Research Article

Statistics in Medicine

Received 1 May 2009, Accepted 5 May 2010 Published online 18 June 2010 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.3984

A model for space–time cluster detection using spatial clusters with flexible temporal risk patterns

Ronald E. Gangnon[*†]

Maps of estimated disease rates over multiple time periods are useful tools for gaining etiologic insights regarding potential exposures associated with specific locations and times. In this paper, we describe an extension of the Gangnon–Clayton model for spatial clustering to spatio-temporal data. As in the purely spatial model, a large set of circular regions of varying radii centered at observed locations are considered as potential clusters, e.g. subregions with a different pattern of risk than the remainder of the study region. Within the spatio-temporal model, no specific parametric form is imposed on the temporal pattern of risk within each cluster. In addition to the clusters, the proposed model incorporates spatial and spatio-temporal heterogeneity effects and can readily accommodate regional covariates. Inference is performed in a Bayesian framework using MCMC. Although formal inferences about the number of clusters could be obtained using a reversible jump MCMC algorithm, we use local Bayes factors from models with a fixed, but overly large, number of clusters to draw inferences about both the number and the locations of the clusters. We illustrate the approach with two applications of the model to data on female breast cancer mortality in Japan and evaluate its operating characteristics in a simulation study. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: cluster detection; spatio-temporal modeling; Markov chain Monte Carlo; Bayes factor; Poisson

1. Introduction

Cluster or anomaly detection in spatial or spatio-temporal data is an important problem in spatial epidemiology and public health surveillance. Cluster detection is the identification of geographically adjacent locations, possibly during a select time period, associated with distinctive disease risks, typically elevated but possibly reduced, relative to normal or background variation in disease rates. Cluster detection, as considered here, is distinct from two related problems: global clustering, e.g. the general tendency for cases to occur near other cases [1], and focused clustering, e.g. the estimation of risk patterns about pre-specified locations [2].

Spatial and spatio-temporal cluster detection problems have been most widely approached within a frequentist hypothesis testing framework. The most popular approaches, the spatial [3, 4] or spatio-temporal [5, 6] scan statistic and their many variants [7–13], are based on the simultaneous evaluation, via Monte Carlo simulation under an assumed null hypothesis of constant risk, of the statistical significance of the largest likelihood ratio test statistic over a series of two-parameter models associated with a large collection of potential clusters of a particular regular geometric form, e.g. circles. As noted by Lawson [14], it is quite difficult to incorporate realistic heterogeneity in background disease risks into these testing procedures.

An alternative approach to cluster detection builds on Bayesian models for spatially or spatio-temporally smoothed rates [15–20]. Although these models do not explicitly incorporate clusters, implicit evidence for hot spot (single cell) clusters can be assessed by examining the so-called exceedence probabilities, either based on residuals (identifying areas at elevated risk relative to model predictions) or posterior estimates (identifying areas at especially elevated

Departments of Biostatistics and Medical Informatics & Population Health Sciences, University of Wisconsin—Madison, Madison, WI, U.S.A.
[*]*Correspondence to: Ronald E. Gangnon, Department of Population Health Sciences, University of Wisconsin, 610 Walnut St., Madison, WI 53726, U.S.A.*
[†]*E-mail: ronald@biostat.wisc.edu*

risk within the specified model) [21]. These *ad hoc* measures do not consider information on plausible cluster shapes or even simple neighborhood structures, nor do they integrate information from both the residuals and the posterior estimates.

A less common, but very attractive, approach to spatial cluster detection uses models for the underlying disease rates that incorporate explicit spatial clusters associated with distinctive, typically elevated, risks [22–24]. These models allow for formal inference regarding the number, locations and risks associated with clusters relative to a model-specified and possibly non-uniform background risk. In the context of spatio-temporal data, Yan and Clayton [25] have previously described an extension of one of these models, the Gangnon–Clayton model [24], to spatio-temporal cluster detection.

In this paper, we describe a different, potentially more flexible, extension of the Gangnon–Clayton model for spatial clustering to spatio-temporal data. In contrast to the Yan–Clayton model, our model utilizes the spatial and temporal structure in constructing the spatial heterogeneity effects and allows for unstructured temporal risk patterns within spatial clusters rather than restricting attention to cylindrical spatio-temporal clusters. We provide formal posterior inferences regarding parameters relevant to cluster detection, including Bayes factors for cluster membership by location and posterior means and standard deviations for cluster-specific log relative risks over time.

In Section 2, we review the Gangnon–Clayton model for spatial clustering and the Yan–Clayton extension of this model to spatio-temporal clustering. We then propose our alternative extension of the Gangnon–Clayton model, a model with spatial clusters with flexible temporal risks. In Section 3, we present two applications of the proposed model using data on female breast cancer mortality in Japan, analyzing both the subset of the data previously analyzed by Yan and Clayton [25] and the full data set. In Section 4, we present a simulation study evaluating cluster detection rates for the Yan–Clayton model and the proposed model with known true models. Finally, in Section 5, we make some concluding remarks.

2. Previous and new model development

2.1. Gangnon–Clayton model for spatial clusters

Since the model for spatial clusters proposed by Gangnon and Clayton [24] serves as the basis for the development of our model for spatio-temporal clusters, we review it in detail here. Consider a study region divided into N subregions or cells. For cell i at spatial location $\mathbf{x}_i$, let y_i be the observed number of events (cases of disease or deaths), E_i be the expected number of events (based on internal or external standardization) and ρ_i be the unknown relative risk, e.g. $E(y_i)=\rho_i E_i$. Assume that y_i, $i=1,2,\ldots,N$, are independent and distributed as Poisson$(\rho_i E_i)$ conditional on the cell-specific relative risk parameters $\rho_1,\rho_2,\ldots,\rho_N$. The log relative risk, $\eta_i=\log\rho_i$, is modeled as

$$\eta_i=\alpha+\varepsilon_i+\sum_{j=1}^{k}\theta_j\mathbf{1}\{d(\mathbf{x}_i,\mathbf{c}_j)\leqslant r_j\} \tag{1}$$

There are three distinct components to this model. The non-spatial component of the model is the intercept α, which is given a flat prior. The spatially unstructured random effects component is ε_i, where $\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_N$ are assumed to be independent and identically distributed (iid) N$(0,1/\pi)$, where π is the unknown random effects precision. π is given a diffuse conjugate gamma prior (mean 100, variance 100).

The spatial clustering component of the model is $\sum_{j=1}^{k}\theta_j\mathbf{1}\{d(\mathbf{x}_i,\mathbf{c}_j)\leqslant r_j\}$, where k is the number of clusters; $\mathbf{c}_j$, r_j are the center and radius of circular (in metric d) cluster j associated with log relative risk θ_j, $j=1,2,\ldots,k$, and $\mathbf{1}\{A\}$ is the indicator of A, which takes the value 1 if A is true and 0 if A is false. The log relative risks $\theta_1,\theta_2,\ldots,\theta_k$ are given a N$(0,\sigma_\theta^2)$ prior. The cluster centers and radii $\mathbf{c}_j$, r_j, $j=1,2,\ldots,k$, are given the so-called 'dartboard' prior [26] in which the center $\mathbf{c}_j$ is chosen with probability proportional to the cell area and the radius r_j is chosen from a uniform distribution of $(0,r_{\max})$. σ_θ^2 and $r_{\max}$ are user-specified parameters and specific values should be chosen based on expert judgment on plausible magnitudes of risk and spatial extent of clusters in a particular application.

