Published in final edited form as:

Environmetrics. 2014 March ; 25(2): 84–96. doi:10.1002/env.2265.

Space-Time Areal Mixture Model: Relabeling Algorithm and Model Selection Issues

M.M. Hossain¹, A.B. Lawson², B. Cai³, J. Choi⁴, J. Liu³, and R. S. Kirby⁵

¹Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

²Division of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina, Charleston, SC, USA

³Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC, USA

⁴Department of Mathematics, Hanyang University, South Korea

⁵Department of Community and Family Health, University of South Florida, Tampa, FL, USA

Abstract

With the growing popularity of spatial mixture models in cluster analysis, model selection criteria have become an established tool in the search for parsimony. However, the label-switching problem is often inherent in Bayesian implementation of mixture models and a variety of relabeling algorithms have been proposed. We use a space-time mixture of Poisson regression models with homogeneous covariate effects to illustrate that the best model selected by using model selection criteria does not always support the model that is chosen by the optimal relabeling algorithm. The results are illustrated for real and simulated datasets. The objective is to make the reader aware that if the purpose of statistical modeling is to identify clusters, applying a relabeling algorithm to the model with the best fit may not generate the optimal relabeling.

Keywords

Space-time mixture model; homogeneous covariate effect; relabeling algorithm; loss function; DIC

1 Introduction

In small area health or environmental investigations, often the interest lies in detecting clusters of high risk areas. Mixture models provide a convenient framework in which homogeneous observations are grouped together into clusters for discriminating areas with various risk levels. Bayesian approaches to mixture modeling provide a parametric framework within which the number of mixing components can be estimated by a reversible jump MCMC algorithm (Richardson and Green, 1997) or a birth-death MCMC algorithm

(Stephens, 2000a). In the nonparametric setting, using Dirichlet process priors may be an alternative method for estimating the number of components (Green and Richardson, 2001). An application of these algorithms was illustrated by Fernandez and Green (2002) by proposing a spatial mixture of Poisson models with a variable number of mixing components (see, Meligkotsidou (2007) for a multivariate extension of Poisson mixtures). The mixing weights in their models were spatially-dependent and estimated from a Markov random field in order to incorporate spatial neighborhood information. The covariates in mixture models can be allowed to have homogeneous or heterogeneous effects. Following Viallefont et al. (2002), the covariate effects are homogeneous if they are the same for all the mixing components but vary with the model assumption (e.g., spatially, temporally or spatio-temporally), otherwise the effects are heterogeneous. Hossain et al. (2012) extended the Fernandez and Green (2002) model to the space-time domain with homogeneous covariate effects. In their model, the weights are space-time dependent, and the authors employed a Markov random field to explain the spatial heterogeneity and an autoregressive-type model for temporal heterogeneity.

Although reversible jump or birth-death MCMC methods are conceptually intriguing approaches when the number of components in a mixture model is considered as unknown, they are computationally intensive; the implementation requires application specific software; the lack of diagnostic tools hamper the wider adoption and ascertaining the validity of the inference; and objectively quantifying the model uncertainty remains challenging (Hastie and Green, 2012). When the number of mixing components is relatively small, a simpler way to estimate the number is by comparing the values for the model selection criteria calculated for various mixture models with fixed number of components (Lawson et al., 2010). Then, the number of components is chosen for which the value of model selection criterion is minimized. In doing this, using the model selection criteria, e.g., deviance information criteria (DIC), in order to find the most parsimonious probabilistic model are often used. For example, Banerjee and Carlin (2003) used the DIC to simplify the spatio-temporal frailty model when applied to Iowa breast cancer data.

The Bayesian implementation of mixture models is often subject to the *label-switching* problem since the likelihood estimate is invariant to the permutation of the labels, i.e. the estimates of mixture components in every iteration of MCMC algorithm is insensitive to the estimates of hidden allocation variable (Jasra et al, 2005). The presence of this problem can be verified if there is any structured pattern or the presence of multiple modes in the posterior samples of mixture components. For example, the density plots of posterior samples for the six mixing components of $\theta = (\theta_1, \dots, \theta_6)$ in Figure 1 show the presence of multimodality in all components and some level of similarity patterns in components 1 and 4 and in components 5 and 6. Given these irregularities in posterior samples, estimates based on ergodic averaging of θ will be unreliable. The data for the Figure 1 was generated from

the model $o_{it} \sim \sum_{l=1}^6 \omega_{itl} \text{Poisson}(\theta_l e_{it})$ for the county and year specific counts of South Carolina low birth weight incidences. The details of this mixture model and the South Carolina low birth weight dataset are provided in sections 2 and 6, respectively. The reason for considering six components in this illustration comes from Table 6 since the model selection criteria are the least for this model.

