Bayesian cluster detection via adjacency modelling

Craig Anderson, Duncan Lee, Nema Dean

Abstract

Disease mapping aims to estimate the spatial pattern in disease risk across an area, identifying units which have elevated disease risk. Existing methods use Bayesian hierarchical models with spatially smooth conditional autoregressive priors to estimate risk, but these methods are unable to identify the geographical extent of spatially contiguous high-risk clusters of areal units. Our proposed solution to this problem is a two stage approach, which produces a set of potential cluster structures for the data and then chooses the optimal structure via a Bayesian hierarchical model. The first stage uses a spatially adjusted hierarchical agglomerative clustering algorithm. The second stage fits a Poisson log-linear model to the data to estimate the optimal cluster structure and the spatial pattern in disease risk. The methodology was applied to a study of chronic obstructive pulmonary disease (COPD) in local authorities in England, where a number of high risk clusters were identified..

1 Introduction

Disease risk varies geographically as a result of many factors, including differences in environmental exposures, and cultural and behavioural differences between the inhabitants of different areas. Within a country such as England, there are substantial inequalities in terms of health and disease risk, with poverty being one of the most important reasons for these differences (Marmot et al. (2010)) Disease maps allow us to illustrate these differences graphically. Such maps are produced by partitioning the study region into n non-overlapping areal units such as electoral wards or census tracts, and then calculating the overall risk of disease for the population living in each areal unit. Health agencies routinely produce such maps for numerous diseases, including cancer (Public Health England (2010)) and cardiovascular disease (Centers for Disease Control and Prevention (2011)). The main use of these maps is that they allow public health officials to visually identify high-risk clusters of areal units, allowing them to focus resources on those areas exhibiting elevated disease risk.

Many different approaches have been proposed for the identification of the spatial extent of high-risk clusters in spatial disease maps, including Bayesian hierarchical modelling (Charras-Garrido et al. (2012)), scan statistics (Kulldorff (1997)) and point process methodology (Diggle et al. (2005)). The first of these is typically based on a Poisson log-linear model, where covariates and/or a set of random effects are used to represent the spatial disease risk pattern. The random effects are included to account for spatial autocorrelation in the response that was not captured by the covariates; and are typically modelled by a conditional autoregressive (CAR) prior. These priors were proposed by Besag et al. (1991) and developed by Leroux et al. (1999), and are a type of Gaussian Markov random field. CAR priors make the naive assumption of global correlation between all pairs of random effects in geographically adjacent areal units, and therefore produce a spatially smooth risk surface. However such smoothing is detrimental to our main aim, which is to identify groups of areas which have much higher (or lower) risks compared with surrounding areas, so an alternative approach is required.

Therefore, this paper outlines new methodology which allows for the estimation of the spatial pattern in disease risk, whilst simultaneously detecting the spatial extent of high or low risk clusters. In doing so the cluster struc-

ture is accounted for when estimating disease risk, so that high risk clusters are not smoothed towards their geographical neighbours that do not exhibit elevated risks. The methodology brings together hierarchical agglomerative clustering techniques and conditional autoregressive models in a two-stage approach. The first stage is a spatially-adjusted hierarchical agglomerative clustering algorithm first proposed in Anderson et al. (2014), which respects the spatial contiguity of the study region. This algorithm is applied to disease data preceding the study period to elicit n candidate cluster configurations containing between 1 and n clusters. The second stage fits an extended Poisson log-linear model to the study data, where Markov Chain Monte Carlo (MCMC) simulation methods are used to estimate both the optimal cluster structure and disease risk.

Applying the clustering algorithm to the study data itself would necessitate the information in the data being used twice, once for eliciting a set of candidate cluster configurations and again for estimating the model parameters. To overcome this issue, a second data set is required for the clustering stage, and emphasis should be placed on obtaining a dataset which is as similar as possible to the study data. Possible choices include data on disease risk in the time period prior to the study period or data on a different disease from the same time period as the study data. This study utilises the former choice, because it is unlikely that there has been any substantial change in the spatial patterns in the population characteristics governing disease risk (such as poverty) unless substantial urban regeneration has taken place. The approach proposed in this paper is thus appropriate for data on chronic diseases whose risk factors are spatially stable, but would be unsuitable for epidemic diseases such as influenza, where the spatial pattern in disease risk in the years prior to an outbreak would be vastly different to the pattern during an outbreak.

The remainder of this paper is organised as follows. Section 2 gives a brief introduction to Bayesian disease mapping, and discusses the existing methods of cluster identification that have been proposed in this context. Section 3 proposes our new methodological extension, while Section 4 establishes its efficacy via simulation. Section 5 presents the motivating application for our methodology, a study of chronic obstructive pulmonary disease (COPD) mortalities in English local authorities in 2010. Finally, Section 6 discusses the implications of this paper and ideas for future work.

2 Bayesian disease mapping

2.1 Study Design and Modelling

The study region $\mathcal{A}$ is partitioned into n non-overlapping areal units $\mathcal{A} = \{\mathcal{A}_1, \ldots, \mathcal{A}_n\}$, and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ and $\boldsymbol{E} = (E_1, \ldots, E_n)$ represent the observed and expected numbers of disease cases in each unit during the study period. The latter are constructed by external standardisation, based on the age and sex demographics of the population living in each areal unit. A Poisson log-linear model is commonly used to estimate disease risk, and a general form is given by

$$\begin{aligned} Y_i|E_i, R_i &\sim \text{Poisson}(E_i R_i) \qquad i = 1, ..., n, \\ \ln(R_i) &= \mathbf{x}_i^T \boldsymbol{\beta} + \phi_i. \end{aligned} \tag{1}$$

