Technical Report RM# 697, Department of Statistics and Probability, Michigan State University

Simultaneous Estimation of Disease Risks and Spatial Clustering: A Hierarchical Bayes Approach

Wenning Feng, Chae Young Lim and Tapabrata Maiti
Department of Statistics and Probability
Michigan State University
{fengwenn,lim,maiti}@stt.msu.edu

Abstract

Detection of clustering and estimation of incidence risks from disease data are important in public health and epidemiological research. The popular models for disease risks such as conditional autoregressive (CAR) models assume known spatial dependence structure. Instead, we consider spatial clusters in which areas are geographically connected. Given spatial clustering idea, we propose a methodology that simultaneously estimates disease risks and detects clusters based on different features of the regression model. The proposed model is flexible in terms of local, regional and global shrinking and in terms of number of clusters, cluster memberships and cluster locations. We develop an algorithm based on the reversible jump Markov chain Monte Carlo (MCMC) method for model estimation. Numerical study shows effectiveness of the proposed methodology.

Keywords: Disease risk estimation; Hierarchical Bayes modeling; Reversible jump MCMC; Shrinkage estimation; Spatial Clustering; Spatial regression.

1 Introduction

Analyzing incidence counts, aggregated over a set of geographically disjoint areas, has been increasingly popular in public health and in spatial epidemiology. An underlying spatial dependence across the "neighboring" areas often plays an important role in statistical analysis of these data. We consider disease mapping to estimate relative risks of each study area and detection of spatial clusters simultaneously in this article. The spatial clusters, determined by the variation in regression parameters, are of our special interest here. Conventional clustering techniques are not suitable if we want areas in each cluster to be geographically connected.

Disease mapping has been proven to be useful to understand disease etiology and has a long history in epidemiology. The standardized mortality ratio (SMR), the ratio of observed cases over the expected number of cases, may be a simple statistics to produce disease maps. Clayton and Kaldor (1987), in their seminal paper, highlighted the drawbacks of using this simple statistics. The generated map could be seriously misleading particularly for rare diseases and sparse populations. Clayton and Kaldor (1987) took an empirical Bayes approach that shrinks the SMR's towards a local or global mean, where the amount of shrinkage may depend on local (spatial) and global variability. The statistical models and methods for disease mapping are well established in the literature. Wakefield (2007) nicely and critically reviewed well known approaches on disease mapping and spatial regression in the context of count data. Instead of repeating the literature, we refer this article and references therein for detail.

Despite the advantages of the models developed in disease mapping literature, most of them, except Knorr-Held and Raßer (2000), used Markov random field (MRF) to model the spatial process and fail to acknowledge spatial discontinuity, i.e., unexpected change in disease risks between adjacent (spatial) areas. This discontinuity is related to unknown spatial distribution of diseases over contiguous geographical areas. As noted by Knorr-Held and Raßer (2000), this is also related to the detection of clusters of elevated (or lowered) risks in diseases. They pointed out the difficulties in detecting discontinuities in the map using an MRF approach taken by other researchers such as Besag *et al.* (1991), Clayton and Bernardinelli (1992) and Mollié (1996).

The discontinuity in high/low disease risk may not reveal the truth when the discontinuity actually exists in the underlying regression model. As a motivating example, we consider the county map of the state of Michigan. One objective is to develop a map of lung cancer risks based on spatial regression. The detail description of the data is given in section 5. Following the existing methods, e.g. Wakefield (2007), one can develop a spatial regression model based on a regressor variable, say, poverty rate. In this case, there will be a single regression surface that represents the whole Michigan. However, if we divide all Michigan counties into three parts (as seen in Figure 1) such as the upper peninsula (Cluster 1), the major lower peninsula (Cluster 2) and the northeastern lower peninsula (Cluster 3), the existence of clusters through the regression lines is clear. Cluster 2 and 3 have similar regression lines, nearly no difference in their positive slopes and negative intercepts. Different from Cluster 2 and 3, Cluster 1 has a negative slope and a positive intercept. This shows the evidence of discontinuities in the disease map but such discontinuity can not be detected when the usual spatial regression with known underlying spatial process is applied.

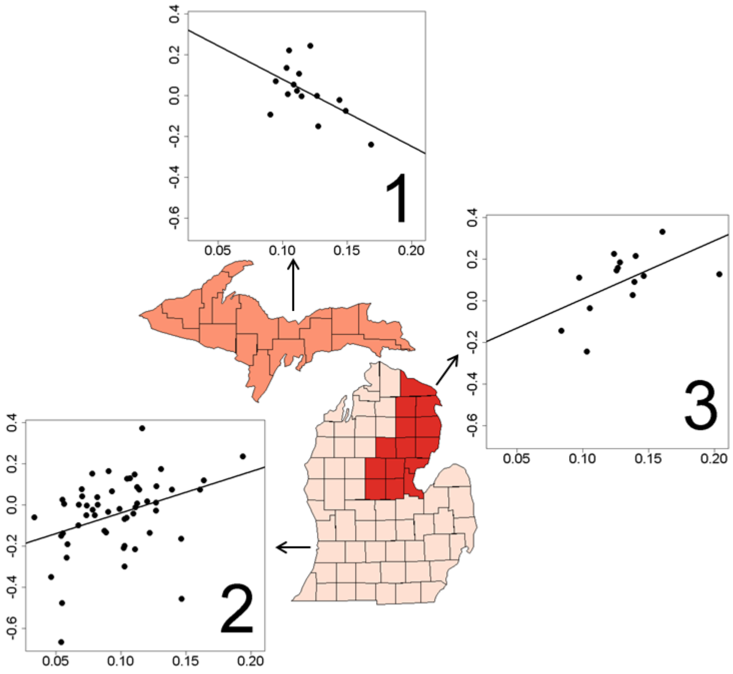


Figure 1: Preliminary Investigation on Michigan Lung Cancer data. Each scatter plot displays the association of log relative risk vs. poverty rate for each cluster.

To detect the regression discontinuity on a spatial domain, the clustering idea can be applied. Clustering technique is a fundamental statistical tool for large data analysis and has a growing literature in recent years. The commonly avaiable techniques, such as K-mean or mixture model based clustering are not suitable in this context since the cluster members under these methods need not necessarily be geographically connected. Knorr-Held and Raßer (2000) first recognized this issue of spatial clustering in the context of disease mapping and proposed a nonparametric Bayesian method. However, their method could be restrictive in many applications including ours. For example, their method considered constant disease risks for all the members in a given cluster, which could lead biased estimates of disease risks for each study area. Also, they mainly focused on disease risk estimation, although the work was motivated for cluster detection. On the other hand, we include covariates to reduce possible bias in the estimates of the disease risks. We like to point out that covariate inclusion in Knorr-Held - Raßer model is not trivial due to change in the dimension of parameter space and use of reversible jump MCMC. We also consider cluster specific random effects in the model, which is similar to the work by Booth *et al.* (2008). Booth *et al.* (2008) developed a regression model based clustering technique, but not in spatial context. Also, their regression model is not applicable for count data.
Our contribution: We propose a new model for disease risk estimation that takes into account not only the local or spatial variation, but also the regional or cluster level variation. The amount of shrinkage depends on local, regional (cluster) and global variation. The spatial clustering is embedded into the regression model so that no prior assumption on underlying spatial structure is needed. The system is data-driven and allows different regression models

in different parts of the map as discontinuities occur.

Our approach is fully Bayesian. A reversible jump Markov chain Monte Carlo (MCMC) algorithm (Green, 1995) has been adopted suitably in this context. The posterior analysis of clustering configuration is not clear in MCMC based Bayesian methods. Thus, we develop innovative methods to obtain the posterior estimate of clustering configuration and model parameters. The complexity in model parameter estimation due to stochastic clustering has been handled via two methods (area-wise and cluster-wise). Through the simulation study, we demonstrate the effectiveness of our method for both estimating the disease risks and detection of adversely affected clusters. The study also establishes superiority of the proposed method over the popularly used Poisson-CAR model in this context.

The rest of the paper is organized as follows: Section 2 defines a spatial clustering model and introduces the cluster specification and the prior distributions. The methods to obtain the posterior estimates of clustering configuration and model parameters are described in Section 3. In Section 4, a simulation study is presented. The simulation has two study designs. Section 5 contains the two real data applications. We conclude our results and make some remarks in Section 7.

2 The Model and Prior Specification

In this section, we describe a hierarchical model that allows cluster-wise varying regression, within cluster variation and between cluster variation. The model consists of three components: the likelihood for disease count data, the latent process model for log relative disease risk and the prior specification for clustering and model parameters.

Suppose there are n geographically contiguous areas on the spatial domain of interest. Let y_i be the observed disease incidence/mortality counts, E_i be the expected population under the risk, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ be the p-dimensional covariate and ν_i be the log relative risk at the i-th area for $i = 1, \cdots, n$. Denote $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)^T$. For the cluster specification, we denote the cluster partition of the n areas as $\mathcal{C} = \{C_1, \ldots, C_k\}$ that consists of k clusters. Denote n_j be the number of areas in C_j and X_j be the corresponding $n_j \times p$ design matrix whose rows are $\{\boldsymbol{x}_i^T : i \in C_j\}$. Note that we observe $\boldsymbol{x}$ and $\boldsymbol{y}$ only and E_i are given. Also, n is known but k, $\mathcal{C}$ and the members in C_j are unknown.

