Environ Ecol Stat (2007) 14:69–82
DOI 10.1007/s10651-006-0007-7

ORIGINAL ARTICLE

Cluster detection using Bayes factors from overparameterized cluster models

Ronald Gangnon · Murray K. Clayton

Received: 1 September 2004 / Revised: 28 February 2005 /
Published online: 23 January 2007

Abstract In this paper, we consider the use of a partition model to estimate regional disease rates and to detect spatial clusters. Formal inference regarding the number of partitions (or clusters) can be obtained with a reversible jump Markov chain Monte Carlo algorithm. As an alternative, we consider the ability of models with a fixed, but overly large, number of partitions to estimate regional disease rates and to provide informal inferences about the number and locations of clusters using local Bayes factors. We illustrate and compare these two approaches using data on leukemia incidence in upstate New York and data on breast cancer incidence in Wisconsin.

Keywords Bayes factor · Cluster detection · Random effects · Reversible jump Markov chain Monte Carlo · Spatial epidemiology

1 Introduction

The analysis of spatial patterns of small area disease rates has been a subject of great interest since the 1980s. Several comprehensive books reviewing much of the work on this problem are now available Elliott et al. (1999), Lawson et al. (1999), Lawson (2001), Lawson and Denison (2002). Partition modeling is one attractive strategy for estimating small area disease rates, particularly for the purpose of spatial cluster detection (Ferreira et al. 2002). In a partition model, the study region is divided into disjoint regions of constant risk. Gangnon and Clayton (2000), Knorr-Held and

R. Gangnon (✉)
Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, WI 53706-1461, USA
e-mail: ronald@biostat.wisc.edu

M. K. Clayton
Departments of Statistics and Plant Pathology, University of Wisconsin–Madison, Madison, WI 53706-1461, USA
e-mail: clayton@stat.wisc.edu

Raßer (2000), and Denison and Holmes (2001) each developed partition models for small area disease rates. The models proposed by Knorr-Held and Raßer (2000) and Denison and Holmes (2001) partition the study region by means of the Voronoi tesselation. The methods differ primarily in the specification of a prior for the regional rates, with Knorr-Held and Raßer (2000) adopting a log-normal prior and Denison and Holmes (2001) adopting a conjugate gamma prior. Ferreira et al. (2002) extend these models to include both covariates and extra-Poisson variation. In these models, the partitions are often not of direct interest, but instead are principally tools for estimating the risk surface.

Gangnon and Clayton (2000) propose an alternative family of partitions, which divide the study region into a large background area and a small number of clusters. Here, the term cluster specifically refers to one of a few isolated areas of locally increased or descreased disease incidence. Other authors use the term *cluster* to refer to the individual partitions in the model, as in traditional clustering algorithms. Our model does not encompass global clustering, e.g., the tendency for cases to occur near other cases. The size and shape of a potential cluster is quite flexible (any connected subset is possible), but is controlled by a user-specified prior. Gangnon and Clayton (2003) consider similar models, but restrict consideration to circular clusters (or to another enumerable set of potential clusters). By restricting the set of potential clusters, one can more easily define a prior for the clusters as well as incorporate covariate effects and extra-Poisson variation. In these clustering models, the location and composition of the clusters are of primary interest.

All of the above approaches require the specification (implicit or explicit) of a prior distribution for the number of partitions (or clusters), and inference is typically obtained with a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm (Green 1995). However, it is often difficult to specify completely a prior distribution for the number of clusters. Instead, one often identifies a plausible upper bound on the number of clusters (a relatively easy task) and adopts a discrete uniform prior. In this article, we consider the merits of drawing inferences from a model with a fixed number of clusters, e.g., the identified upper bound. By doing so, one avoids the more difficult task of prior elucidation for the number of clusters. In addition, by fixing the model dimension, we hope to avoid the complications of the dimension-varying RJMCMC algorithm.

We focus our discussion on the clustering model of Gangnon and Clayton (2003), although the ideas apply equally well to other partition models. In particular, we consider the ability of models with a fixed number of clusters to estimate the small area disease rates and to identify the local evidence for clustering for each small area. The major drawback to this approach is the lack of formal inference about the number of clusters. However, in many applications, we are more interested in local evidence for clustering and less concerned with global inference about the number of clusters.

In Sect. 2, we present the clustering model proposed by Gangnon and Clayton (2003). We also discuss Markov chain Monte Carlo techniques for obtaining posterior inference from models with a fixed number of clusters. In Sect. 3, we re-analyze the well-known New York leukemia data with models having a fixed number of clusters and compare the results to the inferences obtained by Gangnon and Clayton (2003) with a model having a variable number of clusters. In Sect. 4, we present an analysis of breast cancer incidence in Wisconsin. In Sect. 5, we present some concluding remarks and discuss areas of future work.

