Electronic Journal of Statistics
Vol. 14 (2020) 1449–1478
ISSN: 1935-7524
https://doi.org/10.1214/20-EJS1696

A Bayesian approach to disease clustering using restricted Chinese restaurant processes

Claudia Wehrhahn

Department of Statistics, University of California, Santa Cruz, Santa Cruz, U.S.A.
e-mail: cwehrhah@ucsc.edu

Samuel Leonard

Northern California Institute for Research and Education, San Francisco, U.S.A.
e-mail: sleonar2@ucsc.edu

Abel Rodriguez

Department of Statistics, University of California, Santa Cruz, Santa Cruz, U.S.A.
e-mail: abel@soe.ucsc.edu

and

Tatiana Xifara

Airbnb, Inc., San Francisco, U.S.A.
e-mail: xifara@soe.ucsc.edu

Abstract: Identifying disease clusters (areas with an unusually high incidence of a particular disease) is a common problem in epidemiology and public health. We describe a Bayesian nonparametric mixture model for disease clustering that constrains clusters to be made of adjacent areal units. This is achieved by modifying the exchangeable partition probability function associated with the Ewen's sampling distribution. We call the resulting prior the Restricted Chinese Restaurant Process, as the associated full conditional distributions resemble those associated with the standard Chinese Restaurant Process. The model is illustrated using synthetic data sets and in an application to oral cancer mortality in Germany.

Keywords and phrases: Disease clustering, areal data, Chinese restaurant process.

Received May 2019.

1. Introduction

A disease cluster is a higher-than-expected incidence of a particular disease or disorder occurring in close proximity in terms of both time and geography. Although communicable diseases (those that can be spread from one person to another, such as the flu or HIV) often occur in clusters, clusters of non-communicable disease are rare and their presence might indicate the presence

of a harmful environmental factor or other hazard. Therefore, identification of cancer clusters is a key task in epidemiology and public health.

A strand of the statistics literature on disease clustering focuses on methods for confirmatory cluster analysis. Sometimes called *focused tests*, these methods are concerned with determining whether the rate of disease in a pre-specified area (which usually contains some putative health hazard) is higher than expected (e.g., see Stone, 1988, Besag & Newell, 1991, Tango, 1995, Morton-Jones et al., 1999). In contrast, the focus of this paper is on methods for *de novo* identification of disease clusters in datasets in which the presence of such clusters is not known. Methods based on scan statistics (e.g., see Weinstock, 1981, Kulldorff, 1997, Tango & Takahashi, 2005) are well known examples of this type of approaches. Implementations of classical approaches to disease cluster analysis are widely available in a number of platforms. One example is the `R` package `DCluster` (Gómez-Rubio et al., 2005).

Methods for disease clustering can also be classified according to whether they are designed to work with point-referenced or with spatially aggregated (areal) data. In the case of point-referenced data, it is common to distinguish between distance-based methods (Whittemore et al., 1987, Besag & Newell, 1991, and Tango, 1995, among others), which derive tests based on the distribution of the time/distance between locations on which events occurred, and quadrat-based methods (e.g, Openshaw et al., 1987, Kulldorff & Nagarwalla, 1995), which study the variability of case counts in certain subsets of the region of interest (called quadrats). In the case of areal data, frequency tests similar to those used in quadrat-based methods are frequently used (e.g., see Potthoff & Whittinghill, 1966a and Potthoff & Whittinghill, 1966b). Bayesian methods for disease clustering in spatially aggregated data have been proposed by Knorr-Held & Raßer (2000), Gangnon & Clayton (2000), Green & Richardson (2002), Gómez-Rubio et al. (2018), Wakefield & Kim (2013), and Anderson et al. (2014). Other recent contributions to the field include the work of Moraga & Montes (2011), Charras-Garrido et al. (2012), Heinzl & Tutz (2014), and Wang & Rodríguez (2014). Kulldorff et al. (2003), Waller et al. (2006), and Goujon-Bellec et al. (2011) present detailed comparisons of various methods for disease clustering.

It is worth noting that the main goals of disease clustering methods are similar but distinct from those of disease mapping. Typically, disease mapping applications deal with the estimation of smooth covariate-adjusted risk measures, but do not aim at identifying discontinuities in the risk function. On the other hand, the whole point of methods for *de novo* identification of cancer cluster is to pinpoint such discontinuities. Of course, these two objectives are not necessarily opposed (e.g., see Knorr-Held & Raßer, 2000, Green & Richardson, 2002, and Anderson et al., 2014), but most techniques designed for disease mapping are not directly applicable in the context of disease clustering. The literature on disease clustering is also related to, but distinct from, the literature on boundary analysis in areal data (sometimes referred to as "areal wombling", e.g., see Lu & Carlin, 2005; Lu et al., 2007; Fitzpatrick et al., 2010; Li et al., 2015b; Guhaniyogi, 2017).

In this paper we develop a Bayesian approach for *de novo* identification of disease clusters in areal data. Our approach uses a restricted version of the Exchangeable Partition Probability Function (EPPF) associated with a species sampling model (SSM) (Pitman, 1995, 1996) as a prior on the partition of areal units. To simplify our exposition, we focus here on the SSM associated with the Dirichlet process (Ferguson, 1973; Blackwell & MacQueen, 1973; Antoniak, 1974; Lee et al., 2013; Rodríguez & Quintana, 2015), which is sometimes referred to as the Chinese restaurant process (CRP). However, the formulation is more general and our key results (particularly around the form of the full conditional distributions associated with the prior) extend to other SSMs such as the Generalized CRP induced by the two-parameter Poisson-Dirichlet process (Pitman & Yor, 1997).

The restricted prior we introduce in this paper is specifically designed to enforce clusters made of adjacent spatial units (which we call *admissible*). The approach we develop in this paper is related to those developed in Fuentes-García et al. (2010) and Martínez et al. (2014) in the context of time series data. Fuentes-García et al. (2010) consider a special case of our model that assumes ordered observations and uses reversible jump Markov chain Monte Carlo algorithms for inference. More recently, Martínez et al. (2014) propose a change-point model constructed by restricting a Generalized CRP (Pitman, 1995; Gnedin & Pitman, 2006), but their proposal differs from ours in the way the probability associated with inadmissible partitions is redistributed. Our model can be seen as generalizing the ideas in Fuentes-García et al. (2010) and Martínez et al. (2014) to situations in which the EPPF is restricted to partitions driven by general neighborhood graphs. We also show that, for our construction, the full conditional distributions associated with the restricted prior take a simple and appealing form, making the use of reversible jump algorithm unnecessary.

The model we introduce here is also related to the literature on spatially-dependent mixture models. Fernández & Green (2002) consider the use of a Potts model as the joint prior on the cluster indicators of a finite mixture model, leaving the question of how the number of clusters is to be selected open. Loschi & Cruz (2005), Müller et al. (2011), and Page et al. (2016) consider extensions of Hartigan's product partition model (PPM) (Hartigan, 1990) in which the so-called coherence functions account for temporal and/or spatial dependence. These models cannot be easily generalized to other SSMs beyond the CRP, where the prior on the partition cannot be written in terms of the product of coherence functions. Dahl (2008), Blei & Frazier (2011), Ghosh et al. (2011), and Dahl et al. (2017) consider Chinese restaurant processes in which co-clustering probabilities are functions of the distance between observations. In a similar spirit, Li (2015), Li et al. (2013), Li et al. (2014), Li et al. (2015a), Li et al. (2016a), Li et al. (2016b), and Li et al. (2016c) generalize the approach of Blei & Frazier (2011) so that the clustering probabilities depend on side information. A common feature of all these approaches is their focus on soft constraints that encourage nearby areas to cluster together but still allow clusters to be disconnected. In contrast, our focus is on ensuring that clusters are fully con-

nected. This type of constraint is the most natural one in the context of our application to disease clustering, and cannot be easily enforced with any of the models discussed above. For example, Page et al. (2016) note that computational challenges arise when a restricted cohesion function in a PPM is considered in order to assign zero probability to "non-desirable" cluster configurations. Our proposal overcomes these challenges.

The remaining of the paper is organized as follows: Section 2 presents our model and discusses its properties. Section 3 describes our computational approach. Sections 4 and 5 present the analysis of two simulated data sets and an application to oral cancer mortality in Germany, respectively. Finally, Section 6 discusses the limitations of our model as well as future research directions.

2. The model

Suppose that areal data in the form of pairs (y_i, h_i) are available, where y_i records the observed number of cases in region i, and h_i represents the expected number of cases in region i, obtained by internal or external standardization, for $i = 1, \ldots, n$. As is common in the literature, we assume that the counts in region i are independently Poisson distributed and model their rates as a function of h_i and their log-scale relative risk, η_i, log-RR for short. More specifically,

$$y_i \mid \eta_i \sim Pois(h_i e^{\eta_i}), \qquad i = 1, \ldots, n, \tag{1}$$

where $\eta_i = \boldsymbol{x}_i^t \boldsymbol{\theta}_i$, $\boldsymbol{x}_i^t$ is the transpose of a p-dimensional vector of covariates associated to region i, $\boldsymbol{\theta}_i \in \mathbb{R}^p$ is the random effect associated with region i, and $Pois(\lambda)$ denotes the Poisson distribution with rate $\lambda > 0$. When no covariates are available, i.e., $x_{i,j} = 1$, for all i, j, the log-RR reduces to $\eta_i = \theta_i$, with $\theta_i \in \mathbb{R}$.

Our approach assumes that the geographic information associated with the data set is encoded in a known $n \times n$ binary adjacency matrix, $\boldsymbol{W} = [w_{i,i'}]$. This adjacency matrix can be interpreted as defining an unweighted, undirected graph G whose nodes correspond to the different geographical regions under study. In our illustration we focus on first-order neighborhood matrices in which $w_{i,i'}$ is equal to 1 if regions i and i' share a common boundary, and equal to 0 otherwise. However, our methodology applies more generally to higher order neighborhoods, or to other ways to define the (binary) adjacency relationship, e.g., by means of the distances between centroids or of distances based on length of shared boundary.

Recall that our interest is to identify clusters of spatially connected regions that share the same relative log-risk. Therefore, regions i and j can belong to the same cluster k only if G contains a path that connects them for which all the nodes in the path also belong to cluster k. In what follows we will propose a spatially restricted prior distribution for the cluster membership random variable and illustrate how specific adjacency matrices impact the probability mass function of the number of clusters.

To construct our prior on the random effects we borrow ideas from the model-based clustering literature. More specifically, we augment the model in (1) with

TABLE 1
All possible partitions and cluster configurations for a sample of size $n = 3$.