Here R_i represents disease risk in areal unit $\mathcal{A}_i$, and is modelled by a vector of covariates $\mathbf{x}_i^T = (1, x_{i1}, \ldots, x_{ip})$, with coefficients $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$, and a random effect ϕ_i. The random effects $\boldsymbol{\phi} = (\phi_i, \ldots, \phi_n)$ account for the spatial autocorrelation induced into the disease data by factors such as unmeasured confounding, neighbourhood effects and grouping effects. They are modelled by a conditional autoregressive (CAR) prior, which induces spatial autocorrelation via a binary neighbourhood matrix W, where $w_{ij} = 1$ if areal units $(\mathcal{A}_i, \mathcal{A}_j)$ share a common border (denoted $i \sim j$) and $w_{ij} = 0$ otherwise. Note that $w_{ii} = 0$ for all i. CAR priors can be specified as a set of n univariate conditional distributions $f(\phi_i|\boldsymbol{\phi}_{-i})$, where $\boldsymbol{\phi}_{-i} = (\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_n)$. The simplest of these CAR priors was the intrinsic prior proposed by Besag et al. (1991), and this model is given by

$$\phi_i|\boldsymbol{\phi}_{-i} \sim \text{N}\left(\frac{\sum_{j=1}^n w_{ij}\phi_j}{\sum_{j=1}^n w_{ij}}, \frac{1}{\tau(\sum_{j=1}^n w_{ij})}\right) \; i = 1, \ldots, n, \tag{2}$$

where τ is a conditional precision parameter. The conditional expectation of ϕ_i is the mean of the random effects in neighbouring areal units, while the variance is inversely proportional to the number of neighbouring units. This set of conditional distributions correspond to a multivariate Gaussian distribution, with mean zero but an improper precision matrix given by $Q = \text{diag}(W\mathbf{1}) - W$, where $W\mathbf{1}$ is a vector containing the number of neighbours for each areal unit.

2.2 Literature review

Previous research has proposed a number of extensions of the Bayesian hierarchical model outlined above to identify the spatial extent of high or low risk clusters in disease maps. The majority of these treat the elements of the neighbourhood matrix $\{w_{ij}|i \sim j\}$ corresponding to adjacent areas as binary random quantities, where estimating $w_{ij} = 0$ corresponds to identifying a boundary between $(\mathcal{A}_i, \mathcal{A}_j)$ because (ϕ_i, ϕ_j) are conditionally independent and are not smoothed over in the modelling process. One of the first examples of this approach came from Lu et al. (2007), who proposed a logistic regression model for $\{w_{ij}|i \sim j\}$ using a measure of dissimilarity between $(\mathcal{A}_i, \mathcal{A}_j)$ as the covariate. However, this results in an excessively large number of parameters, which led Lee and Mitchell (2012) to treat $\{w_{ij}|i \sim j\}$ as a deterministic function of a small number of parameters and the areal level measure of dissimilarity. The same authors (Lee and Mitchell (2013)) also proposed iteratively re-estimating $\{w_{ij}|i \sim j\}$ and the remaining model parameters conditional on the other until a convergence criterion is reached, where $\{w_{ij}|i \sim j\}$ was updated deterministically based on the other model parameters. Finally, Li et al. (2011) fitted multiple models with different W specifications and thus different potential sets of boundaries to the data, and used the Bayesian Information Criterion (BIC) to choose the best model.

The approaches developed in the above literature produce open boundaries, which are a set of potentially disjoint boundary segments that do not necessarily enclose an areal unit or group of units. In contrast, the aim here is to identify distinct groups of areal units that exhibit substantially different risks compared to their neighbours, and this has the natural consequence that the boundary surrounding them is closed. One of the first and still most widely used cluster detection approaches are scan statistics (Kulldorff (1997)), which identify clusters of areal units that exhibit an elevated risk of disease. Their popularity is in part down to the availability of the SaTScan software, which allows the approach to be easily implemented by others. However, scan statistics merely identify high-risk clusters, and do not simultaneously estimate the spatial pattern in disease risk. This has led to a number of hierarchical modelling approaches such as Knorr-Held (2000), Green and Richardson (2002), Charras-Garrido et al. (2012) and Wakefield and Kim (2013) being proposed, and a first comparison to scan statistics is given by Charras-Garrido et al. (2013). This paper, together with Charras-

Garrido et al. (2012), also assesses the utility of identifying clusters by applying a post processing clustering algorithm to an estimated disease risk map, although the spatial contiguity of the clusters is not guaranteed. Finally, a two-stage approach was proposed by Anderson et al. (2014), where a set of potential cluster structures are identified in the first stage, and then a separate Bayesian hierarchical model is fitted to each of these in turn in stage two, with the best structure being chosen using model comparison techniques. This paper will follow a similar two-stage approach, but will simultaneously estimate the cluster structure and disease risk in a single model. This has the advantages of computational simplicity (only fitting a single model), and correctly allowing the uncertainty in the cluster structure to be incorporated when estimating disease risk.

One major difference between the existing approaches is that Knorr-Held (2000), Wakefield and Kim (2013) and Anderson et al. (2014) force the clusters to be spatially contiguous, while Green and Richardson (2002) and Charras-Garrido et al. (2012) do not. However, in all cases except for Anderson et al. (2014) disease risk is assumed to be constant within a cluster, which allows the relative risk to be partitioned into risk classes/clusters which are easy to interpret for epidemiologists. However, for real data it is likely that disease risk varies within a cluster, and the model proposed here allows for such within cluster variation. The other disadvantage of these approaches is that they involve computationally complex estimation approaches, such as reversible jump Markov chain Monte Carlo algorithms (e.g. Knorr-Held (2000)) or the Monte Carlo Expectation-Maximisation algorithm (e.g. Charras-Garrido et al. (2012)). Such approaches are beyond the scope of most epidemiologists, and no publicly available software exists to allow others to implement them.

3 Method

We propose a two-stage approach for estimating the spatial pattern in disease risk and identifying spatially contiguous clusters that exhibit either elevated or reduced disease risks. In the first stage (Section 3.1) we utilise the spatially adjusted hierarchical agglomerative clustering algorithm proposed by Anderson et al. (2014), and use it to elicit a set of candidate cluster configurations for the data. In the second stage (Section 3.2) we propose a

hierarchical Bayesian model for the disease data which can simultaneously select the optimal cluster configuration from the candidates elicited in Stage 1 and also estimate disease risk. Inference for this model uses MCMC simulation and software to implement both Stage 1 and Stage 2 are provided as supplementary material.

3.1 Stage 1 - Eliciting cluster configurations using hierarchical agglomerative clustering

The method of clustering (for details see Hastie et al. (2001)) involves grouping together objects that are similar whilst separating those that are different, which is appropriate here because we wish to identify groups of areal units with similar disease risks. The clustering algorithm is taken from Anderson et al. (2014) and is applied to disease data preceding the study period, because it is likely to exhibit a similar spatial risk pattern to the study data unless substantial urban regeneration has taken place. If this assumption does not hold, either because of substantial socio-economic changes or vast population migration, then alternative clustering data such as covariate information could be used instead.