2 Statistical model

Suppose that, for n subregions or cells, we observe y_i, the number of cases of disease, and E_i, an expected number of cases calculated by means of internal or external standardization, $i = 1, 2, \ldots, n$. We assume that y_i follows a Poisson distribution with mean $\rho_i E_i$, where ρ_i is the standardized incidence ratio for cell i. Following Gangnon and Clayton (2003), we consider a log-linear model for the standardized incidence ratios, $\log(\rho_i) = \alpha + \phi_i + \epsilon_i$. Here, α is an intercept, ϕ_i is the spatial clustering effect, and ϵ_i is a spatially uncorrelated random effect. Clayton and Kaldor (1987) and Besag et al. (1991) used a similar model for disease mapping. Our model differs in the specification of the spatial effect, which reflects the different goals of our analysis. The intercept α is given a flat prior. The random effects $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are iid $N(0, 1/\tau)$. The random effects precision τ is given a conjugate gamma prior. In the applications, we use a gamma prior with mean 100 and standard deviation 100 so that, with 95% probability, the variance $1/\tau$ falls between 0.003 and 0.40. A variance of 0.40 implies a roughly 12-fold difference in risk between cells at the 2.5th and 97.5th percentiles; a variance of 0.003 implies only a 1.2-fold difference in risk.

We now focus on the spatial clustering effect. In an ideal formulation, the spatial clustering effect is given by $\phi_i = \sum_{j=1}^{k} \theta_j \mathbf{1}_{\{i \in C_j\}}$, where k is the number of clusters; $C_1, C_2, \ldots, C_k$ are the sets of cells belonging to the k clusters; $\theta_1, \theta_2, \ldots, \theta_k$ are the log relative risks associated with each cluster; $\mathbf{1}_A$ is the indicator of A (taking value 1 if A is true and 0 if A is false. $C_1, C_2, \ldots, C_k$ are chosen from a restricted collection of connected subsets with small diameters; circles in a relevant metric are often convenient. In addition, we allow the k clusters to overlap. The choice of the set of potential clusters generally reflects prior knowledge and the goals of analysis. Often, this information is reflected in the choice of the metric used to define circular clusters. However, the methodology is general and will work for any enumerable set of clusters.

Conditioning on the number of clusters k, we adopt the prior specification for the spatial clustering effect given by Gangnon and Clayton (2003). We use a normal prior for the cluster risks, $\theta_1, \theta_2, \ldots, \theta_k$ iid $N(0, \sigma_\theta^2)$. In the examples, we take σ_θ^2 to be 0.355 so that $P(1/4 < e^{\theta_1} < 4) = 0.99$.

For circular clusters centered at the centroids of the n cells, with radii that range from zero to a fixed maximum geographic radius $r_{\max}$, we use the "dartboard" prior (Gangnon and Clayton 2001). The dartboard prior is defined constructively. First, we select one of the n cells by "throwing a dart" at the study region, i.e., with probability proportional to its area. The cluster is centered on the centroid of the selected cell. We then select the radius of the cluster from the uniform distribution on $(0, r_{\max})$. The result is a prior with approximately uniform prior probability of cluster membership for all cells.

Gangnon and Clayton (2003) treated k, the number of clusters, as a parameter to be estimated. To draw inferences from this model, they used a RJMCMC algorithm (Green 1995). Here, we propose an alternative approach to inference based on a fixed number of clusters k. If the true number of clusters, say k_0, is no greater than k, the underlying model is, in fact, correct, albeit possibly overparameterized. That is, if $\phi_i = \sum_{j=1}^{k_0} \theta_j \mathbf{1}_{\{i \in C_j\}}$, then $\phi_i = \sum_{j=1}^{k} \theta_j \mathbf{1}_{\{i \in C_j\}}$, where $\theta_{k_0+1} = \theta_{k+2} = \ldots = \theta_k = 0$

and $C_{k_0+1}, C_{k_0+2}, \ldots, C_k$ are arbitrary. Thus, we would expect similar behavior in the posterior, e.g., concentration of mass on the k_0 true clusters, along with essentially arbitrary placement of the $k - k_0$ excess clusters with cluster risks near 0.

If the cluster locations are known, the foregoing model is a hierarchical Poisson generalized linear model with parameters $\alpha, \theta_1, \theta_2, \ldots, \theta_k$ and $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$. Techniques for sampling from the posterior distribution are described in Gelman et al. (1995) for general models of this type and in Gangnon and Clayton (2003) for this specific model. Specifically, we use a quadratic approximation to the likelihood, which is conjugate to the normal priors, to develop a proposal distribution for a Metropolis-Hastings algorithm (Hastings 1970). Posterior samples for τ are obtained from its (conjugate) full conditional distribution. Posterior samples for $C_1, C_2, \ldots, C_k$, the cluster locations, are obtained from their full conditional distributions. A single iteration of the chain consists of a single update of the cluster locations $C_1, \ldots, C_k$, followed by multiple (in our applications, 10) updates of the other parameters $\alpha, \theta_1, \ldots, \theta_n, \epsilon_1, \ldots, \epsilon_n$, and τ.

3 Example: New York leukemia data

To assess the performance of the fixed k model, we reconsider the New York leukemia data, which have been analyzed previously by means of a variable k model (Gangnon and Clayton 2003). The New York leukemia data set consists of data on leukemia incidence between 1978 and 1982 in eight counties in upstate New York. The eight-county region is divided into 790 cells (census block groups or census tracts). For each block group or tract, the population at risk, the count of incident leukemia cases and the geographic centroid are available. Cases with incomplete location data are fractionally assigned to the possible block groups or tracts in proportion to the populations. Additional background information on the New York leukemia data is available elsewhere (Waller et al. 1994). The observed leukemia rate for each cell is displayed in Fig. 1 according to the Dirichlet tessellation of the cell centroids.