2.1 Likelihood for the count data and the latent model for the log relative risk

We adopt the Poisson likelihood for $\boldsymbol{y}$, that is, $y_i \sim \text{Poisson}(E_i e^{\nu_i})$. The y_i's are assumed to be conditionally independent given $\boldsymbol{\nu}$. Then, the data likelihood is

$$L(\boldsymbol{y} \,|\, \boldsymbol{\nu}) = \prod_{i=1}^{n} \frac{(E_i e^{\nu_i})^{y_i}}{y_i!} \exp(-E_i e^{\nu_i}). \tag{1}$$

We model the log relative risk, $\boldsymbol{\nu}$, with cluster-wise regression coefficients and spatial cluster random effects. For area i in j-th cluster, C_j, the log relative risk is modeled as

$$\nu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}_j + \epsilon_i + u_j, \tag{2}$$

where $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp})^T$ is the regression coefficient and u_j is the random effect for C_j. We assume $u_j \overset{i.i.d.}{\sim} N(0, \sigma^2)$ for $j = 1, \ldots, k$. Thus the between cluster variation is σ^2. The $\epsilon_i' s$ are area specific random effects and are assumed to be $\overset{i.i.d.}{\sim} N(0, \sigma_j^2)$ for $i \in C_j$. This makes the area risks different even after eliminating the differences due to covariates. This also builds shrinkage estimation within a cluster. Unlike common disease mapping models such as a Poisson-CAR model, the proposed model helps detecting clusters, if any.

Using matrix notations, $\boldsymbol{\nu_j} \sim N_{n_j}(X_j^T \boldsymbol{\beta}_j, \Sigma_j)$, where $\boldsymbol{\nu_j}$ is the vector of $\{\nu_i : i \in C_j\}$ and $\Sigma_j = \sigma_j^2 \mathbf{I}_{n_j} + \sigma^2 \mathbf{J}_{n_j}$. Here, $\mathbf{I}_{n_j}$ is the n_j-dimensional identity matrix and $\mathbf{J}_{n_j}$ is the n_j-dimensional square matrix with all entries as unity. Note that Σ_j has the following properties:

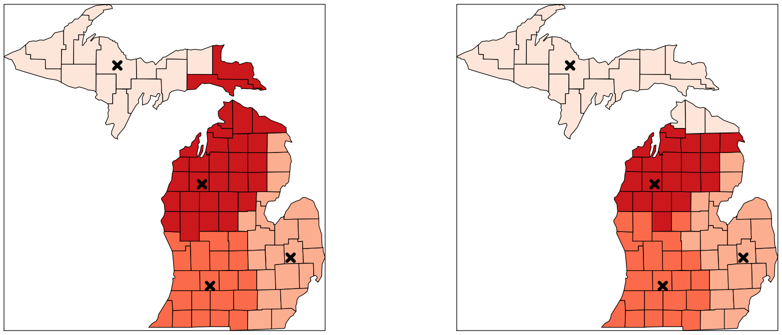
$$\begin{aligned} \Sigma_j^{-1} =& \frac{1}{\sigma_j^2} \mathbf{I}_{n_j} - \frac{\sigma^2}{\sigma_j^2(\sigma_j^2 + n_j \sigma^2)} \mathbf{J}_{n_j}, \\ \det(\Sigma_j) =& (\sigma_j^2)^{n_j} + n_j \sigma^2 (\sigma_j^2)^{n_j - 1} = (\sigma_j^2)^{n_j - 1}(\sigma_j^2 + n_j \sigma^2). \end{aligned} \tag{3}$$

This helps immensely in computing during the implementation since we don't need to invert a n dimensional matrix. Thus, the method is applicable from large to very large data set as long as the maximum cluster size remains moderate. The computational burden per MCMC iteration reduces from $O(n^3)$ to $O(n)$.

The latent model (2) for the log relative risk together with the Poisson likelihood (1) can be viewed as a generalized linear mixed-effect model (GLMM) for disease count data. The proposed model allows varying regression coefficients ($\boldsymbol{\beta}_j$) for different spatial clusters. The spatial cluster random effect u_j captures variability between spatial clusters and the assumption on ϵ_i allows area-specific variability within each cluster.

2.2 Cluster specification and priors

The cluster specification is determined by a prior model. One important feature for clustering spatial data is that, all the areas in a given cluster should be geographically connected. This is a necessary requirement for our clustering model. We call this type of clustering as the *spatial clustering*. It is also named as the partition in Denison and Holmes (2001). The spatial clustering configuration is obtained by the *minimum distance criterion*. For a given vector of the cluster centers $G_k = (g_1, \ldots, g_k)$, which is a subset containing k distinct areas, the clusters are formed by assigning each area to the nearest cluster center. In other words, $C_j = \{i : d(i, g_j) \le d(i, g_l)$ for all $l \ne j\}$, where $d(i, g_j)$ is the *distance* between the area i and the cluster center g_j. The distance between two areas, $d(i_1, i_2)$, can be defined in two ways. Knorr-Held and Raßer (2000) used the minimal number of boundaries to cross from area i_1 to area i_2. The alternative is the Euclidean distance between centroid coordinates of two areas. Figure 2 shows an example of the spatial clustering configuration by these two different distance measures on the Michigan map. Each approach has both advantages and disadvantages: The minimal number of boundaries can be calculated by using the adjacency matrix only, which is usually available in disease mapping data. However, unique assignment may not be possible so that the additional scheme is needed. This can hinder the computational efficiency and the mixing behavior in the Bayesian computation. On the other hand, it is less likely to be an issue for the latter approach since most disease



(a) Euclidean distance (b)Minimal number of boundaries

Figure 2: Clustering Configuration Comparison based on two different distance measures.(Crosses represent the cluster centers)

mapping data have a irregular spatial design. However, the centroid coordinates to represent the area can be misleading for the aggregated data. We want to hold a neutral opinion on these two distance definitions and encourage to apply both approaches in practice. In this paper, the Euclidean distance is used since it performs well on our simulation examples.

Now we introduce a prior model $\pi(\mathcal{C})$ which is decomposed into the prior for the number of clusters, k, and the prior for the cluster centers G_k given the number of clusters. That is, $\pi(\mathcal{C}) = \pi(k)\pi(G_k|k)$ (see Green 1995). Here, we follow Knorr-Held and Raßer's prior specifications on (k, G_k). A prior model on k is assumed to be proportional to $(1-d)^k$, where $d \in [0,1)$ is a constant. When d is small, the prior on k tends to be noninformative. When $d = 0$, the prior becomes a uniform distribution over the areas $\{1, 2, \ldots, n\}$. Knorr-Held and Raßer (2000) recommended to fix the d that is close to 0. In our case, however, the choice of d can be sensitive when σ_j^2's are relatively large. This makes sense since clustering may be difficult or may not even make sense if the within cluster variability is very high. Therefore, we introduce an additional hierarchy with a uniform prior on d to avoid the fixed choice of d. Given the number of clusters, k, the prior for G_k is uniform over all the vectors of areas, i.e.,

$$\pi(G_k \,|\, k) = \frac{1}{\binom{n}{k} k!} = \frac{(n-k)!}{n!}. \tag{4}$$

Prior models for other parameters are $\boldsymbol{\beta}_j \sim N_p(\boldsymbol{\mu}_j, V_j)$ for $j = 1, \ldots, k$, $\sigma^2 \sim \mathrm{IG}(a, b)$ and $\sigma_j^2 \sim \mathrm{IG}(a_j, b_j)$ for $j = 1, \ldots, k$, where IG stands for Inverse Gamma distributions with the pre-specified shape and scale parameters. Even though a flat prior is common for $\boldsymbol{\beta}_j$, the specification of a proper prior is actually crucial in reversible jump MCMC, which requires relative normalizing constants between different subspaces. The prior mean $\boldsymbol{\mu}_j$ can be derived from a preliminary regression analysis, and we recommend a simple *i.i.d.* structure for V_j with relatively large variance so that the normal prior can be close to noninformative. Besides the normal prior, the Cauchy prior can also be employed here. For σ^2 and σ_j^2, the choices of a, b, a_j's and b_j's can follow the same idea by matching the mean from the preliminary regression analysis result and choosing a large variance. As an alternative way to set the variance-covariance parameters, one can consider the reparametrization of σ^2 and

σ_j^2's. For example, let $\sigma_j^2 = \lambda_j \sigma^2$, $j = 1, \ldots, k$. Then one can assume priors on λ_j's instead. As usual, the priors on $\boldsymbol{\beta}_j$'s, σ_j^2's and σ^2 are assumed to be independent.

We consider Bayesian computation using the reversible jump MCMC algorithm for estimation. Detail updating steps are given in the Appendix A.

3 Posterior Analysis

Posterior estimation of clustering configuration and corresponding model parameters is challenging due to the varying structure of clusters at each MCMC iteration. The issue has also been mentioned by Tadesse *et al.* (2005) in the discussion section of their paper. Putting aside as a future research, they made posterior inference conditional on a fixed number of components. We introduce approximation methods for the posterior estimation and investigate the related properties.

3.1 Clustering configuration

The clustering configuration is uniquely determined by (k, G_k). The posterior estimate of k can be obtained by the posterior mode, denoted by $\hat{k}$. However, estimating posterior distribution of $G_{\hat{k}}$ reasonably well, using MCMC samples, may not be feasible since the number of all possible $G_{\hat{k}}$'s is too big. To resolve this issue, we consider a spectral clustering method. We refer the reader to a recent review article by Elavarasi *et al.* (2011). The spectral clustering method includes a group of techniques which makes use of the eigen structure of a similarity matrix. Among all the MCMC iterations which yield $\hat{k}$ clusters, we define a similarity matrix $S_{n\times n}$ as

$$S_{p,q} = \frac{\text{number of times areas } p \text{ and } q \text{ are in the same cluster}}{\text{total number of the MCMC iterations}}.$$

$S_{p,q}$ retains the empirical probability that areas p and q are grouped in the same cluster based on the MCMC samples. Then, with the spectral clustering algorithm (e.g. Shi-Malik algorithm (2000), Ng-Jordan-Weiss algorithm (2002) or Kannan-Vempala-Vetta algorithm (2004)), we can obtain an approximate posterior estimate of the cluster configuration with $\hat{k}$ clusters. The spectral clustering method generates the posterior "mean" estimate of the clustering structure via the pairwise cluster membership linkage with very little additional computational cost.

