BIOMETRICS 56, 922–935
September 2000

# Bayesian Detection and Modeling of Spatial Disease Clustering

**Ronald E. Gangnon**

Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison,
1300 University Avenue, Madison, Wisconsin 53706, U.S.A.
*email:* ronald@biostat.wisc.edu

**and**

**Murray K. Clayton**

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

SUMMARY. Many current statistical methods for disease clustering studies are based on a hypothesis testing paradigm. These methods typically do not produce useful estimates of disease rates or cluster risks. In this paper, we develop a Bayesian procedure for drawing inferences about specific models for spatial clustering. The proposed methodology incorporates ideas from image analysis, from Bayesian model averaging, and from model selection. With our approach, we obtain estimates for disease rates and allow for greater flexibility in both the type of clusters and the number of clusters that may be considered. We illustrate the proposed procedure through simulation studies and an analysis of the well-known New York leukemia data.

KEY WORDS: Bayesian model averaging; Disease clustering; Leukemia; Markov connected component field; Monte Carlo; Poisson; Spatial pattern.

## 1. Introduction

The study of disease clustering is frequently of interest to epidemiologists, statisticians, and the general public. As a statistical problem, disease clustering studies have typically been approached as hypothesis testing problems. The null hypothesis of no clustering, i.e., a common rate of disease across the study region, is fairly clear. The alternative hypothesis of clustering is less well defined. One common definition of clustering is that, when it occurs, cases are closer to other cases than cases are to noncases. This type of clustering can be detected with statistics that measure the average distance between cases. Examples of these distance-based statistics are the studies by Whittemore et al. (1987), Cuzick and Edwards (1990), Ross and Davis (1990), and Selvin, Schulman, and Merrill (1992). Clustering can also be defined as an elevated rate of disease in a small portion of the study area, which is then called the cluster. If the location of a potential cluster is prespecified, models for the clustering process can be developed and cluster risks can be estimated as in Stone (1988), Diggle (1990), and Waller et al. (1992).

In practice, the location of the potential cluster often cannot be specified in advance, and the goal of the study is to determine whether the disease rate is elevated in one of a large number of potential clusters. Openshaw et al. (1988) proposed the geographical analysis machine (GAM) as an exploratory cluster detection method. Turnbull et al. (1990) and Besag and Newell (1991) proposed statistically rigorous alternatives to the GAM based on circles of fixed population radius and circles of fixed case radius, respectively. Kulldorff and Nagarwalla (1995) generalize the previous procedures to arbitrary collections of clusters using likelihood ratio tests.

A well-known data set frequently used to evaluate cluster detection procedures consists 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 one 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 both their own method and the methods of Whittemore et al. (1987), Openshaw et al. (1988), and Turnbull et al. (1990). They concluded that these cluster detection methods do not demonstrate strong evidence of clustering in the New York leukemia data. Kulldorff and Nagarwalla (1995) later analyzed the New York leukemia data using their method and found strong evidence ($p$-value $< 0.0001$) of a cluster in Broome County.

The techniques described above leave several important questions unanswered. They are designed to detect a single cluster; there is no formal assessment of clustering in multiple locations. For example, Kulldorff and Nagarwalla (1995) found evidence suggestive ($p = 0.026$) of a cluster in Cortland County, but they could not formally assess the significance of additional clusters with their single-cluster model. Also, these methods require a specific set of potential clusters be speci-

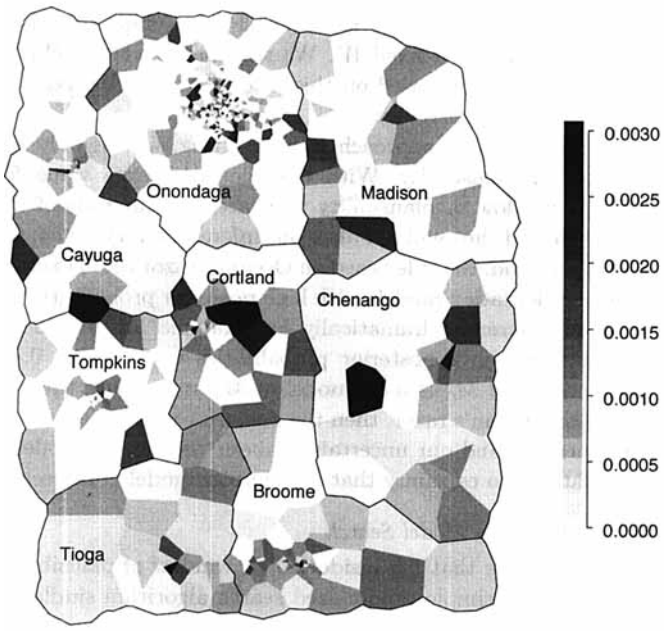


**Figure 1.** Observed cell-specific rates for the New York leukemia data. Region associated with each cell based on Dirichlet tessellation of cell centroids.

fied in advance, e.g., circular clusters centered at the cell centroids. In addition, estimation of disease rates is conducted conditional on the estimated cluster. Such conditional estimates may not accurately reflect the uncertainty about the composition of the cluster.

In this paper, we develop a Bayesian approach to inference about the parameters of a simple model for spatial clustering suitable for analyzing the New York leukemia data. We do not require cluster locations, shapes, or boundaries to be specified. Instead, we require some knowledge of the relative likelihood of various cluster sizes and shapes. This approach allows for multiple clusters and produces estimates of cell-specific rates that reflect the uncertainty in cluster memberships.

An alternative Bayesian approach to the disease clustering problem includes the work of Lawson (1995), who proposes a point process model for detection of cluster locations when exact case locations are known, and of Lawson and Clark (1999), who describe the application of the point process clustering model to cell count data using data augmentation.

Other Bayesian approaches to analyzing spatial disease patterns focus on estimating spatially smoothed disease rates suitable for mapping. Examples of these approaches include Clayton and Kaldor (1987), Bersag, York, and Mollié (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 clustering.

In Section 2, we develop simple models for clustering and suggest a class of Markov connected component field (MCCF) priors for these cluster models. We propose a method for approximating the posterior distribution over cluster models using a randomized variant of backwards elimination to find models with high posterior densities in Section 3. In Section 4, we present our analysis of the New York leukemia data. We present simulation results to illustrate the performance of our proposed inference scheme with known clustering models in Section 5. In Section 6, we provide some concluding remarks.

## 2. Statistical Model

Consider a study region divided into $N$ subregions, or cells. A cell is typically a small geopolitical subregion such as a census tract or block. 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 are interested in drawing inferences about the underlying disease rates $r_i$, $i = 1, 2, \ldots, N$. Although the $O_i$ could be modeled as binomial observations, we assume that the disease is rare and invoke the Poisson approximation, i.e., $O_i \sim \text{Poisson}(r_i n_i)$. The Poisson model also allows for the inclusion of covariate effects by replacing $n_i$ with the expected case count $E_i$ (a situation we do not consider here).

To draw inferences about $\mathbf{r} = (r_1, r_2, \ldots, r_N)$, we propose a Bayesian approach with a hierarchical prior. First, divide the cells into $k + 1$ groups, or components. Call one of these components the background and the other $k$ components clusters. We identify a cluster model with $k$ clusters by its vector of cluster memberships, $\mathbf{c} = (c_1, c_2, \ldots, c_N)$, where $c_i = 0$ if cell $i$ belongs to the background and $c_i = j$ if cell $i$ belongs to cluster $j$, $j = 1, 2, \ldots, k$. The labeling of clusters is unimportant, so we constrain $\mathbf{c}$ to obtain a unique representation for the model: the first nonzero element of $\mathbf{c}$ must be a one, the first element of $\mathbf{c}$ greater than one must be a two, etc. Given a specific cluster model $\mathbf{c}$, we assume that cells belonging to component $j$ share a common disease rate $\lambda_j$ and let $\Lambda = (\lambda_0, \lambda_1, \ldots, \lambda_k)$.

To construct our model, we temporarily assume that $\mathbf{c}$ is known. Given the arbitrary labeling of the clusters, we use an exchangeable prior for $\lambda_1, \lambda_2, \ldots, \lambda_k$. More specifically, we take $\lambda_0 \mid \mathbf{c} \sim \text{gamma}(\alpha_0, \beta_0)$ and $\lambda_j \mid \mathbf{c} \sim \text{gamma}(\alpha, \beta), j = 1, 2, \ldots, k$, where $\text{gamma}(a, b)$ is the gamma distribution with mean $a/b$ and variance $a/b^2$.

Given $\mathbf{O}$ and $\mathbf{c}$, $\lambda_0, \lambda_1, \ldots, \lambda_k$ are independent, $\lambda_0 \mid \mathbf{c}, \mathbf{O} \sim \text{gamma}(\alpha_0 + O_{\cdot 0}, \beta_0 + n_{\cdot 0})$, and $\lambda_j \mid \mathbf{c}, \mathbf{O} \sim \text{gamma}(\alpha + O_{\cdot j}, \beta + n_{\cdot j}), j = 1, 2, \ldots, k$, where $O_{\cdot j} = \Sigma_{i=1}^{N} O_i I_{\{c_i = j\}}$ is the total number of cases in cluster $j$, $n_{\cdot j} = \Sigma_{i=1}^{N} n_i I_{\{c_i = j\}}$ is the total population at risk in cluster $j$, and $I_{\{c_i = j\}} = 1$ if $c_i = j$ and $I_{\{c_i = j\}} = 0$ if $c_i \neq j$. The marginal likelihood of $\mathbf{O}$ given $\mathbf{c}$ is

$$p(\mathbf{O} \mid \mathbf{c}) = \frac{\Gamma(\alpha_0 + O_{\cdot 0})}{\Gamma(\alpha_0)} \cdot \frac{\beta_0^{\alpha_0}}{(\beta_0 + n_{\cdot 0})^{\alpha_0 + O_{\cdot 0}}} \times \prod_{j=1}^{k} \frac{\Gamma(\alpha + O_{\cdot j})}{\Gamma(\alpha)} \cdot \frac{\beta^{\alpha}}{(\beta + n_{\cdot j})^{\alpha + O_{\cdot j}}} \cdot \prod_{i=1}^{N} \frac{n_i^{O_i}}{O_i!}. \quad (1)$$

This closed-form expression for the marginal likelihood allows us to avoid explicit consideration of the disease rates in calculating the posterior distribution over cluster models.