Previously, we reported results from a model using a variable number of clusters k (Gangnon and Clayton 2003). In that analysis, the set of potential clusters consisted of 191,129 circular clusters centered at the block group or tract centroids with $r_{\max} = 20$ km. For the approximate uniform prior on circular clusters described in Sect. 2 and a discrete uniform prior on $0, 1, \ldots, 10$ for k, the posterior distribution for k is given in Fig. 2. The posterior mode for k is 3, and there appears to be strong evidence for at least 2 and no more than five clusters. The posterior probability of fewer than two clusters is 0.063; the posterior probability of more than five clusters is 0.035.

Based on these findings, for models using a fixed number of clusters, we considered four different choices for the numbers of clusters: $k = 3$, the posterior mode; $k = 5$, the plausible upper limit based on the posterior; $k = 10$, the actual upper limit in the prior analysis; and $k = 20$, an even larger number of clusters. Using this series of models, we explored the ability of models with a fixed number of clusters to provide insights into the cluster locations and risks.

For each model, we ran five independent Markov chains, following the advice of Gelman and Rubin (1992). Each chain used a run-in of 100,000 iterations, and 1 million further iterations to obtain the sample of models. Every 100th sample was kept. A subset of the parameters was graphically monitored across the five chains, and all parameters of interest were monitored using Gelman–Rubin statistics. The chains

Fig. 1 Observed cell-specific 5-year leukemia incidence rates for the New York data. Rates are given relative to the overall leukemia rate of 5.5 cases per 10,000 persons. Region associated with each cell based on Dirichlet tessellation of cell centroids. By row from the upper left (northwest) to the lower right (southeast), the eight counties are Cayuga, Onondaga, Madison, Tompkins, Cortland, Chenango, Tioga, and Broome

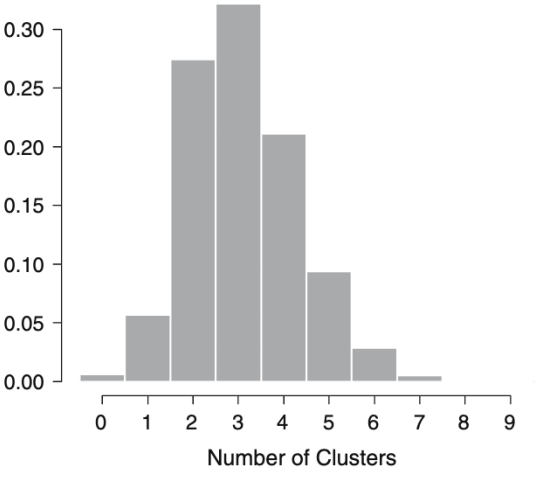

Fig. 2 Posterior distribution of the number of clusters k for the New York leukemia data based on a uniform prior for k (Gangnon and Clayton 2003)

appeared to have converged by that point, and there were no substantial differences in the samples across the chains.

Initially, we focused on inferences about the clustering component of the model. In Fig. 3, we display the prior and posterior probabilities that each cell belongs to one or more clusters, e.g., $P(\phi_i \neq 0) = P(\sum_{j=1}^{k} \mathrm{I}_{\{i \in C_j\}} > 0)$ and $P(\phi_i \neq 0 | y_1, y_2, \ldots, y_n)$, for the four different values of k. Naturally, as the number of clusters k increases, the prior probability that each cell belongs to one or more clusters increases. For $k = 3$, the prior probability ranges from 0.036 to 0.103, with a median of 0.083; for $k = 20$, the