The number of clusters k may be treated either as a parameter to be estimated [24] or as a fixed user-specified constant [27, 28]. In the former case, k is assigned a discrete uniform prior on $0,1,\ldots,k_{\max}$, and inference is based on reversible jump Markov chain Monte Carlo (RJMCMC) techniques as described previously. In the latter case, k should be chosen to be a conservative upper bound on the true number of clusters, e.g. k should be much larger than the anticipated true number of clusters. If the true number of clusters is not greater than the specified k, then the specified model is correct, albeit likely overparameterized.

Consistent inferences about the strength of evidence for clustering at each cell can be obtained using local Bayes factors for clustering [27, 28],

$$\mathrm{BF}_i = \frac{\dfrac{\Pr\left(\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j\} > 0 | \mathbf{y}\right)}{1 - \Pr\left(\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j\} > 0 | \mathbf{y}\right)}}{\dfrac{\Pr\left(\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j\} > 0\right)}{1 - \Pr\left(\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j\} > 0\right)}},$$

the ratio of the posterior odds to prior odds in favor of clustering for each location. For the fixed k model, $\Pr(\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j\} > 0) = 1 - (1 - p_i)^k$, where the probability that cell i belongs to one cluster selected from the prior distribution is p_i. For the variable k model, the prior odds of clustering based on the full prior for k cannot be used as the denominator in the local Bayes factor due to its sensitivity to the often irrelevant choice of $k_{\max}$. That is, if $k_{\max}$ is chosen to be appropriately large, the choice of $k_{\max}$ will not impact the posterior samples and thus should be irrelevant to our assessment of the evidence of clustering. Rather, the relevant reference is determined from the scenario in which we are indifferent to the choice between the correct model (with k clusters) and the same model with an additional random cluster (with $\theta_{k+1} \approx 0$). In this scenario, the prior probability of selecting the model with the additional cluster is $\frac{1}{2}$ and the probability that the additional cluster contains cell i is p_i; hence, the prior probability in favor of clustering at cell i is $p_i/2$.

For both the fixed k and variable k models, we interpret the local Bayes factors as the strength of evidence in favor of clustering using the scale proposed by Jeffreys [29]. A local Bayes factor of 3–10 represents substantial evidence in favor of clustering; a local Bayes factor of 10–30 represents strong evidence in favor of clustering; a local Bayes factor of 30–100 represents very strong evidence in favor of clustering; and a local Bayes factor greater than 100 represents decisive evidence in favor of clustering.

2.2. *Yan–Clayton extension to cylindrical spatio-temporal clusters*

Yan and Clayton [25] proposed the following extension of the Gangnon–Clayton model for spatial clusters to spatio-temporal data based on cylindrical clusters in space–time. Consider a study region divided into N cells and a study time period divided into T time intervals. For space–time cell i, t at spatial location $\mathbf{x}_i$ and time interval t, let y_{it} be the observed number of events, E_{it} be the expected number of events and ρ_{it} be the unknown relative risk. Assume that y_{it}, $i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, T$, are conditionally independent and distributed as Poisson$(\rho_{it} E_{it})$ given $\rho_{11}, \rho_{12}, \ldots, \rho_{NT}$. The log relative risk, $\eta_{it} = \log \rho_{it}$, is modeled as

$$\eta_{it} = \alpha_t + \varepsilon_{it} + \sum_{j=1}^{k} \theta_j \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j, l_j \leqslant t \leqslant u_j\} \tag{2}$$

The intercept α in the Gangnon–Clayton model is replaced with time-varying intercepts α_t, which are given a flat prior. The spatio-temporally unstructured random effects ε_{it}, $i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, T$, are assumed to be iid $\mathrm{N}(0, 1/\pi)$, where π is the unknown random effects precision. π is given a diffuse conjugate gamma prior (mean 100, variance 100).

The spatio-temporal clustering component of the model is $\sum_{j=1}^{k} \theta_j \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j, l_j \leqslant t \leqslant u_j\}$, where k is the number of clusters; $\mathbf{c}_j$, r_j are the center and radius of circle (in metric d) defining the spatial extent of cluster j and l_j, u_j are the lower and upper limits of the time interval defining the temporal extent of cluster j and θ_j is the log relative risk associated with cluster j. The log relative risks $\theta_1, \theta_2, \ldots, \theta_k$ are given a $\mathrm{N}(0, \sigma_\theta^2)$ prior. The centers and radii $\mathbf{c}_j$, r_j, $j = 1, 2, \ldots, k$, are again given the so-called 'dartboard' prior [26] in which the center $\mathbf{c}_j$ is chosen with probability proportional to the cell area and the radius r_j is chosen from a uniform distribution of $(0, r_{\max})$. The lower and upper temporal limits l_j, u_j are given a uniform prior over the set of all temporal intervals with length not more than $L_{\max}$: $\{(l, u) : l \leqslant u, u - l \leqslant L_{\max}\}$.

σ_θ^2, $r_{\max}$ and $L_{\max}$ are user-specified parameters and specific values should be chosen based on expert judgment on plausible magnitudes of risk and spatial/temporal extent of clusters in a particular application. As in the purely spatial model, the number of clusters k may be treated either as a parameter to be estimated or as a fixed user-specified constant.

Evidence for clustering at a given location (at any time) can be assessed using the local Bayes factors for cell i,

$$\mathrm{BF}_i = \frac{\dfrac{\Pr\left(\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j\} > 0 | \mathbf{y}\right)}{1 - \Pr\left(\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j\} > 0 | \mathbf{y}\right)}}{\dfrac{\Pr\left(\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j\} > 0\right)}{1 - \Pr\left(\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j\} > 0\right)}}$$

and evidence for clustering at a given location at a specific time using the local Bayes factor for space–time cell i, t,

$$\mathrm{BF}_{it} = \frac{\dfrac{\Pr\left(\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j, , l_j \leqslant t \leqslant u_j\} > 0 | \mathbf{y}\right)}{1 - \Pr\left(\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j, , l_j \leqslant t \leqslant u_j\} > 0 | \mathbf{y}\right)}}{\dfrac{\Pr\left(\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j, , l_j \leqslant t \leqslant u_j\} > 0\right)}{1 - \Pr\left(\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j, , l_j \leqslant t \leqslant u_j\} > 0\right)}}$$

2.3. Proposed model for spatial clusters with flexible temporal risks

The description of our new spatio-temporal model parallels the description of the prior two models. As in the Yan–Clayton model [25], consider a study region divided into N cells and a study time period divided into T time intervals. For space–time cell i, t at spatial location $\mathbf{x}_i$ and time interval t, let y_{it} be the observed number of events, E_{it} be the expected number of events and ρ_{it} be the unknown relative risk. Assume that y_{it}, $i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, T$, are conditionally independent and distributed as Poisson($\rho_{it} E_{it}$) given $\rho_{11}, \rho_{12}, \ldots, \rho_{NT}$. The log relative risk, $\eta_{it} = \log \rho_{it}$, is modeled as

$$\eta_{it} = \alpha + \tau_t + \varepsilon_i + \gamma_{it} + \sum_{j=1}^{k} \theta_{jt} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j\} \tag{3}$$

The non-clustering component of the model is conceived as two-factor analysis of variance (ANOVA) model with cell and time as the two factors. The intercept α is given a flat prior. The cell main effect parameters ε_i, $i = 1, 2, \ldots, N$, are assumed to be iid N(0, $1/\pi_\varepsilon$), the temporal main effect parameters τ_t, $t = 1, 2, \ldots, T$, are assumed to be iid N(0, $1/\pi_\tau$) and the cell–time interaction parameters γ_{it}, $i = 1, 2, \ldots, N$, $t = 1, 2, \ldots, T$, are assumed to be iid N(0, $1/\pi_\gamma$). The precision parameters π_ε, π_τ, π_γ are each given diffuse conjugate gamma priors (mean 100, variance 100).