### 2.1 *Prior for Cluster Models*

Meaningful inference about cluster models will not be possible without some prior information about clusters. Even frequentist approaches to this problem incorporate a form of prior information through constraints on the sizes and shapes of the clusters considered. We propose using a flexible family of distributions to formalize the prior information about

clusters. From a frequentist point of view, this proposed prior can be viewed as a penalty on cluster size and shape as well as on model complexity.

Our prior formulation is a mixture of point mass on the null (no clustering) model $M_0$ and on a Markov connected component field (MCCF) (Møller and Waagepetersen, 1998) $M_1$. Including point mass on the null model allows for the strong prior belief in the null model typical of many clustering studies. Under an MCCF, the probability of $\mathbf{c}$, a model with $k$ clusters, is given by $p(\mathbf{c}) \propto \exp(-\Sigma_{j=1}^{k} S_j)$, where $S_j$ is a score for cluster $j$ dependent solely on the properties of cluster $j$.

In our example, we construct priors based on four notions: (1) clustering is no more likely in one location than another, (2) the presence of a cluster is less likely than its absence, (3) a small cluster is more likely than a large cluster, and (4) a more circular cluster is more likely than an irregular cluster. These prior beliefs are strictly enforced in many previously published cluster detection approaches (e.g., through the use of circular clusters with fixed radii). With our Bayesian approach, we can potentially detect any type of cluster; it is simply easier to detect *a priori* likely clusters.

We define the size and shape of a cluster using its area $A$ and perimeter $P$ as follows. First, transform $(A, P)$ to $(R_1, R_2) = ((A/\pi)^{1/2}, P/(2\pi))$. Noting that $R_1 \leq R_2$ for any region and that $R_1 = R_2$ if and only if the region is a circle, we use $\varsigma = R_1/R_2$ and $\rho = R_2$ as measures of cluster shape and size. We score each cluster with $S(\rho, \varsigma) = \alpha + S_1(\rho) + S_2(\varsigma)$, where $\alpha$ is a positive constant, $S_1$ is a nondecreasing function of the cluster size $\rho$ with $S_1(0) = 0$, and $S_2$ is a nonincreasing function of the cluster shape $\varsigma$ with $S_2(1) = 0$.

Different scoring functions for clusters could be used to reflect different prior beliefs about both the types of clusters. In practice, specification of the appropriate MCCF prior for a particular setting will require prior knowledge of and expert opinion about spatial patterns of the disease in question and of possible risk factors. In addition, if specific levels of risk are of concern (e.g., a rate ratio of 2.0), evaluations of the marginal likelihood (equation (1)) for an array of hypothetical clusters could be useful in guiding the selection of the MCCF prior.

## 3. Posterior Calculation

### 3.1 *The Window of Plausibility*

Given the large number of potential cluster models, we cannot, in practice, directly evaluate the desired posterior. Instead, we propose using a simple algorithm to calculate an approximation to the posterior. The first component of this algorithm is the window of plausibility, an adaptation of the Occam's window approach to model selection (Madigan and Raftery, 1994). Madigan and Raftery argue that many models are no longer plausible given the observed data and that these implausible models can safely be removed from consideration when calculating the posterior.

Using Occam's window, one determines whether a model remains plausible using two rules. First, using Occam's razor, exclude a model $M$ from further consideration if there exists a simpler submodel $M_s$ of $M$ with $\Pr(M_s \mid \text{data}) > \Pr(M \mid \text{data})$. Second, exclude a model $M$ from consideration if there exists another model $M_m$ such that $\Pr(M \mid \text{data})/\Pr(M_m \mid \text{data}) < 1/W$ for a fixed $W$. We say that models excluded from consideration based on this second rule fall outside the $W$ window of plausibility.

In adapting this approach to our setting, we use only the window of plausibility. With a sufficiently large value for $W$, the window of plausibility truncates the far tails of the posterior and thus will not alter our inferences very much. On the other hand, the rule based on Occam's razor could exclude from consideration models with high posterior probability and alter our inferences dramatically. For example, let two models $M_a$ and $M_b$ have posterior probabilities of 0.51 and 0.49, respectively. If $M_a$ is a submodel of $M_b$ and we use the rule based on Occam's razor, then the posterior probability of $M_a$ becomes one and our uncertainty about the correct model is translated into certainty that the simpler model is correct.

### 3.2 *Randomized Model Search*

To find models that fall inside the $W$ window of plausibility. we propose a simple randomized search algorithm similar to the backwards elimination methods used for variable selection in regression problems. In our approach, we start with a saturated model, i.e., a model with $N - 1$ clusters, and repeatedly merge adjacent components of the current model to produce models with high posterior densities.

Without enumerating all possible sequences of mergers, we will not know with certainty which mergers lead to better models. To address this, let $\mathbf{c}$ be the current model with $k$ clusters, let $i$ and $j$ be adjacent components of $\mathbf{c}$, and let $\mathbf{c}_{ij}$ be the model obtained by merging $i$ and $j$. We will assume that the posterior density of $\mathbf{c}_{ij}$ is a good proxy measure of the likelihood that the merger of $i$ and $j$ leads to good models. To justify this, we note that the posterior density of $\mathbf{c}_{ij}$ measures the posterior likelihood that the true disease rates for $i$ and $j$ are the same, i.e., that the cells in components $i$ and $j$ belong to the same component in the true cluster model.

In practice, we select a merger using the following three steps. First, for each pair $(i, j), i < j$, of adjacent components of $\mathbf{c}$, let $\mathcal{M}_{ij} = p(\mathbf{c}_{ij} \mid \mathbf{O})/p(\mathbf{c} \mid \mathbf{O})$. Note that $\mathcal{M}_{ij}$ depends only on components $i$ and $j$ and not on the models $\mathbf{c}$ and $\mathbf{c}_{ij}$. Thus, after a merger, we need only update $\mathcal{M}_{ij}$ if component $i$ or $j$ is merged with another component. Next, calculate the probability of selecting merger $(i, j)$, $\mathcal{P}_{ij}$, by truncating the $\mathcal{M}_{ij}$ with a $W'$ window of plausibility, i.e., $\mathcal{P}_{ij} \propto \mathcal{M}_{ij}$ if $\mathcal{M}_{ij}/\max_{(i,j)} \mathcal{M}_{ij} \geq 1/W'$ and $\mathcal{P}_{ij} = 0$ otherwise. Finally, merge components $i$ and $j$ of $\mathbf{c}$ and select the new model $\mathbf{c}_{ij}$ with probability $\mathcal{P}_{ij}$. Applying a window to calculate the merger probabilities can prevent many poor mergers from overwhelming a few promising mergers and thus can speed the search process.

The model search consists of repeatedly applying the above merger selection step. To speed up the search, we reduce its scope. We build smaller, but still implausibly large, models called base models or bases. With repeated searches from a smaller base, we can examine many more plausible models in a fixed time. We next outline the steps of a search incorporating bases of one size. The extension to multiple base sizes is straightforward.

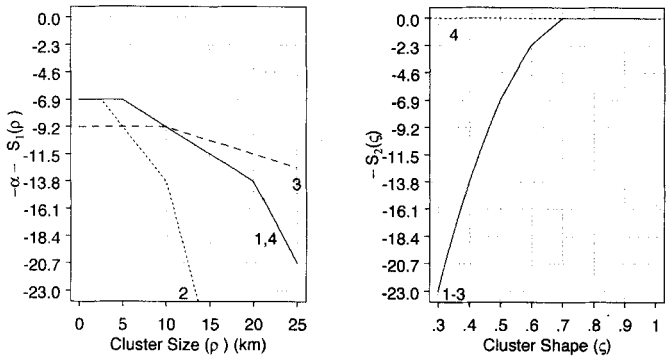


**Figure 2.** Cluster size and cluster shape components of four MCCF priors used in the example and simulation study. Number(s) adjacent to each curve indicate the prior(s) that incorporate that function.

(1) Start with a saturated model with $N - 1$ clusters, denoted $\mathbf{c}_s$, obtained by selecting one of the $N$ cells at random to act as the background. We have found that random selection works well in practice, while data-driven selections do not.
(2) Repeatedly merge adjacent components of $\mathbf{c}_s$ to produce a model with $k_b$ clusters, denoted $\mathbf{c}_b$.
(3) Starting from the base cluster model $\mathbf{c}_b$, repeatedly merge adjacent components to produce a nested series of cluster models with $k_b - 1, k_b - 2, \ldots, 0$ clusters.
(4) Update the current list of plausible cluster models, i.e., models falling within the $W'$ window of plausibility, to reflect the models found in the latest search.
(5) Repeat steps 3 and 4 until a stopping rule is satisfied.
(6) Repeat steps 1–5 until a stopping rule is satisfied.

Possible stopping rules include stopping after a fixed number of iterations or stopping after some number of consecutive failures to find new plausible models. We use an adaptive stopping rule based on sequential determination of whether the success probability $\theta$ in a sequence of Bernoulli trials is zero. With a nontrivial prior and loss function, the optimal Bayes sequential rule is to stop after observing the first success (in this case, we begin a new sequence of searches with a different $\theta$) or when the posterior mean for $\theta$ is less than the cost of the next observation (in this case, we stop the search) (Gangnon, 1998). With good prior choices, we can quickly abandon poor bases and exhaustively sample good bases.

### 3.3 *Approximate Marginal Likelihood of MCCF Model*

In addition to the approximate posterior for the MCCF model, we need the marginal likelihood of the data $\mathbf{O}$ given the MCCF model in order to calculate the posterior for the mixture model. The marginal likelihood for $M_0$ (the null model) is given by equation (1) with $\mathbf{c} = (0, 0, \ldots, 0)$. To estimate the marginal likelihood for $M_1$ (the MCCF model), we express the marginal likelihood as $\Pr(\mathbf{O} \mid M_1) = [\mathrm{E}_{\mathbf{c}|\mathbf{O},M_1}\{\Pr(\mathbf{O} \mid \mathbf{c}, M_1)^{-1}\}]^{-1}$. This expression is implicit in an importance sampling formula given by Newton and Raftery (1994). Applying this result to our approximate posterior provides an estimate of $\Pr(M_0 \mid \mathbf{O})$.