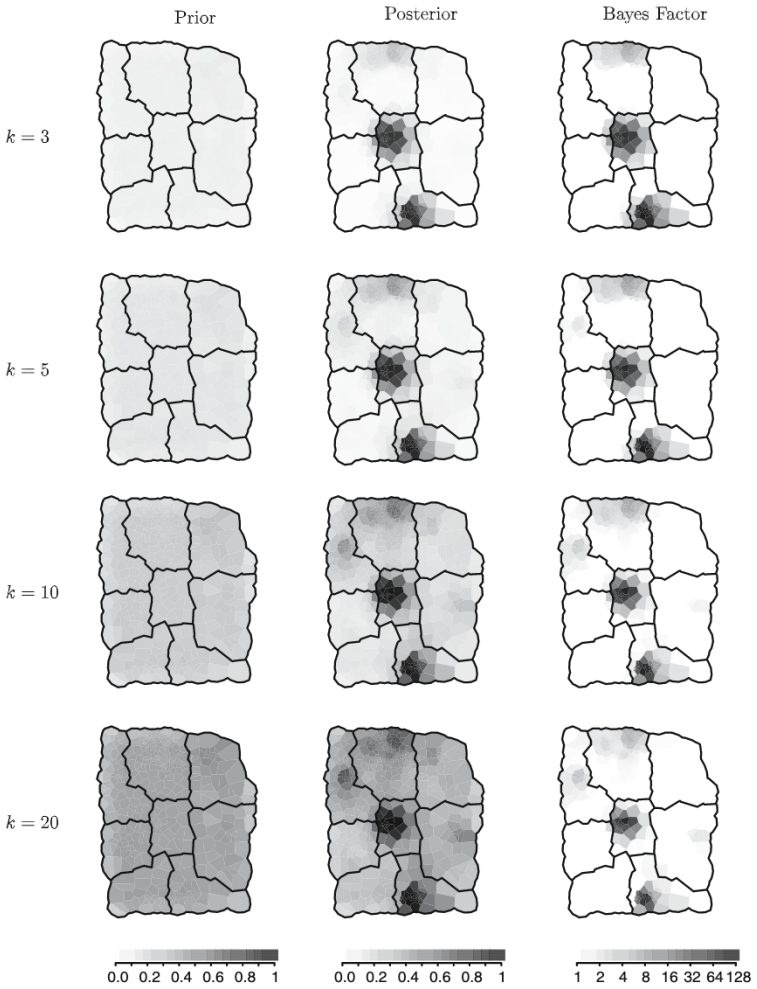

Fig. 3 Inferences about cluster locations in the New York leukemia data using models with fixed $k = 3, 5, 10, 20$: prior probability of cluster membership $P(\phi_i \neq 0)$, posterior probability of cluster membership $P(\phi_i \neq 0|y_1, y_2, \ldots, y_n)$, and Bayes factor for cluster membership

range is 0.22–0.52, with a median of 0.44. Consequently, the posterior probabilities also increase across the entire map, as the number of clusters increases. Nonetheless, in all four plots, we observe high posterior probabilities associated with two areas: a portion of Broome county in the southern section of the map and a portion of Cortland county in the center of the map. In addition, we observe somewhat high-posterior probabilities associated with two additional areas: a large portion of Onondaga County north of Syracuse and a portion of Cayuga County in the northwest section of the map.

To mitigate the influence of the prior and to facilitate comparisons between models with different values for k, we also display the Bayes factor, the ratio of the posterior odds to the prior odds of clustering for each cell. The Bayes factor summarizes the evidence provided by the data in favor of models in which the given cell belongs to one or more clusters, as opposed to models in which the given cell belongs to the

background. In practice, we might hope or even expect that the Bayes factors would be reasonably comparable for different numbers of clusters k.

In Fig. 3, we also display the Bayes factor for clustering at each location. These maps are quite consistent. If we interpret the Bayes factor using the scale proposed by Kass and Raftery (1995), all four maps show strong evidence (a Bayes factor of 20-150) for clustering in Broome and Cortland counties and positive evidence (a Bayes factor of 3-20) for clustering in Onondaga county. For the three larger values of k, there is also positive, albeit weaker, evidence for clustering in Cayuga county. The lack of evidence for this fourth area of clustering in the $k = 3$ model is a natural consequence of assessing the evidence for a fourth cluster in a three cluster model. For models with $k = 10$ and $k = 20$, we also observe weak evidence for clustering in a portion of Chenango county.

Overall, the conclusions about the numbers and locations of clusters to be drawn from the $k = 10$ and $k = 20$ models are quite similar to those drawn from formal inference on k based on RJMCMC. There is compelling evidence for at least two clusters (located in Broome and Cortland counties), substantial evidence for a third cluster (located in Onondaga county), and weaker evidence for one or two additional clusters (located in Cayuga and Chenango counties). There is no evidence in the data for more than five areas of clustering.

In Fig. 4, we display the posterior means for the disease rate in each cell ρ_i, along with the posterior standard deviation for $\log(\rho_i)$. For ease of interpretation, the disease rate is given relative o the overall disease rate of 5.5 cases per 10,000 people. Based on the mapped posterior means, we observe that the areas of clustering in Broome and Cortland counties are associated with elevated leukemia rates, the area of clustering in Onondaga county is associated with lowered lekuemia rates, and the possible clusters in Cayuga and Chenango counties are associated with elevated leukemia rates. Not surprisingly, as the number of clusters in the model is increased, the posterior mean is more variable across cells, e.g., the map is less smooth. There is a corresponding overall increase in the variability of the posterior distributions, particularly in cells that do not belong to the clusters identified above. This probably occurs because, as the number of clusters in the model increases from 5 to 20, our posterior uncertainty about the cluster memberships of these cells increases, e.g., the posterior probability that these cells belong to one or more clusters moves closer to 0.5. Conversely, the posterior standard deviation falls in the areas of clustering in Broome and Cortland counties as our posterior uncertainty about the cluster memberships of these cells decreases, e.g., the posterior probability that these cells belong to one or more clusters increases toward 1. Outside of the apparent clustering, there is relatively little variation in the leukemia rates. The posterior median (95% credible interval) for the random effects standard deviation is 0.11 (0.05,0.30) for $k = 5$, 0.11 (0.05,0.28) for $k = 10$ and 0.11 (0.05, 0.31) for $k = 20$.

4 Application: Wisconsin breast cancer data

We now apply this methodology to a data set on (female) breast cancer incidence in the state of Wisconsin for 1990. Data are available for 716 ZIP code areas. The 1990 ZIP code area defined by ESRI were used, and the geographic centroid and surface area of each ZIP code area were taken from ArcView. The use of ZIP code areas rather than census regions as a unit of analysis is potentially problematic

