Restricted Nonparametric Mixtures models for Disease Clustering

Modelli di mistura non parametrici limitati per la clustering di malattia

Abel Rodríguez and Tatiana Xifara

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 contiguous areal units. This is achieved by modifying the exchangeable partition probability function associated with the Ewen's sampling distribution. The model is illustrated using data on US lung cancer rates.

Abstract *Identificare cluster di malattie (aree con incidenza insolitamente alta di una particolare malattia) è un problema comune in epidemiologia e in materia di sanità pubblica. Descriviamo un modello mistura Bayesiano nonparametrico per il clustering di malattie che forza i cluster ad essere composti da unità di aree contigue. Tale obiettivo è ottenuto modificando la funzione di probabilità di partizione scambiabile associata alla formula di campionamento di Ewens. Il modello è illustrato analizzando i dati sul tasso di incidenza di tumori polmonari negli Stati Uniti.*

Key words: Disease clustering, Areal Data; Nonparametric Bayes; Noisy Exchange Algorithm

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

Abel Rodríguez
Department of Applied Mathematics and Statistics, University of California Santa Cruz, CA, USA, e-mail: abel@soe.ucsc.edu

Tatiana Xifara
Department of Applied Mathematics and Statistics, University of California Santa Cruz, CA, USA e-mail: xifara@soe.ucsc.edu

communicable diseases (those that can be spread from one person to another, such as 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, which 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., Stone, 1988, Tango, 1995, Morton-Jones et al., 1999). See also Besag and Newell, 1991, who call them *focused tests*. In contrast, the focus of this paper is on methods for *de novo* identification of clusters in datasets in which the presence of the clusters is not known. Methods based on scan statistics (e.g., Weinstock, 1981, Kulldorff, 1997, Tango and Takahashi, 2005) a well known example of this type of approaches.

Methods for disease clustering can also be classified according to whether they are designed to work with point-referenced or 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 and 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 and 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 and Whittinghill, 1966a and Potthoff and Whittinghill, 1966b). Bayesian methods for disease clustering in spatially aggregated data have been proposed by Knorr-Held and Raßer (2000), Green and Richardson (2002), Wakefield and Kim (2013) and Anderson et al. (2013). Other recent contributions to the field include the work of Moraga and Montes (2011), Charras-Garrido et al. (2012), Heinzl and Tutz (2014) and Wang and 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 and Raßer, 2000, Green and Richardson, 2002 and Anderson et al., 2013), but most techniques designed for disease mapping are not useful in the context of disease clustering.

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 the Dirichlet process (Ferguson, 1973; Blackwell and MacQueen, 1973; Antoniak, 1974; Lee et al., 2013; Rodríguez and Quintana, 2015) as a prior on the partition of areal units. This re-

stricted prior is designed to enforce clusters made of adjacent spatial units. Our approach is related to those developed in Fuentes-García et al. (2010) and Martínez et al. (2014) in the context of change-point detection in time series analysis. Indeed, our model can be seen as generalizing Fuentes-García et al. (2010) and Martínez et al. (2014) to work for situations in which the EPPF is restricted to partitions driven by general neighborhood graphs.

As motivation, consider data on mortality rates from lung cancer in the 48 contiguous U.S. plus the District of Columbia in 2000 (see Figure 1). These mortality data are based on death certificates that are filed by certifying physicians. They are collected and maintained by the U.S.Ṅational Center for Health Statistics (http://www.cdc.gov/nchs) as part of the U.S. National Vital Statistics System. The data are available from the Surveillance, Epidemiology, and End Results (SEER) program of the National Cancer Institute (http://seer.cancer.gov/seerstat). Figure 1 suggests that mortality rates for lung cancer are particular high across the Appalachia region (the cultural region comprising the central and southern portions of the Appalachian mountain range and extending from the Southern tier of New York to northern Alabama, Mississippi and Georgia). These observed high mortality rates are consistent with the relatively high smoking rates in the region.

The remaining of the paper is organized as follows: Section 2 presents our model and discusses its properties. Section 3 describes our computational approach. Section 4 presents an application of our model to the US lung cancer mortality data presented above. Finally, Section 5 discusses the limitations of our model as well as future research directions.