With a mild effort, one could easily be convinced with the reasoning of a spectral clustering algorithm. In Appendix B, we outline the Ng-Jordan-Weiss algorithm adapted for our similarity matrix and analyze the reasoning under the ideal case where S is block-diagonal. In more general cases, where the off-diagonal blocks are non-zero, we can study the deviation between the spectral clustering result and the true clustering in the light of matrix perturbation theory. Ng *et al.* (2001) gave a theorem of the existence of a "good" spectral clustering under the certain conditions on eigengap, which we won't repeat here. Readers can find details in their paper.

3.2 Model parameters

The challenge in estimating $(\boldsymbol{\beta}_j, \sigma_j^2)$, $j = 1, \dots, k$, comes from the variation in the number of clusters in each MCMC iteration. To overcome the issue, we consider the following two approaches:

M1 Area-wise estimation:
Each area, i, is assigned to one of the clusters at each MCMC iteration so that we can have posterior samples of $\boldsymbol{\beta}, \sigma^2$ for the area i. Then, we can obtain the posterior estimate of $(\boldsymbol{\beta}, \sigma^2)$ for each i. This is easy to implement but this approach produces n sets of posterior estimates for $(\boldsymbol{\beta}, \sigma^2)$. Although we could have similar values for the parameter estimates that are in the same cluster, we lose clustering information.

M2 Cluster-wise estimation:
Alternative approach is to make use of the posterior estimate of clustering configuration given in section 3.1. For each MCMC iteration with $\hat{k}$ clusters, the clustering configuration may be different from the estimated clustering configuration. Then, we find a mapping between cluster labels at the given MCMC iteration and cluster labels at the posterior estimate. Let $(\boldsymbol{\beta}_j^{(l)}, \sigma_j^{2(l)})$ be the parameter sample in the l-th MCMC iteration. If the mapping assigns the g-th cluster in the posterior estimates for the j-th cluster in the l-th MCMC iteration, $(\boldsymbol{\beta}_j^{(l)}, \sigma_j^{2(l)})$ is assigned to be a posterior sample for g-th cluster in the posterior estimate. The mapping procedure is further explained below with a toy example.

Suppose there are 10 areas in total and the posterior estimate has 3 clusters. For a particular MCMC iteration (with 3 clusters only), suppose that the cluster memberships are given as in the Table 1(a). That is, the area 1 is classified into Cluster 1 of the posterior estimate and Cluster 1 of the MCMC sample, and so on. From such information, we can create a square matrix to display the number of common areas shared by different pairs of a posterior cluster and a MCMC sample cluster (Table 1 (b)). Then, the mapping is determined by the order of values in the matrix in Table 1 (b). Since 3 is highest, cluster 3 in the MCMC iteration is mapped into the posterior cluster 1. Next highest value after removing those clusters is 2 so that the cluster 1 is mapped into the posterior cluster 2, etc. The final mapping of cluster labels from MCMC sample to posterior estimate is $(3, 1, 2) \to (1, 2, 3)$, and parameter values at the MCMC iteration is reordered based on the mapping.

In **M2** approach, it may happen that certain posterior samples of model parameters come from MCMC clusters which don't share many common areas with the target cluster in the posterior estimate. Since this may produce a poor estimate, we exclude such samples. For example, in each posterior cluster, we can calculate the posterior estimates of model parameters based on the MCMC samples whose cluster shares at least 90% areas in common.

[1] Three clusters are labeled as 1,2,3 on the row title and column title for MCMC iteration and posterior estimate respectively.

Table 1: Illustration of mapping procedure

Area Index		1	2	3	4	5	6	7	8	9	10
Cluster	Post'r	1	1	1	1	2	2	2	2	2	3
Labels	MCMC	1	3	3	3	1	1	2	3	3	2

(a) Cluster label example

		Post'r		
		1	2	3
MCMC	1	1	2	0
	2	0	1	1
	3	3	2	0

(b) Cluster mapping matrix, $\boldsymbol{L}$.[1]

4 Simulation Study

In the simulation study, we validate the performance of the updating steps of the reversible jump MCMC algorithm based on both the estimation of the clustering configuration and the log relative risks. Since the Scotland lip cancer data (Clayton and Kaldor, 1987) is a popular and extensively studied example for disease risk models, we consider two simulation studies based on the Scotland county configuration which has 56 counties (areas) in total. The first study is designed to focus on the estimation of the clustering configuration while the second study is designed to focus on the estimation of the log relative risks. For each simulation study, we ran 50,000 iterations and the first half was discarded as burn-in period. On a regular dual-core PC, Matlab spent about $2\frac{1}{4}$ hours completing 50,000 MCMC iterations for both designs, which is moderate given a sophisticated algorithm.

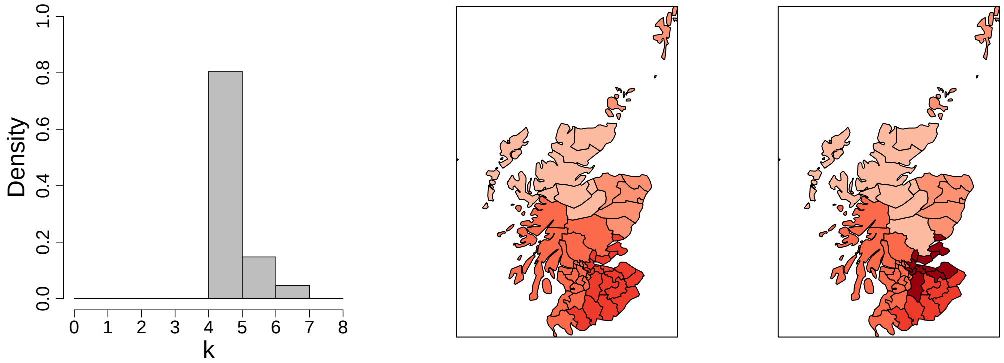
4.1 Design I

In Design I, we assume there are 4 clusters with cluser centers at Ross-Cromarty, Aberdeen, Dumbarton and Tweedale. We consider a covariate x_i simulated from $N(10, 3)$. Without loss of generality, we assume that E_i equals to 1. The parameters $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ are set to be $(-2\log 10, 0.3\log 10)$, $(-2\log 10, 0.5\log 10)$, $(2\log 10, 0.3\log 10)$ and $(2\log 10, 0.5\log 10)$ for the 4 clusters, where β_0 corresponds to the intercept. Note that any two clusters will have the same β_0 or β_1. The range of y_i is quite different for each cluster with given (β_0, β_1). The variance of within-cluster error, σ_j^2, are assumed to be 0.1, 0.2, 0.3 and 0.4, and the between variance, σ^2, is 0.2. The noise level is relatively low so that the differences among clusters due to the different mean functions are visible.

The posterior distribution of k is given in Figure 3 (a). The posterior mode is $\hat{k} = 5$ and the posterior estimate of the clustering configuration is plotted together with the true clustering in Figure 3 (b) and (c). The posterior estimate of the clustering configuration is comparable to the true configuration. The deviation happens at the lower right corner, where a cluster are split into two. Also, the county located at the center is misclassified.

We measure the clustering performance with some similarity statistics. These measures are also used in Booth *et al.* (2008). First we create a 2×2 table with counts $\{n_{ij}\}$ by grouping all $\binom{56}{2} = 1540$ pairs of counties based on whether they are in the same cluster in the posterior clustering configuration and in the true configuration.

Then, we consider the following statistics:



(a) Posterior Distribution of k (b) True clustering (c) Posterior estimate

Figure 3: Posterior Estimate of Clustering under the simulation Design I.

- the ψ-statistic,

$$\psi = 1 - \left(\frac{(n_{11} - n_{1+})^2}{n_{1+} n_{++}} + \frac{(n_{22} - n_{2+})^2}{n_{2+} n_{++}} \right)$$

  The best performance is when $n_{11} = n_{1+}$ and $n_{22} = n_{2+}$, then $\psi = 1$. The worst performance occurs when $n_{12} = n_{1+}$ and $n_{21} = n_{2+}$, where $\psi = 0$.

- Yule's Q association measure

$$Q = \frac{n_{11} n_{22} - n_{12} n_{21}}{n_{11} n_{22} + n_{12} n_{21}}$$

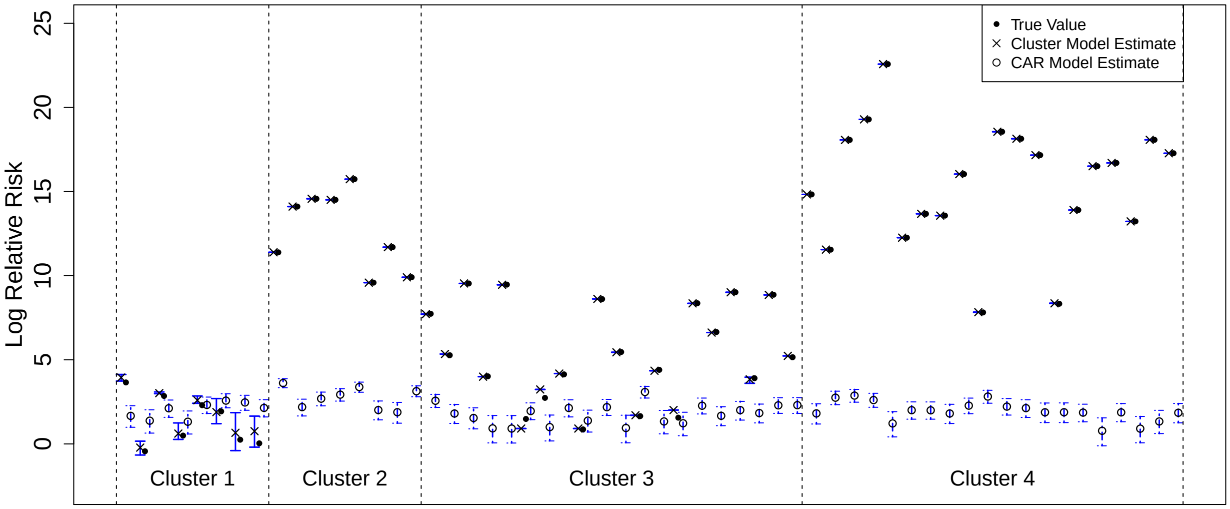
  Q has the range of $[-1, 1]$ and the value close to 1 indicates a good clustering performance.