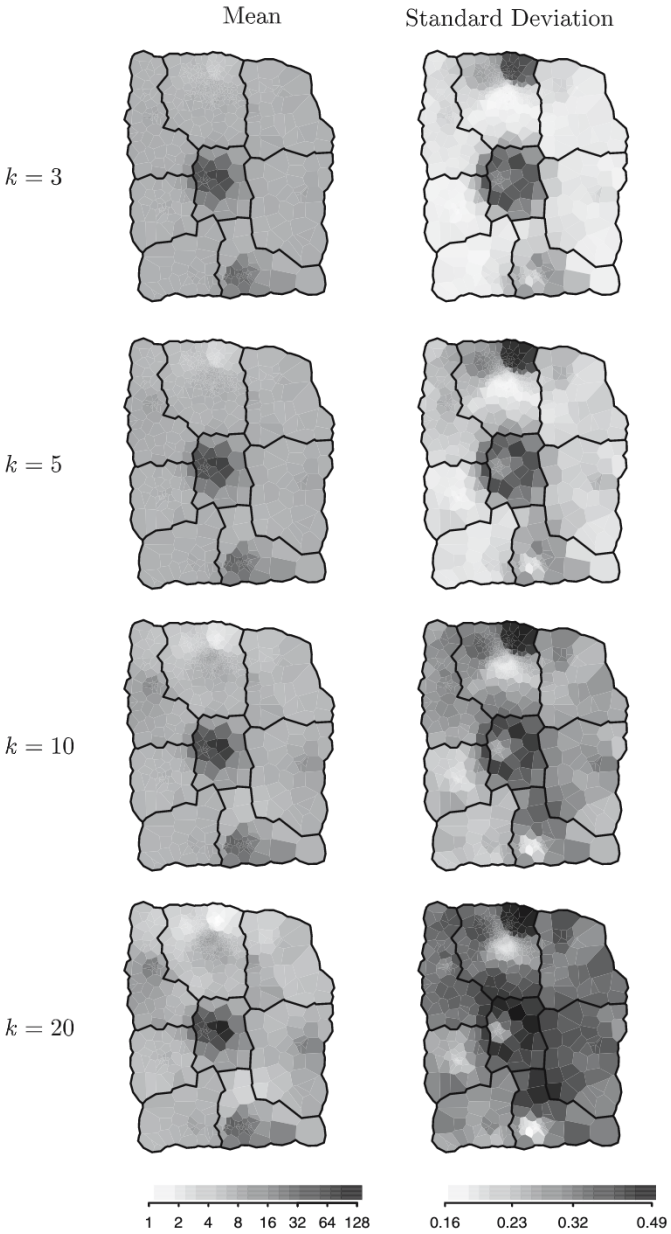

Fig. 4 Inferences about cell-specific leukemia rates for the New York leukemia data using models with fixed $k = 3, 5, 10, 20$: posterior mean for the cell-specific leukemia rate, and posterior standard deviation for the log cell-specific leukemia rate. Rates are given relative to the overall leukemia rate of 5.5 per 10,000 persons

(Krieger et al. 2002, 2003). However, the major concern is the instability of ZIP code boundaries over time. During 1990, there were only a handful of ZIP code changes in Wisconsin. For each ZIP code area, the count of incident breast cancer cases is available from the Wisconsin Cancer Registry. Age-specific female population counts

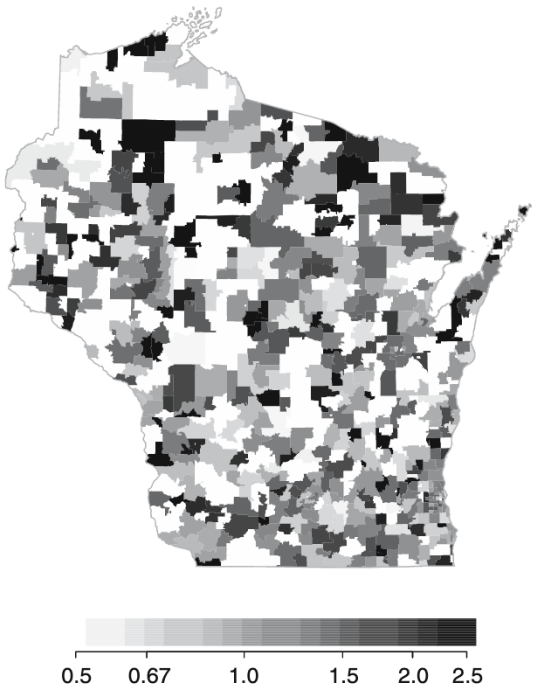

Fig. 5 Observed, age-adjusted standardized incidence ratio for breast cancer by ZIP code area for Wisconsin in 1990

(in 5-year intervals) are available from the Census Bureau. For each ZIP code area, an expected number of breast cancer cases was calculated by means of indirect, internal standardization. The observed standardized incidence ratio for each ZIP code area is displayed in Fig. 5.

For this set of analyses, we used circular clusters centered at the zip code centroids with $r_{\max} = 50\,\text{km}$ as the set of potential clusters, resulting in a total of 29,462 potential clusters. We used the approximate uniform prior on clusters described in Sect. 2. We considered three different choices for the fixed number of clusters in the model: $k = 5$, $k = 10$, and $k = 20$. For each model, we followed the procedures for monitoring convergence described in the previous section.