The earlier works (e.g., Richardson and Green, 1997) to circumvent the problem of label-switching focused on imposing an ordering restriction to the mean parameter, which they referred to as *identifiability constraints*. Stephens (2000b) demonstrated that imposing the identifiability constraint does not always produce the expected results and that the estimates could be misleading (see also, Marin et al., 2005). To resolve this issue, Stephens (2000b) proposed *relabeling algorithms* by post processing the posterior samples from a longer MCMC run generated from a model without restricting the parameter space. The labels are estimated by minimizing the posterior expected loss. The loss function proposed by Stephens (2000b) for the relabeling algorithms depends on the posterior estimates of the mixing weights. Another way to implement the relabeling algorithm is by using the *posterior similarity matrix* as an approximation of true clusters and minimizing the Binder's loss function (Binder, 1978; Binder, 1981) for the estimated and the true clusters, with the posterior similarity matrix defined by pairwise coincidences in MCMC samples. Lau and Green (2007) formulated the optimization of Binder's expected loss as a binary integer programming problem. Fritsch and Ickstadt (2009) extended the adjusted Rand index for relabeling algorithm and compared it with the Binder's loss function approach. They named it as PEAR (posterior expected adjusted Rand).

Given the increasing popularity of mixture models for finding clusters of homogeneous observations, and using a relabeling algorithm for the optimum allocation of labels, it becomes essential to judge which model should be used for the relabeling algorithm. The model selection criteria that provide better predictive ability and for which the estimation of the effective number of parameters for the model with random effects is easy to conduct, may guide model selection. In this paper, we compare the relabeling algorithms with the model selection criteria for a space-time mixture of Poisson regression models with varying number of mixing components and covariates. For simplification, we considered the homogeneous covariate effects. In our simulated data example, we generated two datasets so that the number of mixing components in each dataset is known in advance. The first dataset was generated with two clusters and the second dataset with five clusters (i.e. two and five mixing components, respectively). For both datasets, three simulated covariates with various levels of spatial autocorrelation were introduced. We then compared the optimum expected loss estimates with the model selection criteria. In our real data example, we considered the county level counts of South Carolina low birth weight incidence for the years 1997 to 2007 with a number of demographic and socioeconomic covariates. The motivation for applying the space-time mixture model to this dataset comes from the histogram and density plots of observed county specific standardized incidence ratios for the year 1998 presented in Figure 2. The figure shows the presence of clustering pattern (the details are in Section 6). Since the number of clusters and the set of significant covariates in the real data example are unknown, we used the space-time mixture model in combination with the variable selection approach in determining the significant covariates and follow the values for the model selection criteria to estimate the components number. We then constructed a number of models with various combinations of covariates according to their significance level, computed the values for optimum expected loss, and compared these values with the model selection criteria.

The structure of this article is as follows. In Section 2, we describe the space-time mixture of Poisson regression model, with variable selection and model selection are described in the subsections. The estimation process and the prior distributions are illustrated in Section 3. The problem of label switching and three relabeling algorithms that use the posterior similarity matrix are described in Section 4. Sections 5 and 6 contain the simulated and real data examples, respectively. Section 7 consists of concluding remarks and a discussion of study limitations.

2 Space-Time Mixture of Poisson Regression Model

We assume that disease incidence data are observed as a count data in a fixed set of small areas and for discrete time periods. Small area units are denoted as $i = 1, \dots, n$, and temporal units as $t = 1, \dots, T$. The observed count of incidence is denoted by o_{it} for area i and year t , and the marginal distribution of o_{it} is assumed to follow a mixture of Poisson distributions and is given by

$$o_{it} \sim \sum_{l=1}^L \omega_{itl} \text{Poisson}(\theta_l e_{it} \exp(\eta_{it})),$$

where the offset, e_{it} is the expected incidence at county i and year t , and L is the number of mixing components and is assumed to be unknown. The linear predictor has the form

$$\eta_{it} = \beta_{i0}x_{it0} + \beta_{i1}x_{it1} + \dots + \beta_{ip}x_{itp}, \quad (1)$$

where x_{it0} is equal to one and X_{it} is a vector of order p . The regression coefficients are area-specific; therefore, each area has its own parameter vector $\beta_i = (\beta_{i0}, \beta_{i1}, \dots, \beta_{ip})$. We assume that the covariate effect is homogeneous among all mixing components, only varying spatially. This model is flexible in the sense that it includes two intercept parameters, component specific $\log(\theta_l)$ and area specific β_{i0} , mimicking if there are any inherent differences in the ratios of observed to expected that vary with the components or clusters in addition to the area level intercepts.