The clustering component of the model is $\sum_{j=1}^{k} \theta_{jt} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j\}$, where k is the number of clusters; $\mathbf{c}_j, r_j$ are the center and radius of circle (in metric d) defining the spatial extent of cluster j and θ_{jt} is the log relative risk associated with cluster j for time interval t. The vector of log relative risks $(\theta_{j1}, \theta_{j2}, \ldots, \theta_{jT})$ is given a multivariate normal prior with mean $(0, 0, \ldots, 0)$ and covariance matrix Σ_θ. The centers and radii $\mathbf{c}_j$, r_j, $j = 1, 2, \ldots, k$, are given the so-called 'dartboard' prior in which the center $\mathbf{c}_j$ is chosen with probability proportional to the cell area and the radius r_j is chosen from a uniform distribution of $(0, r_{\max})$.

Σ_θ and $r_{\max}$ are user-specified parameters and specific values should be chosen based on expert judgment on plausible magnitudes of risk and spatial extent of clusters in a particular application. In many cases, it will be convenient to take $\Sigma_\theta = \sigma_\theta^2 \mathbf{I}$, a constant multiple of the identity matrix, allowing the data to fully inform the cluster risk estimates. As in the purely spatial model, the number of clusters k may be treated either as a parameter to be estimated or as a fixed user-specified constant. Evidence for clustering at a given location can be assessed using the local Bayes factor for cell i,

$$\mathrm{BF}_i = \frac{\dfrac{\Pr\left(\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j\} > 0 | \mathbf{y}\right)}{1 - \Pr\left(\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j\} > 0 | \mathbf{y}\right)}}{\dfrac{\Pr\left(\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j\} > 0\right)}{1 - \Pr\left(\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j\} > 0\right)}}$$

There are two key differences between model (3) and the Yan–Clayton model. First, model (3) utilizes the cell and temporal structure of the underlying data to define the heterogeneity effects, e.g. the normal or background variation in

disease rates against which clustering is assessed, whereas the Yan–Clayton model does not. Second, model (3) uses a more expansive definition of the clustering effect. In our model, clustering is defined as any common temporal risk pattern shared by adjacent cells, whereas, within the Yan–Clayton model, clustering is restricted to one common elevated (or decreased) risk for one or more adjacent time periods shared by adjacent cells.

2.4. *Posterior inference*

If the cluster location ($\mathbf{c}_j$) and extent parameters (r_j for the Gangnon–Clayton and proposed models; r_j, l_j, u_j for the Yan–Clayton model) are known, each of the above models is a hierarchical Poisson generalized linear model. Techniques for sampling from the posterior distribution are described in Gelman *et al.* [30]. for general models of this type and in Gangnon and Clayton [24] for the spatial cluster model. Parameters in the linear model are sampled using a Metropolis–Hastings algorithm [31] with proposal distributions based on a quadratic approximation to the likelihood, which is conjugate to the normal priors. Posterior samples for the precision parameters of the normal prior distributions are obtained from the appropriate (conjugate) full conditional distribution.

The joint conditional distribution of the cluster location and extent parameters can be found by direct enumeration of all possible clusters. For efficiency, samples are drawn using a Metropolis–Hastings update based on a truncated version of the full conditional distribution as a proposal distribution [24]. Updates of the chain typically consist of a single update of the cluster location and extent parameters followed by multiple (typically 10) updates of the other parameters.

The largest computational burden in fitting these models is the search over the set of potential cluster location and extent parameters. In the Yan–Clayton model, the size of the search space depends, in a multiplicative fashion, on the number of potential time intervals and the number of potential spatial clusters so that even small increases in the number of time points can dramatically increase the size of the search space. In contrast, within the proposed model, the number of potential clusters that need to be evaluated for a given study region remains fixed regardless of the number of time points under consideration, making our model potentially suitable for much larger data sets.

3. Example: Japan breast cancer mortality data

In this section, we illustrate the application of the proposed model for spatio-temporal clustering to data on breast cancer mortality in Japan from 1975–1994 [32]. Japan is divided into 47 prefectures, and each prefecture is further divided into numerous municipalities. In our examples, we focus on 3201 municipalities within the 46 prefectures on the four main islands of Japan: Hokkaidō, Honshū, Kyūshū and Shikoku. For each municipality, the locations of the municipality offices are available. Approximate municipality borders and areas were obtained using the Dirichlet tessellation [33] of the locations. A map of the prefecture boundaries is provided in Figure 1.

For each municipality, the numbers of deaths due to breast cancer and the size of the female population are available within 5-year age intervals for each year between 1975 and 1994. Following Yan and Clayton [25], we restrict the analysis to females aged 40–74 years, calculate age-standardized expected numbers of deaths based on the overall age-specific breast cancer mortality rates for each municipality and year and aggregate the data into five time periods: 1975–1978, 1979–1982, 1983–1986, 1987–1990 and 1991–1994.

3.1. *Tochigi, Gunma and Saitama prefectures*

To facilitate comparisons of our model with the Yan–Clayton model, we first applied our model to the subset of the data analyzed by Yan and Clayton [25], consisting of municipalities from three prefectures: Tochigi, Gunma and Saitama (shaded black in Figure 1). Here, we consider the same set of potential spatial clusters used by Yan and Clayton: 8960 circular clusters centered at the municipality office locations with $r_{\max}=30$ km. We take $\sigma_\theta^2=0.355$ so that *a priori* the probability that the cluster risk in any time period falls between $\frac{1}{4}$ and 4 is 0.98. We use three different choices for the *fixed* number of clusters in the model: $k=5$, 10, 20. In a sensitivity analysis, we fit models with $k=10$ using two alternative specifications for σ_θ^2, $\frac{0.355}{4}=0.08875$ or $0.355(4)=1.42$. There were no notable differences between models based on the choice of σ_θ^2; hence, only the results for the $\sigma_\theta^2=0.355$ models are presented here.

In Figure 2, we display, for each choice of k, the local Bayes factors in favor of cluster membership for each location. The three maps are very similar. All three maps show decisive evidence for a cluster in the southern portion of the study region and strong evidence for a cluster in the north central portion of the study region.

In Figure 3, we display, for each choice of k, the posterior means, conditional on cluster membership, of the log relative risks for each time period associated with clustering for each municipality with substantial evidence (a Bayes factor above 10) for clustering. As with the maps of the Bayes factors, these sets of maps of estimated cluster risks are

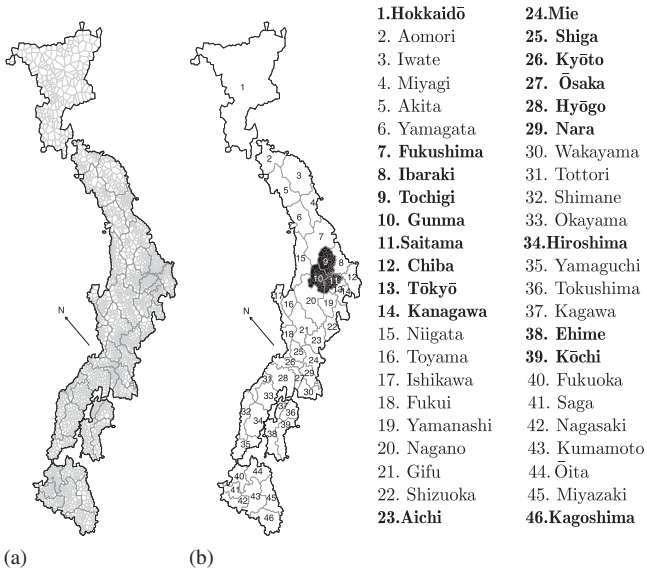

Figure 1. (a) Map of Dirichlet tessellation of 3,201 municipalities on the 4 main islands of Japan. Municipality boundaries in light gray, prefecture boundaries in darker gray, island borders in black. (b) Map of 46 prefectures on the 4 main islands of Japan. Prefecture boundaries in gray. Tochigi, Gunma and Saitama Prefectures shaded in black. Prefectures mentioned in the text are **bolded** in the list of prefecture names.

