect with clustering effects associated with particular areas. A computational algorithm based on reversible jump Markov chain Monte Carlo (RJMCMC) is described. We illustrate our model using the well-known New York leukemia data.

KEY WORDS: Bayesian inference; Clustering; Disease mapping; Generalized linear model; Leukemia; Poisson model; Relative risk; Reversible jump MCMC
1 Introduction

Statistical methods for analyzing 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 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 (cf., Whittemore et al., 1987, Kulldorff and Nagarwalla, 1995). In disease mapping, one typically uses Bayes or empirical Bayes methods to produce smoothed estimates of the cell-specific disease rates suitable for mapping (cf, Clayton and Kaldor, 1987 and Besag et al., 1991). Few efforts have been made to attack simultaneously both the cluster detection problem and the disease mapping problem.

As an example, consider the well-known data set consisting of data on leukemia incidence for a five-year period in an eight-county region of upstate New York. The observed leukemia rates for census blocks (in seven counties) or tracts (in Broome county) are displayed in Figure 1. Waller et al. (1994) provide additional background information about the New York leukemia data as well as analyzes of these data using a number of cluster detection methods, including their own method and the methods of Whittemore et al. (1987), Openshaw et al. (1988), Turnbull et al. (1990), Kulldorff and Nagarwalla (1995) and Gangnon and Clayton (2001) later analyzed the New York leukemia data using cluster detection methods they had developed.

Many Bayesian approaches to analyzing spatial disease patterns focus on mapping spatially smoothed disease rates (for example, Clayton and Kaldor, 1987, Besag et al., 1991 and Waller et al., 1997)). Mapping methods produce stable estimates for cell-specific rates by borrowing strength from neighboring cells. These are most useful for capturing gradual, regional changes in disease rates, and are less useful in detecting abrupt, localized changes indicative of hot spot clustering. The models proposed by Besag et al. (1991) and Waller et al. (1997) incorporate both spatially structured
Figure 1: Observed cell-specific five year leukemia incidence rates for the New York. Region associated with each cell based on Dirichlet tessellation of cell centroids.
(spatial correlation) and unstructured (extra-Poisson variation) heterogeneity in one model. The ability of these models to detect localized clusters is questionable, because they incorporate only a global clustering mechanism. In addition, typically, the spatially structured and unstructured components of the heterogeneity are not separately identifiable by the likelihood (Waller et al., 1997).

A few Bayesian approaches more directly address the disease clustering problem, including Lawson (1995), Lawson and Clark (1999), Gangnon and Clayton (2000), Knorr-Held and Raßer (2000), Lawson (2000) and Denison and Holmes (2001). Lawson (1995) proposes a point process model for detection of cluster locations when exact case (and control) locations are known. Lawson (2000) describes an extension of this model to incorporate both localized clustering and general spatial heterogeneity of disease rates. Lawson and Clark (1999) 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, typically by assuming a uniform spatial distribution within each cell, to produce a point process. One then proposes a clustering model for the point process.

Gangnon and Clayton (2000), Knorr-Held and Raßer (2000) and Denison and Holmes (2001) each consider a relatively nonparametric Bayesian framework for cluster detection in which cells are grouped into clusters. A single, common rate for cells belonging to the same cluster is assumed. The prior specification of Gangnon and Clayton (2000) assumes a large background area and a small number of clusters; the prior probability of a specific set of clusters is based on geographic characteristics of the clusters such as their size and shape. The prior specifications of Knorr-Held and Raßer (2000) and Denison and Holmes (2001) assume clusters are defined by a set of cells chosen as cluster centers (cells belong to the cluster associated with the nearest cluster center); the prior probability of a set of clusters is based on a uniform selection of each cell as a cluster center. All three methods provide very flexible specifications of clusters. The approaches of Knorr-Held and
Raßer (2000) and Denison and Holmes (2001) have some analytic advantages, while the approach of Gangnon and Clayton (2000) more directly models the prior probability of particular clusters. None of these methods include a spatial heterogeneity component in their model.

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 discrete spatial clustering component and a general spatial heterogeneity component to capture extra-Poisson variation. The model requires the specification of a set of potential clusters and a prior distribution on that set of potential clusters. The proposed 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 spatial clustering of disease rates, which includes spatially unstructured random effects to capture extra-Poisson variation in the rates. In Section 3, we describe our implementation of a reversible jump Markov chain Monte Carlo algorithm (RJMCMC) (Green, 1995) sampler for drawing inferences about the model. In Section 4, we analyze the aforementioned data on leukemia incidence in upstate New York using the proposed model. Finally, in Section 5, we close with a discussion of alternative model specifications and 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, i.e., $O_i \sim \text{Poisson}(\rho_i n_i)$, where $\rho_i$ is the disease rate in cell $i$.