K	Partition of data	$\boldsymbol{c}$	$A_k, k = 1, \ldots, K$
1	$\{y_1, y_2, y_3\}$	$(1,1,1)$	$A_1 = \{1,2,3\}$
2	$\{y_1\}, \{y_2, y_3\}$	$(1,2,2)$	$A_1 = \{1\}, A_2 = \{2,3\}$
2	$\{y_1, y_2\}, \{y_3\}$	$(2,2,1)$	$A_1 = \{1,2\}, A_2 = \{3\}$
2	$\{y_1, y_3\}, \{y_2\}$	$(2,1,2)$	$A_1 = \{1,3\}, A_2 = \{2\}$
3	$\{y_1\}, \{y_2\}, \{y_3\}$	$(1,2,3)$	$A_1 = \{1\}, A_2 = \{2\}, A_3 = \{3\}$

a labeling vector $\boldsymbol{c} = (c_1, \ldots, c_n)$ describing to which cluster each observation belongs, hence $c_i \in \{1, \ldots, K\}$, where $K = K(\boldsymbol{c})$ denotes the number of clusters, $K \leq n$, and define $\theta_i = \tilde{\theta}_{c_i}$, where $\tilde{\theta}_k$ is the log-RR of cluster $k = 1, \ldots, K$. Note that the vector $\boldsymbol{c}$ induces a partition of the set of integers $[n] = \{1, \ldots, n\}$ into $A_1, \ldots, A_K$ non null, disjoint sets, whose union is $[n]$. For instance, if $n = 3$, there are 5 possible partitions of the set $[3]$ into $A_1, \ldots, A_K$ clusters, $K \in \{1, 2, 3\}$, induced by all 5 possible configurations of $\boldsymbol{c}$ (see Table 1).

If no spatial information were available (or, alternatively, if $\boldsymbol{W}$ corresponds to the complete graph), it would be convenient to assign $\boldsymbol{c}$ a Chinese restaurant process prior (CRP) (Pitman, 1995),

$$\pi(\boldsymbol{c} \mid \alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \alpha^{K(\boldsymbol{c})} \prod_{k=1}^{K(\boldsymbol{c})} \Gamma\left(n_k(\boldsymbol{c})\right), \tag{2}$$

where $n_k(\boldsymbol{c}) = \sum_{i=1}^{n} \mathbb{1}\left(c_i = k\right)$ is the number of labels having value k or, equivalently, the number of observations in cluster k and $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, \mathrm{d}t$ denotes the Gamma function.

The corresponding full conditionals are given by

$$\pi(c_i = k \mid \boldsymbol{c}_{-i}, \alpha) \propto \begin{cases} n_k(\boldsymbol{c}_{-i}) & k \leq K(\boldsymbol{c}_{-i}), \\ \alpha & k = K(\boldsymbol{c}_{-i}) + 1, \end{cases} \tag{3}$$

where $\boldsymbol{c}_{-i} = (c_1, \ldots, c_{i-1}, c_{i+1}, \ldots, c_n)$. Therefore, each c_i is either an already existing label, with probability proportional to $n_k(\boldsymbol{c}_{-i})$, or a new label, with probability proportional to α. Clearly, the concentration parameter α controls the number of clusters, with larger values of α favoring larger numbers of clusters a priori.

In order to define a spatially restricted prior distribution that enforces connected clusters, we propose to modify (2) by giving zero probability to configurations that involve clusters with non-connected components. More specifically, let G_{A_k} be the subgraph of G involving only the nodes that belong to the set A_k. We call a partition $A_1, \ldots, A_K$ admissible if G_{A_k} is a connected subgraph (but not necessarily complete) for every $k = 1, \ldots, K$. If we define the function $Q(\boldsymbol{c}, \boldsymbol{W})$ as being equal to 1 whenever $\boldsymbol{c}$ is an admissible cluster configuration under $\boldsymbol{W}$, our prior takes the form

$$\pi(\boldsymbol{c} \mid \alpha, \boldsymbol{W}) = \frac{\alpha^{K(\boldsymbol{c})}}{C(\alpha, \boldsymbol{W})} \left\{ \prod_{k=1}^{K(\boldsymbol{c})} \Gamma\left(n_k(\boldsymbol{c})\right) \right\} Q(\boldsymbol{c}, \boldsymbol{W}), \tag{4}$$

where the normalizing constant $C(\alpha, \boldsymbol{W})$ is given by

$$C(\alpha, \boldsymbol{W}) = \sum_{\{\boldsymbol{c}':\ Q(\boldsymbol{c}', \boldsymbol{W})=1\}} \alpha^{K(\boldsymbol{c}')} \left\{ \prod_{k=1}^{K(\boldsymbol{c}')} \Gamma(n_k(\boldsymbol{c}')) \right\}. \tag{5}$$

We call this prior the *Restricted Chinese Restaurant Process* with parameters α and $\boldsymbol{W}$, denoted $\boldsymbol{c} \mid \alpha, \boldsymbol{W} \sim RCRP(\alpha, \boldsymbol{W})$.

In general, there is no closed form expression for $C(\alpha, \boldsymbol{W})$; two exceptions are provided in Appendix A. However, we can still make some general statements. For example, we note that $C(\alpha, \boldsymbol{W})$ is a polynomial function of degree n in α, i.e., we can write

$$C(\alpha, \boldsymbol{W}) = f_1(\boldsymbol{W})\alpha + f_2(\boldsymbol{W})\alpha^2 + \ldots + f_n(\boldsymbol{W})\alpha^n, \tag{6}$$

where the coefficient $f_l(\boldsymbol{W})$ is a weighted sum over the admissible partitions involving l clusters, $f_l(\boldsymbol{W}) = \sum_{\{\boldsymbol{c}':\ K(\boldsymbol{c}')=l\}} \prod_{k=1}^{l} \Gamma\left(n_l(\boldsymbol{c}')\right) Q(\boldsymbol{c}', \boldsymbol{W})$. In particular, $f_1(\boldsymbol{W}) = \Gamma(n)$ for any $\boldsymbol{W}$, $f_n(\boldsymbol{W}) = 1$ for any $\boldsymbol{W}$ that implies a graph G that is connected (and $f_n(\boldsymbol{W}) = 0$ otherwise), and $f_l(\boldsymbol{W}) = |S_{n,l}|$, the unsigned Stirling number of the second kind, when $\boldsymbol{W}$ implies the complete graph.

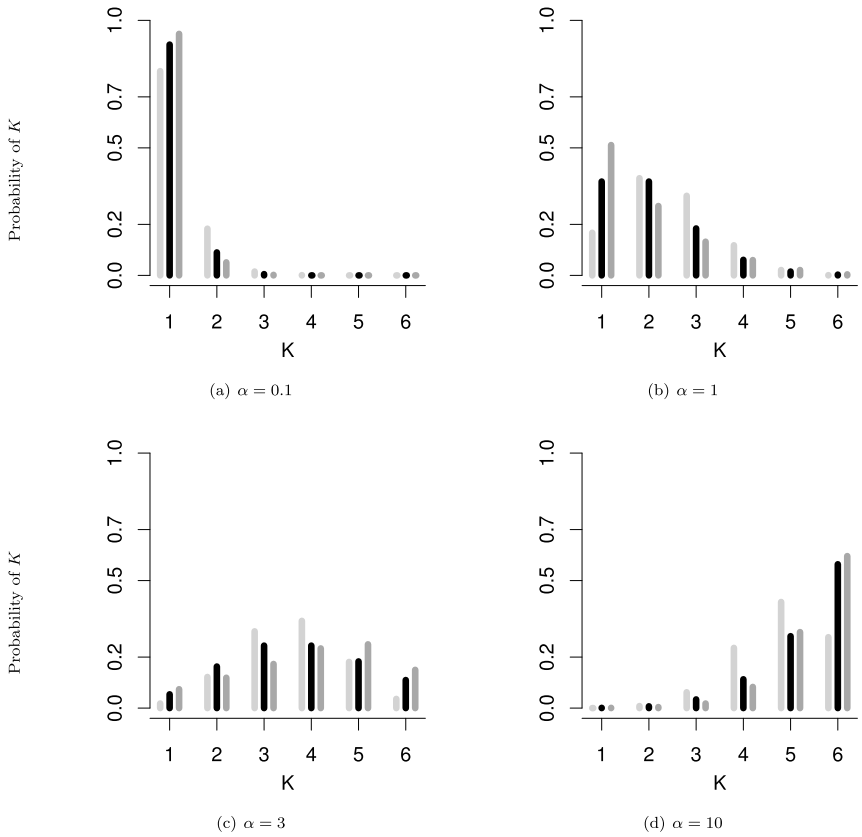
Figure 1 presents the probability associated with the number of clusters $K(\boldsymbol{c})$ for the unrestricted CRP, as well as for the neighborhood structures associated with a star and a linear graphs (see Appendix A), when $n = 6$. Recall that the prior on partitions implied by the model of Fuentes-García et al. (2010) corresponds to the linear graph setting, while the model in Martínez et al. (2014) is constructed to ensure that the prior on $K(\boldsymbol{c})$ matches the one for the unrestricted CRP. It is clear from the graph that the probability of the partitions depends on the underlying adjacency matrix. For both restricted cases, the probability of $K = 1$ and $K = n$ are higher than in the non-restricted case. For smaller values of α ($\alpha = 0.1$ and $\alpha = 1$), the linear graph seems to favor a smaller number of clusters than the star graph. This pattern is reversed for the larger values of α ($\alpha = 3$ and $\alpha = 10$).

While an explicit expression for $C(\alpha, \boldsymbol{W})$ is generally not available, the full conditional distributions associated with (4) take a particularly simple, appealing, and computationally convenient form (see Appendix B):

$$\pi(c_i = k \mid \boldsymbol{c}_{-i}, \alpha, \boldsymbol{W}) \propto \begin{cases} n_k(\boldsymbol{c}_{-i}) & k \leq K(\boldsymbol{c}_{-i}) \text{ and } Q(\boldsymbol{c}, \boldsymbol{W}) = 1, \\ \alpha & k = K(\boldsymbol{c}_{-i}) + 1 \text{ and } Q(\boldsymbol{c}, \boldsymbol{W}) = 1, \\ 0 & Q(\boldsymbol{c}, \boldsymbol{W}) = 0. \end{cases} \tag{7}$$

Hence, when the assignment of a region to a particular cluster leads to an admissible configurations, (7) and (3) agree. On the other hand, for assignments that lead to inadmissible configurations, the full conditional is zeroed out. As before, α controls the number of clusters.