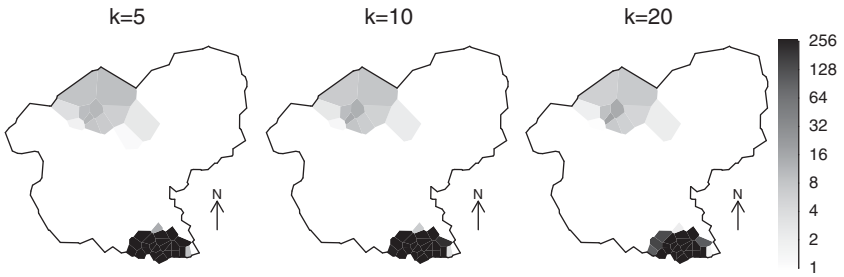

Figure 2. Bayes factor in favor of belonging to a cluster, $\{\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j\} > 0\}$, for each municipality using model (3) with $k = 5$, 10, 20 clusters.

strikingly similar. The cluster in the southern portion of the study region is associated with consistently elevated risks in all time periods (median relative risk across municipalities: 1.42 in 1975–1978, 1.42 in 1979–1982, 1.23 in 1983–1986, 1.29 in 1987–1990 and 1.37 in 1991–1994). The cluster in the north central portion of the study region is associated with intermittently lowered risks (median relative risk across municipalities: 0.66 in 1979–1982, 0.48 in 1983–1986 and 0.42 in 1991–1994).

To compare our results with the results of Yan and Clayton [25], we obtained relevant posterior summary statistics from their model fits from Ping Yan (personal communication). Potential clusters consisted of 89 600 space–time cylinders formed by circles centered at the municipality office locations with $r_{\max} = 30$ km and all possible temporal intervals ($L_{\max} = 5$). The prior variance for cluster risks was $\sigma_\theta^2 = 1$ so that *a priori* the probability that the cluster risk falls between 0.14 and 7.10 is 0.95. The number of clusters k was given a discrete uniform prior on [0, 15], and a reversible jump MCMC algorithm was used for inference.

In Figure 4, we display the local Bayes factor for belonging to a cluster in each time period for each municipality. Inferences drawn from these maps generally match those from our Bayes factor maps in Figure 2. There is decisive evidence, with many local Bayes factors infinite or near infinite, for the cluster identified in the southern portion of the study region in all time periods and strong or very strong evidence for the cluster in the north central portion of the study region in the last four time periods. There is also substantial or strong evidence for a third cluster consisting of a single municipality in the eastern portion of the study region.

In Figure 5, we display the posterior means, conditional on cluster membership, of the log relative risks associated with clustering for each municipality and time period with strong evidence (a local Bayes factor of 10) for clustering. The cluster in the southern portion of the study region is associated with consistently elevated risks in all time periods

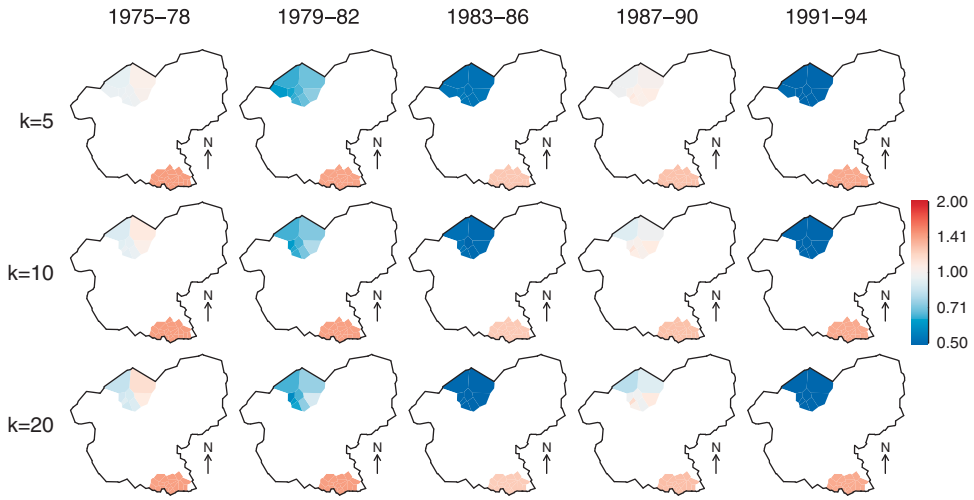

Figure 3. Posterior means, conditional on cluster membership, of the log relative risks associated with clusters, $E(\sum_{j=1}^{k}\theta_{jt}\mathbf{1}\{d(\mathbf{x}_i,\mathbf{c}_j)\leqslant r_j\}|\mathbf{y},\sum_{j=1}^{k}\mathbf{1}\{d(\mathbf{x}_i,\mathbf{c}_j)\leqslant r_j\}>0)$, for each municipality with a Bayes factor for belonging to a cluster greater than 10 using model (3) with $k=5$, 10, 20 clusters.

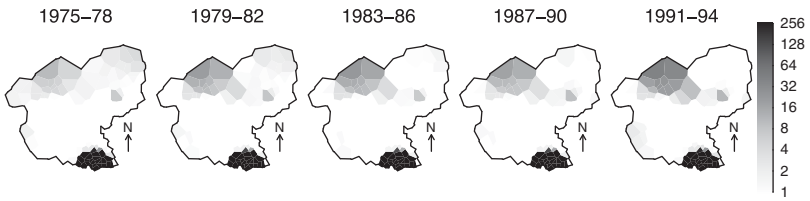

Figure 4. Local Bayes factor for belonging to a cluster in each time period, $P\{\sum_{j=1}^{k}\mathbf{1}\{d(\mathbf{x}_i,\mathbf{c}_j)\leqslant r_j, l_j\leqslant t\leqslant u_j\}>0\}$, for each municipality using model (2) with discrete uniform prior on $[0,15]$ for k from Yan and Clayton.

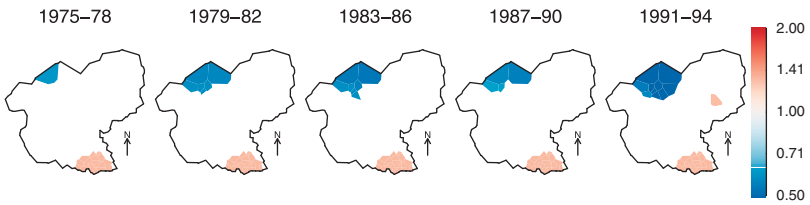

Figure 5. Posterior means, conditional on cluster membership, of the log relative risks associated with clusters, $E(\sum_{j=1}^{k}\sum_{j=1}^{k}\theta_j\mathbf{1}\{d(\mathbf{x}_i,\mathbf{c}_j)\leqslant r_j, l_j\leqslant t\leqslant u_j\}|\mathbf{y},\sum_{j=1}^{k}\mathbf{1}\{d(\mathbf{x}_i,\mathbf{c}_j)\leqslant r_j, l_j\leqslant t\leqslant u_j\}>0)$, for each municipality with a local Bayes factor greater than 10 using model (2) with discrete uniform prior on $[0,15]$ for k from Yan and Clayton.