In Fig. 6, we display the prior and posterior probabilities along with the Bayes factors for cluster membership for each ZIP code area. The behavior of these quantities, as a function of k, is quite similar to the behavior observed in the previous example. As the number of clusters k increases, the prior probability that each cell belongs to one or more clusters increases. For $k = 5$, the prior probability ranges from 0.006 to 0.105, with a median of 0.0.085; for $k = 20$, the range is 0.026 – 0.359, with a median of 0.298. Consequently, the posterior probabilities also increase across the entire map as the number of clusters increases. To ease interpretation, we again focus on the Bayes factors, which are quite consistent for the three values of k. All three models show little, if any, evidence for clustering. There is, at most, positive, but weak, evidence (Bayes factors of 3-5) for two clusters in northwest Wisconsin, one low risk and one high risk.

In Fig. 7, we display the posterior means for the disease rate in each cell ρ_i, along with the posterior standard deviation for $\log(\rho_i)$. As with the New York data, the

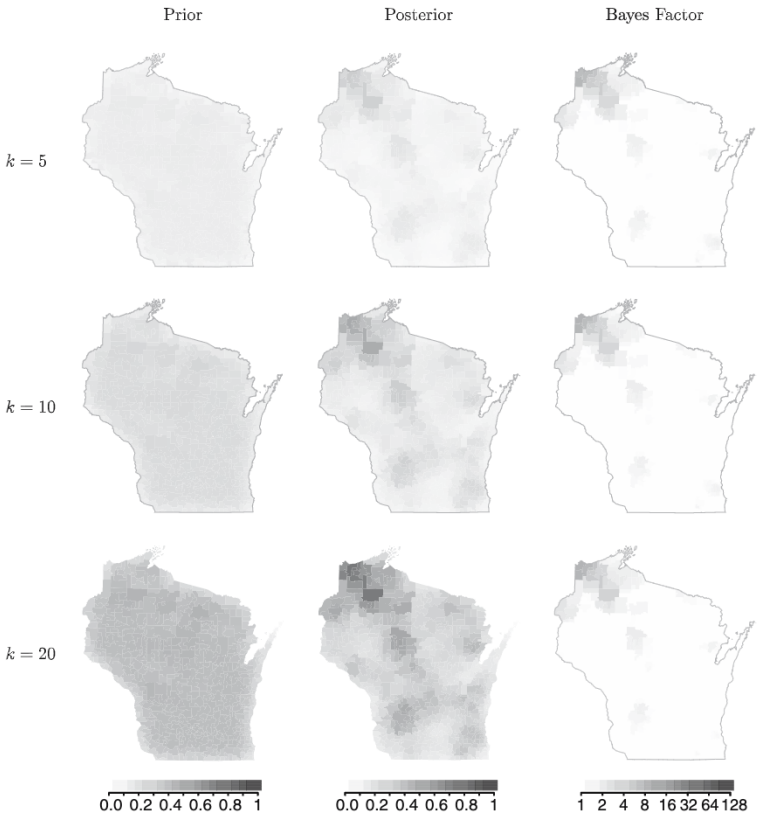

Fig. 6 Inferences about cluster locations in the Wisconsin breast cancer data using models with fixed $k = 5, 10, 20$: prior probability of cluster membership $P(\phi_i \neq 0)$, posterior probability of cluster membership $P(\phi_i \neq 0|y_1, y_2, \ldots, y_n)$, and Bayes factor for cluster membership

map of posterior means is less smooth and the variability of the posterior distributions increases as k increases. Despite the lack of evidence for spatial clustering, there are substantial variations in breast cancer risk across ZIP code areas. The posterior median (95% credible interval) for the random effects standard deviation is 0.25 (0.18,0.32) for $k = 5$, 0.24 (0.17,0.31) for $k = 10$ and 0.24 (0.16,0.31) for $k = 20$.

For comparison, we analyzed these data using the variable k model. We used a discrete uniform prior on $0, 1, \ldots, 10$ for k. The posterior distribution for k is given in Fig. 8; cell-specific posterior summaries are given in Fig. 9. Both the posterior distribution for k and the cell-specific posterior probabilities of cluster membership show, at most, very weak evidence for clustering; the posterior mode for k is 0, and the cell-specific probabilities of cluster membership are all quite small. Overall, the inferences from the variable k model are remarkably similar to the results from the fixed k models.

5 Discussion

In this paper, we revisit the spatial model proposed by Gangnon and Clayton (2003), which includes both spatial clustering and non-spatial random effects. In the prior