- the sensitivity measure n_{11}/n_{1+}:
  The sensitivity measure tells the percentage of the county pairs which are correctly clustered together. The closer to 1 the value is, the better the performance is.

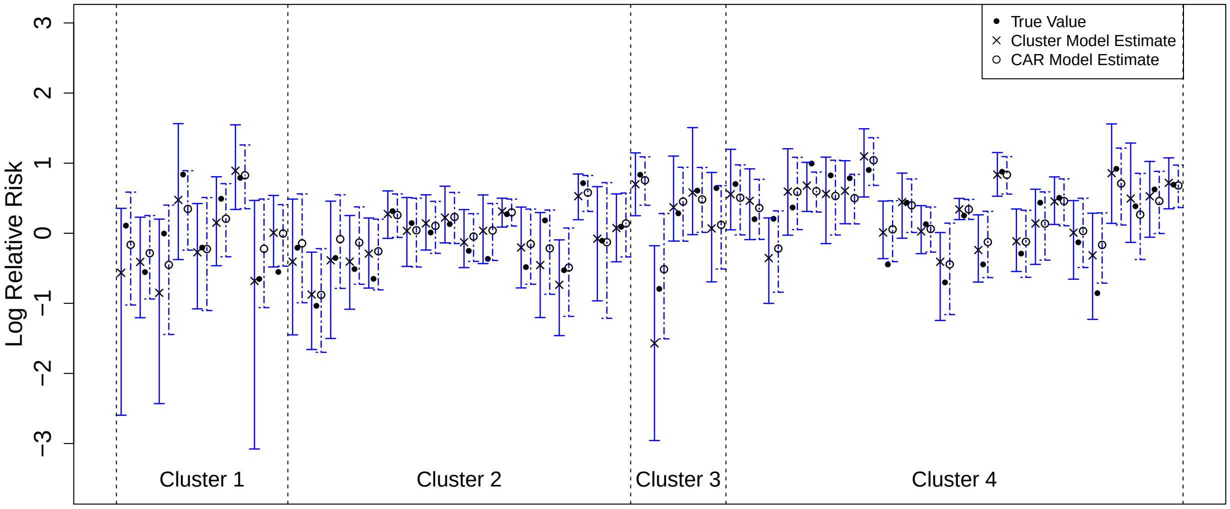
- the specificity measure n_{22}/n_{2+}
  The specificity measure tells the percentage of the county pairs which are correctly seperated. The closer to 1 the value is, the better the performance is.

For Design I, we obtain $\psi = 0.9819$, $Q = 0.9951$, sensitivity $= 0.7477$, specificity $= 0.9928$. All the values are very close to 1 (indicating good clustering performance), except for the sensitivity measure. This is the consequence of the extra cluster. When the original cluster has been splitted into two clusters in the posterior estimate, a number of county pairs, which originally belong together, will be separated and this leads to the decrease in the value of sensitivity measure. However, the specificity measure is still very high, which implies the additional split happens mostly within one original cluster instead of involving several original clusters.

Figure 4 (a) shows posterior estimates and true ν_i with 90% credible intervals. The posterior estimate is calculated by posterior mean. For majority of the counties, the posterior estimates from the cluster model are close to the true values with very narrow 90% credible intervals. On Figure 4 (a), most of the credible intervals from cluster model (solid segments)



(a) Design I



(b) Design II

Figure 4: 90% Credible Interval Plot for Log Relative Risks under the simulation Designs. The symbols and segments are grouped according to the true cluster that each county (area) belongs to. Solid segments represent the credible intervals derived via the cluster model, while the dot-dash segments represent the ones via the Poisson-CAR model.

shrink to single vertical segments because of the narrowness. Counties with wide credible intervals belongs to the cluster 1 with small y_i (less than 5). We also fitted the Poisson-CAR model. The posterior estimates of ν_i are quite different from the true values and the credible intervals also fail to capture the true value for most of the counties, except for the true Cluster 1. All the intervals are generally consistent in their ranges. This is due to the unified parameter estimation from Poisson-CAR model. Under the setting of an apparent clustering tendency and a low noise level, one global regression model, such as the Poisson-CAR model, is unable to provide a flexible and accurate disease risk estimation. To compare the performance of the two models, we consider the following quantity

$$\mathrm{RAD} = \sum_{i=1}^{n} \left| \frac{\hat{\nu}_{i,\mathrm{Cluster}} - \nu_i}{\nu_i} \right| \Bigg/ \sum_{i=1}^{n} \left| \frac{\hat{\nu}_{i,\mathrm{CAR}} - \nu_i}{\nu_i} \right| ,$$

which is the Ratio of averaged Absolute relative Difference between two models. We have RAD = 0.000753 for the Design I which indicates that the cluster model performs better for estimating the log relative risks.

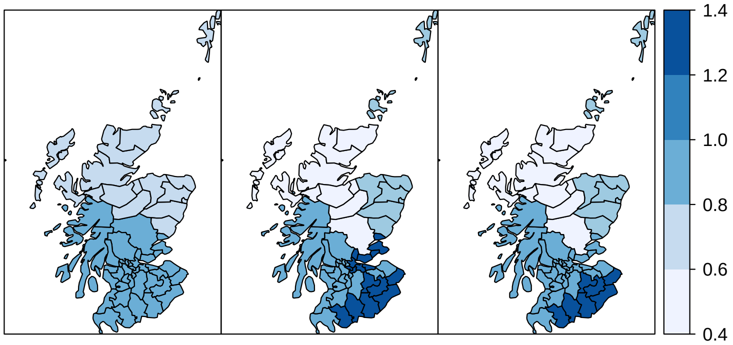
The posterior means for model parameters $\boldsymbol{\beta}_j = (\beta_{0j}, \beta_{1j})'$ are plotted in Figure 5. Under Design I, the difference between the two approaches, **area-wise estimation** and **cluster-wise estimation**, is negligible. This is expected since Design I has quite distinct values of $\boldsymbol{\beta}_j$ for clusters. Posterior estimates of $\boldsymbol{\beta}_j$ from both approaches are close to the true values. The estimation of σ_j^2 (not shown) also behaves similarly and is comparable to the true value.

4.2 Design II

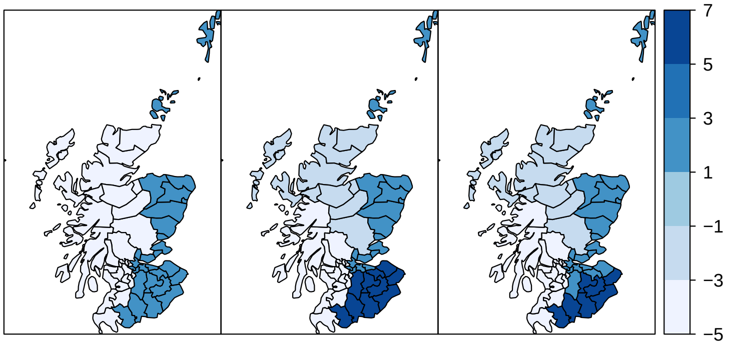
For the second simulation design, we still adopt a 4-cluster design on the Scotland map. We retain the expected risk E_i and x_i from the Scottish Lip cancer data, where $x_i = \mathrm{PcAFF}_i/10$ are scaled percentages of the work force employed in agriculture, fishing and forestry. By controlling the level of $\boldsymbol{\beta}$, we mimic the real data closely but have different levels so that clustering tendency exists. We set the model coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1)$ as $(0.41, -0.38)$, $(0.21, -0.18)$, $(-0.14, -0.18)$ and $(-0.31, 0.38)$. With the same within- and between-cluster variance levels as in Design I, the clustering tendency is much weaker in Design II. This design provides the Poisson-CAR model with a chance to get a better estimation of disease risks, but poses a greater challenge to our model in cluster detection. The question, we want to answer via the second simulation is, whether our model can still estimate the log relative risks well enough in the absence of a strong clustering tendency.

The posterior mode of the number of clusters k is 5 and 90% credible interval is $[4, 8]$. The clustering performance measures are given as $\psi = 0.8720$, $Q = 0.5536$, sensitivy = 0.3931 and specificity = 0.8431. The ψ and specificity measures show our method is generally good and separated the majority of the county pairs correctly. But because of the extra cluster and some misclassifications, the Yule's Q association and sensitivity measures are not very satisfactory.