(relative risk: 1.31). The cluster in the north central portion of the study region is associated with consistently lowered risks after 1978 (median relative risk across cells: 0.55). Although only the estimated risk for the last time period is displayed based on the choice of threshold for the local Bayes factors, the third, single municipality cluster in the eastern portion of the study region is associated with a consistent elevated risk (median relative risk across time periods: 1.30).

Contrasting the inferences based on the two models, there are two notable differences, both of which highlight the advantages of our model formulation. Based on the Yan–Clayton model, the cluster in the north central portion of the study region is associated with a consistently lowered risk from 1979 to 1994, whereas, based on our model, this same region is associated with lowered risk from 1979 to 1986 and from 1991 to 1994, but normal risk from 1987 to 1990. The estimated lowered risk in this region from 1987 to 1990 in the Yan–Clayton model is simply an artifact of the very strong preference, *a priori*, for a single cylindrical cluster bridging the two periods of lowered risk rather than two distinct clusters. It does not indicate actual evidence within the data supporting a lowered risk during that time period.

The other notable difference is the (modest) evidence for a third, single municipality cluster with consistently elevated risks based on the Yan–Clayton model and the lack of evidence for this 'cluster' in our model. This difference reflects the lack of a municipality-level main effect in the specification of the random effects term in the Yan–Clayton

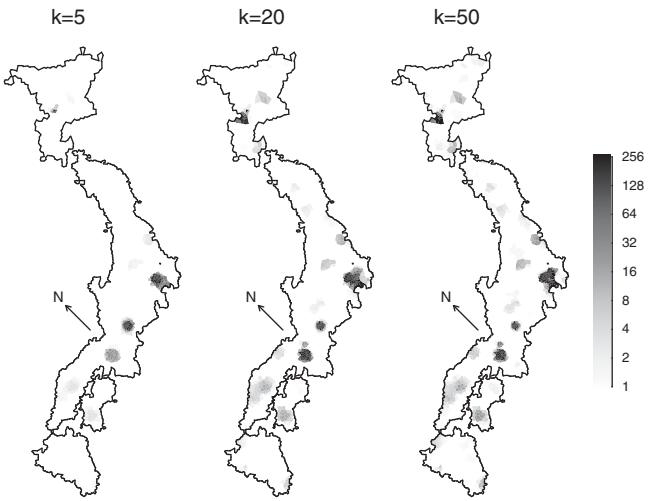

Figure 6. Bayes factor in favor of belonging to a cluster, $\{\sum_{j=1}^{k} \mathbf{1}\{d(\mathbf{x}_i, \mathbf{c}_j) \leqslant r_j\} > 0\}$, for each municipality using model (3) with $k = 5$, 20, 50 clusters.

model. Because of this, the only way to accommodate similarity in risks over time within a single municipality in the Yan–Clayton model is through the clustering effect. Although it is not, strictly speaking, an error to define this behavior as clustering, we believe that it is not desirable to do so. Since one might reasonably expect similarity of risks over time within most, if not all, municipalities, we believe that this behavior should be incorporated into the non-clustering portion of the model. Within our model, the clustering effect is more focused on identifying groups of adjacent locations with similar risk patterns over time and will only identify single locations as clusters if they are associated with very distinctive risk patterns.

3.2. *Four main islands*

We next applied our model to the entire data set consisting of 3201 municipalities on the four main islands of Japan. We again considered circles centered at the municipality office locations with $r_{\max} = 30$ km as potential clusters. There were a total of 99 411 potential clusters. We again take $\sigma_\theta^2 = 0.355$ so that *a priori* the probability that the cluster risk in any time period falls between $\frac{1}{4}$ and 4 is 0.98. We used three different choices for the *fixed* number of clusters in the model: $k = 5$, 20, 50.

In Figure 6, we display, for each choice of k, the local Bayes factors in favor of cluster membership. Based on the $k = 20$ model, there is decisive evidence (a Bayes factor greater than 100) for six distinct areas of clustering, single municipalities in Hokkaidō and Ibraraki prefectures and more extensive areas in Hokkaidō prefecture, in Saitama, Tōkyō and Kanagawa prefectures, in Aichi prefecture and in Ōsaka, Hyōgo and Nara prefectures. In addition, there is strong evidence (a Bayes factor of 10–30) for an additional five areas of clustering in the following prefectures: Fukushima, Ehime and Kōchi, Shiga and Kyōto, Hiroshima, and Kagoshima. There is substantial evidence (a Bayes factor of 3–10) for an additional seven areas of clustering.

Increasing the number of clusters in the model to $k = 50$ produces very similar results. The top 11 distinct areas in terms of evidence for clustering are identical in the $k = 20$ and $k = 50$ models. The only notable changes are increases in the strength of evidence from strong (a Bayes factor of 10–30) to very strong (a Bayes factor of 30–100) for the area of clustering in Shiga and Kyōto prefectures and from substantial (a Bayes factor of 3–10) to strong (a Bayes factor of 10–30) for a third area of clustering in Hokkaidō. Overall, the conclusions about the numbers and locations of clusters to be drawn from the $k = 50$ model are nearly identical to those to be drawn from the $k = 20$ model.

Not surprisingly, results for the $k = 5$ model are somewhat different. Given the evidence for at least 11 distinct areas of clustering from the above models with larger k, it would be unrealistic to expect this underparameterized model to identify all similar areas of clustering. Despite this, the $k = 5$ model does a fairly good job of matching the prior results and demonstrating the need for a larger value of k. There is decisive evidence (a Bayes factor greater than 100) for four of the six areas of clustering previously identified, the single municipalities in Hokkaidō and Ibraraki prefectures and the more extensive areas in Saitama, Tōkyō and Kanagawa prefectures and in Aichi prefecture and very strong evidence (a Bayes factor of 30–100) for a fifth in Ōsaka, Hyōgo and Nara prefectures. There is also substantial evidence for a sixth area of clustering in Hiroshima prefecture. It is somewhat surprising that evidence for clustering is found in this latter region, whereas regions with greater evidence for clustering in the models with larger k, e.g. the second cluster