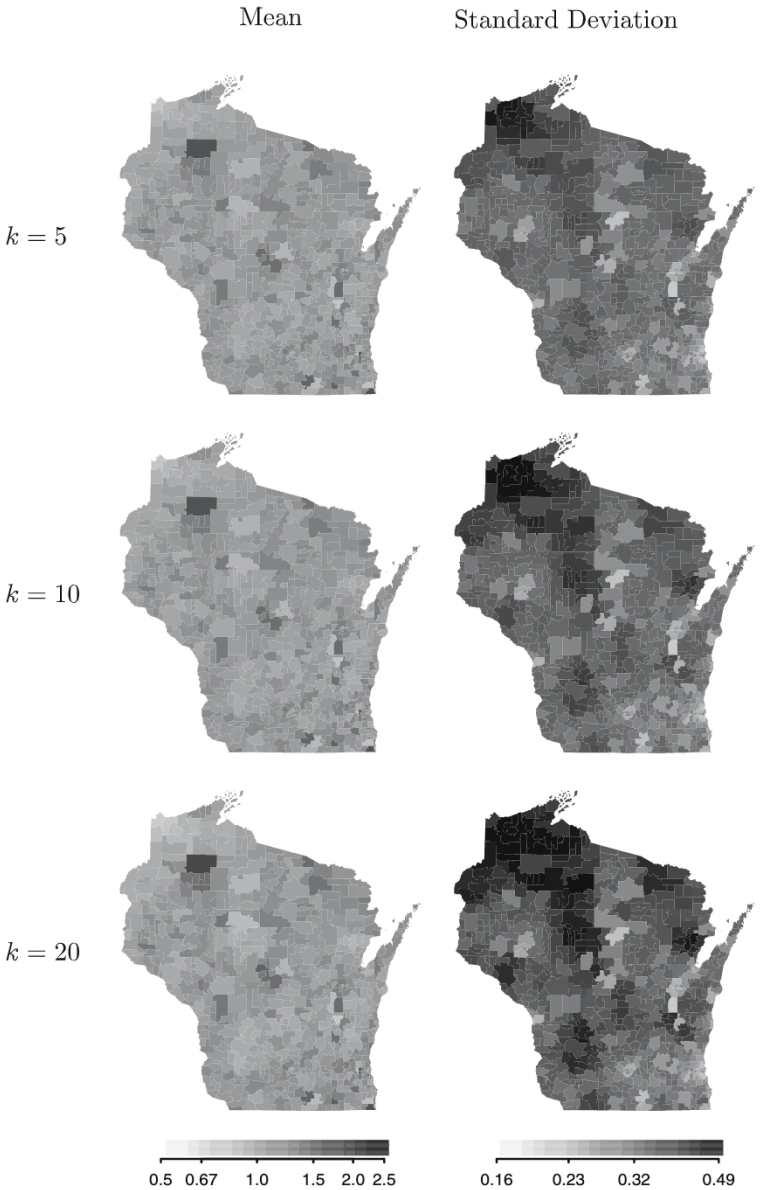

Fig. 7 Inferences about zip code-specific breast cancer risks for the Wisconsin data using models with fixed $k = 5, 10, 20$: Posterior mean for the cell-specific breast cancer risk and posterior standard deviation for the log cell-specific breast cancer risk

work, the number of clusters k was treated as a parameter to be estimated, requiring the use of an RJMCMC algorithm for inference. As an alternative, we consider models with a fixed, but overly large, number of clusters k. Using a fixed value for k, we can estimate the disease risks reasonably well. The identification of clusters is more complicated, because the prior (and hence posterior) probabilities of cluster membership necessarily increase with k. Nonetheless, we can identify the local

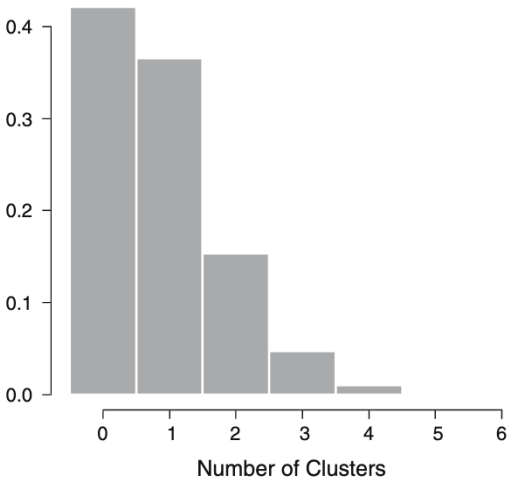

Fig. 8 Posterior distribution of the number of clusters k for the Wisconsin breast cancer data based on a uniform prior for k

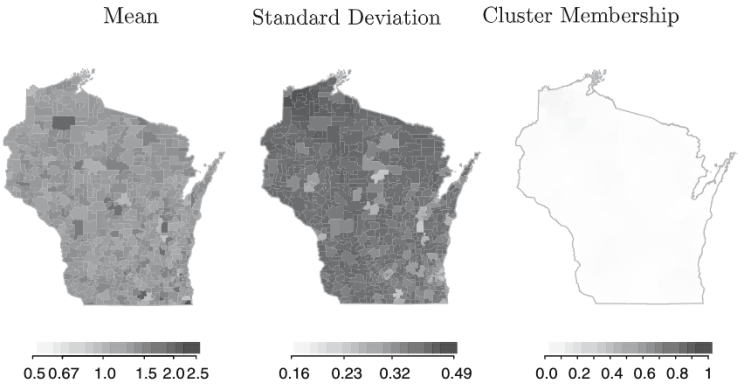

Fig. 9 Inferences about zip code-specific breast cancer risks and cluster locations for the Wisconsin data using model with discrete uniform prior for k: posterior mean for the cell-specific breast cancer risk $E(\rho_i|y_1, y_2, \ldots, y_n)$, posterior standard deviation for the log cell-specific breast cancer risk $V(\log(\rho_i)|y_1, y_2, \ldots, y_n)$, and posterior probability of cluster memebership $P(\phi_i \neq 0|y_1, y_2, \ldots, y_n)$

evidence for clustering using the Bayes factor (the ratio of the posterior odds to the prior odds). There appears to be little dependence of the Bayes factors for clustering on the specific choice of k, as long as the chosen k is greater than the apparent number of clusters.

The two applications illustrate the ability of the fixed k model to assess the strength of evidence for clustering. For both the New York leukemia data and the Wisconsin breast cancer data, there is close agreement between the inferences about numbers and locations of clusters from the fixed k models and from the variable k model. Using the fixed k models, we can easily distinguish between the strong evidence for clustering in the New York data and the lack of evidence for clustering in the Wisconsin data.