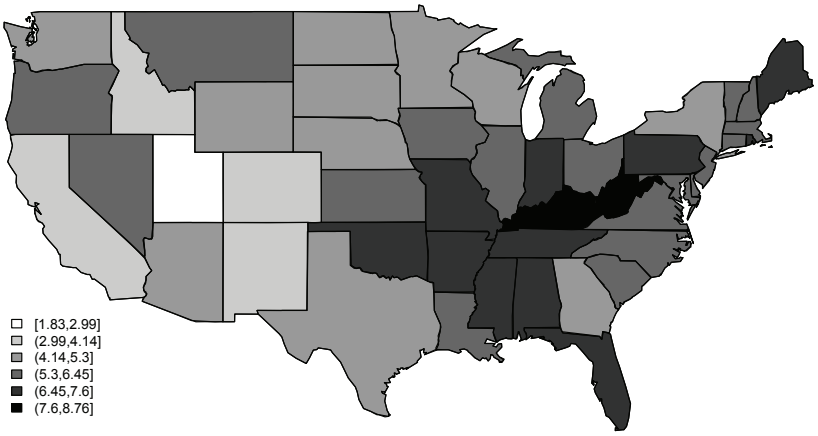


Fig. 1 Observed statewide mortality rates for lung cancer in the US during the year 2000.

2 A spatial clustering model for count data

Let y_i and h_i represent the observed number of cases (e.g. disease incidents or deaths) and the susceptible population in region $i = 1,\ldots,n$, respectively (in our motivating example $n = 49$). As is standard in the literature on disease mapping and clustering, we assume that data comes from a Poisson distribution

$$y_i \mid \lambda_i \sim \mathsf{Poi}(\lambda_i), \qquad i = 1,\ldots,n, \tag{1}$$

where the intensity λ_i is of the form $\log\lambda_i = \log h_i + x_i^T\theta_i$ and x_i is the set of covariates associated with region i. When no covariates are available we simply let $x_i = 1$ for all i, so that $\exp\{\theta_i\}$ corresponds to the disease rate associated with region i.

In addition, we assume that a neighborhood structure among the regions has been defined, and that it is encoded through a known $n \times n$ adjacency matrix $\mathbf{W}$ such that $w_{i,i'} = 1$ if regions i and i' are neighbors and $w_{i,i'} = 0$ otherwise. In this way, the neighborhood of any region i, denoted by ∂i can be easily defined as $\partial i = \{i' : w_{i,i'} = 1\}$. Because we are interested in spatially connected clusters where two regions can belong to a cluster only if they are adjacent, in this paper we focus exclusively on first-order neighborhood structures in which $w_{i,i'} = 1$ if and only if regions i and i' share a border. For example, in our motivating example, our first order neighborhood implies that $w_{i,i'} = 1$ when $i =$ California and $i' =$ Oregon, but $w_{i,i'} = 0$ when $i =$ California and $i' =$ New Mexico.

2.1 A prior model for spatial clustering

We are interested in clustering regions according to their underlying rates. To accomplish this we let $\theta_i = \vartheta_{\xi_i}$, where $\xi = (\xi_1,\ldots,\xi_n)$ is a vector of indicators taking values $\{1,2,\ldots,K\}$ that encode a partition $\rho_n = \{S_1,\ldots S_K\}$ of the n observations into K clusters (e.g., see Table 1), and $\vartheta_1,\ldots,\vartheta_K$ are the parameters associated with each of these K clusters.

K	ξ	Groups in the partition ρ_3
1	$(1,1,1)$	$S_1 = \{y_1,y_2,y_3\}$
2	$(1,1,2)$	$S_1 = \{y_1,y_2\}$, $S_2 = \{y_3\}$
2	$(1,2,1)$	$S_1 = \{y_1,y_3\}$, $S_2 = \{y_2\}$
2	$(1,2,2)$	$S_1 = \{y_1\}$, $S_2 = \{y_2,y_3\}$
3	$(1,2,3)$	$S_1 = \{y_1\}$, $S_2 = \{y_2\}$, $S_3 = \{y_3\}$

Table 1 All possible partitions associated with $n = 3$ geographical areas

A simple set of priors that fit into this framework correspond to setting $\vartheta_1,\ldots,\vartheta_K$ to be an independent and identically distributed sequence (e.g., $\vartheta_k \sim \mathsf{N}(\mu,\Sigma)$) and ξ to follow an Ewen's distribution,

$$p(\xi \mid \alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \alpha^{K(\xi)} \prod_{k=1}^{K(\xi)} \Gamma(m_k(\xi)), \tag{2}$$