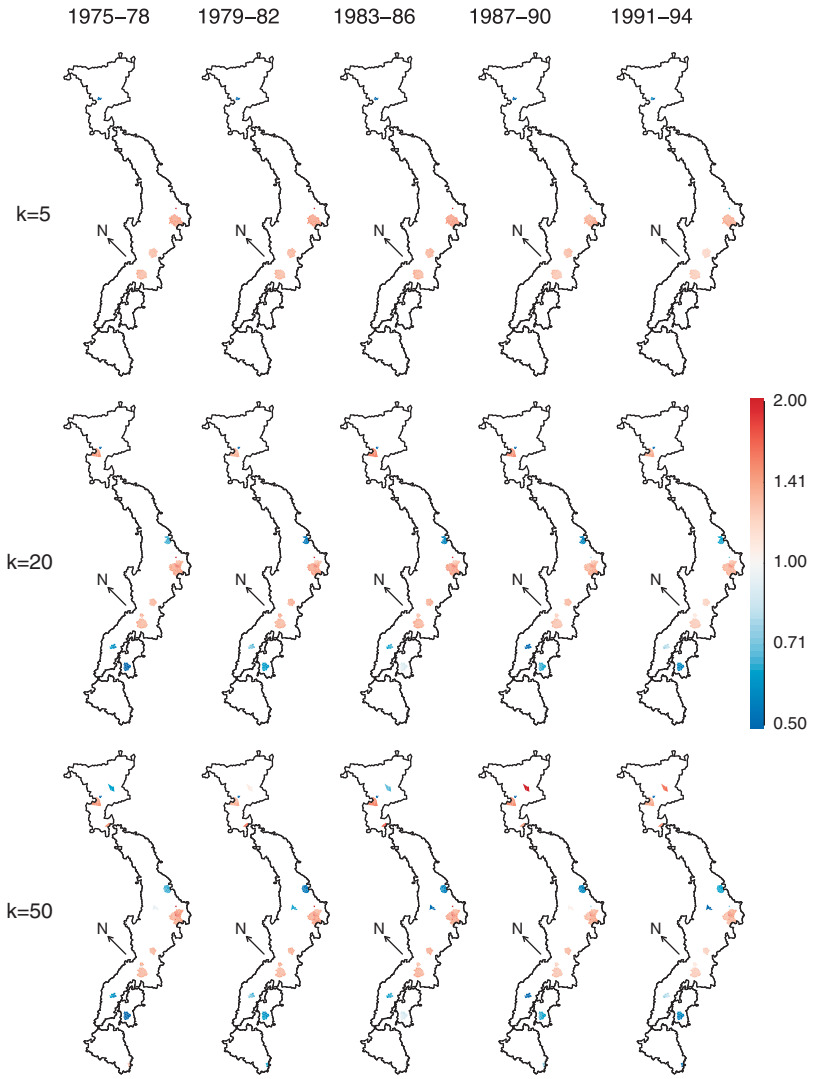

Figure 7. Posterior means, conditional on cluster membership, of the log relative risks associated with clusters, $E(\sum_{j=1}^{k}\theta_{jt}\mathbf{1}\{d(\mathbf{x}_i,\mathbf{c}_j)\leqslant r_j\}|\mathbf{y},\sum_{j=1}^{k}\mathbf{1}\{d(\mathbf{x}_i,\mathbf{c}_j)\leqslant r_j\}>0)$, for each municipality with a Bayes factor for belonging to a cluster greater than 10 using model (3) with $k=5$, 20, 50 clusters.

in Hokkaidō prefecture and the cluster in Shiga and Kyōto prefectures, are not identified. However, given the decisive or very strong evidence for five clusters and substantial evidence for a sixth, one would clearly conclude that $k=5$ is insufficient for our modeling goals.

In Figure 7, we display, for each choice of k, the posterior means, conditional on cluster membership, of the log relative risks for each time period associated with clustering for each location with strong evidence (a Bayes factor above 10) for clustering. These maps are strikingly similar to each other. The only noticeable differences between the three sets of maps reflect the variable impact of hard thresholding based on the Bayes factor for different choices of k. Areas identified in the $k=5$ model tend to have nearly identical risk estimates in the $k=20$ and $k=50$ models. Similarly, areas identified in the $k=20$ model have very similar risk estimates in the $k=50$ model.

In Figure 8, we display, for each choice of k, the posterior mean ± 2 posterior standard deviations, conditional on cluster membership, of the log relative risks for each time period associated with clustering for the single municipality with the largest Bayes factor for clustering in each of the 11 contiguous regions (defined based on geographic proximity and similarity in posterior risk estimates) with Bayes factors for clustering of at least 10 (strong evidence) in the $k=20$ model. As we also observed in Figure 7, the point estimates are very similar across all three models; the standard deviations are also similar. This is true even in cases where the $k=5$ model fails to find evidence for a particular cluster. For the most part, the detected clusters have consistent risks over time. Consistently elevated risks are found in Hokkaidō prefecture (1A), in Saitama, Chiba, Tōkyō and Kanagawa prefectures, in Aichi and Mie prefectures, in Shiga and Kyōto

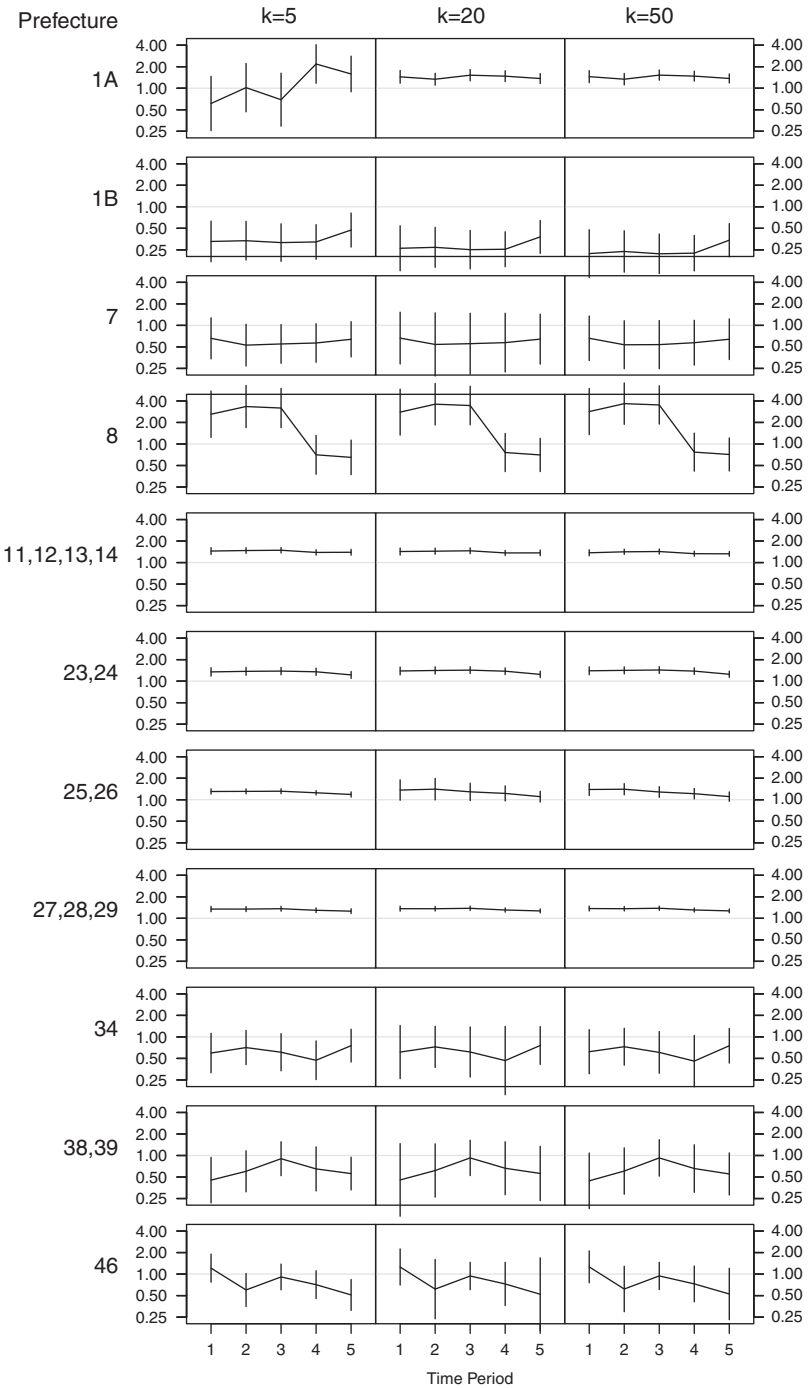

Figure 8. Posterior means ±2 standard deviations, conditional on cluster membership, of the log relative risks associated with clusters, $\{\sum_{j=1}^{k}\theta_{jt}\mathbf{1}\{d(\mathbf{x}_i,\mathbf{c}_j)\leqslant r_j\}|\mathbf{y}, \sum_{j=1}^{k}\mathbf{1}\{d(\mathbf{x}_i,\mathbf{c}_j)\leqslant r_j\}>0\}$, for the municipality with the largest Bayes factor for belonging to a cluster within the 11 contiguous areas of clustering with Bayes factors greater than 10 in model (3) with $k=20$ for model (3) with $k=5$, 20, 50 clusters.