Since the true log relative risks do not show apparent clustering tendency contrast to the Design I, the posterior estimates from both the cluster model and the Poisson-CAR model are comparable (Figure 4 (b)). Indeed, the ratio of averaged absolute relative difference RAD = 0.7117 shows that the cluster model still works better. 90% credible intervals from



(a) β_{1j}



(b) β_{0j}

Figure 5: Posterior estimates of $\boldsymbol{\beta_j}$ under the simulation Design I. Left panel shows the true value; middle panel shows the posterior mean via **M1** area-wise estimation; right panel shows the posterior mean via **M2** cluster-wise estimation

Table 2: Posterior estimates of β_1 based on **M2** under the simulation Design II

Cluster	Posterior Estimate	Credible Interval
1	-0.4834	$[-1.1250, -0.0796]$
2	-0.3145	$[-0.8988, 0.1835]$
3	-0.2577	$[-0.9707, 0.2904]$
4	0.1448	$[0.1401, 0.1497]$
5	0.2056	$[0.0025, 0.4854]$

the Poisson-CAR model are narrower than those of the cluster model for some counties. Similar to Design I, these counties have small y_i, which is nearly 0 in this case. We found that the covariate x_i are relatively large for those counties. Large x_i can amplify a small change in model parameter estimation into a large change in log relative risk estimation. Since cluster models allow different values of model parameters for clusters, this may cause wider credible intervals. On the other hand, we should notice that some of the credible intervals from the Poisson-CAR model fail to include the true values while the corresponding cluster intervals don't. Bayes factor based on the MCMC samples is 27.3145 (Cluster vs. Poisson-CAR) which indicates that the data strongly support the cluster model over the Poisson-CAR model. To conclude, our model can still estimate the disease risks reasonably well even under a weak clustering tendency.

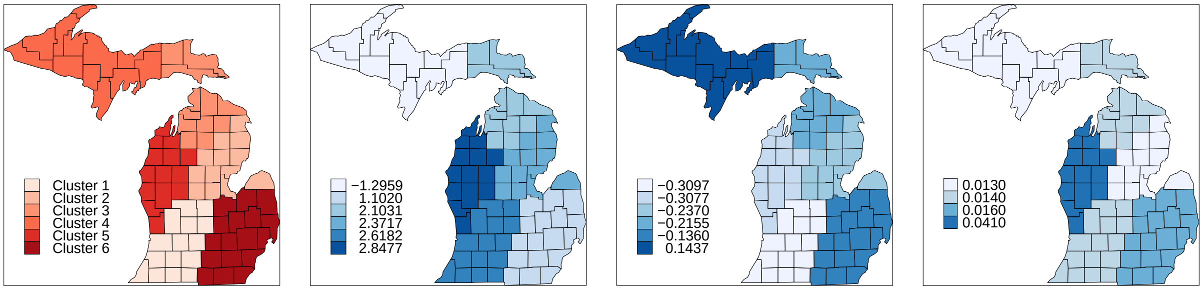
The posterior estimates of model parameters using the area-wise estimation, **M1**, do not show clear difference between clusters because of the averaging of estimation between neighboring clusters. To see whether the cluster model can capture the difference in β_1 among clusters, we compare posterior estimates from the cluster-wise estimation, **M2** with the true β_1. Table 2 shows posterior estimates of β_1 with 90% credible intervals. Note that, the cluster model can detect a significant positive/negative linear association in Cluster 1,4 and 5. On the other hand, the Poisson-CAR model gives the single estimates for model parameters as $\hat{\beta}_1 = -0.11$ with credible interval $[-0.35, 0.12]$, which implies an insignificant linear association and totally ignores the variation among clusters.

5 Applications

We consider two real data examples in this section. The first one is Michigan lung cancer data and the second one is Scotland lip cancer data.

5.1 Michigan lung cancer data

Michigan consists of 83 counties, which are divided into upper and lower peninsulas, separated by the Great Lakes. Due to the geographical discontinuity, there may exist discrepancy on various socio-economical and health variables over Michigan counties. Since the simple correlation between poverty rate and observed SMR of Lung cancer is reasonably high, we consider the poverty rate as the covariate for the analysis. We use lung cancer mortality counts from 2001 to 2005 for 83 Michigan counties available at the SEER database (`seer.cancer.gov`). The expected counts are calculated by the age-adjusted formula as provided by Jin, Carlin and Banerjee (2005), and each county's age distribution is obtained from U.S. Census 2000. The poverty rate at county level, as the covariate, is obtained from the SEER database as well.



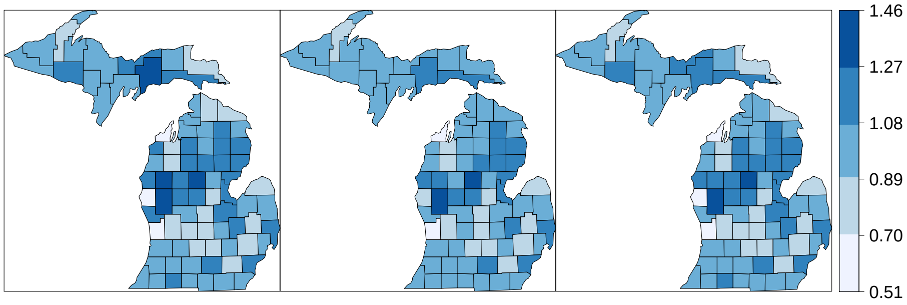
(a) Clustering (b) β_{1j} (c) β_{0j} (d) σ_j^2

Figure 6: Estimation results for Michigan lung cancer data.

The estimated clustering structure is given in Figure 6 (a) with 6 clusters. The clustering configuration coincides with the general picture of Michigan's economic status:

- Cluster 1: Southwestern metropolitan cluster (e.g., Grand Rapids and Kalamazoo)
- Cluster 2: East lake side/chemical industry cluster (e.g., Midland, Bay City and Saginaw)
- Cluster 3: Upper/lower peninsula conjunction cluster (e.g., Gaylord and Mackinaw City)
- Cluster 4: Upper peninsula cluster (e.g., Marquette, Crystal Falls and Negaunee)
- Cluster 5: West lake side cluster (e.g., Traverse City and Muskegon)
- Cluster 6: Auto city cluster (e.g., Flint and Detroit)

Figure 7 shows the estimates of relative risks. Compared to the observed SMR (y_i/E_i), the estimates from the cluster model and the Poisson-CAR model are smoothed, which is expected given the smoothing capability of the models. While two models produce comparable relative risk estimates, the estimates of model parameters show clear difference. The Poisson-CAR model gives a single estimate of $\hat{\beta}_1 = 2.03$ with 90% credible interval $[0.2574, 3.8105]$. The posterior estimates of β_1 from the cluster model show distinct features over clusters (Figure 6 (b)-(d)). For example, the upper peninsula cluster (Cluster 4) has a negative β_1 estimate while the lower peninsula clusters (Cluster 1, 2, 3, 5 and 6) have positive ones. Table 3 shows the significance of β_1 for Cluster 2, 4 and 5. More urbanized areas (lower peninsula) shows that higher poverty rate leads to an increase in disease risk while relatively less urbanized areas (upper peninsula) shows the opposite. When comparing this result with our pre-investigation in Figure 1, the preliminary findings are confirmed. Further investigation with additional covariates and demographic information on Michigan



(a)Observed SMR (b) Cluster (c) Poisson-CAR

Figure 7: Posterior Estimate of relative risks for Michigan Lung Cancer data

Table 3: Posterior estimates of β_1 for Michigan Lung Cancer data

Cluster	Posterior Estimate	Credible Interval
1	2.6182	$[-0.1938, 3.7009]$
2	2.3717	$[0.4878, 4.1826]$
3	2.1031	$[-1.1920, 3.9449]$
4	-1.2959	$[-4.7095, -0.3792]$
5	2.8477	$[0.5525, 4.7126]$
6	1.1020	$[-0.1163, 2.2716]$

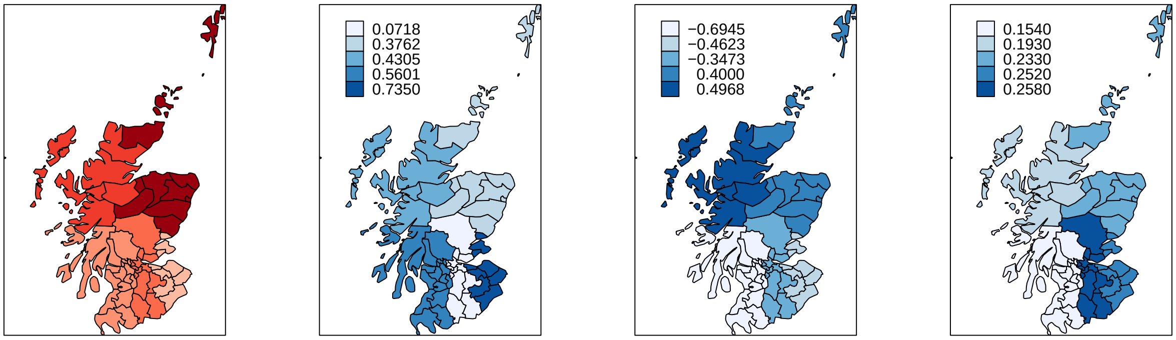
counties is necessary to find thorough reasoning on this discrepancy. One explanation for this apparent contradiction could be due to the fact that the upper peninsula has casinos and people smoke lot more compared to other parts of Michigan. Had we factor in smoking into our regression model, the scenario could have been different. Another clue we can look at is, the correlation between poverty rate and observed log relative risk is -0.4840 for the upper peninsula cluster, which is in accord with the negative slope estimate.

Note that the single estimate from the Poisson-CAR model ignores the underlying differences in different areas. The posterior estimation of β_0 (not shown) shows a similar pattern, which is positive in the upper peninsula cluster but negative in the lower peninsula clusters. This reveals, when there is a very low poverty rate, the log relative risk of lung cancer is higher in upper peninsula than in lower peninsula.

The DIC values show the cluster model is better than the Poisson-CAR model (Cluster=719.3 and CAR=733.8). Also the Bayes factor 746.3 (Cluster vs. CAR) shows a strong support to the cluster model.