where $K(\xi)$ is the number of unique values among $\xi_1, \ldots, \xi_n$ (i.e. the number of clusters in the partition), $m_k(\xi) = \sum_{i=1}^{n} \mathbb{I}(\xi_i = k) = |S_k|$ is the size of the k-th cluster, and $\mathbb{I}(\cdot)$ denotes the indicator function. This specification corresponds to a well-known Dirichlet process mixture of Poisson kernels, which is relatively simple to estimate. Indeed, note that the full conditional distribution for ξ_i given $\xi^{(-i)} = (\xi_1, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_n)$ reduces to

$$p\left(\xi_i \mid \xi^{(-i)}, \alpha\right) \propto \begin{cases} m_k\left(\xi^{(-i)}\right) & k \leq K\left(\xi^{(-i)}\right), \\ \alpha & k = K\left(\xi^{(-i)}\right) + 1, \end{cases} \tag{3}$$

where, in the same spirit as before, $K\left(\xi^{(-i)}\right)$ and $m_k\left(\xi^{(-i)}\right)$ represent the number of partitions and the size of the k-th partition remaining after eliminating observation i from the set. The simple form of the full conditional prior distribution (usually called the Chinese restaurant process) means that deriving a Markov chain Monte Carlo algorithm to estimate the parameters of the model is relatively straightforward. However, this Dirichlet process mixture model ignores the spatial information available through $\mathbf{W}$.

In order to incorporate this spatial information into our clustering procedure we modify the Ewen's distribution in (2) so that partitions that include non-connected clusters receive zero probability a priori. This is done by truncating (2) so that

$$p(\xi \mid \alpha) = \frac{\alpha^{K(\xi)}}{C(\alpha)} \left\{ \prod_{k=1}^{K(\xi)} \Gamma(m_k(\xi)) \right\} \mathbb{I}(Q(\xi) = 0), \tag{4}$$

where

$$Q(\xi) = \sum_{i=1}^{n} \mathbb{I}\left\{ \sum_{i' \neq i} \mathbb{I}(\xi_{i'} = \xi_i) > 0 \right\} \mathbb{I}\left\{ \sum_{i' \in \partial i} \mathbb{I}(\xi_{i'} = \xi_i) = 0 \right\}$$

is the total number of regions such that they are not in a singleton cluster (note that $q_i = \mathbb{I}\left\{\sum_{i' \neq i} \mathbb{I}(\xi_{i'} = \xi_i) > 0\right\}$ equals zero if and only if region i is a singleton) and none of its neighbors belong to the same cluster (note that $z_i = \mathbb{I}\{\sum_{i' \in \partial i} \mathbb{I}(\xi_{i'} = \xi_i) = 0\}$ equals zero if and only if none of its neighbors belong to the same cluster), and

$$C(\alpha) = \sum_{\xi' : Q(\xi') = 0} \alpha^{K(\xi')} \left\{ \prod_{k=1}^{K(\xi')} \Gamma\left(m_k\left(\xi'\right)\right) \right\}$$

is an appropriate normalizing constant. We highlight that $C(\alpha)$ can in principle be written as a polynomial of order K on α, but the exact form of (most of) the poly-

nomial coefficients is generally unknown because of the restriction on the set of partitions that are included in the sum.

Note that, unlike the Ewen's distribution, the truncated prior distribution on ξ described in (4) is not exchangeable. Nonetheless, the full conditional distributions associated with this prior also follow a relatively simple form, simplifying the design of Markov chain Monte Carlo algorithms for posterior inference (see Section 3). This is clearer if we think in terms of the partitions implied by ξ. As with the Dirichlet process, we can find the conditional distribution $\xi_i \mid \xi^{(-i)}$ by removing observation i from the partition $\rho_n = \{S_1, \ldots, S_K\}$ to generate a partition $\rho_{n-1}^{(-i)} = \left\{S_1^{(-i)}, \ldots, S_{K^{(-i)}}^{(-i)}\right\}$, and then reallocating observation i to either one of the existing $K^{(-i)}$ clusters or to a new singleton cluster. (Note that we use the $-i$ superscript to indicate that the i-th observation has been removed). Hence, the prior full conditional probability that observation i is placed in cluster k, $p(\xi_i = k \mid \xi^{(-i)})$, is simply