Using this posterior distribution, we can estimate various cell-specific quantities. For example, a reasonable estimate for the cell-specific disease rates $\mathbf{r}$ is its posterior mean $\hat{r_i} = \mathrm{E}(r_i \mid \mathbf{O}) = \Sigma_{\mathbf{c}} \mathrm{E}(\lambda_{c_i} \mid \mathbf{O}, \mathbf{c}) \cdot \Pr(\mathbf{c} \mid \mathbf{O}), i = 1, 2, \ldots, N$. In contrast to estimates based on a single cluster model, the posterior means smooth the edges of clusters,

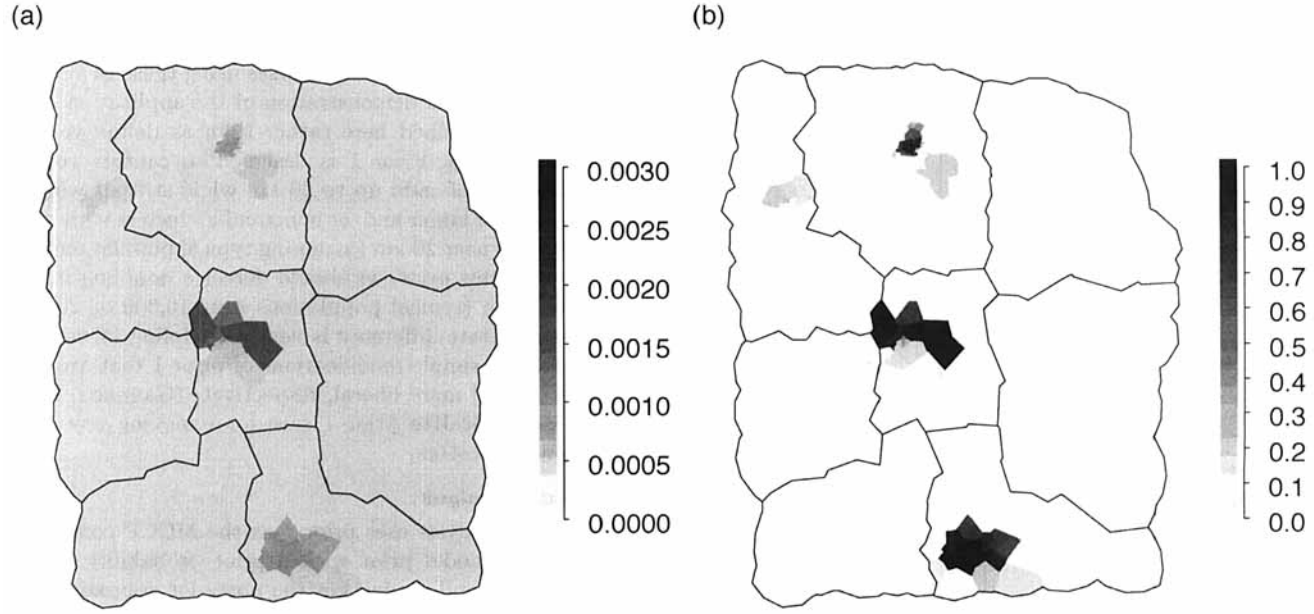


**Figure 3.** Analysis of the New York leukemia data using MCCF prior 1, a gamma(0.739, 1339) prior for the background rate, a gamma(1.478, 1339) prior for cluster rate(s), and prior probability of 0.99 on the null model. Posterior mean cell-specific leukemia rate provided in left panel, and posterior cell-specific cluster membership probability provided in right panel.

**Table 1**

*Summary information from analyses of the New York leukemia data using various MCCF priors in Figure 2. For each analysis, the posterior disease rate (per 10,000 persons) is presented for a pure background cell (posterior cluster membership probability of 0). Areas of clustering evident in at least one analysis (posterior probability greater than 0.01) are labeled by county. Within Onondaga County, potential clusters are labeled by relative position within the county. For each analysis, the posterior mean disease rate and posterior cluster membership probability are presented for the cell(s) in that area with the highest posterior cluster membership probability. If that probability does not exceed 0.01, the possible cluster location is considered to be unseen.*

| | Prior 1 | | Prior 1 replicate analysis | | Prior 2 | | Prior 3 | | Prior 4 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Disease rate (per 10,000) | Posterior probability | Disease rate (per 10,000) | Posterior probability | Disease rate (per 10,000) | Posterior probability | Disease rate (per 10,000) | Posterior probability | Disease rate (per 10,000) | Posterior probability |
| Background | 4.42 | | 4.42 | | 4.99 | | 6.61 | | 4.16 | |
| Possible cluster locations | | | | | | | | | | |
| Broome | 9.58 | 1.00 | 9.58 | 1.00 | 9.97 | 0.93 | NS[a] | NS | 10.35 | 1.00 |
| Cortland | 16.66 | 0.98 | 18.82 | 0.99 | 10.91 | 0.40 | 7.09 | 0.03 | 16.54 | 1.00 |
| North Central Onondaga | 10.05 | 0.96 | 10.20 | 0.98 | 10.98 | 0.69 | NS | NS | 12.24 | 1.00 |
| Cayuga | 5.85 | 0.21 | 6.72 | 0.26 | NS | NS | NS | NS | 8.72 | 0.50 |
| Eastern Onondaga | 5.69 | 0.19 | 6.71 | 0.25 | NS | NS | NS | NS | NS | NS |
| Onondaga/Madison | NS | NS | NS | NS | NS | NS | 2.30 | 1.00 | NS | NS |

[a] NS, not seen.

reflecting the uncertainty in cluster memberships. A grayscale map of these estimates effectively demonstrates both cluster locations and cluster risks. Cluster borders (and uncertainty about cluster borders) are more easily seen if we map the posterior probability that the cell belongs to a cluster, $\Pr(c_i > 0 \mid \mathbf{O}) = \Sigma_{\mathbf{c}:c_i>0} \Pr(\mathbf{c} \mid \mathbf{O}), i = 1, 2, \ldots, N$.

## 4. Example: New York Leukemia Data

We now present an example of the application of our methodology to the New York leukemia data. As mentioned in Section 1, 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 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. We note that two of the original 790 cells have the same centroid (to two decimal places). These two cells were pooled for analysis and display since, for our purposes, they cannot be distinguished.

The New York leukemia data set does not include cell areas and border lengths. Using the cell centroids, we imputed cell areas and borders using a Dirichlet tessellation (Sibson, 1980). We constructed the Dirichlet tessellation using the `S-plus` function `deldir` written by Rolf Turner and available from StatLib. Using this tessellation, we display the observed leukemia rate for each cell in Figure 1. No obvious clusters are evident in this figure.

For our analysis of the New York leukemia data, we use four MCCF priors on cluster models of the form described in Section 2 (see Figure 2). Our analyses using these MCCF priors are intended as a demonstration of the application of the methodology described here rather than as definitive analyses of these data. Prior 1 is designed to capture roughly circular clusters of radii up to 20 km while still allowing for the possibility of larger and/or noncircular clusters with larger risks. For radii near 20 km (assuming typical populations near 90,000), with this prior, we should detect a doubling in risk; at smaller radii (typical populations near 10,000 or 20,000), the detectable rate difference is nearly a tripling in risk. Priors 2 and 3 are simple modifications of prior 1 that are more conservative and more liberal, respectively (Gangnon, 1998). Prior 4 is identical to prior 1 save for removing any shape restriction on clusters.

### 4.1 *Primary Analysis*

Our primary analysis uses prior 1 as the MCCF component of the cluster model prior with a prior probability of 0.99 assigned to the null model. For the posterior approximation, we used a $W = 1000$ window of plausibility; larger values of $W$ produced too many plausible models. For the search, we used two base sizes, 100 clusters (built using $W = 10$) and 11 clusters (built using $W = 100$). Additional details about the search procedure are available in Gangnon (1998).