5.2 Scotland lip cancer data

The Scotland lip cancer data consists of the observed counts and expected counts of the lip cancer cases from 1975 to 1980 in each of its 56 counties. The covariate is $x_i = \text{PcAFF}_i/10$,



(a) Clustering (b) β_{1j} (c) β_{0j} (d) σ_j^2

Figure 8: Estimation results for Scotland lip cancer data.

where PcAFF_i are the percentages of the work force employed in agriculture, fishing and forestry. The posterior mode is $k = 5$ with 90% credible interval $[4, 8]$. The 5-cluster posterior estimate is given in Figure 8 (a). The estimation of relative risks (not shown) indicates that there is not much difference between the cluster model and the Poisson-CAR model. The model parameter estimates based on the Poisson-CAR model are $\hat{\beta}_1 = 0.38([0.16, 0.58])$, $\hat{\beta}_0 = -0.31([-0.95, 0.38])$, where the numbers in the brackets are 90% credible intervals. The estimates from the Poisson-CAR model shows a significant positive linear association between the log relative risks and the covariate. Figure 8 (b)-(d) shows the posterior estimates of model parameters using cluster-wise estimation **M2** for the cluster model. Unlike the single estimate obtained by the Poisson-CAR model, the cluster model is able to detect the differences of model parameters across the clusters. For example, $\hat{\beta}_0$ for southern Scotland is mainly negative while being positive for northern Scotland. The southeastern Scotland tends to have a higher noise level than the rest. Although the general pattern on $\hat{\beta}_1$ shows positiveness, we can see a stronger linear association in southeastern Scotland and a weaker linear association in middle southern part. Table 4 gives the posterior estimates of β_1 with 90% credible interval from which we can see a significant positive linear relationship in two clusters (Cluster 1 and 2). The DIC values are not quite different between the cluster model and Poisson-CAR model. And the Bayes factor (Cluster vs. CAR) equals to 11.4, which

Table 4: Posterior estimates of β_1 for Lip Cancer data

Cluster	Posterior Estimate	Credible Interval
1	0.7350	$[0.1998, 1.1469]$
2	0.5601	$[0.2733, 0.7975]$
3	0.0718	$[-0.7607, 0.3758]$
4	0.4305	$[-0.3538, 1.2129]$
5	0.3762	$[-0.3223, 0.7996]$

indicates a slight preference to the cluster model.

6 Conclusions and Discussions

The model we proposed in this article can perform the dual tasks of spatial clustering detection and regression on the log relative risks in an interactive manner. The discontinuity in the regression modeling will guide the clustering. On the other hand, the spatial clustering structure will improve the knowledge of similarity and dissimilarity of regression models on different areas. Through the simulation study, we successfully showed the effectiveness of our cluster model in the estimation of both clustering structure and log relative risks when a strong clustering tendency presents. Even under a weak clustering setting, our model can still provide a superior estimation of the disease risks over the Poisson-CAR model.

The model design can also be changed based on the research question of interest. The model we described in this paper is a "full" model in the sense that it allows the cluster-wise variation in all the cluster-specific model parameters, slope β_{1j}, intercept β_{0j}, and within-cluster noise level σ_j^2. Such comprehensive design enables dynamic values in the estimation of log relative risks to capture the disease mapping discontinuities. If the main goal is to obtain a good disease risk estimation, we recommend the use of the "full" model. On the other hand, the clustering result contains mixed information as a consequence. The spatial clustering is actually due to the differences in the combination of model parameters $(\beta_{0j}, \beta_{1j}, \sigma_j^2)$, instead of just any single parameter. If the research question cares more about the clustering structure induced by partial model parameters, we can alternatively redesign a "reduced" model correspondingly. For example, if we want to obtain the spatial clustering induced by the difference in slope β_1 only, the "reduced" model we can design is as follow:

$$y_i|\nu_i \overset{\text{indep.}}{\sim} \text{Poisson}(E_i e^{\nu_i})$$

$$\nu_i = \beta_0 + x_i\beta_{1j} + u_j + \epsilon_i \quad \text{for } \forall i \in C_j$$

$$u_j \overset{i.i.d}{\sim} N(0, \sigma_1^2) \;\; \text{and} \;\; \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma_2^2) \;\; i = 1, 2, \ldots, n, \; j = 1, 2, \ldots, k$$

As we can see, the variations of the intercept and the within-cluster noise level are eliminated from the "full" model in the "reduced" model. Under this redesigned model, the spatial clustering result will be purely determined by the dissimilarity in slope only. However, it does also pose an effect on the disease risk estimation because it loses certain flexibilities. Generally speaking, it is a trade-off relation between the spatial clustering determined by concrete information and the delicate estimation of disease risks. The model developed in Knorr-Held and Raßer (2000) is a special "reduced" model under the "full" model. In our model description, one can easily include multiple covariates.

Acknowledgement

The research is partially supported by NSF grants DMS-1106450 and SES-0961649.

References

[1] Besag, J., York, J. C. and Mollié, A. (1991), Bayesian image restoration, with two applications in spatial statistics (with discussion), *Ann. Inst. Statist. Math.*, **43**, 1-59.

[2] Booth, J. G., Casella, G. and Hobert, J. P. (2008), Clustering using objective functions and stochastic search, *J. R. Statist. Soc.* B, **70**, Part 1, 119-139.

[3] Clayton, D. G. and Bernardinelli, L. (1992), Bayesian methods for mapping disease risk, In *Small Area Studies in Geographical and Environmental Epidemiology*, Ed. J. Cuzick and P. Elliott, 205-220, Oxford University Press.

[4] Clayton, D. G. and Kaldor, J. (1987), Empirical Bayes estimates of age-standardized relative risks for use in disease mapping, *Biometrics* **43**, 671-681.

[5] Denison, D. G. T. and Holmes, C. C. (2001), Bayesian partitioning for estimating disease risk, *Biometrics*, **57**, 143-149.

[6] Elavarasi, S. A., Akilandeswari, J. and Sathiyabhama, B. (2011), A survey on partition clustering algorithms, *International Journal of Enterprise Computing and Business Systems* (Online), **1**, Issue 1.

[7] Gelman, A., Carlin J.B., Stern H.S. and Rubin D.B. (2004), “Bayesian Data Analysis”, Publisher: CRC Press, Boca Raton.

[8] Green, P. J. (1995), Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrika*, **82**, 711-732.

[9] Jin, X., Carlin, B. P., Banerjee, S. (2005), Generalized hierarchical multivariate CAR models for areal data, *Biometrics*, **61**, 950-961.

[10] Kannan, R., Vempala, S. and Vetta, A. (2004), On clusterings: Good, bad and spectral, *Journal of the ACM*, **51**, No. 3, 497-515.

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

[12] Mollié (1996), Bayesian mapping of disease, In *Markov Chain Monte Carlo in Practice*, Ed. W. R. Gilks, S. Richardson and D. J. Spiegelhalter, 359-379, Chapman & Hall, New York, NY, USA.

[13] Ng, A., Jordan, M. and Weiss, Y. (2001), On spectral clustering: Analysis and an algorithm, *Advances in Neural Information Processing Systems*, 849-856.

[14] Shi, J. and Malik, J. (2000), Normalized cuts and image segmentation, *IEEE Transactions on PAMI*, **22**, No. 8, 888-905.

[15] Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database, "Mortality - All COD, Aggregated With State, Total U.S. (1969-2007) <Katrina/Rita Population Adjustment>", National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released June 2010. Underlying mortality data provided by NCHS (www.cdc.gov/nchs).

[16] Tadesse, M., Sha, N. and Vannucci, M. (2005), Bayesian variable selection in clustering high-dimensional data, *Journal of the American Statistical Association*, **100**, 602-617.

[17] Wakefield, J. (2007), Disease mapping and spatial regression with count data, *Biostatistics*, **8**, 158-183.

A Reversible jump MCMC algorithm

Due to the clustering structure, the dimension of the parameter space in our hierarchical model changes along with the number of the clusters k. Green (1995) proposed a reversible jump MCMC which can handle the dimension changing problem. To fulfill the dimension-matching requirement, the auxiliary variable(s) are generated. The acceptance probability α is formulated in the following format

$$\alpha = \min(1, \mathcal{L} \times \mathcal{A} \times \mathcal{P} \times \mathcal{J}), \tag{5}$$

where $\mathcal{L}$ is the likelihood ratio, $\mathcal{A}$ is the prior ratio, $\mathcal{P}$ is the proposal ratio for the augmenting variable(s), and $\mathcal{J}$ is the Jacobian for the change of parameters.

We consider five update steps: *Birth Step* that adds one more cluster, *Death Step* that removes one cluster, *Latent Step* that updates $\boldsymbol{\nu}$, *Parameter Step* that updates $(\beta_1, \ldots, \beta_k, \sigma_1^2, \ldots, \sigma_k^2, \sigma^2)$ and *Shift Step* that improves the mixing behavior. *Birth Step* and *Death Step* involve dimension-chaning updates.

A.1 Birth Step

Suppose we have k clusters at the current stage with $G_k = (g_1, \ldots, g_k)$ as the cluster centers. The *Birth Step* includes the following components:

(1) Determine a new cluster center by uniformly select one area from the $(n-k)$ areas which are not cluster centers. Insert the new cluster center into G_k to form the G_{k+1} by uniformly select one position among the $(k+1)$ positions. Then, a new cluster partition will been formed via the minimum distance criterion. We denote the extra cluster as C^*.