$$
\begin{aligned}
&p\left(\xi_i = k \mid \xi^{(-i)}, \alpha\right) \propto \\
&\begin{cases}
\frac{\alpha^{K\left(\xi^{(-i)}\right)}}{C(\alpha)} \prod_{k'=1}^{K\left(\xi^{(-i)}\right)} \Gamma\left(m_{k'}\left(\xi^{(-i)}\right) + \mathbb{I}(k'=k)\right) & k \leq K\left(\xi^{(-i)}\right), \sum_{i' \in \partial i} \mathbb{I}(\xi_{i]} = k) > 0, \\
0 & k \leq K\left(\xi^{(-i)}\right), \sum_{i' \in \partial i} \mathbb{I}(\xi_{i]} = k) = 0, \\
\frac{\alpha^{1+K\left(\xi^{(-i)}\right)}}{C(\alpha)} \prod_{k=1}^{K\left(\xi^{(-i)}\right)} \Gamma\left(m_k\left(\xi^{(-i)}\right)\right) & k = K\left(\xi^{(-i)}\right) + 1,
\end{cases}
\end{aligned}
$$

which simplifies to

$$
p(\xi_i = k \mid \xi^{(-i)}, \alpha) \propto \begin{cases}
m_k\left(\xi^{(-i)}\right) & k \leq K\left(\xi^{(-i)}\right), \sum_{i' \in \partial i} \mathbb{I}(\xi_{i'} = k) > 0, \\
0 & k \leq K\left(\xi^{(-i)}\right), \sum_{i' \in \partial i} \mathbb{I}(\xi_{i'} = k) = 0, \\
\alpha & k = K\left(\xi^{(-i)}\right) + 1.
\end{cases} \quad (5)
$$

Note that this expression is very similar to the full conditional distribution in (3), with the main difference being that region has zero probability of being allocated to any non-singleton cluster that are not represented among its neighbors. Hence, we call this prior a restricted Chinese Restaurant process. As in the regular Chinese Restaurant process, α controls the a priori expected number of clusters, with larger values of α favoring an allocation in which all observations are assigned to singleton clusters, and values of α close to zero favoring a single cluster. Because α plays such a critical role in the behavior of the model in the sequel we treat it as an unknown hyperparameter and assign it a Gamma prior distribution with parameters a_α and b_α.

3 Computation

Even with the constraints associated with connected clusters, the number of possible partitions grows very fast with n, making explicit evaluation of the posterior distribution unfeasible in most realistic scenarios. Hence, we focus on developing Markov chain Monte Carlo (MCMC) algorithms Robert and Casella (2005) for posterior inference in our spatial clustering model.

For simplicity we focus on the case where no covariates are available (i.e., $x_i = 1$ for all i) and a Gaussian prior on $\vartheta_k \sim \mathsf{N}(\mu, \sigma^2)$ with hyperpriors $\mu \sim \mathsf{N}(\kappa, \phi^2)$ and $\sigma^2 \sim \mathsf{IGam}(a_\sigma, b_\sigma)$, and fixed hyperparameters κ, ϕ^2, a_σ and b_σ. Firstly, note that the full conditional posterior for ξ_i is

$$p\left(\xi_i = k \mid y_i, \xi^{(-i)}, \alpha, \{\vartheta_k\}, \mu, \sigma^2\right) \propto \begin{cases} m_k\left(\xi^{(-i)}\right) p(y_i \mid \vartheta_k) & k \leq K\left(\xi^{(-i)}\right), \sum_{i' \in \partial i} \mathbb{I}(\xi_{i'} = k) > 0, \\ 0 & k \leq K\left(\xi^{(-i)}\right), \sum_{i' \in \partial i} \mathbb{I}(\xi_{i'} = k) = 0, \\ \alpha p\left(y_i \mid \vartheta_{K\left(\xi^{(-i)}\right)+1}\right) & k = K\left(\xi^{(-i)}\right) + 1, \end{cases}$$

where $\vartheta_{K\left(\xi^{(-i)}\right)+1} \sim N(\mu, \sigma^2)$ and

$$p(y_i \mid \vartheta) = \frac{\exp\{-h_i \exp(\vartheta)\} (h_i \exp(\vartheta))^{y_i}}{y_i!}.$$

Secondly, the parameters $\vartheta_1, \ldots, \vartheta_{K(\xi)}$ are conditionally independent a posteriori with

$$p(\vartheta_k \mid \mathbf{y}, \xi, \mu, \sigma) \propto \exp\left\{ -\frac{\left(\vartheta_k - (\mu + \sigma^2 \sum_{i:\{\xi_i = k\}} y_i)\right)^2}{2\sigma^2} - \exp\{\vartheta_k\} \sum_{i:\{\xi_i = k\}} h_i \right\}.$$