Let $(\boldsymbol{Y}^{(1)}, \boldsymbol{E}^{(1)}), \ldots, (\boldsymbol{Y}^{(q)}, \boldsymbol{E}^{(q)})$ denote the observed and expected disease counts for the q time intervals (usually years) preceding the study period. These earlier data are used to elicit a set of n potential cluster configurations for the study data, which are denoted here by $\{\mathcal{C}_1, \ldots, \mathcal{C}_n\}$. Here $\mathcal{C}_k = \{\mathcal{C}_k(1), \ldots, \mathcal{C}_k(k)\}$ partitions the n areal units $\mathcal{A} = \{\mathcal{A}_1, \ldots, \mathcal{A}_n\}$ into k spatially contiguous groups, where $\mathcal{C}_k(j)$ is the jth cluster. The set of all possible spatially contiguous cluster configurations for the study region $\mathcal{A}$ is very large, so we use this clustering step to vastly reduce the number of potential cluster structures to be considered in stage 2.

The data are clustered on the log standardised incidence ratio scale, that is $\ln(\boldsymbol{Y}^{(j)}/\boldsymbol{E}^{(j)})$, because this corresponds to the linear predictor scale in (1). Let $\boldsymbol{\psi} = [\ln(\boldsymbol{Y}^{(1)}/\boldsymbol{E}^{(1)}), \ldots, \ln(\boldsymbol{Y}^{(q)}/\boldsymbol{E}^{(q)})]$ be the $n \times q$ matrix whose columns comprise $\ln(\boldsymbol{Y}^{(j)}/\boldsymbol{E}^{(j)})$ for $j = 1, \ldots, q$, and denote the ith row by $\boldsymbol{\psi}_i = [\ln(Y_i^{(1)}/E_i^{(1)}), \ldots, \ln(Y_i^{(q)}/E_i^{(q)})]$, the vector of q values for areal unit $\mathcal{A}_i$. The data are clustered using a modified hierarchical agglomerative clustering algorithm, which initially considers each data point as its own

singleton cluster, and then joins together the two least dissimilar clusters at each stage to form a larger cluster. This process is repeated until only one cluster containing all data points remains. For a configuration with k clusters the dissimilarity, d_{ij}, between clusters i ($\mathcal{C}_k(i)$) and j ($\mathcal{C}_k(j)$) can be measured by a number of metrics called linkage methods, but based on the results of Anderson et al. (2014) we will proceed with centroid linkage in this paper.

Centroid linkage measures the dissimilarity as the Euclidean distance between the average of the two clusters, that is $d_{ij} = ||\bar{\mathcal{C}}_k(i) - \bar{\mathcal{C}}_k(j)||$, where $\bar{\mathcal{C}}_k(i) = (1/n_i) \sum_{f:\mathcal{A}_f \in \mathcal{C}_k(i)} \boldsymbol{\psi}_f$, and n_i is the number of areal units in cluster $\mathcal{C}_k(i)$. The hierarchical agglomerative clustering algorithm described above is extended so that it produces spatially contiguous clusters, which is achieved by only allowing clusters containing two areal units which share a common border to be merged at each step. The algorithm produces a set of candidate cluster structures $\{\mathcal{C}_1, \ldots, \mathcal{C}_n\}$ as follows:

Algorithm

1. Construct $\mathcal{C}_n = \{\mathcal{C}_n(1), \ldots, \mathcal{C}_n(n)\}$, an initial cluster structure where each areal unit is in its own singleton cluster.

2. Repeat the following steps for $h = n, \ldots, 2$, where step h produces $\mathcal{C}_{h-1}$ from $\mathcal{C}_h$.

   (a) Compute the $h \times h$ distance matrix D, whose klth element is given by

   $$D_{kl} = \begin{cases} |d_{kl}| & \text{if } k \sim l \,\&\, k > l \\ \infty & \text{otherwise,} \end{cases}$$

   where d_{kl} is the distance between clusters $(\mathcal{C}_h(k), \mathcal{C}_h(l))$ as measured by centroid linkage, and $k \sim l$ means that the clusters contain at least one pair of areas that share a common border.

   (b) Set $\{i, j\} = \arg\min(D_{kl})$, that is the identifiers of the two clusters that have the minimum dissimilarity. In case of ties, $\{i, j\}$ is randomly selected from these.

(c) Compute

$$\mathcal{C}_{h-1} = \{\mathcal{C}_h(1), \ldots, \mathcal{C}_h(i-1), \mathcal{C}_{h-1}(i), \mathcal{C}_h(i+1), \ldots, \mathcal{C}_h(j-1), \mathcal{C}_h(j+1), \ldots, \mathcal{C}_h(h)\},$$

where $\mathcal{C}_{h-1}(i) = \mathcal{C}_h(i) \cup \mathcal{C}_h(j)$.

3.2 Stage 2 - A model for estimating the cluster structure and disease risk

The study data are denoted by $(\boldsymbol{Y}, \boldsymbol{E})$, and the best cluster structure for these data from the set of n candidates $\{\mathcal{C}_1, \ldots, \mathcal{C}_n\}$ elicited from stage 1 is estimated together with disease risk by extending the Poisson log-linear CAR model given by (1) and (2) in two main ways. This approach takes advantage of the natural ordering of the cluster structures to allow the number of clusters to be considered as a univariate parameter within the model. The mechanism for implementing a given cluster structure is the neighbourhood matrix W, which is altered so that w_{ij} only equals one if areal units $(\mathcal{A}_i, \mathcal{A}_j)$ share a border and are in the same cluster. Thus if two adjacent areal units are in the same cluster their random effects are partially autocorrelated and are smoothed over in the modeling, while if they are in different clusters they are conditionally independent and are not smoothed over. Thus there is a one-to-one relationship between the number of clusters and the value of W, and the n candidate values of W are denoted by $(W_1, \ldots, W_n)$. Here W_1 corresponds to a single cluster and thus equals W, the original adjacency structure of the region. This value thus enforces strong spatial smoothing across the region, as no high or low risk clusters have been identified. In contrast, W_n corresponds to all n areal units being assigned to their own cluster of size one, and thus W_n is the zero matrix. This value thus corresponds to independent random effects with no spatial smoothing constraints.