The mixing weight, ω_{itl} is assumed to be space, time and component dependent. The weights

satisfy two conditions, $0 < \omega_{itl} < 1$ and $\sum_{l=1}^L \omega_{itl} = 1$, and are determined by a hidden (or,

$$\{Z_{it}\}$$

unobserved) allocation (or, cluster) variable, $i=1, \dots, n$
 $t=1, \dots, T$ as $\omega_{itl} = p(Z_{it} = l)$. It is

modeled by a logistic transformed function $\omega_{itl} = \exp(h_{itl}/\phi) / \sum_{l=1}^L \exp(h_{itl}/\phi)$ to ensure that the above two conditions are satisfied (Fernandez and Green, 2002). Note that, by making the mixing weights vary with space and time as opposed to the traditional mixing weights where they depend only on the components, the model for the cluster variable, Z_{it} , allows borrowing information from neighboring areas and preceding times by incorporating spatial and temporal dependence functions in the specification of h_{itl} so that similar outcomes close in space and time can be assigned with similar cluster label.

The function h_{itl} in its simplest form will include spatial and temporal random effects (Hossain and Lawson, 2010); a complicated structure could include the space-time non-separable interaction effect (Knorr-Held, 2000). The parameter $\phi (> 0)$ is a scaling factor. In our application to real and simulated datasets, we considered a simpler model similar to Hossain and Lawson (2010), where $h_{itl} = \kappa_{il} + \gamma_{tl}$, the κ_l 's are the structured spatial random effects, γ_l 's are the structured temporal random effects, and the scaling factor ϕ was set to one. Fernandez and Green (2002) explained the implication of ϕ for two limiting cases on mixing weights, and mentioned that the estimation of ϕ is sensitive to the precision parameter of their Markov random field model. We also observed a similar problem and it appears that setting the value for ϕ to one is a reasonable choice.

2.1 Variable Selection

To investigate whether any particular set of covariates in (1) is more significant than others in finding the clusters with the appropriate label, applying the variable selection method in a Bayesian framework is one of the straight forward approaches. The variable selection method facilitates both quantifying the uncertainty associated with each covariate in its selection, and combining with the model selection criteria in finding the appropriate number of mixing components while using the mixture models. Following this approach may be useful in finding the most parsimonious probabilistic model with the estimate of uncertainty attached to each covariate.

Recently, O'Hara and Sillanpaa (2009) provided an in-depth summary of all the variable selection methods proposed so far in this area of research. Most commonly used methods use an auxiliary indicator variable in the linear predictor model to indicate the presence or absence of each covariate in the model, e.g., Gibbs variable selection or stochastic search variable selection. The exception is the model space approach where the reversible jump MCMC algorithm was used for the variable model dimension. Depending on model complexities and correlation between the covariates each method has its own pros and cons. By using the auxiliary indicator variable approach, the linear predictor in (1) can be written as

$$\eta_{it} = \beta_{i0}x_{it0} + \tau_1\beta_{i1}^*x_{it1} + \cdots + \tau_p\beta_{ip}^*x_{itp},$$

where τ_q is an auxiliary indicator variable to indicate the presence ($\tau_q = 1$) or absence ($\tau_q = 0$) of covariate $\mathbf{X}_q$ ($q = 1, \dots, p$).