Since this posterior distribution does not belong to a tractable family, we sample ϑ_k using a slice sampler (Damien et al., 1999). The algorithm proceeds by introducing independent exponentially distributed auxiliary variables $u_1, \ldots, u_{K(\xi)}$. The full conditional posterior density for u_k reduces to a truncated exponential distribution

$$p(u_k \mid \vartheta_k) \propto \exp\{-u_k\} \mathbb{I}\left(u_k > \exp\{\vartheta_k\} \sum_{i:\{\xi_i = k\}} h_i\right),$$

while the full conditional posterior for ϑ_k from

$$p(\vartheta_k \mid u_k, \mu, \sigma^2) \propto$$
$$\exp\left\{-\frac{\left(\vartheta_k - (\mu + \sigma^2 \sum_{i:\{\xi_i=k\}} y_i)\right)^2}{2\sigma^2}\right\} \mathbb{I}\left(\vartheta_k < \log u_k - \log \sum_{i:\{\xi_i=k\}} h_i\right).$$

The full conditional posterior for μ which reduces to a normal distribution

$$\mathsf{N}\left(\frac{\sigma^2\kappa + \phi^2 \sum_{k=1}^{K(\xi)} \vartheta_k}{\sigma^2 + \phi^2 K(\xi)}, \frac{\sigma^2\phi^2}{\sigma^2 + \phi^2 K(\xi)}\right)$$

and the full conditional of σ^2 which reduces to an inverse Gamma distribution,

$$\mathsf{IGam}\left(a_\sigma, b_\sigma + \frac{\sum_{k=1}^{K(\xi)} (\vartheta_k - \mu)^2}{2}\right).$$

Finally, we consider sampling the hyperparameter α. From (4) the posterior for α is given by

$$p(\alpha \mid \xi) \propto p(\alpha) \frac{\alpha^{K(\xi)}}{C(\alpha)} \left\{\prod_{k=1}^{K(\xi)} \Gamma\left(m_k(\xi)\right)\right\} \mathbb{I}(Q(\xi) = 0), \tag{6}$$

where $p(\alpha)$ is a Gamma prior with fixed parameters a_α and b_α. This posterior distribution is doubly intractable: not only it does not belong to any well-known family, but it cannot even be evaluated in closed form because the normalizing constant $C(\alpha)$ involves a sum over an exponentially large number of terms. To address this difficulty we use the Noisy Exchange Algorithm (NEA) (Alquier et al., 2016) to allow inference on this doubly intractable distribution.

Our implementation of NEA uses a random walk Metropolis-Hastings algorithm with log-normal proposals, $\log\{\alpha^*\} \sim \mathsf{N}\left(\log\{\alpha\}, \omega^2\right)$, and replaces the ratio of intractable constants $\frac{C(\alpha)}{C(\alpha^*)}$ in the acceptance probability with an unbiased estimator obtained using Bridge Sampling (Gelman and Meng, 1998),

$$\frac{C(\alpha)}{C(\alpha^*)} \approx \frac{1}{N} \sum_{j=1}^{N} \frac{[\alpha]^{K(\xi'_j)}}{[\alpha^*]^{K(\xi'_j)}}.$$

The samples $\xi'_1, \ldots, \xi'_N$ used in this approximation are dependent but approximately identically distributed samples from (4), and they are obtained by running (for each iteration of our algorithm) a second MCMC algorithm based on the full conditionals in (5).

4 Illustration

We applied our model to the U.S. lung cancer mortality data presented in Figure 1. Our analysis uses the following values for the hyperparameters, $\kappa = 5, \phi = 0.1$, $a_\sigma = b_\sigma = 2$ and $a_\alpha = b_\alpha = 1$. Figure 2 shows the prior distribution on the number of clusters implied by our restricted hierarchical model. All estimates are based on 30,000 samples obtained after a burn-in period of 10,000 iterations. When sampling α we used $\omega^2 = 0.25$ as the variance of the random walk (leading to an acceptance rate of roughly 44%), and $N = 100$ samples obtained after a short burnin period of 100 iterations to estimate the ratio of normalizing constants. Convergence was checked by examining trace plots of the log-posterior distribution, the number of clusters represented in the data, and the hyperparameter α. There is no evidence of lack of convergence, although the autocorrelation for some of the parameters is relatively high.