However, the intrinsic CAR prior outlined in (2) is not appropriate here, since our model could produce a neighbourhood matrix W in which an areal unit has no neighbours due to it being a singleton cluster. If this was areal unit i, this would cause $\sum_{j=1}^{n} w_{ij} = 0$, yielding an infinite mean and variance in (2). Instead, our random effects are modelled via the localised CAR model outlined in Lee et al. (2014), where an extended random effects vector $\widetilde{\boldsymbol{\phi}} = (\boldsymbol{\phi}, \boldsymbol{\phi}^*)$ is used, with $\boldsymbol{\phi}^*$ being the global random effect which is potentially common to all areas and prevents the infinite mean and variance problem

outlined above. An extended (n+1) × (n+1) neighbourhood matrix $\widetilde{W}$ is specified for this vector, which takes the form

$$\widetilde{W} = \begin{pmatrix} W & \boldsymbol{w}_* \\ \boldsymbol{w}_*^T & 0 \end{pmatrix}$$

where $\boldsymbol{w}_* = (w_1*, \ldots, w_n*)$ and $w_i* = I[\sum_{i\sim j}(1 - w_{ij}) > 0]$. Here, $I[.]$ denotes an indicator function, which sets $w_i* = 1$ if any entry in row i of the neighbourhood matrix W is changed from a 1 to a 0 due to a neighbouring area being in a different cluster. Otherwise, $w_i^* = 0$. Based on this extended neighbourhood matrix, $\widetilde{\boldsymbol{\phi}}$ is modelled as $\widetilde{\boldsymbol{\phi}} = \mathrm{N}(\mathbf{0}, \tau^2 Q(\widetilde{W}, \epsilon)^{-1})$ with the precision matrix

$$Q(\widetilde{W}, \epsilon)^{-1} = \mathrm{diag}(\widetilde{W}1) - \widetilde{W} + \epsilon 1. \tag{3}$$

This corresponds to the intrinsic CAR model for the extended random effects vector $\widetilde{\phi}$, with a small positive constant added to the diagonal of the precision matrix to ensure that it is invertible. The invertibility of $Q(\widetilde{W}, \epsilon)^{-1}$ is required as its determinant is computed when updating W, and Lee and Mitchell (2013) suggest that the results are insensitive to ϵ and set $\epsilon = 0.001$. The full conditionals of this extended CAR model are given by

$$\begin{aligned} \phi_i|\widetilde{\boldsymbol{\phi}}_{-i} &\sim \mathrm{N}\left(\frac{\sum_{j=1}^n w_{ij}\phi_j + w_{i*}\phi_*}{\sum_{j=1}^n w_{ij} + w_{i*} + \epsilon}, \frac{\tau^2}{\sum_{j=1}^n w_{ij} + w_{i*} + \epsilon}\right), \\ \phi_*|\widetilde{\boldsymbol{\phi}}_{-*} &\sim \mathrm{N}\left(\frac{\sum_{j=1}^n w_{j*}\phi_j}{\sum_{j=1}^n w_{j*} + \epsilon}, \frac{\tau^2}{\sum_{j=1}^n w_{j*} + \epsilon}\right). \end{aligned} \tag{4}$$

This means that the conditional expectation is a weighted average of the random effects in neighbouring areas and the global random effect ϕ^*, with binary weights based on the current choice of W matrix. Here, $\widetilde{\boldsymbol{W}} = (\widetilde{W}_1, \ldots, \widetilde{W}_n)$ is the set of extended neighbourhood matrices related to the set of cluster structures elicited in (3.1), where $\widetilde{W}_j$ is the matrix corresponding to the cluster structure with j clusters. Given this extended CAR prior the overall Bayesian hierarchical model we propose is given by

$$\begin{aligned} Y_i|E_i, R_i &\sim \mathrm{Poisson}(E_i R_i) \text{ for } i = 1, \ldots, n, \\ \ln(R_i) &= \beta_0 + \phi_i, \end{aligned}$$

$$
\begin{aligned}
\widetilde{\boldsymbol{\phi}} &\sim N(\mathbf{0}, \tau^2 Q(\widetilde{W}, \epsilon)^{-1}), \\
\widetilde{W} &\sim \text{Discrete}(\widetilde{W}_1, \ldots, \widetilde{W}_n; \pi_1, \ldots, \pi_n), \\
\pi_j &= \frac{\exp(-j\theta)}{\sum_{i=1}^{n} \exp(-i\theta)}, \\
\beta_0 &\sim N(0, 1000) \text{ for } j = 1, \ldots, p, \\
\theta &\sim \text{Uniform}(0, 1), \\
\tau^2 &\sim \text{Uniform}(0, 1000).
\end{aligned} \tag{5}
$$

Initially, a discrete uniform prior was considered for $\widetilde{W}$, but it may not be appropriate to give equal weighting to structures with extremely large numbers of clusters as the spatial autocorrelation present in the data suggests the number of clusters will be relatively small. Therefore our prior probabilities for $(\widetilde{W_1}, \ldots, \widetilde{W_n})$ are given by $(\pi_1, \ldots, \pi_n)$, with an additional parameter θ being introduced to control the strength of the weights. When $\theta = 0$ a discrete uniform prior is assumed for $\widetilde{W}$, while $\theta = 1$ corresponds to a scaled exponential weighting which gives larger prior weight to values of W corresponding to fewer clusters.

Inference for this model is carried out using a Markov-Chain Monte Carlo (MCMC) algorithm, which produces posterior distributions for each of the model parameters. The estimated number of clusters in the data should be chosen as the average value from the posterior distribution of $\widetilde{W}$, with uncertainty estimated via a 95% credible interval. Here we will select the number of clusters using the posterior mode, because it is the most commonly occuring cluster structure in the MCMC algorithm, but the median could also be used. The mean would not be sensible because the number of clusters follows a discrete distribution and requires an integer value.

4 Simulation study

A simulation study was conducted to establish the efficacy of the two-stage modelling approach outlined in the previous section. The template for the study was based on the set of 324 local authorities in England, which is also the study region for the motivating application presented in Section 5. A study was conducted comparing the two-stage approach proposed here with existing alternatives, and the results are summarised below.

4.1 Data Generation

Clustered disease data were generated according to the template shown in Figure 1. The template consists of 15 clusters of different sizes, which include one large cluster shaded in light grey and 14 smaller clusters shaded in either white (low risk clusters) or dark grey (high risk clusters), some of which are singletons. Disease data were generated under this template from model (1), with the simplification that no covariates were included. The random effects were generated from a multivariate Gaussian distribution with a spatially correlated precision matrix defined by the CAR model proposed by Leroux et al. (1999). The intrinsic model (2) is not used for the data generation because its precision matrix is singular. Clustered disease data were obtained by specifying a piecewise constant mean function for $\boldsymbol{\phi}$, which follows the template shown in Figure 1. The values in Figure 1 are multiplied by C, with larger values of C representing larger differences between the clusters, which should thus be easier to identify. Values of $C = 0, 0.5, 1$ are used in this study, with $C = 0$ corresponding to a spatially smooth risk surface which is equivalent to having a single cluster covering the entire study region. For the analyses described in this section the expected disease counts are set equal to 100 for each area.