(2) Generate the parameters (β^*, σ_*^2) for C^*. Denote the number of areas, the design matrix, the response vector, the variance-covariance matrix and the log relative risks in the new cluster C^* as n^*, X^*, $\boldsymbol{y}^*$, Σ^* and $\boldsymbol{\nu}^*$ respectively. And the priors are

$\beta^* \sim N_p(\mu^*, V^*)$, $\sigma_*^2 \sim \mathrm{IG}(a^*, b^*)$. The joint proposal density is

$$f(\beta^*, \sigma_*^2 \,|\, \boldsymbol{\nu}^*, \sigma^2) \propto \\ \frac{1}{(\sigma_*^2)^{(n^*-1)/2}} \cdot \frac{1}{(\sigma_*^2 + n^*\sigma^2)^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\nu}^* - X^{*\prime}\beta^*)'\Sigma^{*-1}(\boldsymbol{\nu}^* - X^{*\prime}\beta^*)\right) \\ \times \exp\left(-\frac{1}{2}(\beta^* - \mu^*)'V^{*-1}(\beta^* - \mu^*)\right)\left(\frac{1}{\sigma_*^2}\right)^{a^*+1} \exp\left(-\frac{b^*}{\sigma_*^2}\right) \tag{6}$$

We can generate (β^*, σ_*^2) in a hierarchical way. First, use the Greedy-Grid search to generate σ_*^2 from the following density:

$$f(\sigma_*^2|\boldsymbol{\nu}^*, \sigma^2) \propto \\ \left(\frac{1}{\sigma_*^2}\right)^{a^*+(n^*+1)/2} \exp\left(-\frac{b^*}{\sigma_*^2}\right) \cdot \frac{1}{(\sigma_*^2 + n^*\sigma^2)^{1/2}} \cdot \frac{1}{\det(A)^{1/2}} \cdot \exp\left(-\frac{1}{2}R + \frac{1}{2}B'A^{-1}B\right)$$

where

$$A = X^*\Sigma^{*-1}X^{*\prime} + V^{*-1}$$
$$B = X^*\Sigma^{*-1}\boldsymbol{\nu^*} + V^{*-1}\mu^*$$
$$R = \boldsymbol{\nu^{*\prime}}\Sigma^{*-1}\boldsymbol{\nu^*} + \mu^{*\prime}V^{*-1}\mu^*$$

Then we can generate β^* from the following Gaussian density with additionally conditioning on σ_*^2: $\beta^*|\sigma_*^2, \boldsymbol{\nu}^*, \sigma^2 \sim N_p(A^{-1}B, A^{-1})$.

(3) Update the log relative risks in C^*. Let's denote the updated log relative risks in the new cluster C^* as $\tilde{\boldsymbol{\nu}}^*$. The proposal density is

$$f(\tilde{\boldsymbol{\nu}}^*|\boldsymbol{y}^*, \beta^*, \sigma_*^2, \sigma^2) \propto \exp\left(-\frac{1}{2}(\tilde{\boldsymbol{\nu}}^* - X^{*\prime}\beta^*)'\Sigma^{*-1}(\tilde{\boldsymbol{\nu}}^* - X^{*\prime}\beta^*)\right)\exp\left(\sum_{i\in C^*} y_i\tilde{\nu}_i - \sum_{i\in C^*} E_i e^{\tilde{\nu}_i}\right)$$

Here, we approximate $e^{\tilde{\nu}_i}$ by its Taylor expansion up to the second order around the log relative risk value in the previous iteration ν_i, i.e. $e^{\tilde{\nu}_i} \approx e^{\nu_i} + (\tilde{\nu}_i - \nu_i)e^{\nu_i} + \frac{1}{2}(\tilde{\nu}_i - \nu_i)^2 e^{\nu_i} = (e^{\nu_i} - \nu_i e^{\nu_i})\tilde{\nu}_i + \left(\frac{1}{2}e^{\nu_i}\right)\tilde{\nu}_i^2 + c$, where c is a constant free of $\tilde{\nu}_i$. Alternatively, one can follow the idea of the Laplace approximation by expanding around the mode. In the algorithm, it needs an extra step to search for the mode, for example via the penalized iterative least squares (PIRLS) algorithm presented in Bates (2011b). For an efficient algorithm design, we prefer the former expansion. Then, the proposal density to obtain the updated value $\tilde{\boldsymbol{\nu}}^*$ is given by

$$\tilde{\boldsymbol{\nu}}^*|\boldsymbol{y}^*, \beta^*, \sigma_*^2, \sigma^2 \sim N_{n^*}((\Sigma^{*-1} + Q^*)^{-1}(\Sigma^{*-1}X^{*\prime}\beta^* + P^*), (\Sigma^{*-1} + Q^*)^{-1}) \tag{7}$$

where P^* is a n^*-dimensional vector with the entries of $\{y_i - E_i e^{\nu_i} + E_i\nu_i e^{\nu_i} : i \in C^*\}$ and the Q^* is a $n^* \times n^*$ diagonal matrix with the diagonal entries as $\{E_i e^{\nu_i} : i \in C^*\}$.

For convention, we denote the log relative risks of all areas in the previous iteration (i.e. before the update) as $\boldsymbol{\nu}$ and those after the update as $\tilde{\boldsymbol{\nu}}$. From the description of the updating procedure, it's clear that the only differences between $\boldsymbol{\nu}$ and $\tilde{\boldsymbol{\nu}}$ are the log relative risks in the new cluster C^*.

(4) Calculate the acceptance probability α. In the birth step, the state transits from $(k, G_k, \boldsymbol{\theta}_k)$ to $(k+1, G_{k+1}, \boldsymbol{\theta}_{k+1})$ where $\boldsymbol{\theta}_k = (\boldsymbol{\nu}, \beta_1, \ldots, \beta_k, \sigma^2, \sigma_1^2, \ldots, \sigma_k^2)$ and $\boldsymbol{\theta}_{k+1} = (\tilde{\boldsymbol{\nu}}, \beta_1, \ldots, \beta_k, \beta^*, \sigma^2, \sigma_1^2, \ldots, \sigma_k^2, \sigma_*^2)$. By taking the auxiliary variables $\mathbf{U} = (\beta^*, \sigma_*^2, \tilde{\boldsymbol{\nu}}^{\boldsymbol{*}})$ and $\mathbf{U}^* = \boldsymbol{\nu}^*$, the invertible deterministic function $\mathbf{q}_{k,k+1}(\boldsymbol{\theta}_k, \mathbf{U}) = (\boldsymbol{\theta}_{k+1}, \mathbf{U}^*)$ maintains the dimensionality during the Markov chain transition.

Based on the format in equation (5), we have the likelihood ratio

$$\mathcal{L} = \frac{L(\boldsymbol{y}|k+1, G_{k+1}, \boldsymbol{\theta}_{k+1})}{L(\boldsymbol{y}|k, G_k, \boldsymbol{\theta}_k)} = \prod_{i \in C^*} \exp\left[(\tilde{\nu}_i - \nu_i)y_i - E_i(e^{\tilde{\nu}_i} - e^{\nu_i})\right],$$

the prior ratio

$$\begin{aligned}\mathcal{A} &= \frac{\pi(k+1, G_{k+1}, \boldsymbol{\theta}_{k+1})}{\pi(k, G_k, \boldsymbol{\theta}_k)} \\ &= \frac{\pi(k+1)}{\pi(k)} \frac{\pi(G_{k+1}|k+1)}{\pi(G_k|k)} \frac{p(\tilde{\boldsymbol{\nu}}|k+1, G_{k+1}, \boldsymbol{\theta}_{k+1})}{p(\boldsymbol{\nu}|k, G_k, \boldsymbol{\theta}_k)} \pi(\beta^*)\pi(\sigma_*^2) \\ &= \frac{1-d}{n-k} \cdot \frac{p(\tilde{\boldsymbol{\nu}}|k+1, G_{k+1}, \boldsymbol{\theta}_{k+1})}{p(\boldsymbol{\nu}|k, G_k, \boldsymbol{\theta}_k)} \pi(\beta^*)\pi(\sigma_*^2), \text{ and}\end{aligned}$$

the proposal ratio

$$\begin{aligned}\mathcal{P} &= \frac{g(k, G_k|k+1, G_{k+1}) h(\mathbf{u}^*|k+1, G_{k+1}, \boldsymbol{\theta}_{k+1}, k, G_k)}{g(k+1, G_{k+1}|k, G_k) h(\mathbf{u}|k, G_k, \boldsymbol{\theta}_k, k+1, G_{k+1})} \\ &= (n-k) \frac{P(\mathit{Death\ Step})}{P(\mathit{Birth\ Step})} \frac{1}{f(\beta^*, \sigma_*^2|\boldsymbol{\nu}^*, \sigma^2)} \frac{p_{\tilde{\boldsymbol{\nu}}^*}(\boldsymbol{\nu}^*)}{p_{\boldsymbol{\nu}^*}(\tilde{\boldsymbol{\nu}}^*)},\end{aligned}$$

where $f(\beta^*, \sigma_*^2|\boldsymbol{\nu}^*, \sigma^2)$ is given by equation (6) and $p_{\cdot}(\cdot)$ is the density of the multivariate normal distribution in (7).

The Jacobian is

$$\mathcal{J} = |\mathbf{J}| = \left| \left. \frac{\mathrm{d}\mathbf{q}_{k,k+1}(\boldsymbol{\theta}, \mathbf{u})}{\mathrm{d}(\boldsymbol{\theta}, \mathbf{u})} \right|_{(\boldsymbol{\theta}, \mathbf{u}) = (\boldsymbol{\theta_k}, \mathbf{U})} \right| = 1.$$

A.2 Death Step

In the *Death Step*, the status moves from $(k+1, G_{k+1}, \boldsymbol{\theta}_{k+1})$ to $(k, G_k, \boldsymbol{\theta}_k)$. Following procedures are involved in this step.

(1) A discrete unform random variable j over $\{1, \ldots, k+1\}$ is generated to determine the cluster center g_j and the corresponding β_j and σ_j^2 which will be removed. Then the cluster C_j disappears and is merged to the rest of the k clusters according to the *minimum distance criterion*.