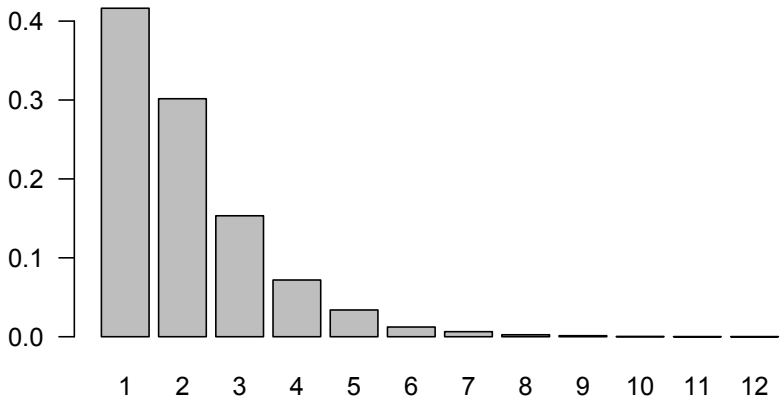


Fig. 2 Prior distribution on the number of unique clusters, $p(\xi) = \int p(\xi \mid \alpha) p(\alpha \mid a_\alpha, b_\alpha)$. Values were obtained by simulation.

Figure 3(a) presents our Monte Carlo estimates of $p(\xi_i = \xi_{i'} \mid \mathbf{y})$, the posterior probability that any pair of states belong to the same cluster (note that states have been reordered to facilitate interpretation). The posterior distribution over partitions is quite concentrated and favors somewhere between 8 and 12 clusters. Some aspects of the partition in which there is substantial uncertainty include whether Idaho should belong to a "southern" cluster together with Utah, Colorado, New Mexico

Texas, or to a "northern" cluster with Wyoming, Wisconsin, the Dakotas, Montana, Nebraska, Minnesota, Kansas and Michigan, whether California and Arizona belong to a cluster with Nevada, Oregon and Washington, or to a cluster of their own, and whether Delaware should be included in a small cluster with Pennsylvania, Ohio and Indiana, or with a bigger one including most of the North West and New England.

The probabilities in 3(a) can be used to find a point estimate of the partition, $\hat{\xi} = (\hat{\xi}_1, \ldots, \hat{\xi}_n)$, by minimizing the expected cost function $\hat{U}(\hat{\xi}) = \mathsf{E}\left\{U(\hat{\xi}, \xi) \mid \mathbf{y}\right\}$, where

$$U(\hat{\xi}, \xi) = \sum_{i=1}^{n} \sum_{i'=i+1}^{n} \eta_1 \mathbb{I}(\xi_i = \xi_{i'}) \mathbb{I}(\hat{\xi}_i \neq \hat{\xi}_{i'}) + \eta_2 \mathbb{I}(\xi_i \neq \xi_{i'}) \mathbb{I}(\hat{\xi}_i = \hat{\xi}_{i'})$$

(see Lau and Green, 2007 for further details). The ratio η_1/η_2 controls the relative cost of separating states that in truth belong to the same cluster and the cost of placing together two states that in truth belong to different clusters. In our analysis we take $\eta_1/\eta_2 = 1$, which yields a point estimate with 9 clusters (see Figure 3(b)). One of those clusters involves West Virginia and Kentucky (two states that we had identified as having particularly high mortality rates for lung cancer), with the rest of the Appalachian region being clustered with other Southern states. Following on our discussion on uncertainty, note that this point estimates sets California and Arizona with Nevada, Oregon and Washington, Idaho with the southern cluster of states, and Delaware with most of the North East and New England.

Finally, Figure 4 presents the posterior means for the mortality rates generated by our model. Although there are similarities with Figure 1, our estimates provide further smoothing of the rates.

5 Discussion

We have introduced a new model for disease clustering based on a restricted Ewen's distribution, and derived a Metropolis-Hastings algorithm to estimate it. Aside for disease clustering, the model can potentially be applied in other settings where clustering of lattice data is desired (e.g., in image segmentation).

There are two extensions of this model that we plan to pursue elsewhere. The first one is to extend the construction to models based on restrictions of other stick-breaking priors (such as the Pitman-Yor process, e.g., see Pitman, 1996). The second one is to consider for general forms for the prior $p(\vartheta_1, \ldots, \vartheta_K \mid \xi)$ that allows us to incorporates the spatial information encoded in $\mathbf{W}$ not only on the partition structure, but also on the coefficients associated with each cluster.

References

Alquier, P., Friel, N., Everitt, R., and Boland, A. (2016). Noisy monte carlo: convergence of markov chains with approximate transition kernels. *Statistics and Computing*, 26(1):29–47.

Anderson, C., Lee, D., and Dean, M. (2013). Identifying clusters in bayesian disease mapping. Technical report, arXiv:1311.0660.

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, **2**:1152–1174.

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

Blackwell, D. and MacQueen, J. B. (1973). Ferguson distribution via Pólya urn schemes. *The Annals of Statistics*, **1**:353–355.