The CRP prior has sometimes been criticized in the context of clustering applications because of their tendency to create clusters of unbalanced size. In disease clustering applications, where the disease clusters can usually be



(a) $\alpha = 0.1$ (b) $\alpha = 1$

(c) $\alpha = 3$ (d) $\alpha = 10$

FIG 1. *Prior probability associated with the number of clusters $K(\boldsymbol{c})$ under non-restricted (light grey), star (black) and linear (dark grey) RCRPs, for different values of α.*

expected to be rare and small a priori, this behavior (which will carry out to the restricted model we discuss below) is an appealing feature of the model.

Having defined the spatially restricted prior distribution for the labeling random variables, the rest of the model is specified by $\tilde{\theta}_k \overset{iid}{\sim} N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$ are the mean and variance of the normal distribution. Additionally, we consider hyperprior distributions for α, μ, and σ^2. Our model can finally be written as

$$y_i \mid \tilde{\boldsymbol{\theta}}, c_i \sim Poi(h_i e^{\tilde{\theta}_{c_i}}), \quad \boldsymbol{c} \mid \alpha \sim RCRP(\alpha), \quad \tilde{\theta}_k \mid \mu, \sigma^2 \overset{iid}{\sim} N(\mu, \sigma^2), \tag{8}$$

with one more level for the hyperpriors

$$\mu \sim N(\kappa, \phi^2), \qquad \sigma^2 \sim IG(a, b), \qquad \alpha \sim \pi, \tag{9}$$

where the parameters κ, ϕ^2, a, and b are fixed, $IG(a, b)$ denotes the inverse gamma distribution with shape and scale parameters $a > 0$ and $b > 0$, respectively, and π is a distribution on $\mathbb{R}^+$.

3. Computational aspects

We use Markov Chain Monte Carlo (MCMC) algorithms (Smith & Roberts, 1993; Robert & Casella, 2005) to generate samples from the posterior distributions associated with the proposed model. This Section describes the full conditional distributions involved.

The full conditional distribution for the components of the indicator vector $\boldsymbol{c}$ takes the relatively simple form:

$$\pi(c_i = k \mid y_i, \boldsymbol{c}_{-i}, ...) \propto \begin{cases} n_k(\boldsymbol{c}_{-i}) f(y_i \mid \tilde{\theta}_k) & k \leq K(\boldsymbol{c}_{-i}) \text{ and } Q(\boldsymbol{c}, \boldsymbol{W}) = 1, \\ \alpha f(y_i \mid \tilde{\theta}_{K(\boldsymbol{c}_{-i})+1}) & k = K(\boldsymbol{c}_{-i}) + 1 \text{ and } Q(\boldsymbol{c}, \boldsymbol{W}) = 1, \\ 0 & Q(\boldsymbol{c}, \boldsymbol{W}) = 0, \end{cases} \tag{10}$$

where $\tilde{\theta}_{K(\boldsymbol{c}_{-i})+1} \sim N(\mu, \sigma^2)$ and $f(y_i \mid \tilde{\theta}_k)$ denotes the probability mass function of a Poisson distribution with rate parameter $h_i e^{\tilde{\theta}_k}$. This resembles algorithm 8 from Neal (2000). Note that there is an implicit and very standard relabeling of vector $\boldsymbol{c}$ every time a cluster becomes empty (i.e., $n_k = 0$, for some k), as in the "no gap" algorithm of MacEachern & Müller (1998). The Markov chain that results from cycling through these full conditional distributions is irreducible: we can move from between any two admissible configurations by first breaking each cluster (one region at a time, starting with the "periphery" to ensure that admissibility is preserved), and then reassembling the new clusters.

While the need to repeatedly check on the admissibility of the configurations might suggest that the computational cost of implementing this algorithm would be high, that is not the case. Using the fact that the current configuration must be admissible, a careful implementation of the algorithm only requires that, for each i, we check that the cluster currently containing observation i remains fully connected if that observation is removed (which, in the worst case scenario, can be done in quadratic time on the size of that cluster). Then c_i is updated with the label of any of its neighbours (which can be directly identified from $\boldsymbol{W}$ in linear time) or with a new label according to the probabilities given by (10).

As with a standard mixture model, posterior distributions for log-RR parameters $\tilde{\theta}_1, \ldots, \tilde{\theta}_{K(\boldsymbol{c})}$ are conditionally independent and take the form

$$\pi(\tilde{\theta}_k \mid \boldsymbol{y}, ...) \propto \exp\left\{ -\frac{1}{2\sigma^2} \left(\tilde{\theta}_k - \left[\mu + \sigma^2 y_k^+ \right] \right)^2 - e^{\tilde{\theta}_k} h_k^+ \right\}, \qquad k = 1, \ldots, K(\boldsymbol{c}), \tag{11}$$

where $y_k^+ = \sum_{\{i:c_i=k\}} y_i$ and $h_k^+ = \sum_{\{i:c_i=k\}} h_i$. Since these posterior distributions do not belong to any tractable family of distributions, one must resort to algorithms such as random walk Metropolis-Hastings (M-H) or Hamiltonian Monte Carlo to sample from them. However, these algorithms require that the user selects a number of tuning parameters (such as the variance of the random walk in random walk M-H algorithms, or the size and number of steps in Hamiltonian Monte Carlo algorithms). Instead, we resort to slice samplers algorithms (Damien et al., 1999), which do not require the selection of any tuning param-

eters and therefore facilitate the use of our approach by practitioners. More specifically we introduce unit-rate exponentially distributed auxiliary random variables, u_k, leading to truncated exponential and truncated normal conditional distributions for u_k and $\tilde{\theta}_k$, respectively,

$$\pi(u_k \mid \tilde{\theta}_k, ...) \propto Exp(u_k \mid 1)\mathbb{1}\left(u_k \geq \exp\{\tilde{\theta}_k\}h_k^+\right),$$
$$\pi(\tilde{\theta}_k \mid u_k, ...) \propto N\left(\tilde{\theta}_k \mid \mu + \sigma^2 y_k^+, \sigma^2\right)\mathbb{1}\left(\tilde{\theta}_k \leq \log(u_k) - \log(h_k^+)\right),$$

where $Exp(\cdot \mid \lambda)$ denotes the exponential distribution with rate λ. On the other hand, the hyperparameters μ and σ^2 are sampled from their conjugate posterior distributions

$$\mu \mid ... \sim N\left(\frac{\sigma^2\kappa + \phi^2\sum_{k=1}^K \tilde{\theta}_k}{\sigma^2 + \phi^2 K}, \frac{\sigma^2\phi^2}{\sigma^2 + \phi^2 K}\right),$$
$$\sigma^2 \mid ... \sim IG\left(a + \frac{K}{2}, b + \frac{1}{2}\sum_{k=1}^K(\tilde{\theta}_k - \mu)^2\right).$$

Finally, we discuss the process of generating samples from the full conditional distribution of α. This step is particularly difficult because the full conditional

$$\pi(\alpha \mid ...) \propto \pi(\alpha)\frac{\alpha^K}{C(\alpha, \boldsymbol{W})}\left\{\prod_{k=1}^K (n_k - 1)!\right\} Q(\boldsymbol{c}, \boldsymbol{W}), \tag{12}$$

is double intractable: additionally to the posterior not belonging to any known distribution, it involves the computation of the intractable normalizing constant $C(\alpha, \boldsymbol{W})$.

In this paper we use the noisy exchange algorithm (NEA), proposed by Alquier et al. (2016), for sampling from (12). The NEA updates α using a M-H step replacing the ratio of normalizing constants $C(\alpha, \boldsymbol{W})/C(\alpha^*, \boldsymbol{W})$ by an unbiased importance sampling estimator. Note that, as a consequence of importance sampling,

$$\begin{aligned}\frac{C(\alpha, \boldsymbol{W})}{C(\alpha^*, \boldsymbol{W})} &= \sum_{\{\boldsymbol{c}': Q(\boldsymbol{c}', \boldsymbol{W})=1\}}\left(\frac{C(\alpha, \boldsymbol{W})}{C(\alpha^*, \boldsymbol{W})}\frac{\pi(\boldsymbol{c}' \mid \alpha, \boldsymbol{W})}{\pi(\boldsymbol{c}' \mid \alpha^*, \boldsymbol{W})}\right)\pi(\boldsymbol{c}' \mid \alpha^*, \boldsymbol{W}) \\ &= \mathsf{E}_{\boldsymbol{c}' \mid \alpha^*, \boldsymbol{W}}\left(\frac{\alpha^{K(\boldsymbol{c}'_j)}}{\alpha^{*K(\boldsymbol{c}'_j)}}\right),\end{aligned}$$

where $\mathsf{E}_{\boldsymbol{c}' \mid \alpha, \boldsymbol{W}}$ denotes the expectation under $\boldsymbol{c}' \mid \alpha, \boldsymbol{W}$. Therefore, the ratio of normalizing constants is approximated by

$$\frac{C(\alpha, \boldsymbol{W})}{C(\alpha^*, \boldsymbol{W})} \approx \frac{1}{N}\sum_{j=1}^N \frac{\alpha^{K(\boldsymbol{c}'_j)}}{\alpha^{*K(\boldsymbol{c}'_j)}}, \tag{13}$$

where $\boldsymbol{c}'_1, \ldots, \boldsymbol{c}'_N$ are samples from (4), obtained by running a second MCMC algorithm based on the full conditionals given by (7). In order to speed up computations, we considered a discrete prior distribution for α with support on a

relatively small number of point masses, and computed the ratios of normalizing constants in advance.

4. Simulated data

We conduct two simulation studies to ascertain the performance of our model. Both scenarios assume that the spatial association is given by the first-order neighborhood structure of the counties in the U.S. state of Ohio, and that $h_i = 100$ for every $i = 1, \ldots, 88$. Scenario I involves eight snake-shaped clusters with log-RR $\theta_i \in \{-2, 0, 2\}$ (see Figure 2, top row). Note that in this case some clusters share same disease risk. Scenario II is formed by 4 round-shaped connected clusters with log-RR $\theta_i \in \{-1.5, -0.5, 0.5, 1.5\}$, so each cluster has a different disease risk (see Figure 2, bottom row).