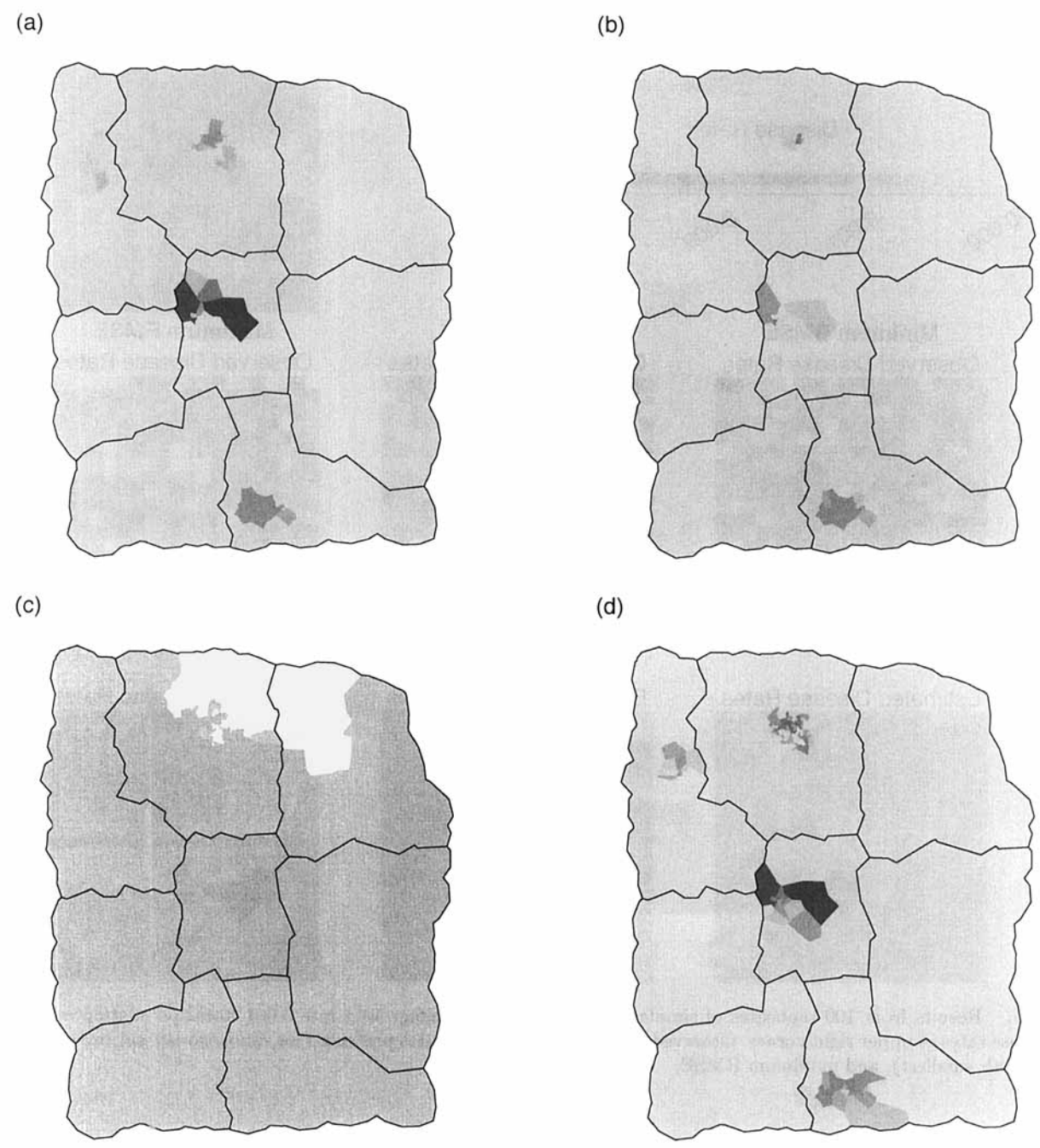


**Figure 4.** Posterior mean cell-specific leukemia rates from four additional analyses of the New York leukemia data. (**a**) Replication of the analysis using MCCF prior 1 presented in Figure 3. (**b**) An analysis using the more conservative MCCF prior 2. (**c**) An analysis using the more liberal MCCF prior 3. (**d**) An analysis using the MCCF prior 4, which does not penalize cluster shape.

In Figure 3, we display the estimated (posterior mean) disease rate for each cell, i.e., the mean cell-specific disease rate from the composite posterior, and the estimated posterior probability that each cell belongs to a cluster. The evidence in favor of the clustering model is overwhelming; the posterior probability of the null model is $1.23 \times 10^{-8}$. Given this support in the data for clustering, the prior probability assigned to the null model is effectively irrelevant. In Figure 3, we observe clear evidence for three areas of clustering—in Broome County, Cortland County, and (north central) Onondaga County. Disease rates and posterior probabilities associated with these clusters as well as potential clusters in Cayuga County and in (eastern) Onondaga County are presented in Table 1. The evidence for clustering is strongest in Broome

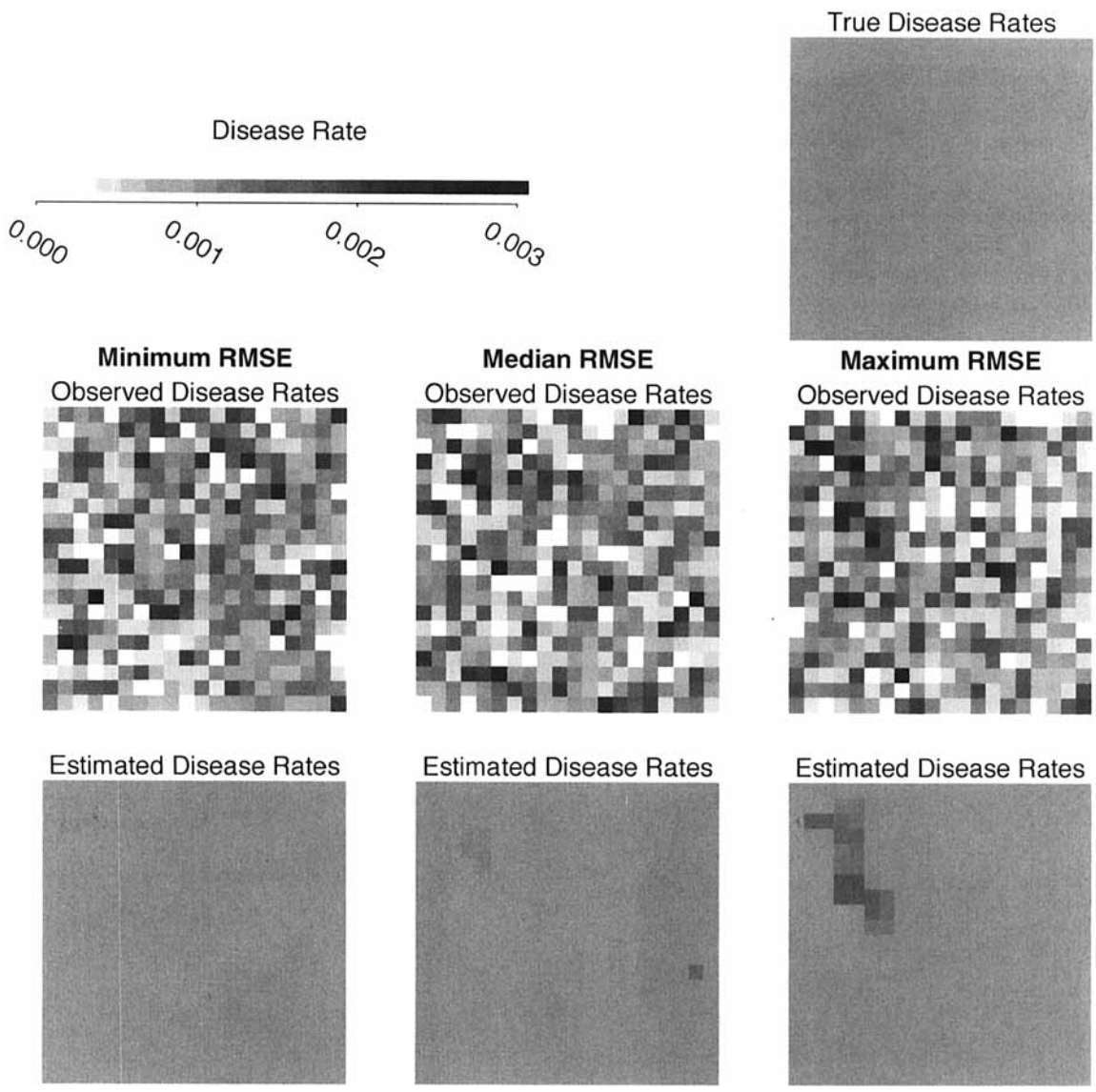


**Figure 5.** Results from 100 replicates of simulation 1—null model, background rate 0.001, total population $\sim$ 1 million. True disease rates in upper right corner. Observed and estimated disease rates presented for simulations resulting in minimum, median (50th smallest), and maximum RMSE.

County, and that cluster involves the largest population. The highest risk is associated with the cluster in Cortland County.

In the analyses reported by Waller et al. (1994) using both general and focused methods and Kulldorff and Nagarwalla (1995), there is some evidence for the clusters in Broome County and Cortland County. Previous analyses do not find strong evidence for the third cluster in Onondaga County. Our procedure finds strong evidence for three clusters because our model incorporates multiple clusters simultaneously, a feature not present in most previously described cluster detection techniques.

### 4.2 *Supplementary Analyses*

In Figure 4 and Table 1, we present the estimated rates from four additional analyses of the New York leukemia data. Figure 4a shows a reanalysis of the New York leukemia data using the same prior as before and a different random number seed for the model search. There is very little, if any, overlap between the sample of 2183 models found here and the sample of 4588 models found in the primary analysis. Despite this, the two approximate posteriors are quite similar, suggesting that our procedure is robust.

Figure 4b displays the results from an analysis using prior 2, a more conservative prior emphasizing small clusters. Under this prior, there is still evidence of clustering, but it is less overwhelming than in the primary analysis. The posterior probability of the null model is now 0.021, and there is strong evidence for the Broome County cluster. The presence of clusters in (north central) Onondaga County and in Cortland County is uncertain.