Charras-Garrido, M., Abrial, D., De Goër, J., Dachian, S., and Peyrard, N. (2012). Classification method for disease risk mapping based on discrete hidden markov random fields. *Biostatistics*, 13(2):241–255.

Damien, P., Wakefield, J., and 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(2):331–344.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**:209–230.

Fuentes-García, R., Mena, R. H., and Walker, S. G. (2010). A probability for classification based on the dirichlet process mixture model. *Journal of classification*, 27(3):389–403.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185.

Goujon-Bellec, S., Demoury, C., Guyot-Goubin, A., Hémon, D., Clavel, J., et al. (2011). Detection of clusters of a rare disease over a large territory: performance of cluster detection methods. *International journal of health geographics*, 10(1):53.

Green, P. J. and Richardson, S. (2002). Hidden markov models and disease mapping. *Journal of the American statistical association*, 97(460):1055–1070.

Heinzl, F. and Tutz, G. (2014). Clustering in linear-mixed models with a group fused lasso penalty. *Biometrical Journal*, 56(1):44–68.

Knorr-Held, L. and Raßer, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56(1):13–21.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and methods*, 26(6):1481–1496.

Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine*, 14(8):799–810.

Kulldorff, M., Tango, T., and Park, P. J. (2003). Power comparisons for disease clustering tests. *Computational Statistics & Data Analysis*, 42(4):665–684.

Lau, J. W. and Green, P. J. (2007). Bayesian Model-Based clustering procedures. *Journal of Computational and Graphical Statistics*, 16(3):526–558.

Lee, J., Quintana, F. A., Müller, P., and Trippa, L. (2013). Defining predictive probability functions for species sampling models. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 28(2):209.

Martínez, A. F., Mena, R. H., et al. (2014). On a nonparametric change point detection model in markovian regimes. *Bayesian Analysis*, 9(4):823–858.

Moraga, P. and Montes, F. (2011). Detection of spatial disease clusters with lisa functions. *Statistics in medicine*, 30(10):1057–1071.

Morton-Jones, T., Diggle, P., and Elliott, P. (1999). Investigation of excess environmental risk around putative sources: Stone's test with covariate adjustment. *Statistics in medicine*, 18(2):189–197.

Openshaw, S., Charlton, M., Wymer, C., and Craft, A. (1987). A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information System*, 1(4):335–358.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. In Ferguson, T. S., Shapeley, L. S., and MacQueen, J. B., editors, *Statistics, Probability and Game Theory. Papers in Honor of David Blackwell*, pages 245–268. Hayward, CA:IMS.

Potthoff, R. F. and Whittinghill, M. (1966a). Testing for homogeneity: I. the binomial and multinomial distributions. *Biometrika*, 53(1-2):167–182.

Potthoff, R. F. and Whittinghill, M. (1966b). Testing for homogeneity: Ii. the poisson distribution. *Biometrika*, pages 183–190.

Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods*. Springer, New York, second edition edition.

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

Stone, R. A. (1988). Investigations of excess environmental risks around putative sources: statistical problems and a proposed test. *Statistics in medicine*, 7(6):649–660.

Tango, T. (1995). A class of tests for detecting 'general'and 'focused'clustering of rare diseases. *Statistics in Medicine*, 14(21-22):2323–2334.

Tango, T. and Takahashi, K. (2005). A flexibly shaped spatial scan statistic for detecting clusters. *International journal of health geographics*, 4(1):11.

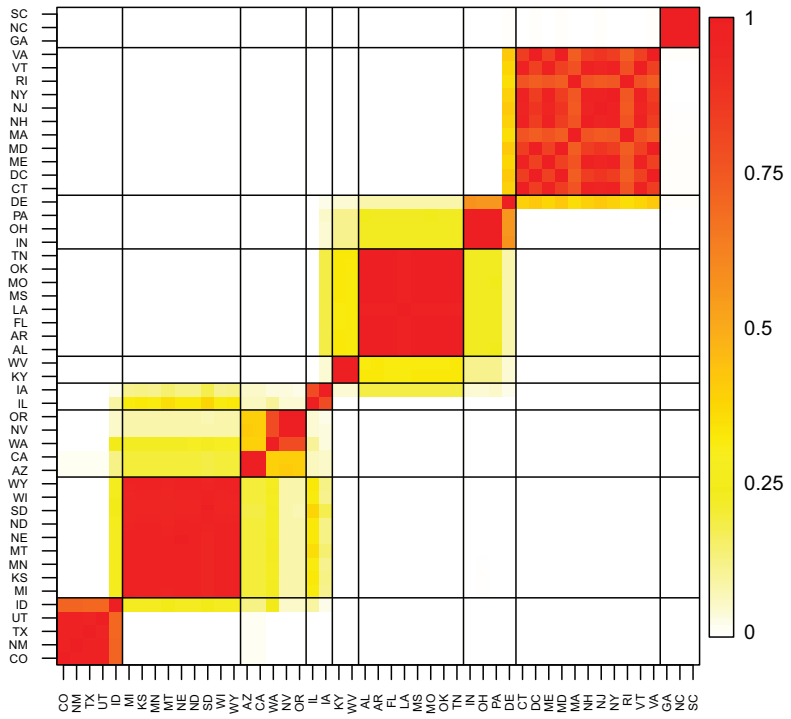
Wakefield, J. and Kim, A. (2013). A bayesian model for cluster detection. *Biostatistics*, 14(4):752–765.