Each of the simulated data sets consist of the study data plus three sets of "prior" data, with the "prior" data being used for the clustering step. To allow for the fact that the log risk surfaces for the study and prior data sets are unlikely to be identical, uniform random noise was added to the random effects from the three prior data sets, which corresponds to multiplicative random noise on the risk scale. To provide a suitable analogue with real data across three different years, different levels of noise were added to the three prior data sets, with larger noise added to the data which were further away in time. This noise is added to account for variation from year to year in the real data, with the assumption that these differences are larger as the time differences increase. The uniform random noise for the three prior data sets were on the following intervals $[-0.05, 0.05]$, $[-0.1, 0.1]$ and $[-0.15, 0.15]$, and were chosen to approximate the correlations between the study and prior data sets in the motivating application in Section 5. 200 datasets were generated for each of the three scenarios ($C = 0, 0.5, 1$), and the model proposed here was compared against two alternatives. The first was the proposal of Anderson et al. (2014) which used fixed effects to model the clusters, and

the second was the Besag-York-Mollié (Besag et al. (1991)) model, hereafter known as BYM, which is commonly used in disease mapping. In the case of the BYM model, the posterior classification approach described in Charras-Garrido et al. (2012) and Charras-Garrido et al. (2013) was implemented to identify the clusters, which is based on model-based clustering with Gaussian mixtures (Fraley et al. (2012)). However, this approach does not produce spatially contiguous clusters, so a further post-processing step was implemented to partition the clusters identified into spatially contiguous groups. We note that we have not compared our approach to a method such as Knorr-Held (2000) or Charras-Garrido et al. (2012), because software to implement these complex estimation methods is not publicly available.

4.2 Results

The results of the study are summarised in Figure 2, which displays a comparison of the relative performances of our approach and the two alternatives using three different metrics. The accuracy of the risk surfaces estimated by each approach is quantified by their root mean square error (RMSE), while the correctness of the estimated cluster structures is quantified by both the number of clusters identified and the Rand Index (Rand (1971)) between the true and estimated cluster structures. The latter is a measure of the similarity between two cluster structures and lies in the interval $[0, 1]$. It is computed as the proportion of pairs of areal units classified either in the same or in different clusters by both methods, that is the proportion of pairwise agreements between the two methods. A value of one indicates complete agreement between the two cluster configurations, while a value of zero indicates that no pair of areal units are classified in the same way under both configurations.

The top panel of Figure 2 shows boxplots of the numbers of clusters estimated by each method in the 200 simulated data sets, where the true values of 1 (when $C = 0$) and 15 (when $C = 0.5, 1$) are represented by dashed lines. The middle panel displays boxplots of the Rand index for all simulated data sets, while the bottom panel shows the RMSE values for the estimated risk surface.

The top panel shows that when $C = 0$ all three methods estimate the correct number of clusters on average, but our method has the lowest stan-

dard deviation with a value of 0.25 compared with 1.39 for the fixed effect model and 5.06 for the BYM. When $C = 0.5$ both our model and the random effects model estimate the correct number of clusters, while the BYM overestimates the number of clusters, with a median of 17 clusters observed. Again, our model provides the most precise estimates, with a standard deviation of 1.45 compared to 3.15 for the fixed effect approach. When $C = 1$, all three models estimate the correct number of clusters on average, but again the model proposed here has the lowest standard deviation (0.25) compared with the fixed effect (1.30) and BYM (2.41) approaches. This suggests that our model estimate the correct number of clusters with less error than the other methods.

A median Rand Index of 1 is obtained for all three models when $C = 0$ or $C = 1$, while when $C = 0.5$ we obtained a medians of 0.983 for the BYM model and 1 for both our model and the fixed effect model. Our model has the best standard deviation for $C = 0$, with a value of 0.0018 compared to 0.0092 for the fixed effect and 0.0655 for the BYM. For $C = 0.5$ and $C = 1$, our model outperforms the BYM in terms of standard deviation, but has higher values than the fixed effect model; for $C = 0.5$, the fixed effect model has a value of 0.0069 compared to 0.0512 for our approach, and for $C = 1$ the fixed effect approach has a value of 0.0009 compared to 0.0024 for our model. Under each scenario, our model obtained the correct cluster structure more often than either of the other approaches; for $C = 0$ we obtained a Rand Index of 1 on 187 occasions, compared to 161 for the fixed effect and 185 for the BYM. For $C = 0.5$, we obtained the correct cluster structure on 142 occasions, compared to 122 for the fixed effect and 1 for the BYM, and when $C = 1$ we correctly identified the cluster structure on 190 occasions, compared to 149 for the fixed effect approach and 185 for the BYM. It therefore appears that our approach performs best of the three in terms of correctly identifying the true cluster structure.

Finally, the bottom panel shows that the model proposed here performs the best of the three in terms of RMSE when $C = 0$, with a median of 0.042 compared with 0.051 and 0.074 for the fixed effects and BYM approaches respectively, but performs poorest of the three for RMSE when $C = 1$, with a median of 0.161 compared with 0.066 and 0.105. It seems likely that our model performs worse than the BYM model in this respect because our model has a single set of random effects, while the BYM model has two sets to share

the modelling burden in this extreme case and thus the independent random effects are able to capture the jumps in risk between clusters. It appears that our model can provide the best model fit when $C = 0$ and there is no underlying clustering present, which means we are less likely than the other methods to identify a 'false positive' result. In the cases where $C = 0.5$ and $C = 1$ we are able to do a better job than the other methods at estimating the cluster structures, but the model fit is slightly affected in these cases.