4.1. A comparison model

We compare the performance of our model against the boundary distance (BD) model proposed by Knorr-Held & Raßer (2000). The BD model uses a Poisson likelihood similar to ours, but assigns a different prior distribution on the partition indicators $c_1, \ldots, c_n$ that is inspired by K-means clustering. More specifically, their prior is specified hierarchically through a prior distribution on the number of clusters, K, a uniform prior distribution on the set of cluster centers, denoted $(g_1, \ldots, g_K)$, with $g_k \in \{1, \ldots, n\}$, and a measure of distance between regions, which in their case is given by the minimum number of boundaries that need to be crossed to move between the two regions. Given K and $(g_1, \ldots, g_K)$, each region i is assigned to the cluster whose center is closest. If one region is equally distant from two cluster centers, then it is assigned to the cluster with the smallest index position. As we will show in our simulations, this prior strongly favors round shaped cluster configurations.

In terms of the prior for the log-RR, Knorr-Held & Raßer (2000) make a choice that is similar to ours. In particular, they set $\theta_i = \tilde{\beta}_{c_i}$, with $\log \tilde{\beta}_k \sim N(\mu, \sigma^2)$. Their posterior sampling MCMC scheme is based on a reversible jump MCMC iterating over *birth*, *death*, *shift*, *switch*, *height* and *hyper* steps. This algorithm tends to mix very slowly, and requires a large number of iterations (in the order of millions of samples) to produce approximations with a reasonably small Monte Carlo error.

4.2. Prior specification and comparison criteria

In our simulation studies, we compare the performance of the RCRP model and BD model, under different specifications of the prior distributions. For the RCRP model, we fix $\alpha = 4$ (resulting on a prior distribution on the number of clusters centered roughly around $K = 5$), and a product of independent priors for (μ, σ^2),

$$\pi_1(\mu, \sigma^2) = N(\mu \mid q_{0.5}, s_n^2/2) IG(\sigma^2 \mid 2, s_n^2/2), \tag{14}$$

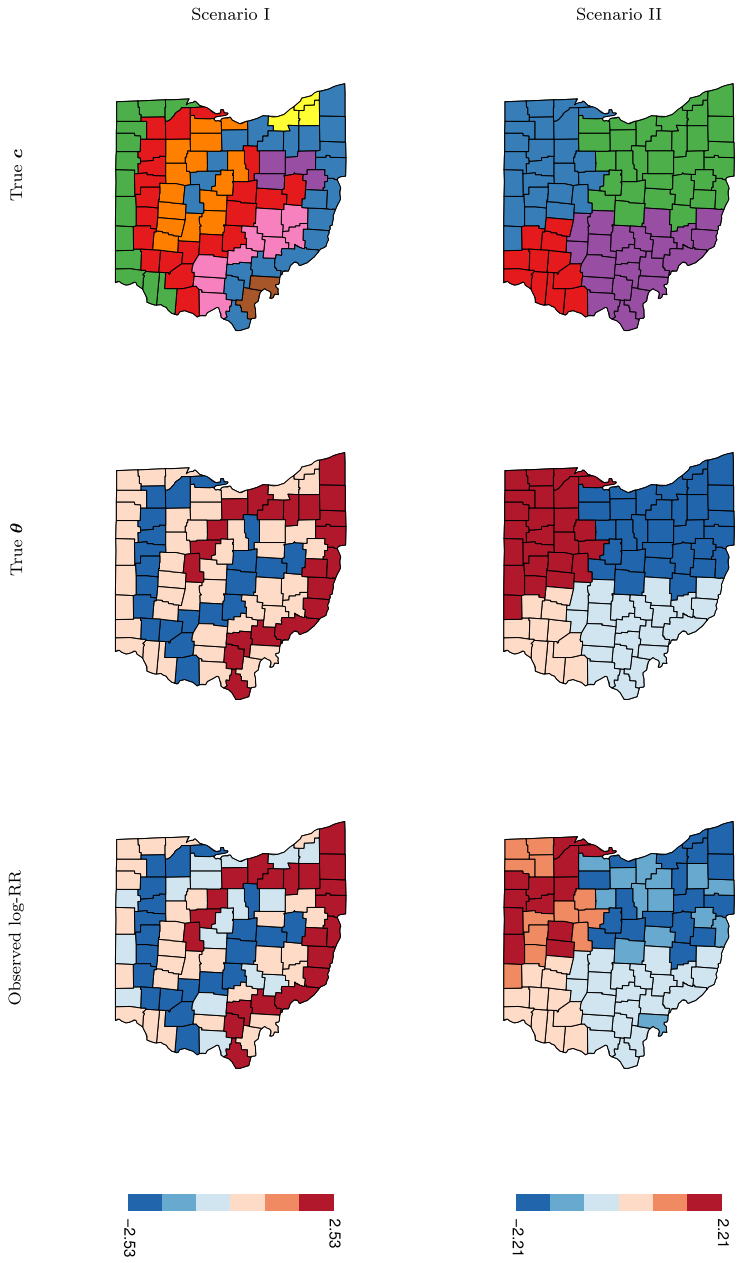


FIG 2. *Simulation Scenarios I and II: True cluster configuration, true relative risk in log scale, and observed relative risk in log scale for Scenarios I and II.*

where $q_{0.5}$ and s_n^2 denote the median and unbiased sample variance of $\log(y_i/h_i)$. The hyperparameters $\kappa = q_{0.5}$ and $a = 2$ were chosen such that the prior mean of the log RR are centered at $q_{0.5}$ and the prior variance of σ^2 is infinity. A sensitivity analysis involving three more combinations of hyperparameters for ϕ^2 and b is included in the supplementary material (Wehrhahn et al., 2020). Results appeared to be robust to the different prior specifications.

To ensure that the comparison between models is fair, we slightly modify the hyperpriors for the BD model from those originally used in Knorr-Held & Raßer (2000). In particular, we assign (μ, σ^2) the same hyperpriors as the RCRP model. Additionally, rather than the original geometric prior used by Knorr-Held & Raßer (2000), we consider two slightly different prior distributions for K that more closely resemble the prior implied by our model. These two priors correspond to a truncated Poisson and a truncated Negative Binomial, respectively.

$$\pi_1(K) = Poi(K \mid \lambda = 5)\mathbb{1}(K \geq 1),$$
$$\pi_2(K) = NB(K \mid p = 0.2, r = 1)\mathbb{1}(K \geq 1),$$

so that $E(K) = 5.03$ under π_1 and $E(K) = 5.1$ under π_2.

For each of the two models, we report point estimates for the cluster configuration, $\hat{\boldsymbol{c}}$, heat maps of the posterior probability of two regions belonging to the same cluster, $\pi(c_i = c_j \mid \boldsymbol{y})$, $i \neq j$ (which provide a measure of the uncertainty associated with the point estimates), and a comparison between the prior and posterior distributions over the number of clusters K. The point estimate $\hat{\boldsymbol{c}}$ is obtained by minimizing (using iterative componentwise optimization) a slightly modified version of the expected loss function discussed in Lau & Green (2007),

$$U(\hat{\boldsymbol{c}}) = Q(\hat{\boldsymbol{c}}, \boldsymbol{W}) \sum_{i=1}^{n} \sum_{j=i+1} \mathbb{1}\left(\hat{c}_i = \hat{c}_j\right) \left[\frac{w_2}{w_1 + w_2} - \pi(c_i = c_j \mid \boldsymbol{y})\right].$$

Note that the ratio w_1/w_2 controls the relative loss of incorrectly clustering or separating a pair of regions, and the multiplier $Q(\hat{\boldsymbol{c}}, \boldsymbol{W})$ ensures that our point estimate corresponds to an admissible partition. In our illustrations we set $w_1/w_2 = 1$.

We evaluate the ability of the models to identify clusters using the adjusted random index (ARI, Hubert & Arabie, 1985) of the posterior cluster configurations. The ARI evaluates the agreement in cluster assignment between two cluster configurations. It ranges between -1 and 1, larger values indicating agreement between cluster configurations. On the other hand, estimates of the log-RR are evaluated through the mean squared error (MSE) of the posterior mean of the log-RR.

4.3. Results

For each model, a single Markov chain was generated. In all cases, the inferences presented below are based on 10,000 samples obtained after burn-in and

thinning. The amount of burn-in varied; we discarded 40,000 samples in both instances of the RCRP model, 200,000 for the BD model in Scenario I, and 300,000 for the BD model in Scenario II. Convergence was evaluated by standard convergence tests, as implemented in the `CODA R` package (Plummer et al., 2009), and by examining the trace plot of the log-posterior distribution, the number of clusters $K(\boldsymbol{c})$, and the hyperparameters μ and σ^2.

Figure 3 displays point estimates for the cluster configurations for Scenario I and Scenario II. Under Scenario I, the BD model struggles even though the rates associated with the clusters are quite well differentiated. In particular, note that the BD model breaks down the 8 snake-like clusters into a large number of very small, round clusters. Under scenario II, the BD model improves its performance substantially, but still tends to slightly overestimate the number of clusters. In particular, note that the upper left and bottom right clusters are being broken down by the BD model into two subclusters each. In contrast, the RCRP model is able to recover the true cluster structure in both scenarios.

Figure 4 provides further insight into the estimates of the cluster structure by displaying the heat maps of the posterior probability of two regions belonging to the same cluster. To facilitate visualization, regions are ordered according to the true cluster configuration. In general, there is very little uncertainty associated with the point estimates presented in Figure 3, particularly for the RCRP model. Along similar lines, Figure 5, displays boxplots of the posterior distribution of the ARI. We can see that, while the posterior distribution of the ARI for the RCRP model is concentrated around 1 (further confirming that the model places high probability on the true clustering configuration), the values for the BD model tend to be much smaller, particularly in Scenario I. It is also worth noting that, for the BD model, $\pi_2(K)$ leads to much higher variability in the quality of the cluster estimates.

Finally, we compute the MSE of the log-RR with respect to the true log-RR. Under Scenario I the MSE for the RCRP model and $\alpha = 4$, BD model and $\pi_1(K)$, and BD model and $\pi_2(K)$ were 0.00145, 0.00579, and 0.00731, respectively. On the other hand, under Scenario II, the respective MSEs were 0.00057, 0.00073 and 0.00074. For both scenarios the RCRP model has the best performance; in the best case scenario, the MSE of the BD model, was 3.99 and 1.28 times bigger that the MSE for the RCRP model in Scenario I and Scenario II, respectively. Also, note that, under the BD model, prior $\pi_1(K)$ shows the best performances.

Further results comparing the performance of the models regarding the number of clusters and the log-RRs can be found in the online supplementary material (Wehrhahn et al., 2020).