Waller, L. A., Hill, E. G., and Rudd, R. A. (2006). The geography of power: statistical performance of tests of clusters and clustering in heterogeneous populations. *Statistics in Medicine*, 25(5):853–865.

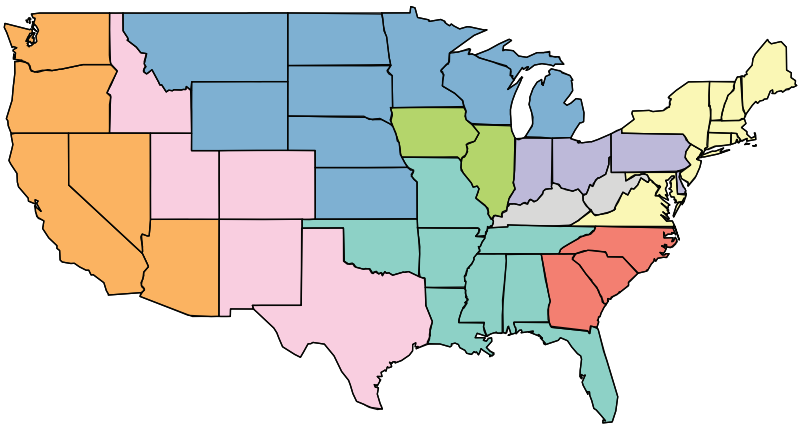
Wang, H. and Rodríguez, A. (2014). Identifying pediatric cancer clusters in florida using log-linear models and generalized lasso penalties. *Statistics and Public Policy*, 1(1):86–96.

Weinstock, M. A. (1981). A generalised scan statistic test for the detection of clusters. *International Journal of Epidemiology*, 10(3):289–293.

Whittemore, A. S., Friend, N., Brown, B. W., and Holly, E. A. (1987). A test to detect clusters of disease. *Biometrika*, 74(3):631–635.



(a) Posterior probability that two states are assigned to the same cluster



(b) Point estimate of the cluster structure

Fig. 3 Posterior estimates of the cluster structure associated withe U.S. lung cancer mortality data. Vertical lines in panel (a) correspond to the point estimate presented in panel (b).

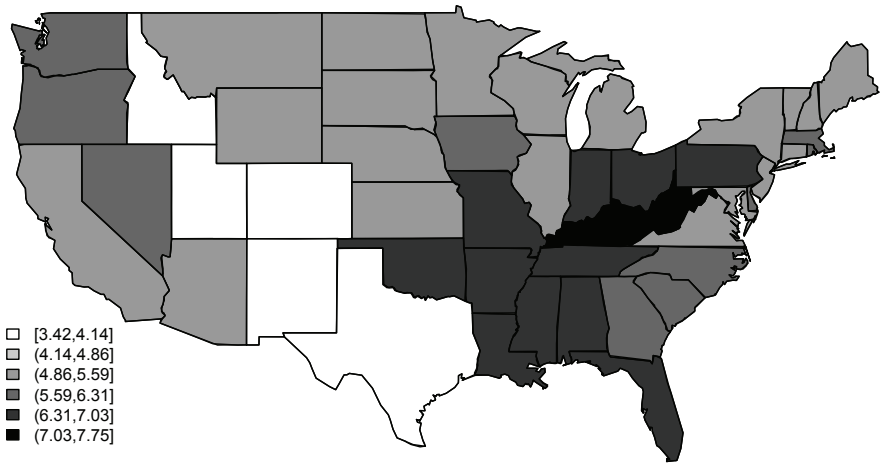


Fig. 4 Posterior means for the disease rates, $\exp\{\theta_i\}$, for the U.S. lung cancer mortality data.