There is obviously a loss of efficiency associated with the inclusion of excess clusters in the model, which is clearly reflected in the posterior standard deviations. Although we do not, in general, view this as a serious problem, we could minimize the impact

of the excess clusters by adopting a two-stage procedure. At the first stage, perform an initial fit with a large k to identify an apparent upper bound for k. At the second stage, perform a final fit using the apparent upper bound for k identified in the first stage.

There are computational advantages associated with the fixed k model. It is easier to implement the simple Gibbs steps for changing clusters than the dimension-changing transitions required in an RJMCMC algorithm. It is also easier to monitor convergence with the fixed k model, and, in our experience, the chains appeared to converge more quickly. Most importantly, it is also generally easier to identify an upper bound for k than to specify a full prior distribution for k. Although we have not performed formal comparisons of computational speed, in our experience, there have not been substantial differences in computational speed that would lead one to favor one model over the other.

In the examples, we utilized an approximately uniform (spatially neutral) prior for the clusters, which is ideal for exploratory studies and routine surveillance. However, despite our efforts to achieve neutrality, there are still substantial variations in the prior probability of cluster memberships across the study cells in Figs. 2 and 4. We believe that the use of local Bayes factors advocated here has the potential to minimize the impact of a non-uniform prior on the assessment of evidence for or against clustering, because the Bayes factor naturally factors out the assumed prior. Here, we observed minimal impact of different priors based on changing the value of k. In future work, we plan to explore the robustness of Bayes factors in this setting using different specifications of the cluster prior.

Acknowledgements The authors would like to thank Jane McElroy, Bethany Hendrickson, John Hampton, Patrick Remington, and Amy Trentham-Dietz for their assistance with the Wisconsin breast cancer dataset. Partially supported by grant CA82004 from the National Cancer Institute.

References

Besag J, York J, Mollié A (1991) Bayesian image restoration, with two applications in spatial statistics (with discussion). Ann Inst Stat Math 43:1–59

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

Denison D, Holmes C (2001) Bayesian partitioning for estimating disease risk. Biometrics 57:143–149

Elliott P, Wakefield J, Best N, Briggs D (1999) Spatial epidemiology: methods and applications. Oxford University Press, New York

Ferreira JTAS, Denison DGT, Holmes CC (2002) Partition modelling. In: Lawson AB, Denison DGT (eds) Spatial cluster modelling. Chapman & Hall/CRC, Boca Raton, Florida pp 125–145

Gangnon RE, Clayton MK (2000) Bayesian detection and modeling of spatial disease clustering. Biometrics 56:922–935

Gangnon RE, Clayton MK (2001) A weighted average likelihood ratio test for spatial clustering of disease. Stat Med 20:2977–2987

Gangnon RE, Clayton MK (2003) A hierarchical model for spatially clustered disease rates. Stat Med 22:3213–3228

Gelman A, Carlin JB, Stern HS, Rubin DB (1995) Bayesian data analysis. Chapman & Hall, London

Gelman A, Rubin D (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7:457–472

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

Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773–795

Knorr-Held L, Raßer G (2000) Bayesian detection of clusters and discontinuities in disease maps. Biometrics 56:13–21
Krieger N, Waterman P, Chen JT, Soobader MJ, Subramanian SV, Carson R (2002) Zip code caveat: bias due to spatiotemporal mismatches between zip codes and US census defined geographic areas—the Public Health Disparities Geocoding Project. Am J Public Health 92:1100–1102
Krieger N, Watermen PD, Chen JT, Soobader MJ, Subramanian SV (2003) Monitoring socioeconomic inequalities in sexually transmitted infections, tuberculosis, and violende: geocoding and choice of area-based socioeconomic measures—the Public Health Disparities Geocoding Project. Public Health Rep 118:240–260
Lawson AB (2001) Statistical methods in spatial epidemiology. John, New York
Lawson AB, Bohning D, Biggeri A, Viel J-F, Bertollini R (1999) Disease mapping and risk assessment for public health. John, New York
Lawson AB, Denison DGT (eds) (2002) Spatial cluster modelling. Chapman & Hall/CRC, Boca Raton, Florida
Waller LA, Turnbull BW, Clark LC, Nasca P (1994) Spatial patten analyses to detect rare disease clusters. In: Lange N, Ryan L, Billard L (eds) Case studies in biometry. John, New York pp 3–22

Biographical sketches

Dr. Ronald Gangnon is Associate Scientist in the Department of Biostatistics and Medical Informatics at the University of Wisconsin-Madison. He received his bachelor's degree in 1992 from the University of Minnesota–Duluth and his Ph.D. in 1998 from the University of Wisconsin-Madison. His research interests are in spatial epidemiology, with particular emphasis on cluster detection and, more generally, in the application of statistics to medical data.

Dr. Murray Clayton is Professor in the Departments of Statistics and Plant Pathology at the University of Wisconsin-Madison. He received his bachelor's degree in 1979 from the University of Waterloo in Canada, and his Ph.D. in 1983 from the University of Minnesota. His research interests generally involve the application of statistics to problems in agricultural, biological, and environmental sciences. His recent work has focused especially on the detection of spatial clustering of human diseases and on the spatial analysis of categorical data.