5 Motivating application

5.1 Study design

The study region is the country of England, which is the largest of the four constituent nations of the United Kingdom and has a population of approximately 53 million people. The country is divided into $n = 324$ local authorities, containing populations of between 7338 and 1061074 people with a median value of 124781. The disease data are the numbers of mortalities with a primary diagnosis of chronic obstructive pulmonary disease (COPD) in each local authority in 2010. The expected mortality numbers were calculated using external standardisation, based on age and sex adjusted rates for the whole of England. The top panel of Figure 3 displays the Standardised Incidence Ratio (SIR) for COPD mortalities, which is the ratio of the observed to the expected numbers of mortalities. The figure identifies regions of high risk in the north of the country, which includes deprived regions such as Northumberland and Merseyside. Additionally, there are high risk areas identified in the south east of England, including areas of Essex and Kent.

5.2 Results

The two-stage clustering model proposed in Section 3 was applied to these data, where the clustering step used respiratory disease data from 2007 to 2009. The fitted risk surfaces for these data sets exhibit similar spatial patterns to the 2010 study data, with Pearson's correlation coefficients of 0.97 for each year in turn. Markov-Chain Monte Carlo inference was used to obtain these results, with 5000 samples used for burn-in and a further 5000 used for the inference. Figure 4 displays the posterior probabilities for the different cluster structures, and the optimal cluster structure was chosen to

be that corresponding to the mode cluster number, which in this case was 40. Our method has the advantage of being able to quantify the uncertainty in the number of clusters identified, and a 95% credible interval for this ranges between 31 and 47. In addition, the median cluster number was 39. Note that due to the agglomerative nature of the clustering algorithm in Section 3.1, the 31 cluster structure could be formed by merging together 16 of the clusters in the structure containing 47 clusters.

The estimated risk surface (greyscale) and cluster structure (white dots) for the configuration with 40 clusters are displayed in the bottom panel of Figure 3, which has the same scale as the SIR plot in the top panel of that figure. In the majority of cases, there do appear to be differences in risks between neighbouring clusters. In particular, there appears to be a suggestion of a "north-south" divide, with two high risk clusters identified in the north of the country - one containing Northumberland and Cumbria in the far north, and the other containing the cities of Manchester and Liverpool as well as their surrounding areas. In contrast, the large cluster covering most of the south of the country, has a much lower risk of COPD. There is another distinct high-risk cluster in the south-east of the country, covering Essex and the east of London. The clusters appear to be based around grounds of socio-economic deprivation, which is well known to be linked with disease risk. The high risk areas in Figure 3 are generally areas with high levels of deprivation, while the lower risk areas are more affluent.

6 Discussion

The main aim of this paper was to develop statistical methodology to simultaneously estimate the spatial pattern in disease risk and identify clusters of areas exhibiting high (and low) risk. To achieve this aim a new methodology has been developed which fuses together spatial agglomerative hierarchical clustering techniques with an extended conditional autoregressive model, with inference based on Markov-Chain Monte Carlo simulation. This approach allows us to identify an optimal cluster structure which best describes the data, and it extends Anderson et al. (2014) by quantifying the uncertainty in the cluster structure. The clustering techniques are applied to disease risk data prior to our study period, allowing us to elicit candidate cluster structures for the study data. The prior clustering approach is

intended for use on chronic diseases in socio-economically stable regions; if this is not the case alternative clustering data such as covariate information should be used. These candidate structures have a natural ordering in terms of the number of clusters, which allows them to be considered as a univariate parameter in our Bayesian hierarchical model. This model estimates disease risk directly via the random effects, allowing for correlation between neighbouring areal units which are in the same cluster, but not enforcing it for areas in different clusters. Our approach differs from that used in Anderson et al. (2014), where the cluster structure was fixed when estimating the remaining model parameters. Here we are able to produce a credible interval for the number of clusters and can identify other potential alternative cluster structures which are supported by the data. The model outlined here does not contain covariate information because our aim was to identify clusters in the risk surface rather than in the residual error surface after accounting for known covariates.

The simulation study in Section 4 shows that our model generally outperforms the BYM model with the posterior classification step, with improved performances for both risk estimation and cluster identification. This is unsurprising, since our model attempts to estimate the cluster structure in the data, while the BYM approach estimates a smooth risk surface and then attempts to identify clusters in this smooth surface. One of our key aims was to develop a method which picked out extreme clusters; our model obtained a median Rand Index of 1 all three simulated cases, which indicates that in most cases it is able to correctly pick out all of the extreme clusters. Our model also performs well compared with the fixed effects approach proposed by Anderson et al. (2014), in terms of identifying the correct number of clusters. However our risk estimates do not appear to be quite as accurate as those obtained under the fixed effects model in cases where there is clustering present in the data. This is because the fixed effects approach has extra parameters in the mean model, while our approach accounts for clusters in the correlation structure of the random effects. However, in the case where no clustering is actually present, our model performs the best of the three. This is important to note, since in many cases it will not be known beforehand whether clusters actually exist within the data.

There is scope to extend this method into the spatio-temporal domain, thus allowing us to identify changes in the risk surface over time. This would

allow the authorities to identify whether a particular health intervention has had the desired effect in terms of reducing disease risk in a high-risk cluster. It would also allow for identification of clusters where the disease risk has increased over time, thus allowing health authorities to investigate the possible causes for any such deterioration in health.

Supplementary material

The supplementary material available online contains software which allows the user to run this model on the simulated data outlined in Section 4 of this paper.

Acknowledgements

The work of the first author was funded initially by the Carnegie Trust and then by the Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS).

References

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

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Annals of the Institute of Statistical Mathematics 43, 1–20.

Centers for Disease Control and Prevention, 2011. National Cardiovascular Disease Surveillance System. Technical Report. http://www.cdc.gov/dhdsp/ncvdss/.

Charras-Garrido, M., Abrial, D., de Goer, J., 2012. Classification method for disease risk mapping based on discrete hidden Markov random fields. Biostatistics 13, 241–255.

Charras-Garrido, M., Azizi, L., Forbes, F., Doyle, S., Peyrard, N., Abrial, D., 2013. On the difficulty to delimit disease risk hot spots. Journal of Applied Earth Observation and Geoinformation 22, 99–105.

Diggle, P., Rowlingson, B., Su, T., 2005. Point process methodology for on-line spatio-temporal disease surveillance. Environmetrics 16, 423–434.

Fraley, C., Raftery, A.E., Murphy, T.B., Scrucca, L., 2012. mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation.

Green, P., Richardson, S., 2002. Hidden Markov models and disease mapping. Journal of the American Statistical Association 97, 1055–1070.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer New York Inc.. chapter 14.3.

Knorr-Held, L., 2000. Bayesian modelling of inseparable space-time variation in disease risk. Statistics in Medicine 19, 2555–2567.