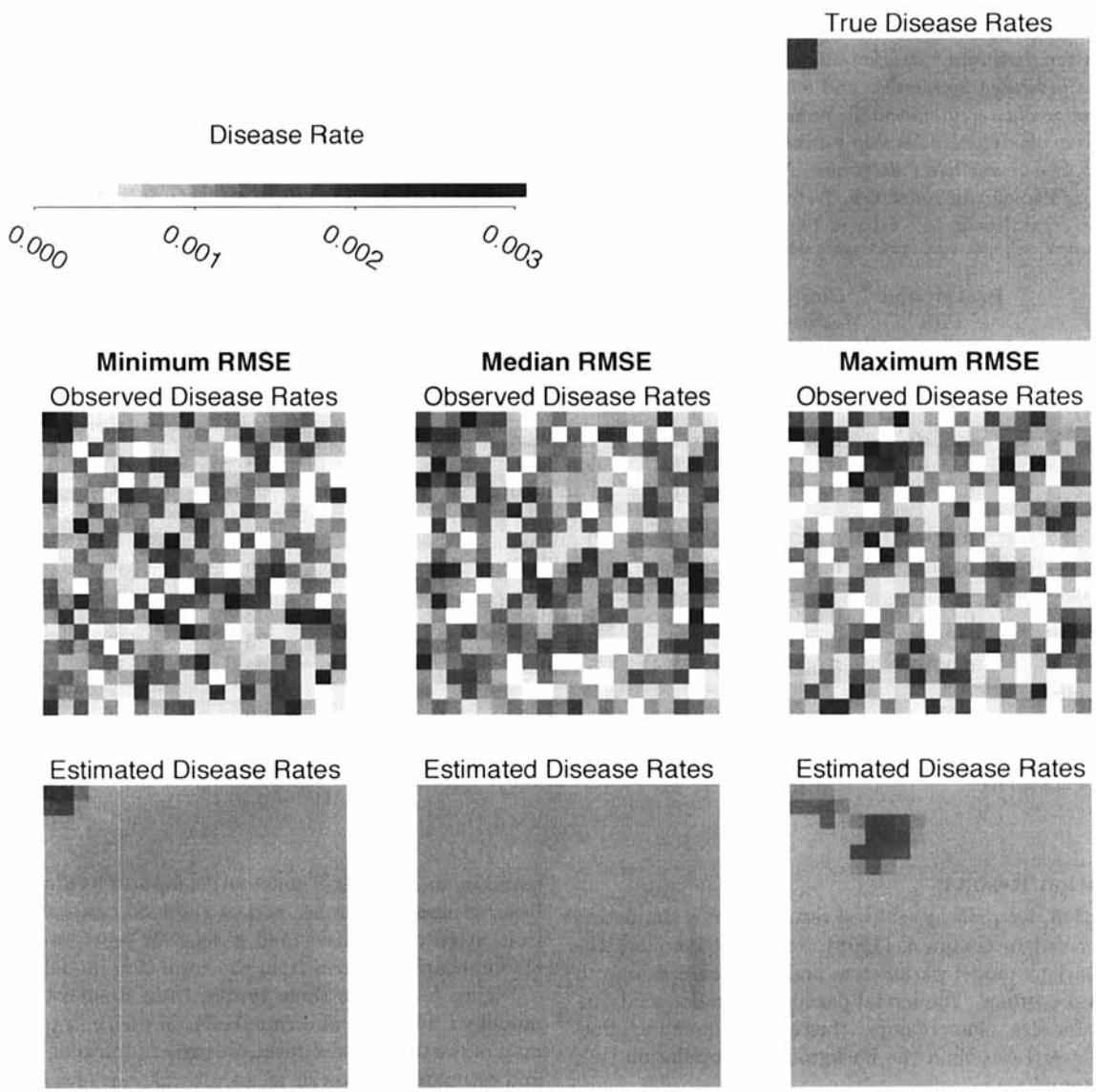


**Figure 6.** Results from 100 replicates of simulation 2—2 × 2 cluster model, background rate 0.001, cluster rate 0.002, total population ~ 1 million. True disease rates in upper right corner. Observed and estimated disease rates presented for simulations resulting in minimum, median (50th smallest), and maximum RMSE.

In Figure 4c, we present the results of an analysis using prior 3, which places less prior weight on small clusters and more prior weight on large clusters. Now, instead of three positive (high rate) clusters, there is strong evidence for a single negative (low rate) cluster ranging across northern portions of Onondaga County and Madison County. A small amount of posterior probability (3%) is associated with a positive cluster in Cortland County.

Finally, Figure 4d displays the results of an analysis in which the shape component is flat. The impact of ignoring cluster shape in the prior is immediately evident. Although the general locations of clustering are the same as in the primary analysis, cells *seem arbitrarily* included or excluded from clusters producing very jagged cluster shapes.

Synthesizing these analyses, we find evidence for three distinct clusters in the New York leukemia data located in Broome County, Cortland County, and Onondaga County. The sources of these clusters are uncertain, and the apparent clusters may simply be artifacts reflecting demographic differences across the study region. If potential clustering of these data is a current public health concern, we recommend further studies of leukemia incidence in this region of New York using data from later (or earlier) time periods that includes information on type of leukemia and demographic variables. In addition to analyses to detect general clustering, we also suggest focused analyses (of these new data) centered on the three clusters discovered here.

**Table 2**

*Cluster detection rates for sets of* 100 *simulations from known clustering models. Simulations also presented in figures indicated by number and letter. True cluster detection defined as one or more cells in the true cluster having posterior cluster membership probability greater than* 0.5. *False cluster detection defined as one or more cells with posterior cluster membership probability greater than* 0.5 *not belonging to the true cluster and not connected by such cells to a true cluster detection. Note that it is possible for a simulation to result in both a true cluster detection and a false cluster detection. Detections of* 1/2 *cells in cluster (respectively, all cells in cluster) defined as at least* 1/2 *(respectively, all) cells in the true cluster having posterior cluster membership probabilities greater than* 0.5.

| Fig. | Background rate | Cluster rate | Cluster | Approximate total population | False cluster detections | True cluster detections | Detections of 1/2 cells in cluster | Detections of all cells in cluster |
|---|---|---|---|---|---|---|---|---|
| Null model | | | | | | | | |
| 5 | 0.0010 | NA[a] | NA | 1 million | 3 | NA | NA | NA |
| Cluster size | | | | | | | | |
| 6 | 0.0010 | 0.0020 | 2 × 2 | 1 million | 3 | 8 | 7 | 2 |
| 7a | 0.0010 | 0.0020 | 3 × 3 | 1 million | 8 | 58 | 51 | 12 |
| 7b | 0.0010 | 0.0020 | 4 × 5 | 1 million | 5 | 98 | 90 | 6 |
| Population size | | | | | | | | |
| | 0.0010 | 0.0020 | 2 × 2 | 2 million | 9 | 48 | 47 | 21 |
| 7c | 0.0010 | 0.0020 | 2 × 2 | 5 million | 11 | 96 | 96 | 84 |
| Negative cluster | | | | | | | | |
| | 0.0010 | 0.0005 | 3 × 3 | 1 million | 4 | 3 | 2 | 1 |
| | 0.0010 | 0.0005 | 3 × 3 | 2 million | 5 | 23 | 22 | 8 |
| 7d | 0.0010 | 0.0005 | 3 × 3 | 5 million | 8 | 94 | 92 | 51 |
| Cluster shape | | | | | | | | |
| | 0.0010 | 0.0020 | 1 × 20 | 1 million | 8 | 50 | 2 | 0 |
| 7e | 0.0010 | 0.0040 | 1 × 20 | 1 million | 11 | 100 | 100 | 20 |

[a] NA, not applicable.

## 5. Simulation Results

In this section, we present selected results from a simulation study described in Gangnon (1998) designed to explore the effects of various model parameters and prior choices on our estimation procedure. The model parameters under study included cluster size, cluster shape, cluster risk (including clusters with lower rates than the background), population size, and background rate.

For these simulations, data were generated from known clustering models on a $20 \times 20$ grid of square cells. Populations for the 400 cells were generated as uniform random variates to produce total populations of approximately 1 million, 2 million, 5 million, and 10 million. (The actual total populations were 998,335, 2,000,189, 5,004,439, and 10,019,630). One hundred realizations from each model were generated using the Poisson random number generator `rpois` in `S-plus`.

For analysis of the simulations presented here, we used the following prior. If the true background disease rate was $\lambda_b$, we took a gamma$(\lambda_b 1000, 1000)$ prior for the background rate and a gamma$(2\lambda_b 1000, 1000)$ prior for cluster rates. For the MCCF component of the cluster model prior, we selected prior 1 in Figure 2 and assigned prior probability of 0.9 to the null model.

For the posterior approximation, we used a $W = 100$ window of plausibility. For the model search, we used two bases of 100 clusters and 11 clusters. For building the bases, we used a $W = 10$ window for merger selection; for searching from the 11 cluster base, we used a $W = 100$ window for merger selection. Adaptive stopping rules were chosen to ensure at least 2 searches and at most 9 consecutive failures from an 11-cluster base, at least 9 searches and at most 30 consecutive failures from a 100-cluster base, and at least 32 searches and at most 89 consecutive failures from the saturated model.

Figure 5 displays some results from simulations of a null model with background rate 0.001. In addition to a grayscale map of the true disease rates, we present maps of the observed and estimated (posterior mean) disease rates for each of three simulated data sets. We chose these data sets based on the root mean squared error, RMSE $= (\Sigma_{i=1}^N n_i(\hat{r_i} - r_i)^2 / \Sigma_{i=1}^N n_i)^{1/2}$. The simulations presented here have the minimum, the 50th smallest (labeled median), and the maximum RMSE values.

Here we observe little evidence for clustering, as is appropriate. The evidence favors the correct null model, i.e., $\max_i \Pr(c_i > 0 \mid \mathbf{O}) < 1/2$, for 97 of the simulated datasets. The estimates from the simulations resulting in the minimum and median RMSE are quite accurate and are typical of the majority of the simulations. The simulation resulting in the maximum RMSE, showing evidence of a large cluster, is a rare exception. Overall, using prior 1 as the MCCF prior, the proposed procedure estimates the null model well.

In Figure 6, we display results from simulations of a model with a single $2 \times 2$ cluster in the upper left corner of the grid. The background rate is 0.001, and the cluster rate is 0.002. We found evidence for the true cluster, i.e., at least one cell in the cluster with $\Pr(c_i > 0 \mid \mathbf{O}) > 1/2$, in 8 of the 100 simulations. The false cluster detection rate in these simulations was 3%, identical to the rate for the null model.