5. An application to oral cancer in Germany

As a second illustration, this Section reports our analysis of the oral cancer mortality data discussed in Knorr-Held & Raßer (2000) (see Figure 6). The

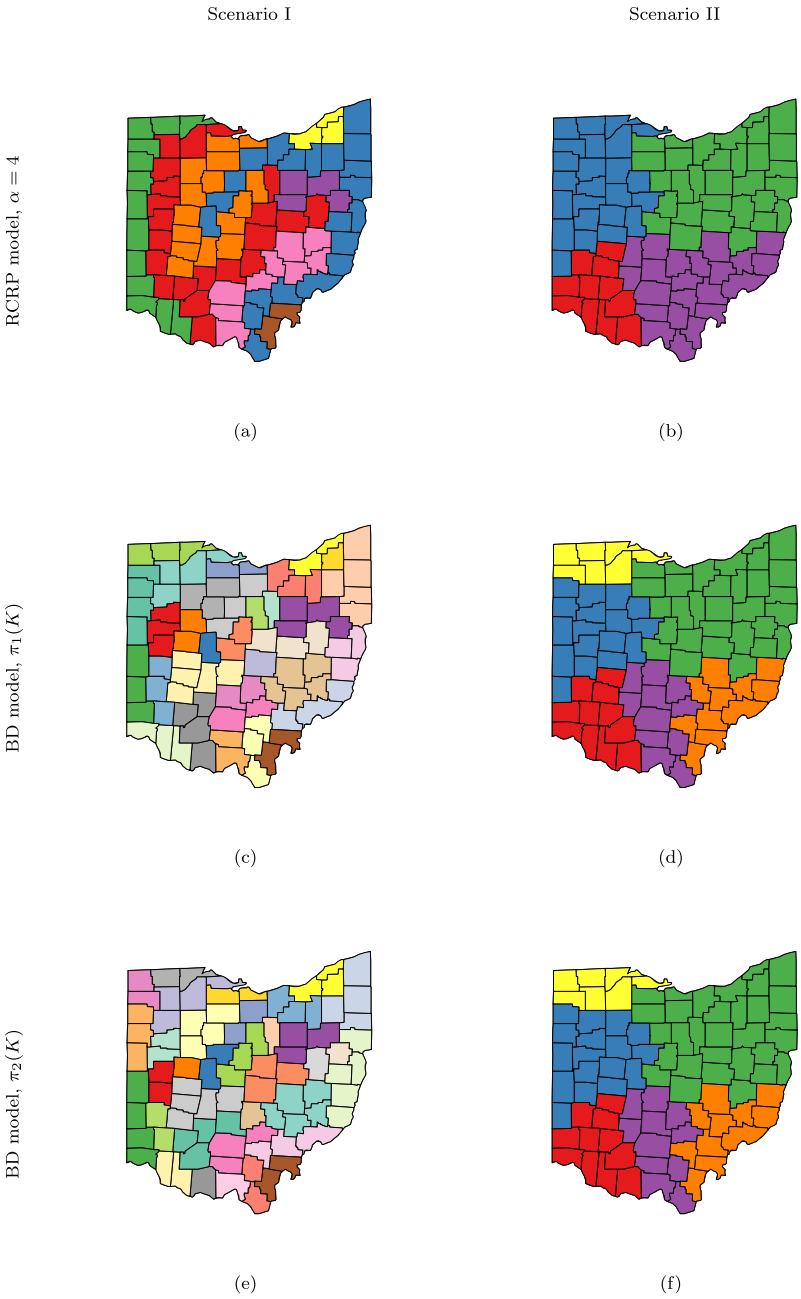


FIG 3. *Cluster configuration estimates under Scenario I and Scenario II: minimum expected loss cluster configuration for RCRP model and BD model under priors π_1 and π_2 for K.*

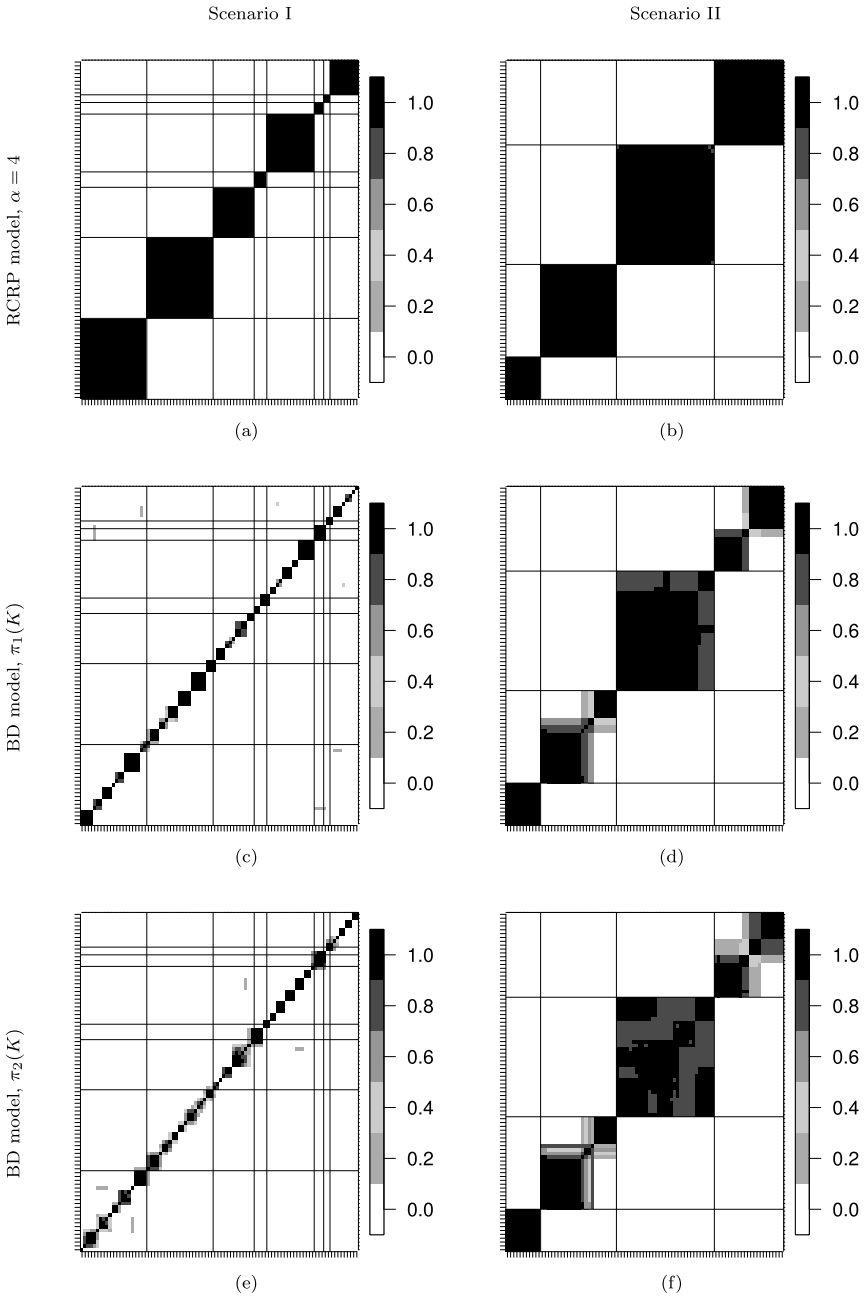


FIG 4. *Heat map of two regions belonging to the same cluster under Scenario I and Scenario II: heat map for RCRP model and BD model under priors π_1 and π_2 for K.*

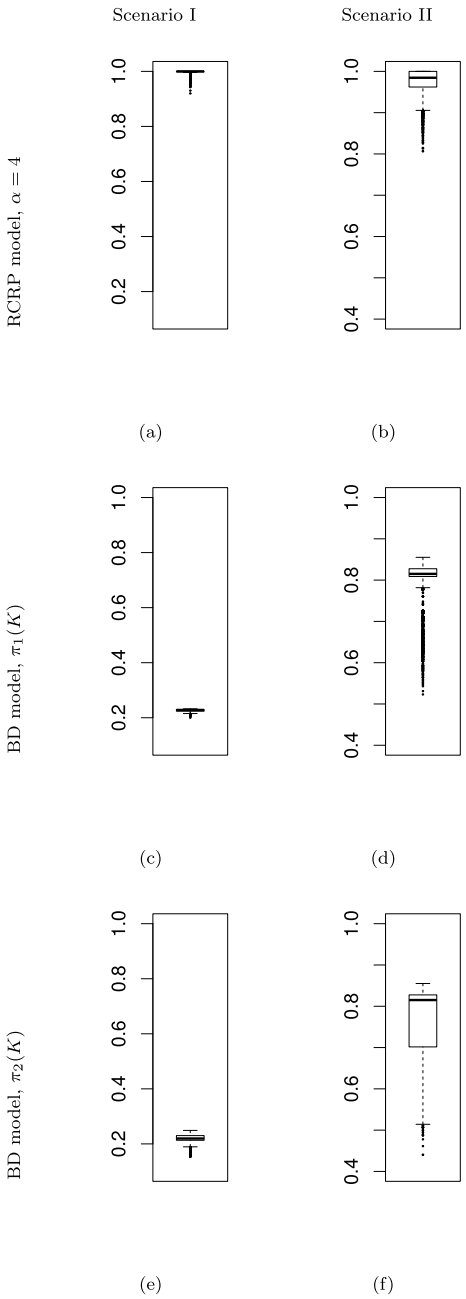
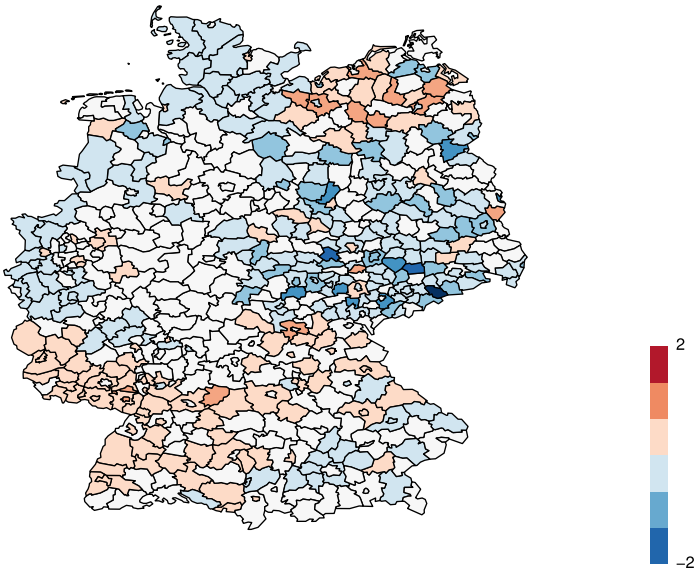


FIG 5. *Posterior distribution of ARI under Scenario I and Scenario II: Boxplot of posterior distribution of ARI for RCRP model and BD model under priors π_1 and π_2 for K.*



(a) Observed log RR

FIG 6. *Observed log relative risk for Germany data.*

data set registers the observed and expected number of deaths during the period 1986–1990 across 544 administrative districts in Germany. As in the simulation studies, this analysis emphasizes a comparison between the RCRP and the BD models.