We model the cell-specific disease rates using a log-linear model $\log(\rho_i) = \mu + \sum_{j=1}^{k} \theta_j I_{i \in c_j} + \epsilon_i$. There are three basic components in this model: a non-spatial component ($\mu$), a spatial clustering
component \(\sum_{j=1}^{k} \theta_j I(i \in c_j)\), and a spatial heterogeneity (or extra-Poisson variation) component \(\epsilon_i\). Our primary interest lies in a prior specification for the spatial clustering component of the model; fairly standard priors are available for the other two components.

In our development here, the non-spatial component of the model consists of a single parameter \(\mu\). 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 effects of covariates such as age and sex.

For the spatial heterogeneity effects \(\epsilon_i\), we follow other authors (cf. Waller et al. (1997)) in proposing an exchangeable normal prior for the \(\epsilon_i\)'s; that is, \(\epsilon_i \sim N(0, 1/\tau)\). For the parameter \(\tau\), we use a proper, but relatively weak, gamma prior distribution for \(\tau\). In Section 4, a gamma distribution with mean 1 and variance \(1/4\) is used.

The spatial clustering component of the model is based on the following parameters: \(k\), the number of clusters; \(c_1, c_2, \ldots, c_k\), the sets of cells belonging to the \(k\) clusters; and \(\theta_1, \theta_2, \ldots, \theta_k\), the log relative risks associated with each cluster. We develop a prior for the spatial clustering component of the model by successively conditioning on parameters. Given \(k, c_1, c_2, \ldots, c_k\) (i.e., the number of clusters and their locations), we assign an exchangeable normal prior for \(\theta_1, \theta_2, \ldots, \theta_k\); that is, \(\theta_j \sim N(0, \sigma_\theta^2)\). The prior variance \(\sigma_\theta^2\) must be chosen in advance. Given the relatively small number of clusters in most settings, the data will not provide enough information to reliably estimate \(\sigma_\theta^2\). In the example, we take \(\sigma_\theta^2\) to be 0.355 so that, a priori, the relative risk associated with a cluster falls between 0.25 and 4.00 with 99% probability.

Next, given \(k\) (i.e., the number of clusters), we select \(c_1, c_2, \ldots, c_k\) exchangeably (independently) from a prior distribution on the space of possible clusters; denote this distribution by \(p(c)\). Note that, under such a specification, one of the clusters may overlap or, in the extreme, even duplicate another cluster. To make this 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 (2001).

We consider circular clusters centered at the cell centroids as potential clusters. We center clusters at the centroids to avoid empty clusters. The radius of the circles varies continuously from zero up to a fixed maximum radius, $r_{\text{max}}$. 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 $0 = r_{i,(1)} < r_{i,(2)} < \ldots < r_{i,(m_i)} \leq r_{\text{max}}$ be the ordered distances from the centroid of cell $i$ to the centroids of all cells, truncated at $r_{\text{max}}$. (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 centered 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 cluster from the study region. Specifically, we first select a cluster center and then, conditional on that center, select a cluster radius. We first select a point from a uniform distribution over the study area and making the centroid of the cell to which the point belongs the cluster center. The radius of the circle is then selected at random from a uniform distribution on $[0, r_{\text{max}}]$. Thus, the prior probability of selecting cluster $i,j$ is

\[
p(i,j) = \frac{a_i}{A} \cdot \frac{r_{i,j+1} - r_{i,j}}{r_{\text{max}}},
\]

where $a_i$ is the area of cell $i$, $A$ is the area of the study region, and $r_{i,m_i+1} = r_{\text{max}}$.

Finally, we select a prior distribution for $k$, the number of clusters. One possibility would be distributions 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 $0, 1, 2, \ldots, k_{\text{max}}$. 

7
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 analyzing GLMs are discussed in Gelman et al. (1995), and we follow their approach. The normal prior distributions for $\mu$, $\theta_1, \theta_2, \ldots, \theta_k$, and $\epsilon_1, \epsilon_2, \ldots, \epsilon_N$ are conjugate to a normal approximation to the Poisson likelihood. For this normal approximation, we can easily find the full conditional distributions, and Gibbs sampling would be appropriate. To correct for the approximation, the Gibbs sampler is used as a proposal distribution in a Metropolis-Hastings algorithm (Hastings, 1970).