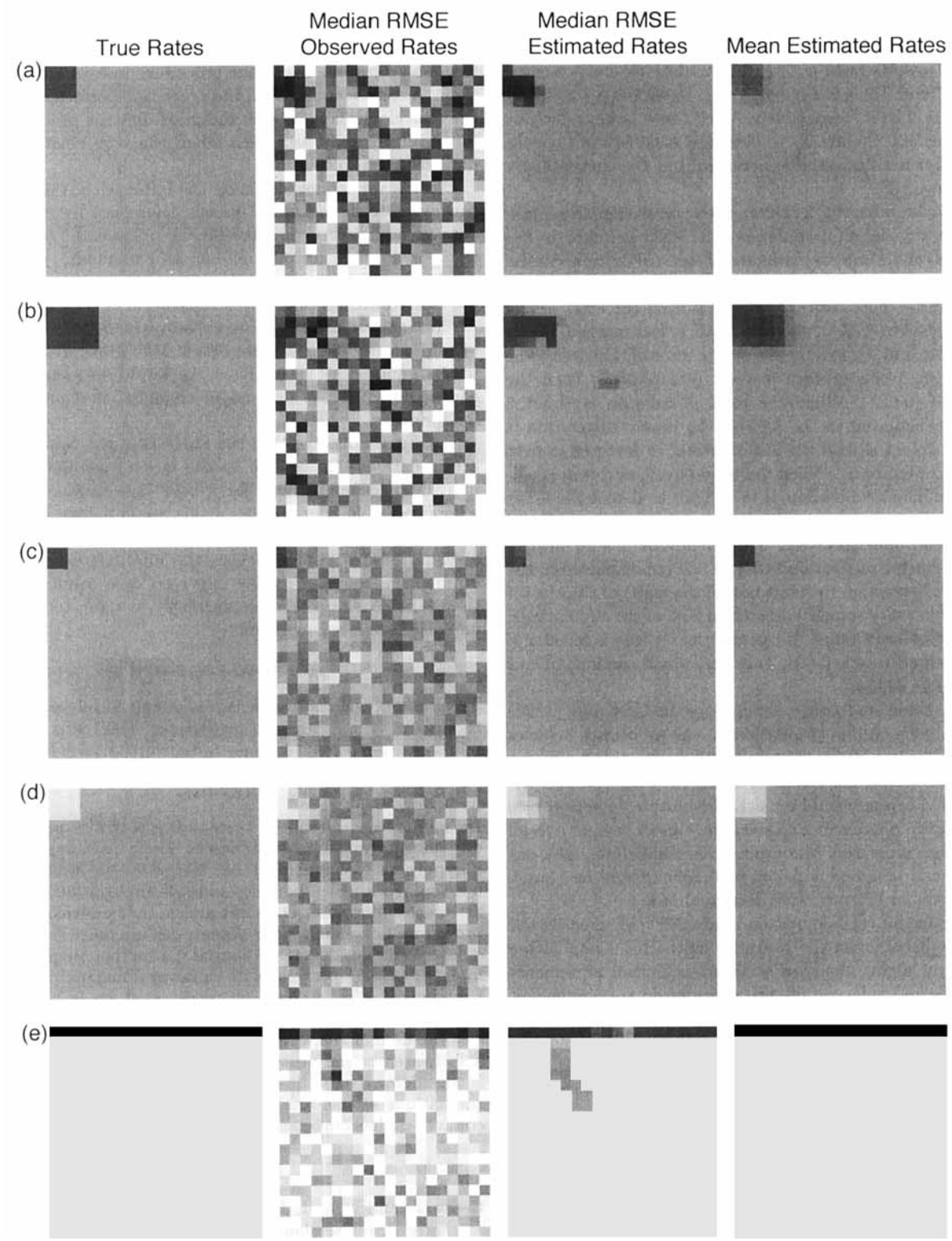


**Figure 7.** Results from 100 replicates of five additional simulations, all using a background rate of 0.001. (**a**) Results from a $3 \times 3$ cluster with rate 0.002 and total population of about 1 million. (**b**) Results from a $4 \times 5$ cluster with rate 0.002 and total population of about 1 million. (**c**) Results from a $2 \times 2$ cluster with rate 0.002 and total population of about 5 million. (**d**) Results from a $3 \times 3$ cluster with rate 0.0005 and total population of about 5 million. (**e**) Results from a $1 \times 20$ cluster with rate 0.004 and total population of about 1 million.

Selected results from additional simulations are presented in Figure 7 and Table 2. Figure 7a and 7b show a $3 \times 3$ cluster and a $4 \times 5$ cluster, respectively. With these large, relatively circular clusters, the ability to detect at least 1/2 of the cells in the cluster is substantially increased, but the entire cluster is rarely detected.

Figure 7c shows a $2 \times 2$ cluster with population size 5 million. With this larger population, the entire cluster is frequently detected, and the estimated rates are quite accurate. With a population of approximately 2 million, the improvement in cluster detection is still substantial (at least 1/2 of the cells detected 47% of the time) but is less marked.

Figure 7d and 7e involves *a priori* unusual clusters. The first is a $3 \times 3$ cluster with a lower rate (0.0005) than the background (0.001). When the total population is about 5 million (as displayed in the figure), the lower cluster rate is frequently detected, and the entire cluster is detected in over half of the simulations. When the total population is approximately 1 million or 2 million, it is difficult to detect the lower cluster rate.

The second unusual cluster is a noncircular $1 \times 20$ cluster. When the cluster rate is doubled to 0.004 (as displayed in the figure), a large section (at least half of the cells) of the cluster is detected in every simulation, and, in 20% of the simulations, the entire cluster is found. When the cluster rate is 0.002 and the background rate is 0.001, only very small sections of this cluster can be found.

Overall, these and other simulations in Gangnon (1998) show that, with certain prior choices, our procedure behaves reasonably for a null model while still being able to detect an *a priori* likely, small cluster with an elevated disease rate. Additionally, as one would expect, the cluster detection rate increases if the population, cluster size, cluster risk, or overall disease rate is increased. Moreover, given compelling data, our procedure can also find *a priori* unlikely clusters, i.e., large, linear clusters or clusters with decreased risk.

The findings of this simulation study are confirmed by the asymptotic behavior of the posterior as the total population in the study region approaches infinity. Formal statements of theorems and their proofs are provided in the Appendix. Applying these asymptotic results, one can demonstrate that the ability to detect clusters will increase with increases in population, cluster size, cluster risk, and/or overall disease rate, as was observed in our simulation study. Additionally, with probability approaching one, the posterior probability of any cluster model with incorrect clusters approaches zero. In this sense, the posterior distribution consistently estimates the true cluster model.

## 6. Conclusions

The combination of our search algorithm and the window of plausibility produces a robust approximation to the posterior. Even when there is virtually no overlap between search samples, the model-averaged estimates can be remarkably similar, as we see in Figures 3a and 4a. However, as one should expect, the choice of an MCCF prior can dramatically affect our estimates for the disease rates, e.g., the contrast between Figure 3a and Figure 4c. In addition, simulation studies demonstrate the good performance of our procedure in a variety of situations. As one would expect, the procedure performs well when the true cluster agrees with the prior, but the procedure performs surprisingly well even when the true cluster is quite unlikely *a priori*.

Clearly, we cannot eliminate the influence of the prior. We can evaluate its influence on our inferences by analyzing a data set with several different MCCF priors and by replicating the model search. In our opinion, at a minimum, one should conduct the following analyses: (1) a primary analysis using the desired MCCF prior, (2) a replication of the first analysis, (3) an analysis using a more conservative MCCF prior, and (4) an analysis using a more liberal MCCF prior. In addition to the minimal set of analyses, we would also suggest using additional MCCF priors and/or changing the gamma priors for the rates.

Finally, we observe that the analysis of the New York leukemia data is limited due to the lack of available data on individual leukemias and for specific age strata. There are several possible adaptations of this methodology that would allow for the inclusion of covariate strata and/or multiple diseases. A simple approach uses the methods described here but replaces populations by expected case counts based on the available strata. We are currently working to implement more sophisticated methods.

## Acknowledgements

The authors wish to thank two anonymous referees for their thoughtful comments and suggestions. REG was supported by the National Eye Institute (grant NIH-EY07119).

## Résumé

De nombreuses méthodes statistiques actuelles pour l'étude de zones de maladies sont basées sur le paradigme du test d'hypothèse. Typiquement ces méthodes ne produisent pas d'estimateurs utiles pour les taux de maladie ou les risques attachés à une zone. Dans cet article, nous développons une procédure bayesienne pour obtenir des inférences sur des modèles spécifiques de zonage spatial. La méthodologie proposée intègre des idées provenant de l'analyse d'images, de la notion bayesienne de moyennage de modèles, et de la sélection de modèles. Avec notre approche, nous obtenons des estimations de taux de maladie, et permettons une plus grande souplesse à la fois en terme de type de zones et de nombre de zones à considérer. Nous illustrons la procédure proposée par des simulations et une analyse des données New Yorkaises bien connues de leucémie.

## References

Besag, J. and Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Series A* **154,** 143–155.

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

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

Cuzick, J. and Edwards, R. (1990). Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society, Series B* **52,** 73–104.

Diggle, P. J. (1990). A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point. *Journal of the Royal Statistical Society, Series A* **153,** 349–362.

Gangnon, R. E. (1998). *Disease Rate Mapping via Cluster Models.* Madison, Wisconsin: University of Wisconsin.

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine* **14,** 799–810.

Lawson, A. B. (1995). Markov chain Monte Carlo methods for putative pollution source problems in environmental epidemiology. *Statistics in Medicine* **14,** 2473–2486.

Lawson, A. B. and Clark, A. (1999). Markov chain Monte Carlo methods for putative sources of hazard and general clustering. In *Disease Mapping and Risk Assessment for Public Health,* A. B. Lawson, D. Bohning, A. Biggeri, J.-F. Viel, and R. Bertollini (eds), Chapter 9. Chichester: Wiley.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* **89,** 1535–1546.

Møller, J. and Waagepetersen, R. P. (1998). Markov connected component fields. *Advances in Applied Probability* **30,** 1–35.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society, Series B* **56,** 3–48.

Openshaw, S., Craft, A. W., Charlton, M., and Birch, J. M. (1988). Investigation of leukaemia clusters by use of a geographical analysis machine. *Lancet* **1,** 272–273.

Ross, A. and Davis, S. (1990). Point pattern analysis of the spatial proximity of residences prior to diagnosis of persons with Hodgkin's disease. *American Journal of Epidemiology* **132,** S53–S62.

Selvin, S., Schulman, J., and Merrill, D. W. (1992). Distance and risk measures for the analysis of spatial data: A study of childhood cancers. *Social Science and Medicine* **34,** 769–777.

Sibson, R. (1980). The Dirichlet tessellation as an aid in data analysis. *Scandinavian Journal of Statistics* **7,** 14–20.

Stone, R. A. (1988). Investigations of excess environmental risks around putative sources: Statistical problems and a proposed test. *Statistics in Medicine* **7,** 649–660.

Turnbull, B. W., Iwano, E. J., Burnett, W. S., Howe, H. L., and Clark, L. C. (1990). Monitoring for clusters of disease: Application to leukemia incidence in upstate New York. *American Journal of Epidemiology* **132,** S136–S143.

Waller, L. A., Turnbull, B. W., Clark, L. C., and Nasca, P. (1992). Chronic disease surveillance and testing of clustering of disease and exposure: Application to leukemia incidence and TCE-contaminated dumpsites in upstate New York. *Environmetrics* **3,** 281–300.