Kulldorff, M., 1997. A Spatial Scan Statistic. Communications in Statistics 26, 1481–1496.

Lee, D., Mitchell, R., 2012. Boundary detection in disease mapping studies. Biostatistics 13, 415–426.

Lee, D., Mitchell, R., 2013. Locally adaptive spatial smoothing using conditional autoregressive models. Journal of the Royal Statistical Society Series C 62, 593–608.

Lee, D., Rushworth, A., Sahu, S., 2014. A Bayesian localised conditional auto-regressive model for estimating the health effects of air pollution. Biometrics 70, 419–429.

Leroux, B., Lei, X., Breslow, N., 1999. Estimation of disease rates in small areas: A new mixed model for spatial dependence, in: Statistical Models in Epidemiology, the Environment, and Clinical Trials. Springer-Verlag, New York, pp. 135–178.

Li, P., Banerjee, S., McBean, A., 2011. Mining boundary effects in areally referenced spatial data using the Bayesian information criterion. Geoinformatica 15, 435–454.

Lu, H., Reilly, C., Banerjee, S., Carlin, B., 2007. Bayesian areal wombling via adjacency modelling. Environmental and Ecological Statistics 14, 433–452.

Marmot, M., Atkinson, T., Bell, J., Black, C., Broadfoot, P., Cumberlege, J., Diamond, I., Gilmore, I., Ham, C., Meacher, M., Mulgan, G., 2010. Fair Society, Healthy Lives: The Marmot Review. Technical Report.

Public Health England, 2010. Cancer e-Atlas. Technical Report. http://www.ncin.org.uk/cancer information tools/eatlas/.

Rand, W., 1971. Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association 66, 846–850.

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

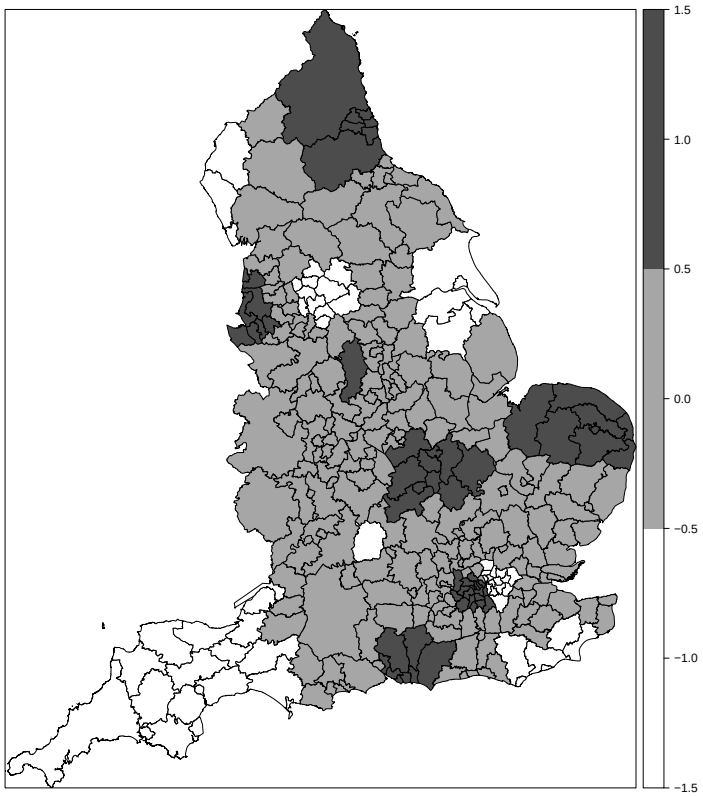

Figure 1: Plot of the simulated cluster structure in the English local authorities.