5.1. Prior specification

Under the RCRP model, we employ a discrete uniform prior distribution for α. This prior has support on the set $\{16, 20, 24, 28, 32\}$ and is denoted $\pi_1(\alpha)$. This choice results in a prior distribution for the number of clusters centered around 16. As we discussed before, the use of a discrete prior enables additional flexibility while containing the computational cost of the MCMC algorithm (by allowing us to pre-compute approximations to the ratio of intractable normalizing constants used in the noisy exchange algorithm). As in the case of the simulated data, the hyperprior on (μ, σ^2) is given by (14).

For the BD model, two prior specifications for K were considered:

$$\pi_3(K) = NB(K \mid p = 0.65, r = 92.84)\mathbb{1}(K \geq 1),$$
$$\pi_4(K) = NB(K \mid p = 0.01, r = 1)\mathbb{1}(K \geq 1).$$

Both prior distributions are centered around 50, with $\pi_4(K)$ being more dispersed than $\pi_3(K)$. Knorr-Held & Raßer (2000) used $\pi_4(K)$ in their analysis. In

an attempt to reproduce their results, for the hyperparameters (μ, σ^2) we also follow Knorr-Held & Raßer (2000) and use independent priors with $p(\mu) \propto 1$ and $\sigma^2 \sim IG(1, 0.01)$.

5.2. Results

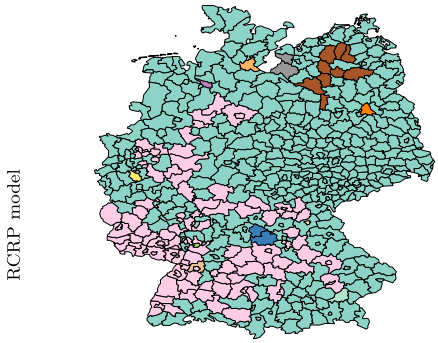
As before, we generate a single Markov chain for each model specification. For the RCRP model, one sample of size 10,000 was generated, obtained by saving 1 out of every 10 iterations and after a burn-in period of 20,000 samples. In order to approximate the intractable ratio of normalizing constants, one sample of size 10,000 was generated for each pair of values of α in the support of the discrete prior, after a burn-in period of 15,000 iterations. Based on these samples, the ratios of normalizing constants, used in the M-H updating step of α, were estimated as in (13). For the BD model, we also generated posterior samples of size 10,000. However, following Knorr-Held & Raßer (2000), in this case the samples were obtained after a burn-in period of 1,000,000 iterations by saving 1 out of every 10,000 iterations. As in the case of the simulated data, convergence of these chains was evaluated by standard convergence tests, as implemented in the CODA R package (Plummer et al., 2009), and by examining the trace plot of the log-posterior distribution, the number of clusters in the data, and the hyperparameters μ and σ^2.

Figure 7 presents our point estimate of the partition structure under each model. The differences are again striking. The RCRP model identifies five clusters: a main one, formed by the vast majority of regions, two small ones (formed by 110 and 9 regions, respectively), and a couple of singleton clusters. In contrast, the BD model identifies 84 and 98 clusters under $\pi_4(K)$ and $\pi_3(K)$, respectively. The shape of the clusters suggests that we might be in a situation that is similar to the one in our first simulated data sets, where the BD model artificially splits large, non-circular clusters into a large number of smaller, roughly circular ones.

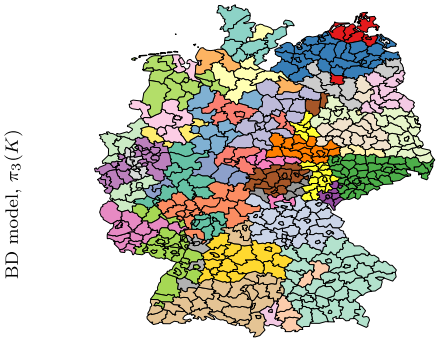
To further emphasize the difference in the reported partition structure, we present in Figure 8 heat maps of the posterior probability of two regions belonging to the same cluster. From these graphs we can see that, while the level of uncertainty in the point estimates is somewhat larger in this case when compared to the simulation studies, it is clear that all models are quite certain about the main features of their reported partition estimates.

Finally, Figure 9 displays the posterior mean log-RR estimate under both models. There are some clear similarities in the estimated rates. For example, both the RCRP and the BD models estimate a higher incidence of the disease in the Southwest corner of Germany, and a lower incidence in the East of the country. However, it is clear that the RCRP model smooths the rates much more than the BD model. This is not surprising given the very different estimates of the partition structures induced by these models.

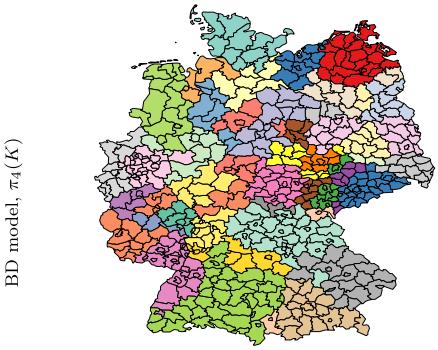
Further results comparing the performance of the models regarding the number of clusters can be found in the supplementary material (Wehrhahn et al., 2020).



(a)



(b)



(c) BD model, $\pi_4(K)$

FIG 7. *Minimum posterior expectation loss cluster configuration estimate for Germany data.*

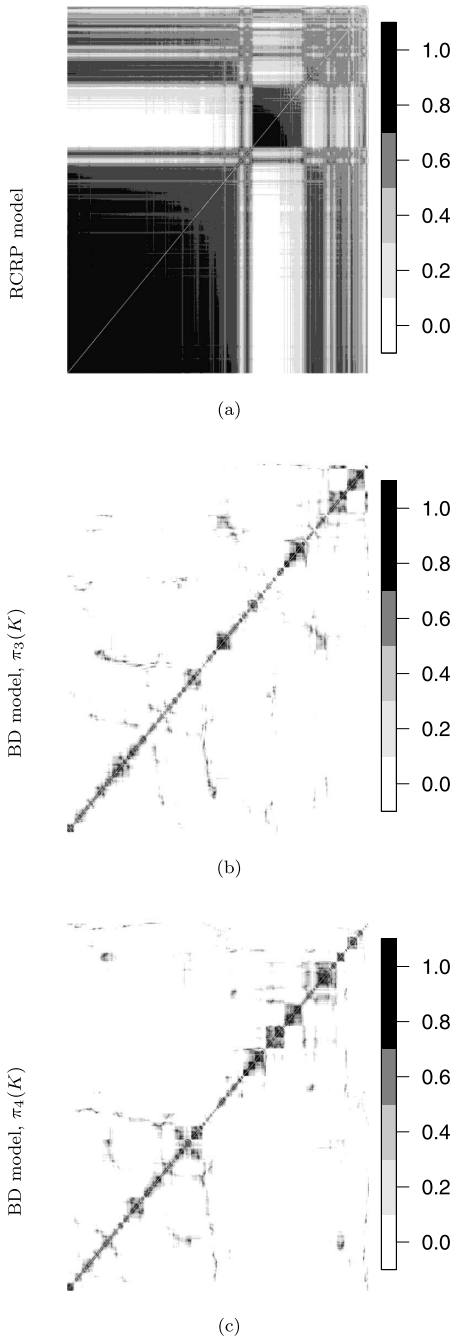


FIG 8. *Heat map of probability of two regions belonging to the same cluster, Germany data.*

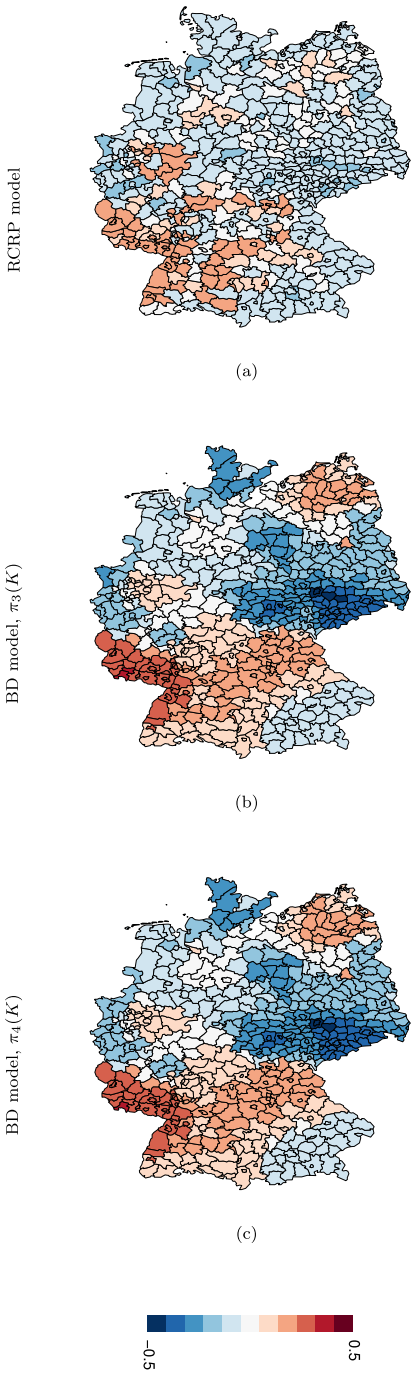


FIG 9. *Posterior mean log relative-risk, Germany data.*

6. Discussion

We have proposed a restricted mixture model for detecting clusters of non communicable diseases. The restriction is imposed in the CRP prior for the cluster membership vector, constraining the space of possible configurations only to those resulting in connected clusters, i.e., those where there is a path joining any pair of regions that includes only regions that belong to the cluster. We show that the model is very flexible and less computationally demanding that the alternative BD model.