Waller, L. A., Turnbull, B. W., Clark, L. C., and Nasca, P. (1994). Spatial pattern analyses to detect rare disease clusters. In *Case Studies in Biometry,* N. Lange, L. Ryan, and L. Billard (eds), 3–22. New York: John Wiley and Sons.

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

Whittemore, A., Friend, N., Brown, B. W., and Holly, E. A. (1987). A test to detect clusters of disease. *Biometrika* **74,** 31–35.

*Received January* 1999. *Revised November* 1999. *Accepted January* 2000.

## APPENDIX

### *Asymptotic Results*

Here we consider the behavior of the posterior on cluster models as $n_t = \Sigma_{i=1}^N n_i$, the total population at risk, approaches infinity. Since the population only influences the posterior through the marginal likelihood, we primarily examine asymptotic behavior of equation (1).

Let $r_i$ be the true disease rate for cell $i$, $i = 1, 2, \ldots, N$. Assume that (1) $n_i/n_t \to \pi_i > 0$ as $n_t \to \infty$ and (2) $r_i > 0$ for $i = 1, 2, \ldots, N$. Given a model $\mathbf{c}$ with $k$ clusters, let $|\mathbf{c}| = k+1$ be the number of components. For $j = 0, 1, \ldots, k$, define $O_j^{\mathbf{c}} = \Sigma_{i=1}^N O_i I_{\{c_i=j\}}$, the number of cases in component $j$; $n_j^{\mathbf{c}} = \Sigma_{i=1}^N n_i I_{\{c_i=j\}}$, the population at risk in component $j$; $\pi_j^{\mathbf{c}} = \Sigma_{i=1}^N \pi_i I_{\{c_i=j\}}$, the asymptotic proportion of the population in component $j$; and $\lambda_j^{\mathbf{c}} = \Sigma_{i=1}^N r_i \pi_i I_{\{c_i=j\}}/\pi_j^{\mathbf{c}}$, the asymptotic pooled disease rate for component $j$. The asymptotic behavior of the marginal likelihood $p(\mathbf{O} \mid \mathbf{c})$ is given by the following theorem.

THEOREM 1: *Let* $\mathbf{c}_1$ *and* $\mathbf{c}_2$ *be arbitrary cluster models. Let* $(a_j, b_j) = (a, b)$ *for* $j \geq 1$. *Let* $g(\cdot, a, b)$ *represent the density of a gamma random variable with mean* $a/b$ *and variance* $a/b^2$. *Under the assumptions given above, with probability one, as* $n_t \to \infty$,

$$\frac{p(\mathbf{O} \mid \mathbf{c}_1)}{p(\mathbf{O} \mid \mathbf{c}_2)} \sim \frac{\left(\sqrt{2\pi}\right)^{|\mathbf{c}_1|-|\mathbf{c}_2|}}{n_t^{\frac{1}{2}(|\mathbf{c}_1|-|\mathbf{c}_2|)}} \cdot \prod_{i=1}^{N} \left( \frac{O_{c_{1i}}^{\mathbf{c}_1}/n_{c_{1i}}^{\mathbf{c}_1}}{O_{c_{2i}}^{\mathbf{c}_2}/n_{c_{2i}}^{\mathbf{c}_2}} \right)^{O_i}$$
$$\times \frac{\prod_{j=0}^{|\mathbf{c}_2|-1} \left(\lambda_j^{\mathbf{c}_2} \pi_j^{\mathbf{c}_2}\right)^{\frac{1}{2}}}{\prod_{j=0}^{|\mathbf{c}_1|-1} \left(\lambda_j^{\mathbf{c}_1} \pi_j^{\mathbf{c}_1}\right)^{\frac{1}{2}}} \cdot \frac{\prod_{j=0}^{|\mathbf{c}_1|-1} g(\lambda_j^{\mathbf{c}_1}, a_j, b_j)}{\prod_{j=0}^{|\mathbf{c}_2|-1} g(\lambda_j^{\mathbf{c}_2}, a_j, b_j)}.$$

To prove Theorem 1, we first establish the following two lemmas.

LEMMA 1: *Let* $a_0, a_1, \ldots, a_N$ *be arbitrary constants. Let* $C$ *be any nonempty subset of* $1, 2, \ldots, N$. *Let* $O_C = \Sigma_{i \in C} O_i$, $\pi_C = \Sigma_{i \in C} \pi_i$, *and* $\lambda_C = \Sigma_{i \in C} \lambda_i \pi_i / \pi_C$. *Under the assump-*

*tions of Theorem 1, with probability one, as* $n_t \to \infty$,

$$\frac{\Gamma(a_0+O_C)}{\prod_{i\in C}\Gamma(a_i+O_i)} \sim \frac{n_t^{a_0-1/2}}{\prod_{i\in C} n_t^{a_i-1/2}} \cdot \frac{\pi_C^{a_0-1/2}}{\prod_{i\in C}\pi_i^{a_i-1/2}} \cdot \frac{{\lambda_C}^{a_0-1/2}}{\prod_{i\in C}\lambda_i^{a_i-1/2}} \times \prod_{i\in C}\left(\frac{O_C}{O_i}\right)^{O_i} \cdot \frac{\sqrt{2\pi}}{\prod_{i\in C}\sqrt{2\pi}}.$$

*Proof.* As $n_t \to \infty$,

$$\begin{aligned}
&\frac{\Gamma(a_0+O_C)}{\prod_{i\in C}\Gamma(a_i+O_i)} \\
&\quad\sim \frac{(O_C+a_0-1)^{O_C+a_0-1/2}\cdot e^{-(O_C+a_0-1)}\cdot\sqrt{2\pi}}{\prod_{i\in C}\left\{(O_i+a_i-1)^{O_i+a_i-1}\cdot e^{-(O_C+a_i-1)}\cdot\sqrt{2\pi}\right\}} \\
&\quad\sim \frac{O_C^{a_0-1/2}}{\prod_{i\in C} O_i^{a_i-1/2}} \cdot \frac{O_C^{O_C}}{\prod_{i\in C} O_i^{O_i}} \\
&\qquad\times \frac{\left(1+\frac{a_0-1}{O_C}\right)^{O_C}\cdot e^{-(a_0-1)}}{\prod_{i\in C}\left(1+\frac{a_i-1}{O_i}\right)^{O_i}\cdot e^{-(a_i-1)}} \cdot \frac{\sqrt{2\pi}}{\prod_{i\in C}\sqrt{2\pi}} \\
&\quad\sim \frac{n_t^{a_0-1/2}}{\prod_{i\in C} n_t^{a_i-1/2}} \cdot \frac{(\lambda_C\pi_C)^{a_0-1/2}}{\prod_{i\in C}(\lambda_i\pi_i)^{a_i-1/2}} \\
&\qquad\times \prod_{i\in C}\left(\frac{O_C}{O_i}\right)^{O_i} \frac{\sqrt{2\pi}}{\prod_{i\in C}\sqrt{2\pi}}.
\end{aligned}$$

The first formula is obtained by direct application of Stirling's formula. The second formula follows by grouping similar terms and noting that, by the Strong Law of Large Numbers, $O_i + a_i \sim O_i$ and $O_C + a_0 \sim O_C$ as $n_t \to \infty$. The final expression follows since $O_i \sim \lambda_i\pi_i n_t$ and $O_C \sim \lambda_C\pi_C n_t$ as $n_t \to \infty$ by the Strong Law of Large Numbers and $\lim_{x\to\infty}(1+a/x)^x = e^a$.

LEMMA 2: *Let* $a_0, a_1, \ldots, a_N$ *and* $b_0, b_1, \ldots, b_N$ *be arbitrary constants. Let* $O_C$, $n_C$, $\pi_C$, *and* $\lambda_C$ *be as in Lemma 1. Under the assumptions of Theorem 1, with probability one, as* $n_t \to \infty$,

$$\frac{\prod_{i\in C}(b_i+n_i)^{a_i+O_i}}{(b_0+n_C)^{a_0+O_c}} \sim \frac{\prod_{i\in C} n_t^{a_i}}{n_t^{a_0}} \cdot \frac{\prod_{i\in C}\pi_i^{a_i}}{\pi_C^{a_0}} \cdot \prod_{i\in C}\left(\frac{n_i}{n_C}\right)^{O_i} \cdot \frac{e^{-\lambda_C b_0}}{\prod_{i\in C} e^{-\lambda_i b_i}}.$$

*Proof.* As $n_t \to \infty$,

$$\begin{aligned}
&\frac{\prod_{i\in C}(b_i+n_i)^{a_i+O_i}}{(b_0+n_C)^{a_0+O_c}} \\
&\quad\sim \frac{\prod_{i\in C} n_i^{a_i}}{n_C^{a_0}} \cdot \frac{\prod_{i\in C} n_i^{O_i}}{n_C^{O_C}} \cdot \frac{\prod_{i\in C}\left(1+\frac{b_i}{n_i}\right)^{O_i}}{\left(1+\frac{b_0}{n_C}\right)^{O_C}} \\
&\quad\sim \frac{\prod_{i\in C} n_t^{a_i}}{n_t^{a_0}} \cdot \frac{\prod_{i\in C}\pi_i^{a_i}}{\pi_C^{a_0}} \cdot \prod_{i\in C}\left(\frac{n_i}{n_C}\right)^{O_i} \cdot \frac{\prod_{i\in C} e^{\lambda_i b_i}}{e^{\lambda_C b_0}}.
\end{aligned}$$

The first expression follows by grouping similar terms and noting that $n_i + b_i \sim n_i$ and $n_C + b_0 \sim n_C$ as $n_t \to \infty$. The second expression follows since $n_i \sim \pi_i n_t$ and $n_C \sim \pi_C n_t$ by assumption and $\lim_{x\to\infty}(1+a/x)^{f(x)} = e^{af}$ if $\lim_{x\to\infty} f(x)/x = f$.