To make this discussion more concrete, we explicitly describe the Metropolis-Hastings steps for updating $\mu$. To propose a new value for $\mu$, we need $O_t$ (the total case count in the study region), $E_t$ (the current value for the expected number of cases in the entire study region), and $\mu_0$, the current value for $\mu$ in addition to the prior mean and variance for $\mu$, denoted by $\eta$ and $\sigma^2_{\mu}$. A proposed new value of $\mu$, denoted $\mu'$, is drawn from a normal distribution with mean

$$
\eta_p = \frac{E_t}{E_t + 1/\sigma^2_{\mu}} \mu_0 + \frac{1/\sigma^2_{\mu}}{E_t + 1/\sigma^2_{\mu}} \eta + \frac{O_t - E_t}{E_t + 1/\sigma^2_{\mu}}
$$

and variance

$$
\sigma^2_p = \frac{1}{E_t + 1/\sigma^2_{\mu}}.
$$
The new value $\mu'$ is accepted with probability

$$
\min \left\{ 1, \frac{\phi(\mu', \eta_p', \sigma_p^2) \phi(\mu', \eta, \sigma^2)}{\phi(\mu', \eta_p, \sigma_p^2) \phi(\mu, \eta, \sigma^2)} \frac{l(O_t, E_t')}{l(O_t, E_t)} \right\};
$$

otherwise, the current value $\mu$ is retained. In this equation, $\phi(\cdot; \mu, \sigma^2)$ is the density of a normal random variable with mean $\mu$ and variance $\sigma^2$ and $l(y, \mu)$ is the likelihood of a Poisson random variable with observed count $y$ and mean $\mu$. The definitions of Metropolis-Hastings steps for the other parameters $\theta_1, \theta_2, \ldots, \theta_k$ and $\epsilon_1, \epsilon_2, \ldots, \epsilon_N$ follow the same template.

The gamma prior distribution for $\tau$ (the inverse of the prior variance for $\epsilon_1, \epsilon_2, \ldots, \epsilon_N$) is also conjugate, so samples for $\tau$ can be obtained using the Gibbs sampler. To be concrete, if the prior distribution for $\tau$ follows a $\text{gamma}(a, b)$ distribution (mean $a/b$ and variance $a/b^2$), the full conditional distribution of $\tau$ is $\text{gamma}(a + N, b + \sum_{i=1}^{N} \epsilon_i)$.

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 (Green, 1995).

In addition to the steps for fixed clusters described above, we propose the following three steps.

1. ADD: Propose a new cluster $c_{k+1}$ and its associated parameter $\theta_{k+1}$ 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 $\theta$).

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 $c_{k+1}$. 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 $\hat{\theta}_c = (O_c - E_c)/(E_c + 1/\sigma^2_\theta)$, where $O_c$ is the number of cases in the cluster, $E_c$ is the current value for the expected number of cases in the cluster and $\sigma^2_\theta$ is the prior variance for $\theta$'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 $c$ with probability

$$p_{select}(c) = \frac{p(c)\phi(\hat{\theta}_c, 0, \sigma^2_\theta) l(O_c, e^{\hat{\theta}_c} E_c)}{\sum_c p(c)\phi(\theta_c, 0, \sigma^2_\theta) l(O_c, e^{\theta_c} E_c)}.$$ 

After the cluster $c_{k+1}$ is selected, a value for its associated log relative risk, $\theta_{k+1}$, is proposed using the normal approximation described earlier, i.e., sampled from a normal distribution with mean $\hat{\theta}_c$ and variance $1/(E_c + 1/\sigma^2_\theta)$.

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 $c = c_{k+1}$),

$$\min \left\{ 1, \frac{p_d(k+1)p(k+1)}{p_a(k)p(k)} \frac{1}{k+1} \frac{p(c)}{p_{select}(c)} \phi(\theta_{k+1}, 0, \sigma^2_\theta) \frac{l(O_c, e^{\theta_{k+1}} E_c)}{l(O_c, E_c)} \right\}.$$
For the DROP step (letting $c = c_k$),

$$
\min \left\{ 1, \frac{p_a(k-1)p(k-1)}{p_d(k)} \frac{p_{\text{select}}(c)}{p(c)} \frac{\phi(\theta_k, \hat{\theta}_c, 1/(E_c + 1/\sigma^2_\phi))}{\phi(\theta_k, 0, \sigma^2_\phi)} \frac{l(O_c,e^{-\theta_k}E_c)}{l(O_c,E_c)} \right\}.
$$