A number of extensions of this model are possible. For example, while we have focused here on restrictions of a CRP prior, the basic approach could be used to restrict any other exchangeable SSM. Our key result around the structure of the full conditional distribution of the restricted model should extend in a straightforward fashion. Similarly, the model could easily accommodate different likelihood functions and other more general definitions of the binary neighborhood matrix $\boldsymbol{W}$. Furthermore, while this paper has focused on applications to disease clustering, it is clear that our approach can be extended to applications in time series and image segmentation. Another extension would involve using dependent priors for the cluster specific parameters $\tilde{\theta}_1, \ldots, \tilde{\theta}_K$. Indeed, we might expect that nearby clusters have more similar rates than clusters located far away. For example, we could consider a (proper) conditionally autoregressive prior for $\tilde{\theta}_1, \ldots, \tilde{\theta}_K$. However, such an approach introduces complex identifiability issues that make prior elicitation complex. In particular, note that a model in which there is a single cluster is equivalent to the (limit) model in which we allow for any number of clusters but make the spatial correlation in the prior goes to 1. This means that the prior distribution would have an even more critical effect on the model, one that would be hard to measure a priori. In that sense, the use of independent priors can be seen as a kind of "maximum separation" prior that will maximize the ability of the model to identify a disease cluster. Finally, an anonymous referee suggested the use of a more general form for $Q(\boldsymbol{c}, \boldsymbol{W})$ that would allow for continuous values on $[0, 1]$.

Our model works by restricting an EPPF, which leads to a model that is partially exchangeable. That is, the probability distribution implied by the model is invariant to simultaneous permutations of the observations *and* the rows and columns of the neighborhood matrix $\boldsymbol{W}$. A referee pointed out that using an *exchangeable* PPF as the starting point is not required. While this is true as a general modeling strategy, the appropriateness of such an approach will depend on the application at hand. For example, in applications to time series data (where there is a natural ordering to the observations), the use of such a non-exchangeable PPF as our starting point might make sense. However, in spatial applications such as the ones considered in this paper, partial exchangeability would seem to be the right assumption: why would one want to have a different model when counties are listed alphabetically in the database vs. when they are listed by, say, population size?

One shortcoming of our model is that it uses a single set of random effects to capture both overdispersion in the Poisson model and spatial dependence. One

way to address this limitation is to use two sets of random effects: one set in which they are independent (meant to capture over-dispersion), and another set in which they are dependent and modeled using our restricted CRP (therefore capturing the spatial effects in the data). See, for example, Banerjee et al. (2014). Such an approach could be easily incorporated into our disease clustering model.

Appendix A: Two special restrictions

In what follows we consider two special adjacency graphs for which the normalizing constant $C(\alpha, \boldsymbol{W})$ can be computed in closed form: (1) when the underlying graph G is the star graph, and (2) when the underlying graph G is linear. The former is of interest for its simplicity, and the latter because of its potential application to modeling ordered/time series data.

The analysis of the these two special graphs is also useful in terms of highlighting the differences between our model and that of Martínez et al. (2014). Indeed, if the approach in Martínez et al. (2014) was extended beyond time series models to accommodate general adjacency graphs, it would lead to exactly the same prior distribution on partitions under either graphs. That is because the number of admissible partitions for any size happens to be the same under both graphs. In contrast, our model leads to quite different specifications for these two graphs because our prior weights partitions according to the size of the clusters.

A.1. Restriction under a star graph

Under this adjacency structure, in any cluster configuration with l clusters there will be exactly $l-1$ singleton clusters, and one cluster with $n-l+1$ observations (which must include the root node). This is because any cluster of size greater than one must necessarily include the root node in order to be formed of adjacent regions. See the top row of Table 2 for an example with $n = 4$. Hence,

$$f_l\left(\boldsymbol{W}_{\text{star}}\right) = \binom{n-1}{n-l}\Gamma(n-l+1) = \frac{(n-1)!}{(l-1)!},$$

and

$$C\left(\alpha, \boldsymbol{W}_{\text{star}}\right) = \sum_{l=1}^{n} \frac{(n-1)!}{(l-1)!}\alpha^l.$$

A.2. Restriction under a linear graph

In a linear graph, there are also a total of $\binom{n-1}{l-1}$ configurations involving exactly l clusters. Indeed, note that choosing l adjacent clusters is equivalent to picking $l-1$ breakpoints out in $n-1$ possible positions for them. However, unlike the star case, each of these configurations involves clusters of different sizes (see the bottom row of Table 2).

TABLE 2
All possible partitions and count of cluster sizes, $\boldsymbol{a}$ for a sample of size $n = 4$ under star and linear graphs.

Graph		$K = 1$	$K = 2$	$K = 3$	$K = 4$
linear	$A_k, k = 1, \ldots, K$	$\{1, 2, 3, 4\}$	$\{\{1\}, \{2, 3, 4\}\}$ $\{\{1, 2, 3\}, \{4\}\}$ $\{\{1, 2\}, \{3, 4\}\}$	$\{\{1\}, \{2\}, \{3, 4\}\}$ $\{\{1\}, \{2, 3\}, \{4\}\}$ $\{\{1, 2\}, \{3\}, \{4\}\}$	$\{\{1\}, \{2\}, \{3\}, \{4\}\}$
star	$A_k, k = 1, \ldots, K$	$\{1, 2, 3, 4\}$	$\{\{1, 2, 3\}, \{4\}\}$ $\{\{1, 2, 4\}, \{3\}\}$ $\{\{1, 3, 4\}, \{2\}\}$	$\{\{1, 2\}, \{3\}, \{4\}\}$ $\{\{1, 3\}, \{2\}, \{4\}\}$ $\{\{1, 4\}, \{2\}, \{3\}\}$	$\{\{1\}, \{2\}, \{3\}, \{4\}\}$

If we interpret the indexes $i_1, \ldots, i_{l-1} \in \{1, \ldots, n-1\}$ as the sizes of the clusters (organized in sequential order), we have:

$$f_l\left(\boldsymbol{W}_{\text{linear}}\right) = \begin{cases} \Gamma(n) & l = 1, \\ \displaystyle\sum_{i_1=1}^{n+l-1} \sum_{i_2=1}^{n-i_1-1} \cdots \sum_{i_{l-1}=1}^{n-i_{l-2}^{+}-1} \Gamma(n - i_{l-1}^{+}) \prod_{r=1}^{l-1} \Gamma(i_r) & l \geq 2, \end{cases}$$

where $i_l^{+} = \sum_{r=1}^{l} i_r$. Therefore,

$$C\left(\alpha, \boldsymbol{W}_{\text{linear}}\right) = \Gamma(n)\alpha + \sum_{l=2}^{n} \left\{ \sum_{i_1=1}^{n+l-1} \sum_{i_2=1}^{n-i_1-1} \cdots \sum_{i_{l-1}=1}^{n-i_{l-2}^{+}-1} \Gamma(n - i_{l-1}^{+}) \prod_{r=1}^{l-1} \Gamma(i_r) \right\} \alpha^l.$$

The computation of f_l can be efficiently implemented using a simple recursive algorithm similar to the one used to enumerate all possible combinations of n elements into l groups. Since $\boldsymbol{W}$ is fixed in our model, this computation only needs to be done once at the beginning of the MCMC algorithm.

Appendix B: Derivation of the full conditional distributions for the restricted Chinese restaurant process

Recall that the spatially restricted joint distribution for $\boldsymbol{c}$ is given by

$$\pi(\boldsymbol{c} \mid \alpha, \boldsymbol{W}) = \frac{\alpha^{K(\boldsymbol{c})}}{C(\alpha, \boldsymbol{W})} \left\{ \prod_{k=1}^{K(\boldsymbol{c})} \Gamma\left(n_k(\boldsymbol{c})\right) \right\} Q(\boldsymbol{c}, \boldsymbol{W}), \tag{15}$$

where $C(\alpha, \boldsymbol{W})$ is the normalizing constant. From (15), it follows that

$$\pi(c_i = l \mid \boldsymbol{c}_{-i}, \alpha, \boldsymbol{W}) \propto \alpha^{K(\boldsymbol{c})} \left\{ \prod_{k=1}^{K(\boldsymbol{c})} \Gamma(n_k(\boldsymbol{c})) \right\} Q(\boldsymbol{c}, \boldsymbol{W}),$$

where $l = 1, \ldots, K(\boldsymbol{c}_{-i})+1$. If $c_i = l$ leads to an inadmissible configuration, then $\pi(c_i = l \mid \boldsymbol{c}_{-i}, \boldsymbol{W}, \alpha) \propto 0$. On the other hand, if $c_i = l$ leads to an admissible

configuration, then we must consider two cases. For $l \leq K(\boldsymbol{c}_{-i})$,

$$\pi(c_i = l \mid \boldsymbol{c}_{-i}, \alpha, \boldsymbol{W}) \propto \alpha^{K(\boldsymbol{c}_{-i})} \prod_{k=1}^{K(\boldsymbol{c}_{-i})} \Gamma\left(n_k(\boldsymbol{c}_{-i}) + \mathbb{1}(k = l)\right),$$
$$= n_l(\boldsymbol{c}_{-i}) \left[\alpha^{K(\boldsymbol{c}_{-i})} \prod_{k=1}^{K(\boldsymbol{c}_{-i})} \Gamma\left(n_k(\boldsymbol{c}_{-i})\right)\right]. \tag{16}$$

This follows from the fact that

$$\Gamma\left(n_k(\boldsymbol{c}_{-i}) + \mathbb{1}(k = l)\right) = \begin{cases} \Gamma\left(n_k(\boldsymbol{c}_{-i})\right) & k \neq l, \\ n_k(\boldsymbol{c}_{-i})\Gamma\left(n_k(\boldsymbol{c}_{-i})\right) & k = l. \end{cases}$$

On the other hand, if $l = K(\boldsymbol{c}_{-i}) + 1$,

$$\pi(c_i = l \mid \boldsymbol{c}_{-i}, \alpha, \boldsymbol{W}) \propto \alpha^{K(\boldsymbol{c}_{-i})+1} \prod_{k=1}^{K(\boldsymbol{c}_{-i})} \Gamma\left(n_k(\boldsymbol{c}_{-i})\right),$$
$$= \alpha \left[\alpha^{K(\boldsymbol{c}_{-i})} \prod_{k=1}^{K(\boldsymbol{c}_{-i})} \Gamma\left(n_k(\boldsymbol{c}_{-i})\right)\right]. \tag{17}$$

The simplified expression in (7) are obtained by noting that both (16) and (17) include a term of the form

$$\alpha^{K(\boldsymbol{c}_{-i})} \prod_{k=1}^{K(\boldsymbol{c}_{-i})} \Gamma\left(n_k(\boldsymbol{c}_{-i})\right),$$

which can be treated as part of the proportionality constant.

The previous derivation assumes that there is at least one value of c_i in the set $\{1, \ldots, K(\boldsymbol{c}_{-i}) + 1\}$ that leads to an admissible configuration. Otherwise the normalizing constant is zero and the full conditional is not well defined. Since our MCMC algorithm must start in an admissible configuration, and the full conditionals maintain admissibility, this is not an issue for our purposes.