prefectures, and in Ōsaka, Hyōgo and Nara prefectures. Consistently lowered risks are found in Hokkaidō prefecture (1B), in Fukushima prefecture and in Hiroshima prefecture. Three areas show evidence of time-varying cluster risks. In Ibaraki prefecture, risks are elevated in the first 3 time periods and return to normal levels thereafter. In Ehime and Kōchi prefectures, risks appear to be reduced in the initial and final time periods and close to normal during the middle time period. In Kagoshima prefecture, risks are normal in the initial time period and appear to progressively decline over time.

4. Simulation study

To explore the differences in the operating characteristics of the proposed model and the Yan–Clayton model in terms of cluster detection, we performed a small simulation study using the underlying geography and population structure (expected breast cancer case counts) from the Tochigi, Gunma and Saitama prefecture subset of the Japan breast cancer data. For the simulation study, we considered seven scenarios, one null model with no clusters and six single cluster models. To create the six single cluster models, we first selected three circular clusters with total expected case counts of approximately 70 cases (cluster #1: center: municipality id 11 368, radius: 10 km; cluster #2: center: municipality id 9387, radius: 12.5 km; cluster #3: center: municipality id 9205, radius: 11.5 km). We then assigned one of two temporal risk patterns: (1) $\theta_1 = \theta_2 = 0$, $\theta_3 = \theta_4 = \theta_5 = \log 2$ or (2) $\theta_1 = \theta_3 = \theta_5 = \log 2$, $\theta_2 = \theta_4 = 0$. The first temporal risk pattern is a cylindrical space–time cluster that is more parsimoniously expressed in terms of model (2), the Yan–Clayton model, whereas the second temporal risk pattern is a seasonal (or alternating) risk pattern that is more parsimoniously expressed in terms of model (3).

For each of the seven scenarios, we simulated 100 data sets in which the expected case counts were the expected case count from the Japan breast cancer data for cells outside the cluster and the expected case count from the Japan breast cancer data multiplied by the appropriate temporal cluster risk parameter for cells inside the cluster. For each simulated data set, we obtained 5000 posterior samples (after a 5000 sample burn-in) from models (2) and (3) with $k = 10$ using the prior specifications used for the data analysis in Section 3.1. Using these 5000 posterior samples, we then calculated the local Bayes factors in favor of clustering for each location, BF_i, and the maximum local Bayes factor across all locations for the null (no cluster) model and across all locations in the cluster for the other scenarios as the overall or global strength of evidence for clustering in each data set.

In Figure 9, we display, for both models applied to each of the seven scenarios, the proportion of simulated data sets for which the maximum local Bayes factor for clustering across locations exceeds the evidence thresholds of 3 (substantial), 10 (strong), 30 (very strong) and 100 (decisive) suggested by Jeffreys [29]. Based on the null scenario, an evidence threshold of 3 for the local Bayes factor appears to be excessively liberal with false detection rates around 50 per cent for both models. An evidence threshold of 10 produces more acceptable false detection rates of 7 per cent for model (3) and 11 per cent for model (2), whereas evidence thresholds of 30 or 100 result in very low false detection rates (1–2 per cent).

Using the threshold of 10 for the local Bayes factor, the true detection rates for the two models are reasonably close for clusters with the first (cylindrical) temporal risk pattern (cluster #1: 96 per cent for model (3) vs 98 per cent for model (2); cluster #2: 87 vs 90 per cent; cluster #3: 96 vs 98 per cent), although, not surprisingly, the detection rates are consistently higher for higher for model (2). In contrast, there are more substantial differences in the detection rates in favor of model (3) for the second (alternating) temporal risk pattern (cluster #1: 93 per cent for model (3) vs

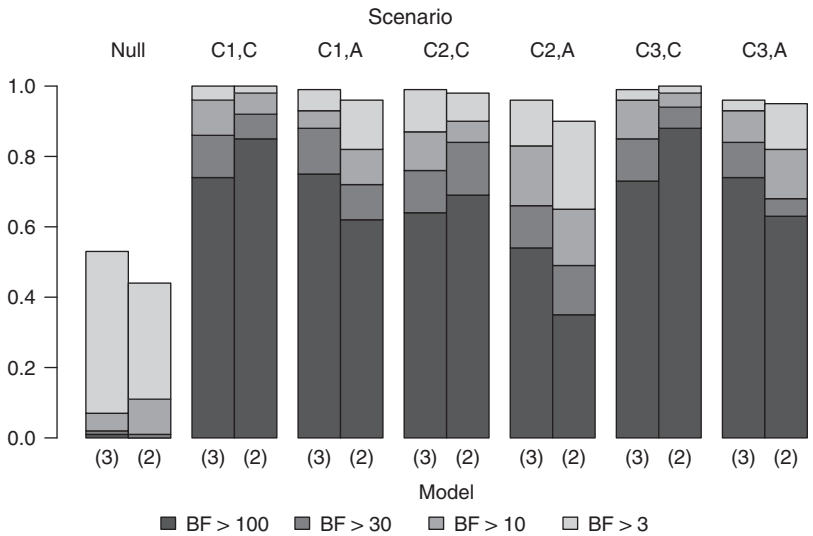

Figure 9. Proportion of simulated data sets for which the maximum local Bayes factor for clustering across locations exceeds 3, 10, 30 and 100 for the proposed model (3) and the Yan–Clayton model (2) for the seven scenarios. Null indicates the null model with no clusters; C1 indicates cluster #1, C2 indicates cluster #2, C3 indicates cluster #3, C indicates the cylindrical temporal risk pattern and A indicates the alternating temporal risk pattern.

82 per cent for model (2); cluster #2: 83 vs 65 per cent; cluster #3: 93 vs 82 per cent). Findings with thresholds of 30 or 100 are similar, although the advantage of model (2) for detecting cylindrical clusters is greater.

Overall, the observed differences in performance between the two models are not unexpected. The more flexible model (3) is less powerful in scenarios in which the more constrained model is correct and more powerful in other scenarios in which the more constrained model is incorrect (or less parsimonious). However, even in the most favorable case for the Yan–Clayton model (2), the reduction in the detection rate is fairly small. More importantly, we observed large advantages in terms of computational speed for model (3) over the Yan–Clayton model, even for these relatively small data sets. Analyses using model (3) were completed in approximately $\frac{1}{3}$ of the time required for analyses from the Yan–Clayton model.

5. Discussion

In this paper, we have presented a novel extension of the Gangnon–Clayton model for spatial clustering [24] to spatio-temporal data. In contrast to the previous extension of this model by Yan and Clayton [25], our model utilizes the spatial and temporal structure of the underlying data in constructing the baseline heterogeneity effects and allows for unstructured temporal risk patterns within spatial clusters rather than restricting attention to cylindrical spatio-temporal clusters. The analyses of the Japan female breast cancer mortality rates presented here illustrate many of the appealing features of the proposed model, including capturing time-varying risk patterns within clusters in a relatively parsimonious fashion and distinguishing expected similarity in risk over time within a single municipality from clustering while still identifying truly unusual temporal risk patterns in a single municipality.