Note that, without loss of generality, we can assume the $k$th cluster is chosen to be dropped. If 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 $\theta$. Again, without loss of generality, we may assume cluster $k$ 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 $\theta_k$). The probability that cluster $c$ is selected as the new cluster $k$ is then given by

$$
p_{\text{select}}(c) = \frac{p(c)\phi(\hat{\theta}_k, 0, \sigma^2_\phi)l(O_c,e^{\theta_k}E_c)}{\sum_c p(c)\phi(\theta_k, 0, \sigma^2_\phi)l(O_c,e^{\theta_k}E_c)}.
$$

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

4 Example: New York Leukemia Data

We now present an example of the application of our methodology. The New York leukemia data set consists of data on leukemia 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 choice of leukemia was based on its “remarkably uniform” distribution (Turnbull et al., 1990).

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 leukemia 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 leukemia data is available in Waller et al. (1994) and Gangnon and Clayton (2000). The observed leukemia rate for each cell in Figure 1 using the Dirichlet tessellation of the cell centroids. No obvious clusters are evident in this figure. (Insertion point for Figure 1)

For our analysis of the New York leukemia data, we utilized the prior described in Section 2. Following Gelman and Rubin (1992), we ran five independent Markov chains. Each chain used a run-in of 10,000 iterations, and samples from the next 10,000 iterations were used for inference. The chains appeared to converge by that point, and there were not substantial differences in the samples across chains.

In Table 1, we present the posterior distribution of the number of clusters $k$ included in the model. Based on this distribution, there does not appear to be strong evidence about the correct number of clusters in the model. A model with no clusters has a posterior probability of 0.10. Higher posterior probabilities are associated with the one cluster (0.28), two cluster (0.32) and three cluster (0.19) models. The posterior probability of a model with more than three clusters is approximately 0.11. Thus, we have fairly strong evidence of clustering in the data, but equivocal evidence for the correct number of clusters (1, 2, or 3).

In Figure 2, we display the posterior means for the cluster risks associated with each cell $\exp(\sum_{j=1}^{k} \theta_j I_{i \in c_j})$, separately for each value of the number of clusters $k = 1, 2, \ldots, 7$ and averaged across these models ($k > 0$). In Figure 3, we display the posterior probability that a cell belongs to a cluster $\Pr(\sum_{j=1}^{k} \theta_j I_{i \in c_j} \neq 0)$, separately for each value of the number of clusters $k = 1, 2, \ldots, 7$ and averaged across these models ($k > 0$).

These figures show convincing evidence for three areas of clustering in the New York leukemia data. The term "areas of clustering" is used instead of "clusters" to indicate that the data support
<table>
<thead>
<tr>
<th>$k$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flat prior</td>
<td>0.10</td>
<td>0.28</td>
<td>0.32</td>
<td>0.19</td>
<td>0.08</td>
<td>0.02</td>
<td>0.01</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Geometric prior</td>
<td>0.29</td>
<td>0.40</td>
<td>0.22</td>
<td>0.07</td>
<td>0.01</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
<td>&lt; 0.01</td>
</tr>
</tbody>
</table>

Table 1: Posterior distribution of the number of clusters $k$ based on 50,000 Markov chain Monte Carlo simulations. The first line is the posterior distribution based on a flat prior for $k$. The second line is the posterior distribution using a geometric distribution with success probability $1/2$ as a prior for $k$. 
Figure 2: Posterior mean of the cluster risk associated with each cell $\exp(\sum_{j=1}^{k} \theta_j I_{i \in C_j})$, separately for each value of the number of clusters $k = 1, 2, \ldots, 7$ and averaged across these models ($k > 0$) based on 50,000 Markov chain Monte Carlo simulations.
Figure 3: Posterior probability that each cell belongs to a cluster $\Pr(\sum_{j=1}^{k} \theta_j I_{i \in C_j} \neq 0)$, separately for each value of the number of clusters $k = 1, 2, \ldots, 7$ and averaged across these models ($k > 0$) based on 50,000 Markov chain Monte Carlo simulations.
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 leukemia risk. This area includes the city of Binghamton. The second area of clustering is located in Cortland County in the center of the study region and is also associated with an increased leukemia risk. The third area of clustering is located in Onondaga County, north of Syracuse, and is associated with a decreased risk of leukemia. As the number of clusters in the model increases, the estimated risks or benefits associated with these clusters increases.