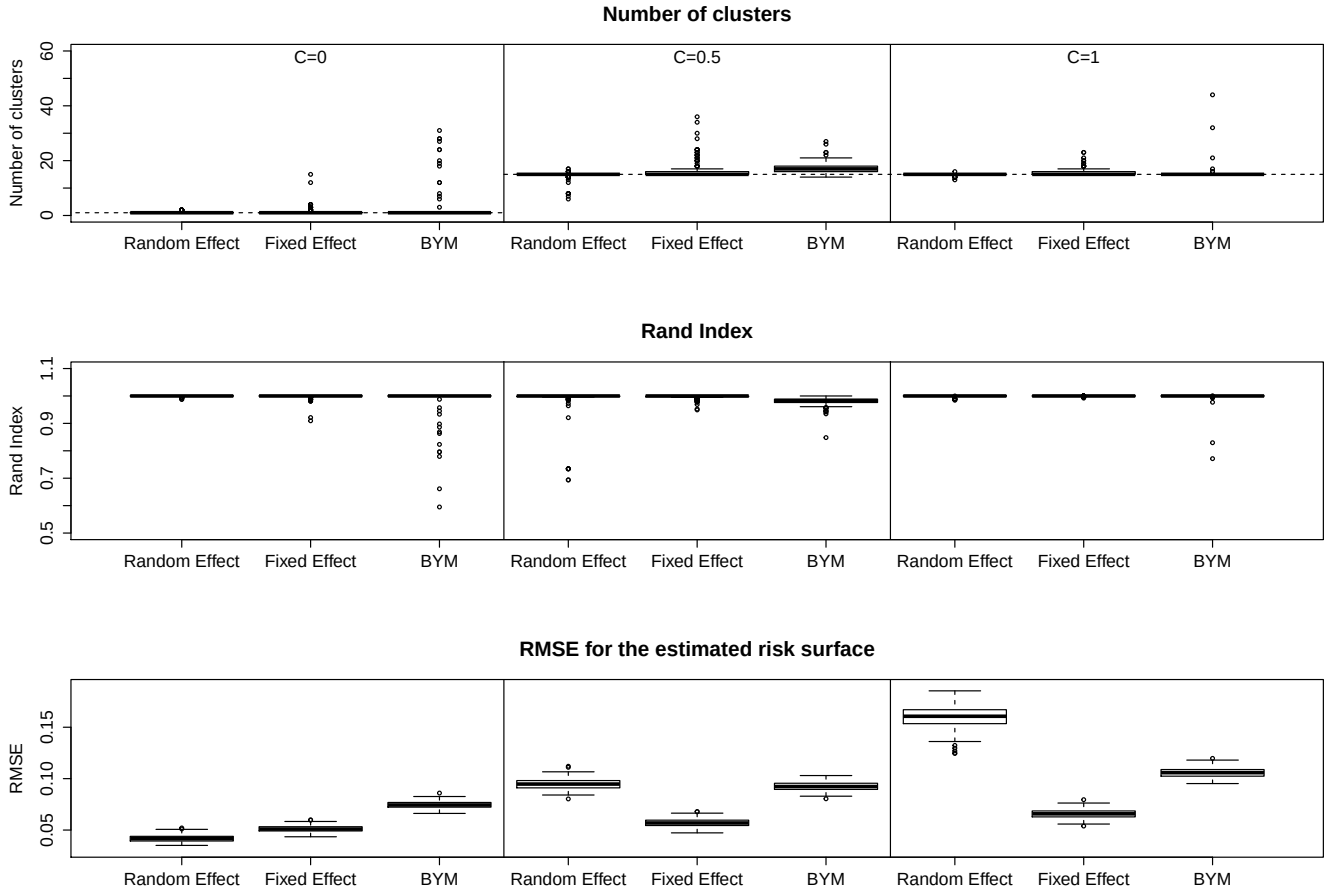

Figure 2: Summary of the simulation study results. The top, middle and bottom panels display boxplots of the estimated number of clusters, the Rand Index and the root mean square error of the estimated risk surface for each model in turn. The results relate to $C = 0$ (left panels), $C = 0.5$ (middle panels) and $C = 1$ (right panels). Within each panel, the boxplots relate to our proposed random effects model (left), fixed effects model (middle) and BYM (right). In the top panel the dashed lines represent the true number of clusters.

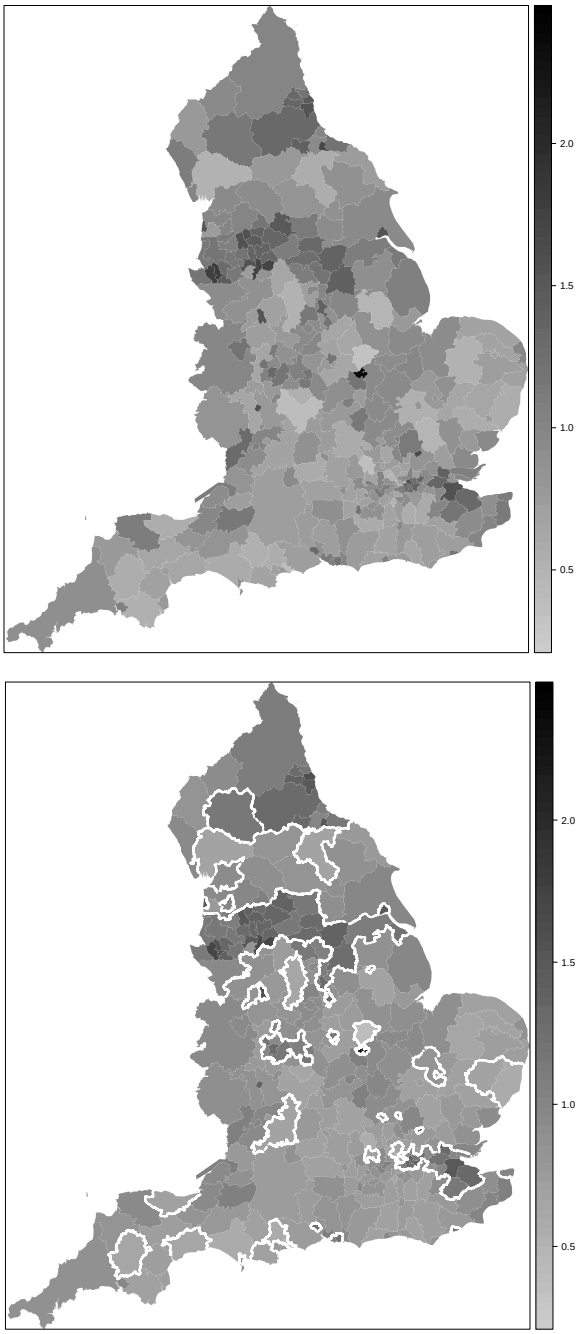

Figure 3: The top panel displays the standardised incidence ratio (grey-scale) for COPD mortalities in English local authorities in 2010. The bottom plot displays the estimated risk surface (grey-scale) from the model with 40 clusters (white dots).

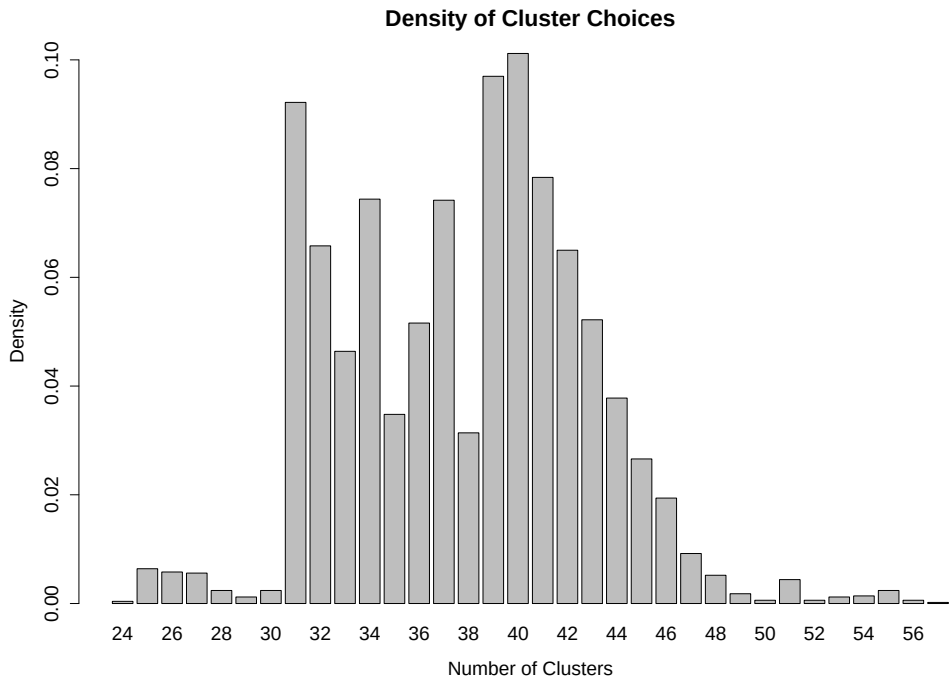

Figure 4: Plot of the posterior probability of each cluster configuration.