*Proof of Theorem 1.* Let $\mathbf{c}^* = (0, 1, \ldots, N-1)$. We can directly apply Lemmas 1 and 2 to the components of $\mathbf{c}_1$ and $\mathbf{c}_2$ to establish the desired result for $p(\mathbf{O} \mid \mathbf{c}_1)/p(\mathbf{O} \mid \mathbf{c}^*)$ and $p(\mathbf{O} \mid \mathbf{c}_2)/p(\mathbf{O} \mid \mathbf{c}^*)$. The result for $p(\mathbf{O} \mid \mathbf{c}_1)/p(\mathbf{O} \mid \mathbf{c}_2)$ then follows by considering the ratio of the previous two expressions.

To develop an asymptotic expression for the posterior, we classify cluster models as follows: A cluster model $\mathbf{c}$ is valid if $r_i = \lambda_{c_i}^{\mathbf{c}}$ for $i = 1, 2, \ldots, N$; otherwise, a cluster model $\mathbf{c}$ is invalid. A cluster model $\mathbf{c}$ is a minimal valid cluster model if it is valid and no other valid cluster model has fewer components. All minimal valid cluster models can be found from one minimal valid cluster model by selecting, in turn, each component of the model as the background. The following theorem specifies the asymptotic behavior of the posterior distribution.

THEOREM 2: *Let* $\mathbf{c}$ *be a cluster model,* $\mathcal{C}$ *be the collection of all cluster models with connected components,* $\mathcal{V}$ *be the collection of all valid cluster models, and* $\mathcal{M}$ *be the collection of all minimal valid cluster models. Under the assumptions given above,*

(i) *If* $\mathbf{c} \in \mathcal{C} \setminus \mathcal{V}$, $p(\mathbf{c} \mid \mathbf{O}) \to 0$ *as* $n_t \to \infty$ *with probability one.*
(ii) *If* $\mathbf{c} \in \mathcal{V} \setminus \mathcal{M}$, $p(\mathbf{c} \mid \mathbf{O}) \to 0$ *in probability as* $n_t \to \infty$.
(iii) *If* $\mathbf{c} \in \mathcal{M}$,

$$p(\mathbf{c} \mid \mathbf{O}) \to \frac{p(\mathbf{c}) \cdot \prod_{j=0}^{|\mathbf{c}|-1} g\left(\lambda_j^{\mathbf{c}}, a_j, b_j\right)}{\sum_{\mathbf{c}\in\mathcal{M}} p(\mathbf{c}) \cdot \prod_{j=0}^{|\mathbf{c}|-1} g\left(\lambda_j^{\mathbf{c}}, a_j, b_j\right)}$$

*in probability as* $n_t \to \infty$.

To establish Theorem 2, we define $\gamma_{\mathbf{c}} = \Pi_{i=1}^N (\lambda_{c_i}^{\mathbf{c}}/\lambda_i)^{\lambda_i\pi_i}$. Note that $\gamma_{\mathbf{c}} \leq 1$. An alternative characterization of valid cluster models based on $\gamma_{\mathbf{c}}$ is given by the following lemma.

LEMMA 3: *A cluster model,* $\mathbf{c}$, *is invalid if and only if* $\gamma_{\mathbf{c}} < 1$.

*Proof.* Let $\mathbf{c}$ be a cluster model and $\mathbf{\Lambda_c} = (\lambda_{c_1}^{\mathbf{c}}, \lambda_{c_2}^{\mathbf{c}}, \ldots, \lambda_{c_N}^{\mathbf{c}})$. By definition, the cluster model $\mathbf{c}$ is valid if and only if $\mathbf{\Lambda_c} = \mathbf{\Lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_N)$. Now consider the function $L(\mathbf{x}) = \Sigma_{i=1}^N \lambda_i\pi_i \log x_i - \pi_i x_i$. Standard maximization techniques using first and second derivatives of $L$ can be used to establish that $L$ has a unique maximum at $\mathbf{x} = \mathbf{\Lambda}$. Thus, $D(\mathbf{x}) = L(\mathbf{x}) - L(\mathbf{\Lambda}) \leq 0$ and $D(\mathbf{x}) = 0$ if and only if $\mathbf{x} = \mathbf{\Lambda}$. Since $\log\gamma_{\mathbf{c}} = D(\mathbf{\Lambda_c})$, the desired result holds.

LEMMA 4: *Let* $\mathbf{c}_1$ *be an invalid cluster model and* $\mathbf{c}_2$ *be a valid cluster model. Under the assumptions of Theorem 1, with probability one, for any* $\epsilon > 0$,

$$\frac{p(\mathbf{O} \mid \mathbf{c}_1)}{p(\mathbf{O} \mid \mathbf{c}_2)} = o\left(\frac{(\gamma_{\mathbf{c}_1} + \epsilon)^{n_t}}{n_t^{\frac{1}{2}(|\mathbf{c}_1| - |\mathbf{c}_2|)}}\right) \quad \text{as } n_t \to \infty.$$

*Proof.* Note that, by the Strong Law of Large Numbers, with probability one,

$$\prod_{i=1}^{N} \left(\frac{O_{c_1 i}^{\mathbf{c}_1} / n_{c_1 i}^{\mathbf{c}_1}}{O_{c_2 i}^{\mathbf{c}_2} / n_{c_2 i}^{\mathbf{c}_2}}\right)^{O_i / n_t} \to \gamma_{\mathbf{c}_1} \quad \text{as } n_t \to \infty.$$

Thus, with probability one, for any $\epsilon > 0$, there exists an $N(\epsilon)$ such that

$$\prod_{i=1}^{N} \left(\frac{O_{c_1 i}^{\mathbf{c}_1} / n_{c_1 i}^{\mathbf{c}_1}}{O_{c_2 i}^{\mathbf{c}_2} / n_{c_2 i}^{\mathbf{c}_2}}\right)^{O_i} \leq (\gamma_{\mathbf{c}_1} + \epsilon/2)^{n_t}, \qquad n_t > N(\epsilon).$$

It follows immediately that, with probability one, for any $\epsilon > 0$,

$$\frac{1}{(\gamma_{\mathbf{c}} + \epsilon)^{n_t}} \prod_{i=1}^{N} \left(\frac{O_{c_1 i}^{\mathbf{c}_1} / n_{c_1 i}^{\mathbf{c}_1}}{O_{c_2 i}^{\mathbf{c}_2} / n_{c_2 i}^{\mathbf{c}_2}}\right)^{O_i / n_t} \to 0$$

as $n_t \to \infty$. The desired result now is a direct consequence of Theorem 1.

LEMMA 5: *Let* $\mathbf{c}_1$ *be a valid cluster model. Let* $\mathbf{c}_2$ *be a minimal valid cluster model. Let* $g(\cdot, a, b)$ *represent the density of a gamma random variable with mean* $a/b$ *and variance* $a/b^2$. *Under the assumptions of Theorem 1,*

(i) *If* $| \mathbf{c}_1 | > | \mathbf{c}_2 |$, *then*

$$\frac{p(\mathbf{O} \mid \mathbf{c}_1)}{p(\mathbf{O} \mid \mathbf{c}_2)} = O_p\left(\frac{1}{n_t^{\frac{1}{2}(|\mathbf{c}_1| - |\mathbf{c}_2|)}}\right) \qquad \text{as } n_t \to \infty.$$

(ii) *If* $| \mathbf{c}_1 | = | \mathbf{c}_2 |$, *then, with probability one,*

$$\lim_{n_t \to \infty} \frac{p(\mathbf{O} \mid \mathbf{c}_1)}{p(\mathbf{O} \mid \mathbf{c}_2)} = \frac{\prod_{j=0}^{|\mathbf{c}_1|-1} g\left(\lambda_j^{\mathbf{c}_1}, a_j, b_j\right)}{\prod_{j=0}^{|\mathbf{c}_2|-1} g\left(\lambda_j^{\mathbf{c}_2}, a_j, b_j\right)}.$$

*Proof.* (i) First, since $\mathbf{c}_2$ is a minimal valid cluster model, $\mathbf{c}_2$ must be nested within $\mathbf{c}_1$. Thus, the likelihood ratio test statistic

$$\prod_{i=1}^{N} \left(\frac{O_{c_1 i}^{\mathbf{c}_1} / n_{c_1 i}^{\mathbf{c}_1}}{O_{c_2 i}^{\mathbf{c}_2} / n_{c_2 i}^{\mathbf{c}_2}}\right)^{O_i}$$

represents a test of two nested hypotheses. The standard regularity conditions are satisfied so the test statistic converges in distribution and hence is $O_p(1)$. The desired result then follows directly from Theorem 1.

(ii) Since both $\mathbf{c}_1$ and $\mathbf{c}_2$ are minimal valid cluster models, they are identical except for the selection of the background component. Only the final term in the expression given in Theorem 1 depends on the labeling of the components. Thus, the desired result holds.

*Proof of Theorem 2.* All three results are immediate consequences of the previous two lemmas. To demonstrate (i) and (ii), note that, if $\mathbf{c}_1$ is a minimal valid cluster model, then

$$p(\mathbf{c} \mid \mathbf{O}) \leq \frac{p(\mathbf{c}) \cdot p(\mathbf{O} \mid \mathbf{c})}{p(\mathbf{c}_1) \cdot p(\mathbf{O} \mid \mathbf{c}_1)}.$$

The desired convergences are then obtained by direct application of Lemma 4 for (i) and Lemma 5(i) for (ii). To demonstrate (iii), consider that

$$\frac{1}{p(\mathbf{c} \mid \mathbf{O})} = \sum_{\mathbf{c}_1 \in \mathcal{C}} \frac{p(\mathbf{c}_1) \cdot p(\mathbf{O} \mid \mathbf{c}_1)}{p(\mathbf{c}) \cdot p(\mathbf{O} \mid \mathbf{c})}$$

$$\to \sum_{\mathbf{c}_1 \in \mathcal{M}} \left\{ \frac{p(\mathbf{c}_1) \cdot \prod_{j \in L(\mathbf{c}_1)} g\left(\lambda_j^{\mathbf{c}_1}, a_j, b_j\right)}{p(\mathbf{c}) \cdot \prod_{j \in L(\mathbf{c})} g\left(\lambda_j^{\mathbf{c}}, a_j, b_j\right)} \right\}$$

in probability as $n_t \to \infty$ by direct application of Lemmas 4 and 5 since the summation involves only a finite number of terms.