To summarize the risks associated with these apparent clusters, we use the posterior expected risk associated with the cell in each area with the largest estimated probability of belonging to a cluster. The estimated risk associated with the area of clustering in Cortland County is 1.40. The estimated risk associated with the area of clustering in Broome County is 1.19. The estimated risk associated with the area of clustering in Onondaga County is 0.81.

In examining these figures, we observe that even models with only one or two clusters show evidence of three areas of clustering in the study region. We believe that the ability of these apparently under-parameterized models to identify the effects of multiple clusters accounts for the difficulty in formally identifying the number of clusters through its posterior distribution. For example, a model assigning posterior probability of 0.50 to each of two single cluster models may provide a very good approximation to a two cluster model.

To this point, we have only considered the clustering component of our model. We now examine the spatial heterogeneity component of the model ($\epsilon_i$). The posterior expected value of $\tau$ across all values of $k$ is 0.28 with a 95% posterior probability interval of (0.20, 0.39). For $k = 0$, the posterior expected value of $\tau$ is slightly larger (0.29); as the value of $k$ increases, the posterior expected value of $\tau$ decreases (to roughly 0.27 for $k \geq 3$), indicating that the clustering component of the model is explaining some of the apparent heterogeneity in disease rates. However, a large portion of the variation in disease rates remains unexplained.
In Figure 3, we present the posterior means for the disease rate in each cell, first without a clustering component ($k = 0$) and second with an unspecified number of clusters ($k > 0$). Although the two maps seem almost indistinguishable at a glance, some evidence of the three areas of clustering is apparent upon study. For the representative cell in Cortland County, the posterior mean rate is 13.7 per 10,000 persons with a clustering component versus 11.4 per 10,000 persons without a clustering component. For Broome County, the rates are 7.1 per 10,000 and 6.6 per 10,000, respectively. For Onondaga County, the rates are 3.6 per 10,000 and 4.5 per 10,000 respectively.

In truth, a flat 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 flat prior provide an importance sample for the posterior based on the geometric prior; the importance sampling weights for a model with $k$ clusters is proportional to $0.5^k$. The posterior for $k$ based on this second prior is provided in Table 1. Compared with the posterior based on a flat prior, this distribution is shifted substantially towards models with $k = 0, 1$ or 2. There is little support for a model with $k > 3$ (posterior probability < 0.10).

Despite this shift in the posterior for $k$, the resulting posterior for the cluster risks (and associated probabilities of belonging to a cluster) still shows evidence of the three areas of clustering. Under this posterior, the estimated risks associated with the representative cells described above are 1.31 in Cortland County, 1.14 in Broome County and 0.83 in Onondaga County. This again demonstrates the ability of the one and two cluster models to capture, at least partially, the risks associated with three clusters.

Many previous analyses of the New York leukemia data have been published. Most of the previous analyses have been based on hypothesis testing methods and solely aimed at detecting a single cluster with an elevated risk of disease. These methods have generally detected clustering
Figure 4: Posterior mean of the disease rate associated in each cell, first without a clustering component ($k = 0$) and second with an unspecified number of clusters ($k > 0$) based on 50,000 Markov chain Monte Carlo simulations.
in either Broome County or Cortland County (Waller et al., 1994). Some methods such as that of Kulldorff and Nagarwalla (1995) 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 leukemia data was described by Gangnon and Clayton (2000). 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 parameter $k$) and the ability to incorporate spatial heterogeneity effects into the model.

Gangnon and Clayton (2000) found evidence for three clusters associated with an increased risk of leukemia: 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 (2000) places relatively larger weight on the many small, overlapping clusters within Syracuse than the more uniform prior used in our analysis. A recent analysis by Denison and Holmes (2001) produced an estimated risk surface that shows apparent evidence for all four features described above. They found compelling evidence for elevated leukemia 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 spatial clustering and spatial heterogeneity effects in 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 for heterogeneity effects incorporates non-localized extra-Poisson variation in disease rates.
In addition to formal posterior inference on the number of clusters, the explanatory power of the clustering effects can be assessed by the percent reduction in the variance of the heterogeneity effects associated with their inclusion. We conclude by briefly commenting on two extensions of this work.

In our presentation, we have focused on a “flat prior” for the clusters. Likewise, inference about the possibility of certain prespecified clusters can be evaluated using a “flat” prior for the clusters. On the other hand, prior knowledge of cluster locations can be incorporated into these models in one of two ways. 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 \( k_{\text{max}} \). 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 \( \mu \) with the linear predictor \( \mu + \beta'x \). One could also easily extend the model to incorporate interactions between the covariate effects and the clusters.

REFERENCES


WHO.