(2) Update $\boldsymbol{\nu_j} = \{\nu_i : i \in C_j\}$. Note that entries in $\boldsymbol{\nu}_j$ are reassigned to one of remaining clusters. Thus, we update $\boldsymbol{\nu}_j$ with new cluster membership. Let $\boldsymbol{\nu}_{j(i)}$ denote the subset of $\boldsymbol{\nu}_j$ that is merged into C_i for $i = 1, \ldots, j-1, j+1, \ldots, k+1$. Without loss of

generality, we assume $\boldsymbol{\nu}_{j(i)}$ is non-empty. The updated value is denoted as $\tilde{\boldsymbol{\nu}}_{j(i)}$. The size of $\boldsymbol{\nu}_{j(i)}$, the design matrix, the response vector are denoted as $n_{j(i)}$, $X_{j(i)}$, $\boldsymbol{y}_{j(i)}$ respectively. Then, we have

$$\tilde{\boldsymbol{\nu}}_{j(i)}|\boldsymbol{\nu}_i, \beta_i, \sigma_i^2, \sigma^2 \sim N(\overline{\mu}, \overline{\Sigma}),$$

where $\overline{\mu} = X'_{j(i)}\beta_i + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{\nu}_i - X'_i\beta_i)$ and $\overline{\Sigma} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ with $\Sigma_{11} = \sigma_i^2\mathbf{I}_{n_i} + \sigma^2\mathbf{J}_{n_i}$, $\Sigma_{12} = \sigma^2\mathbf{J}_{n_i\times n_{j(i)}}$, $\Sigma_{21} = \sigma^2\mathbf{J}_{n_{j(i)}\times n_i}$.

Similar to (7), the proposal density is given by

$$\tilde{\boldsymbol{\nu}}_{j(i)}|\boldsymbol{y}_{j(i)}, \beta_i, \sigma_i^2, \sigma^2 \sim N_{n_{j(i)}}((\overline{\Sigma}^{-1} + Q_{j(i)})^{-1}(\overline{\Sigma}^{-1}\overline{\mu} + P_{j(i)}), (\overline{\Sigma}^{-1} + Q_{j(i)})^{-1})$$

where P and Q matrices have the same definitions as before.

(3) The acceptance probability α is the reciprocal of the one in the *Birth Step*.

A.3 Latent Step and Parameter Step

In the *Latent Step*, we update the $\boldsymbol{\nu}$. Eventually, we update $\boldsymbol{\nu}$ values in each cluster independently based on the proposal density given by (7). Furthermore, when we determine the acceptance probability, we also do it cluster-wisely. The cluster-wise operation can help us improve the convergence performance of the MCMC algorithm. The acceptance probability for the C_j is

$$\alpha = \min(1, \mathcal{L}\cdot\mathcal{A}\cdot\mathcal{P})$$

where the likelihood ratio $\mathcal{L}$ remains the same as before, and the prior ratio $\mathcal{A}$ and the proposal ratio $\mathcal{P}$ are

$$\mathcal{A} = \frac{p(\tilde{\boldsymbol{\nu}}_j|\beta_j, \sigma^2, \sigma_j^2, k, G_k)}{p(\boldsymbol{\nu}_j|\beta_j, \sigma^2, \sigma_j^2, k, G_k)}, \qquad \mathcal{P} = \frac{p_{\tilde{\boldsymbol{\nu}}_j}(\boldsymbol{\nu}_j)}{p_{\boldsymbol{\nu}_j}(\tilde{\boldsymbol{\nu}}_j)}$$

$(\beta_1, \ldots, \beta_k, \sigma_1^2, \ldots, \sigma_k^2, \sigma^2)$ are updated in the *Parameter Step.* A hybrid Gibbs sampling can be applied here via the following proposals. For $j = 1, \ldots, k$,

$$\beta_j|\cdot \sim N_p(A_j^{-1}B_j, A_j^{-1})$$

where $A_j = X_j\Sigma_j^{-1}X'_j + V_j^{-1}$ and $B_j = X_j\Sigma_j^{-1}\boldsymbol{\nu}_j + V_j^{-1}\mu_j$

$$\begin{aligned}\sigma_j^2|\cdot \propto & \left(\frac{1}{\sigma_j^2}\right)^{\frac{n_j-1}{2}+a_j+1} \exp\left(-\frac{1}{\sigma_j^2}\left[\frac{1}{2}(N_j - X'_j\beta_j)'(N_j - X'_j\beta_j) + b_j\right]\right) \\ & \cdot \frac{1}{(\sigma_j^2 + n_j\sigma^2)^{1/2}} \exp\left(\frac{\sigma^2}{2\sigma_j^2(\sigma_j^2 + n_j\sigma^2)}(N_j - X'_j\beta_j)'\boldsymbol{J}_{n_j}(N_j - X'_j\beta_j)\right)\end{aligned}$$

and

$$\begin{aligned}\sigma^2|\cdot \propto & \left(\frac{1}{\sigma^2}\right)^{a+1} \exp\left(-\frac{b}{\sigma^2}\right) \\ & \cdot \prod_{j=1}^{k} \frac{1}{(\sigma_j^2 + n_j\sigma^2)^{\frac{1}{2}}} \exp\left(\frac{\sigma^2}{2\sigma_j^2(\sigma_j^2 + n_j\sigma^2)}(N_j - X'_j\beta_j)'\boldsymbol{J}_{n_j}(N_j - X'_j\beta_j)\right)\end{aligned}$$

Since the proposal densities for σ_j^2 and σ^2 are not from the known distributions, we consider the Greedy-grid sampling algorithm.

Also, we update d from the posterior density of d which depends only on k.

A.4 Shift Step

As noted in Knorr-Held and Raßer (2000), moving one cluster center to its neighborhood can improve the mixing performance of the MCMC.

(1) Suppose we have k clusters at the current stage. Among the k cluster centers, there are $n(G_k)$ of them having at least one neighborhood who is not a cluster center. Choose one cluster center uniformly out of the $n(G_k)$ cluster centers, and denote it as g_j. Suppose g_j has $m(g_j)$ neighborhood areas who are not cluster centers. Then choose one area uniformly out of the $m(g_j)$ areas as the new cluster center g_j^* to replace the original g_j in G_k. Denote the new set of the cluster centers as $\tilde{G}_k$. Note that, even though the order of the cluster centers is not changed, $(k, \tilde{G}_k)$ may still determine a different cluster partition via the *minimum distance criterion.*

(2) The acceptance probability α is

$$\alpha = \min(1, \mathcal{L} \cdot \mathcal{A} \cdot \mathcal{P})$$

The likelihood ratio $\mathcal{L} = 1$. The prior ratio is

$$\mathcal{A} = \frac{\pi(k, \tilde{G}_k, \boldsymbol{\theta_k})}{\pi(k, G_k, \boldsymbol{\theta_k})} = \frac{p(\boldsymbol{\nu}|k, \tilde{G}_k, \beta_1, \ldots, \beta_k, \sigma^2, \sigma_1^2, \ldots, \sigma_k^2)}{p(\boldsymbol{\nu}|k, G_k, \beta_1, \ldots, \beta_k, \sigma^2, \sigma_1^2, \ldots, \sigma_k^2)},$$

and the proposal ratio is

$$\mathcal{P} = \frac{g(k, G_k|k, \tilde{G}_k)}{g(k, \tilde{G}_k|k, G_k)} = \frac{n(G_k)m(g_j)}{n(\tilde{G}_k)m(g_j^*)}.$$

B Adapted Ng-Jordan-Weiss algorithm and analysis under the ideal situation

B.1 Adapted Ng-Jordan-Weiss algorithm

Suppose we want to obtain k clusters on n areas. $S_{n\times n}$ is the similarity matrix defined in Section 3.1. The algorithm contains the following steps:

1. Define D as a diagonal matrix whose ith diagonal entry is the sum of S's ith row. Then create the matrix $L = D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$.

2. Find the k largest eigenvalues of L and stack the corresponding eigenvectors in columns to form a matrix. When there are duplicate eigenvalues, choose the eigenvectors to be orthogonal to each other. Normalize each row of the matrix to have unit length. Denote the finalized matrix as $X \in \mathcal{R}^{n\times k}$.

3. Treat each row of X as a point in $\mathcal{R}^k$, and cluster them into k clusters via K-means. The cluster assignment of each row indicates the clustering membership of the respective area.

B.2 Analysis of algorithm under ideal situation

For simplicity of discussion, let's assume the first n_1 areas belong to Cluster 1, the next n_2 areas belong to Cluster 2 and so on. The ideal situation occurs when

$$S_{p,q} = \begin{cases} 1 & \text{if area } p \text{ and } q \text{ belong to the same cluster} \\ 0 & \text{otherwise} \end{cases}$$

Then

$$S = \begin{pmatrix} S_{11} & & \text{\huge 0} \\ & \ddots & \\ \text{\huge 0} & & S_{kk} \end{pmatrix} \quad \text{and} \quad D = \begin{pmatrix} D_{11} & & \text{\huge 0} \\ & \ddots & \\ \text{\huge 0} & & D_{kk} \end{pmatrix}$$

where S_{ii} is $n_i \times n_i$ matrix with all entries as 1 and D_{ii} is $n_i \times n_i$ diagonal matrix with all diagonal entries as n_i, $i = 1, \dots, k$.

$$L = D^{-\frac{1}{2}} S D^{-\frac{1}{2}} = \begin{pmatrix} L_{11} & & \text{\huge 0} \\ & \ddots & \\ \text{\huge 0} & & L_{kk} \end{pmatrix}$$

with $L_{ii} = n_i^{-1} S_{ii}$.

For the block-diagonal matrix L, the set of its eigenvalues is the union of the eigenvalues of each block matrix L_{ii}. Also, It is easy to see that the eigenvalues of L_{ii} are 1 and $(n_i - 1)$ 0's. Therefore, the k largest eigenvalues of L are k 1's. Then $X = [x_1 x_2 \dots x_k]$ where $x_i \in \mathcal{R}^n$ is defined as a vector of zeroes except that the entries at between $(\sum_{j=1}^{i-1} n_j + 1)$ and $(\sum_{j=1}^{i} n_j)$ are 1 for $i = 1, \dots, k$.

With the nice and discrete form of X, the final step k-means will yield the true clustering for the algorithm.