Acknowledgements

The authors would like to thank the Editor and three anonymous referees for helpful comments and suggestions. Claudia Wehrhahn was partially supported by award NSF-DMS 1622444 and Abel Rodriguez was partially supported by award NSF-DMS 1738053 and 1740850.

Supplementary Material

Supplementary material to: "Bayesian approach to Disease Clustering using restricted Chinese restaurant processes"
(doi: 10.1214/20-EJS1696SUPP; .pdf).

References

Alquier, P., Friel, N., Everitt, R. & Boland, A. (2016). Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Statistics and Computing* **26** 29–47. MR3439357

Anderson, C., Lee, D. & Dean, N. (2014). Identifying clusters in Bayesian disease mapping. *Biostatistics* **15** 457–469.

Antoniak, C. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *Annals of Statistics* **2** 1152–1174. MR0365969

Banerjee, S., Carlin, B. P. & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC. MR3362184

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

Blackwell, D. & MacQueen, J. B. (1973). Ferguson Distribution via Pólya Urn Schemes. *The Annals of Statistics* **1** 353–355. MR0362614

Blei, D. M. & Frazier, P. I. (2011). Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research* **12** 2461–2488. MR2834504

Charras-Garrido, M., Abrial, D., De Goër, J., Dachian, S. & Peyrard, N. (2012). Classification method for disease risk mapping based on discrete hidden Markov random fields. *Biostatistics* **13** 241–255.

Dahl, D. B. (2008). Distance-based probability distribution for set partitions with applications to Bayesian nonparametrics. *JSM Proceedings. Section on Bayesian Statistical Science, American Statistical Association, Alexandria, Va.*

Dahl, D. B., Day, R. & Tsai, J. W. (2017). Random partition distribution indexed by pairwise information. *Journal of the American Statistical Association* **112** 721–732. MR3671765

Damien, P., Wakefield, J. & Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61** 331–344. MR1680334

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *Annals of Statistics* **1** 209–230. MR0350949

Fernández, C. & Green, P. J. (2002). Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the royal statistical society: series B (Statistical methodology)* **64** 805–826. MR1979388

Fitzpatrick, M. C., Preisser, E. L., Porter, A., Elkinton, J., Waller, L. A., Carlin, B. P. & Ellison, A. M. (2010). Ecological boundary detection using Bayesian areal wombling. *Ecology* **91** 3448–3455.

Fuentes-García, R., Mena, R. H. & Walker, S. G. (2010). A probability for classification based on the Dirichlet process mixture model. *Journal of classification* **27** 389–403. MR2748990

Gangnon, R. E. & Clayton, M. K. (2000). Bayesian detection and modeling of spatial disease clustering. *Biometrics* **56** 922–935.

Ghosh, S., Ungureanu, A. B., Sudderth, E. B. & Blei, D. M. (2011). Spatial distance dependent Chinese restaurant processes for image segmentation. In *Advances in Neural Information Processing Systems*. 1476–1484.

Gnedin, A. & Pitman, J. (2006). Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical sciences* **138** 5674–5685. MR2160320

Gómez-Rubio, V., Ferrándiz-Ferragud, J. & López-Quílez, A. (2005). Detecting clusters of disease with R. *Journal of Geographical Systems* **7** 189–206.

Gómez-Rubio, V., Molitor, J. & Moraga, P. (2018). Fast Bayesian classification for disease mapping and the detection of disease clusters. In *Quantitative Methods in Environmental and Climate Research*. Springer, 1–27.

Goujon-Bellec, S., Demoury, C., Guyot-Goubin, A., Hémon, D. & Clavel, J. (2011). Detection of clusters of a rare disease over a large territory: performance of cluster detection methods. *International journal of health geographics* **10** 53.

Green, P. J. & Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American statistical association* **97** 1055–1070. MR1951259

Guhaniyogi, R. (2017). Bayesian nonparametric areal wombling for small-scale maps with an application to urinary bladder cancer data from Connecticut. *Statistics in medicine* **36** 4007–4027. MR3713645

Hartigan, J. A. (1990). Partition models. *Communications in statistics-Theory and methods* **19** 2745–2756. MR1088047

Heinzl, F. & Tutz, G. (2014). Clustering in linear-mixed models with a group fused lasso penalty. *Biometrical Journal* **56** 44–68. MR3152702

Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of classification* **2** 193–218.

Knorr-Held, L. & Rasser, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics* **56** 13–21.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and methods* **26** 1481–1496. MR1456844

Kulldorff, M. & Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine* **14** 799–810. URL http://dx.doi.org/10.1002/sim.4780140809.

Kulldorff, M., Tango, T. & Park, P. J. (2003). Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis* **42** 665–684. MR1977177

Lau, J. W. & Green, P. J. (2007). Bayesian model-based clustering procedures. *Journal of Computational and Graphical Statistics* **16** 526–558. MR2351079

Lee, J., Quintana, F. A., Müller, P. & Trippa, L. (2013). Defining predictive probability functions for species sampling models. *Statistical science: a review journal of the Institute of Mathematical Statistics* **28** 209. MR3112406

Li, C., Phung, D., Rana, S. & Venkatesh, S. (2013). Exploiting side information in distance dependent chinese restaurant processes for data clustering.

In *2013 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

Li, C., Rana, S., Phung, D. & Venkatesh, S. (2014). Regularizing topic discovery in EMRS with side information by using hierarchical Bayesian models. In *2014 22nd International Conference on Pattern Recognition*. IEEE, 1307–1312.

Li, C., Rana, S., Phung, D. & Venkatesh, S. (2015a). Small-variance asymptotics for Bayesian nonparametric models with constraints. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 92–105.

Li, C., Rana, S., Phung, D. & Venkatesh, S. (2016a). Data clustering using side information dependent Chinese restaurant processes. *Knowledge and information systems* **47** 463–488.

Li, C., Rana, S., Phung, D. & Venkatesh, S. (2016b). Dirichlet Process Mixture Models with Pairwise Constraints for Data Clustering. *Annals of data science* **3** 205–223.

Li, C., Rana, S., Phung, D. & Venkatesh, S. (2016c). Hierarchical Bayesian nonparametric models for knowledge discovery from electronic medical records. *Knowledge-Based Systems* **99** 168–182.

Li, C. Y. (2015). *Exploiting side information in Bayesian nonparametric models and their applications*. Ph.D. thesis, Deakin University.

Li, P., Banerjee, S., Hanson, T. A. & McBean, A. M. (2015b). Bayesian models for detecting difference boundaries in areal data. *Statistica Sinica* 385–402. MR3328821

Loschi, R. H. & Cruz, F. R. (2005). Extension to the product partition model: computing the probability of a change. *Computational Statistics & Data Analysis* **48** 255–268. MR2133587

Lu, H. & Carlin, B. P. (2005). Bayesian areal wombling for geographical boundary analysis. *Geographical Analysis* **37** 265–285.

Lu, H., Reilly, C. S., Banerjee, S. & Carlin, B. P. (2007). Bayesian areal wombling via adjacency modeling. *Environmental and Ecological Statistics* **14** 433–452. MR2405556

MacEachern, S. N. & Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7** 223–338.

Martínez, A. F., Mena, R. H. et al. (2014). On a nonparametric change point detection model in Markovian regimes. *Bayesian Analysis* **9** 823–858. MR3293958

Moraga, P. & Montes, F. (2011). Detection of spatial disease clusters with LISA functions. *Statistics in medicine* **30** 1057–1071. MR2767842

Morton-Jones, T., Diggle, P. & Elliott, P. (1999). Investigation of excess environmental risk around putative sources: Stone's test with covariate adjustment. *Statistics in medicine* **18** 189–197.

Müller, P., Quintana, F. & Rosner, G. L. (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics* **20** 260–278. MR2816548

Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9** 249–265. MR1823804

Openshaw, S., Charlton, M., Wymer, C. & Craft, A. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information System* **1** 335–358.

Page, G. L., Quintana, F. A. et al. (2016). Spatial product partition models. *Bayesian Analysis* **11** 265–298. MR3465813

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability theory and related fields* **102** 145–158. MR1337249

Pitman, J. (1996). Some developments of the blackwell-macqueen urn scheme. *Lecture Notes-Monograph Series* 245–267. MR1481784

Pitman, J. & Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25** 855–900. MR1434129

Plummer, M., Best, N., Cowles, K. & Vines, K. (2009). *CODA: Output analysis and diagnostics for MCMC*. R package version 0.13-4.

Potthoff, R. F. & Whittinghill, M. (1966a). Testing for homogeneity: I. the binomial and multinomial distributions. *Biometrika* **53** 167–182. MR0216642

Potthoff, R. F. & Whittinghill, M. (1966b). Testing for homogeneity: Ii. the Poisson distribution. *Biometrika* 183–190. MR0216643

Robert, C. P. & Casella, G. (2005). *Monte Carlo statistical methods (Springer Texts in Statistics)*. Secaucus, NJ, USA: Springer-Verlag. MR2080278

Rodríguez, A. & Quintana, F. A. (2015). On species sampling sequences induced by residual allocation models. *Journal of statistical planning and inference* **157** 108–120. MR3279480

Smith, A. F. & Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)* 3–23. MR1210421

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

Tango, T. (1995). A class of tests for detecting 'general' and 'focused' clustering of rare diseases. *Statistics in Medicine* **14** 2323–2334.

Tango, T. & Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International journal of health geographics* **4** 11.

Wakefield, J. & Kim, A. (2013). A Bayesian model for cluster detection. *Biostatistics* **14** 752–765.

Waller, L. A., Hill, E. G. & Rudd, R. A. (2006). The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations. *Statistics in Medicine* **25** 853–865. MR2225167

Wang, H. & Rodríguez, A. (2014). Identifying Pediatric Cancer Clusters in Florida Using Log-Linear Models and Generalized Lasso Penalties. *Statistics and Public Policy* **1** 86–96.

Wehrhahn, C., Leonard, S., Rodriguez, A. & Xifara, T. (2020). Supplementary material to: "Bayesian approach to Disease Clustering using restricted Chinese restaurant processes". DOI: 10.1214/20-EJS1696SUPP.

Weinstock, M. A. (1981). A generalised scan statistic test for the detection of clusters. *International Journal of Epidemiology* **10** 289–293.

Whittemore, A. S., Friend, N., Brown Jr, B. W. & Holly, E. A. (1987). A test to detect clusters of disease. *Biometrika* **74** 631–635. MR0909368