First, we note the computational advantages of our proposed model for spatio-temporal clustering compared with models using cylindrical space–time clusters, confirmed by the differences in computational speed observed in the simulation study. The primary computational burden in fitting these models is the search over the set of potential clusters. In the cylindrical space–time model, the size of the search space depends, in a multiplicative fashion, on the number of potential time intervals and the number of potential spatial clusters so that even small increases in the number of time points can dramatically increase the size of the search space. In contrast, within our model, the number of potential clusters that need to be evaluated for a given study region remains fixed regardless of the number of time points under consideration, making our model potentially suitable for much larger data sets.

Rather than formally estimating the number of clusters within a RJMCMC algorithm [22, 25], we indirectly estimate the numbers of clusters and provide direct assessments of the evidence for and risks associated with clusters using local Bayes factors from models with a fixed, but overly large number of clusters. As found previously [23, 24], inferences, both in terms of cluster memberships and cluster risk parameters, are very consistent across models with different numbers of clusters, as long as the chosen number of clusters is sufficiently large. Robustness of cluster risk estimates is even evident in models with an insufficient number of clusters.

Although we use circles as potential clusters for illustration, adaptations of model to other cluster shapes, such as rectangles or ellipses, are straightforward. Adaptations to irregularly shaped clusters [8] are somewhat more complicated, but feasible. Given the exploratory nature of these cluster detection problems, the relative simplicity of circular clusters may be advantageous in many settings, even those in which the true clusters are not circular [25].

In our applications, we have focused on the use of an independence prior for the risk parameters within a cluster over time. One could easily incorporate additional prior information about expected cluster risk patterns through the mean of the multivariate normal prior, its variance–covariance matrix or both. For example, one could use an increasing linear trend for the mean and a compound symmetry or autoregressive correlation structure for the variance–covariance matrix. In most settings, we believe that this use of prior information is unnecessary, as the cluster risk parameters are likely to be well-identified by the data given that they reflect average risk across multiple cells, and, in some settings, it may even be counterproductive.

References

1. Besag J, Newell J. The detection of clusters of rare diseases. *Journal of the Royal Statistical Society*, *Series A* 1991; **154**:143–155.
2. Kulldorff M. Statistical methods for spatial epidemiology: tests for randomness. In *GIS and Health for Europe*, Gatrell A, Løytønen M (eds). Taylor & Francis: London, 1998; 49–62.
3. Kulldorff M. A spatial scan statistic. *Communications in Statistics*, *Part A* 1997; **26**:1481–1496.
4. Kulldorff M, Nagarwalla N. Spatial disease clusters: detection and inference. *Statistics in Medicine* 1995; **14**:799–810.
5. Kulldorff M, Athas W, Feuer E, Miller B, Key C. Evaluating cluster alarms: a space–time scan statistic and brain cancer in Los Alamos. *American Journal of Public Health* 1998; **88**:1377–1380.
6. Kulldorff M. Prospective time-periodic geographical disease surveillance using a scan statistic. *Journal of the Royal Statistical Society*, *Series A* 2001; **164**:61–72.

7. Duczmal L, Assuncao R. A simulated annealing strategy for the detection of arbitrarily shaped spatial clusters. *Computational Statistics and Data Analysis* 2004; **45**:269–284.
8. Gangnon RE, Clayton MK. Likelihood-based tests for detecting spatial clustering of disease. *Environmetrics* 2004; **15**:797–810.
9. Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. *International Journal of Health Geographics* 2005; **4**:11.
10. Kulldorff M, Huang L, Pickle L, Duczmal L. An elliptic spatial scan statistic. *Statistics in Medicine* 2006; **25**:3929–3943.
11. Assuncao R, Costa M, Tavares A, Ferreira S. Fast detection of arbitrarily shaped disease clusters. *Statistics in Medicine* 2006; **25**:723–742.
12. Takahashi K, Kulldorff M, Tango T, Yih K. A flexibly shaped space–time scan statistic for disease outbreak detection and monitoring. *International Journal of Health Geographics* 2008; **7**:14.
13. Gangnon RE. Local multiplicity adjustments for spatial cluster detection. *Environmental and Ecological Statistics* 2010; **17**(1):55–71.
14. Lawson A. Disease cluster detection: a critique and a Bayesian proposal. *Statistics in Medicine* 2005; **25**:897–916.
15. Besag J, York J, Mollie A. Bayesian image restoration with applications in spatial statistics (with Discussion). *Annals of the Institute of Mathematical Statistics* 1991; **43**:1–59.
16. Clayton D, Bernardinelli L. Bayesian methods for mapping disease risk. In *Geographical and Environmental Epidemiology*: *Methods for Small Area Studies*, Elliot P, Cuzick J, English D, Stern R (eds). Oxford University Press: Oxford, 1992; 205–220.
17. Bernardinelli L, Clayton D, Pascutto C, Montomoli C, Ghislandi M, Songini M. Bayesian analysis of space–time variation in disease risk. *Statistics in Medicine* 1995; **14**:2433–2443.
18. Waller LA, Carlin BP, Xia H, Gelfand AE. Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association* 1997; **92**:607–617.
19. Knorr-Held L, Besag J. Modeling risk from a disease in time and space. *Statistics in Medicine* 1998; **17**:2045–2060.
20. Knorr-Held L. Bayesian modeling of inseparable space–time variation in disease risk. *Statistics in Medicine* 2000; **19**:2555–2567.
21. Richardson S, Thomson A, Best N, Elliott P. Interpreting posterior relative risk estimates in disease-mapping studies. *Environmental Health Perspectives* 2004; **112**:1016–1025.
22. Gangnon RE, Clayton MK. Bayesian detection and modeling of spatial disease clustering. *Biometrics* 2000; **56**:922–935.
23. Clark AB, Lawson AB. Spatio-temporal cluster modeling of small area health data. In *Spatial Cluster Modeling*, Lawson AB, Denison DGT (eds). Chapman & Hall: New York, 2002; 235–258.
24. Gangnon RE, Clayton MK. A hierarchical model for spatially clustered disease rates. *Statistics in Medicine* 2003; **22**:3213–3228.
25. Yan P, Clayton M. A cluster model for space–time disease counts. *Statistics in Medicine* 2006; **25**:867–881.
26. Gangnon RE, Clayton MK. A weighted average likelihood ratio test for spatial clustering of disease. *Statistics in Medicine* 2001; **20**:2977–2987.
27. Gangnon RE, Clayton MK. Cluster detection using Bayes factors from overparameterized cluster models. *Environmental and Ecological Statistics* 2007; **14**:69–82.
28. Gangnon RE. Impact of prior choice on local Bayes factors for cluster detection. *Statistics in Medicine* 2006; **25**:883–895.
29. Jeffreys H. *The Theory of Probability*. Oxford University Press: Oxford, 1961; 432.
30. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. Chapman & Hall: London, 1995.
31. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970; **57**:97–109.
32. Ohtaki M, Kawasaki H, Satoh K, Nakayama T, Yanagihara H, Yamaguchi N. Visualization of time–geographical distribution of cancer mortality in Japan. *International Statistical Symposium and Bernoulli Society EAPR Conference*, Taipei, Taiwan, 2002.
33. Sibson R. The Dirichlet tessellation as an aid in data analysis. *Scandinavian Journal of Statistics* 1980; **7**:14–20.