We employed the method proposed by Kuo and Mallick (1998) to select the significant variables. The method assumes a priori that the indicator for q^{th} variable being selected or not is independent of covariate effects, so $p(\tau_q\beta_{iq}^*) = p(\tau_q)p(\beta_{iq}^*)$. Thus, independent prior distributions can be assigned to each τ_q and β_{iq}^* . The original proposal of Kuo and Mallick (1998) also suggested to use $\tau_q \sim \text{Bernoulli}(0.5)$ for the prior distribution of the auxiliary indicator variable. We preferred a hierarchical approach where $\tau_q \sim \text{Bernoulli}(a_{\tau q})$, and $a_{\tau q} \sim \text{Beta}(1,1)$, $q = 1, \dots, p$, instead of fixing $a_{\tau q}$ to a non-informative guess (i.e., 0.5). The prior specification for β_{iq}^* ($i=1, \dots, n$) will be elaborated in a later section.

2.2 Model Selection

The Bayesian implementation of mixture models provides a comprehensive framework to estimate the number of mixing components by applying the reversible jump or birth-death MCMC algorithm. However, these methods are computationally intensive, and ascertaining the validity of the inference and model uncertainty seems difficult. We chose to use model selection criteria instead, and estimated the number of mixing components for which the value of model selection criteria is the minimum. Since our space-time mixture of Poisson model is implemented in a Bayesian framework and includes random effects, we used the model selection criteria that are commonly recommended in these settings. For example, DIC (Spiegelhalter et al, 2002), DIC₃ (Celeux et al., 2006), and mean square prediction error (MSPE) (Gelfand and Ghosh, 1998). As a cautionary note, model selection criteria that are based on predictive ability (e.g., DIC) do not always guarantee greater accuracy of cluster detection since the overestimation of risks in the high risk areas often leads to superior predictive measure of overall fit. For example, Best et al. (2005) demonstrated for a variety of disease mapping models with limited simulation studies that the model selected by the smallest DIC often fails to detect clusters of high risk areas. Nonetheless, we preferred to use DIC because of its wider applications in Bayesian setting, and ease of computation.

Below we give a brief illustration for each of these model selection criteria, and the notation M was used to indicate a specific model. In our context, the model space was varying with the number of components and the set of covariates. For example, in the real data example, M was first used to indicate a model with a fixed number of clusters or components, from 2 to 7 components, and each model incorporates the variable selection approach. When the number of components was determined to a specific value (say, L), M was used again to indicate models with varying sets of covariates. In the simulated data example, since the number clusters was known a priori, M was used to indicate models with varying sets of covariates.

The DIC for the model M is defined as $DIC_M = \overline{D(\Theta_M)} + p_M$, where Θ_M is the set of all parameters under model M , $\overline{D(\Theta_M)}$ is the posterior mean deviance and p_M is the effective number of parameters measuring the model complexity. In mixture models the allocation variables indicating the group membership are discrete and the label changes over the MCMC samples, so the conventional approach of estimating deviance at $\hat{\Theta}_M$ for estimating p_M is not effective. To overcome this issue, we follow the Gelman et al (2004) proposal to measure p_M which is defined as half the posterior variance of the deviance. Celeux et al. (2006) recommended using the DIC₃ for mixture models. The DIC₃ for model M is defined as $DIC_{3M} = \overline{D(\Theta_M)} + \left[\overline{D(\Theta_M)} + 2 \log \hat{p}(\mathbf{o} | \Theta_M) \right]$, where $\log p(\mathbf{o} | \Theta_M)$ can be computed from the MCMC samples. For mixture model, it is also straightforward to use MSPE as a model selection criterion. The MSPE for model M is calculated by

$$MSPE_M = \sum_{g=1}^G \sum_{i=1}^n \sum_{t=1}^T (o_{it} - o_{it}^{(g,M)})^2 / nTG$$

, where $o_{it}^{(g,M)}$ is the predicted o for the model M at g^{th} iteration and G is the MCMC sample size. As it appears, MSPE does not separate out any specific parameter for measuring the model complexity, although Gelfand and Ghosh

(1998) illustrated that for increasing model complexity MSPE increases and has a similar performance like other model selection information criteria.

3 Posterior Sampling and the Prior Distribution Specification

The mixture of Poisson regression models was implemented with a variable number of mixing components L , and the value of it was determined by using the model selection criteria, i.e., by the minimum values of DIC, DIC₃, and MSPE. Once this number was determined, the estimation of all other parameters follows the profile likelihood approach. Appendix A provides the posterior distributions for all the parameters. A mixture of Gibbs and Metropolis-Hasting algorithms can be applied to generate posterior samples and from which all the posterior estimates can be derived. We use the WinBUGS package (Spiegelhalter et al., 2003) for generating posterior samples for each parameter and for the derivation of posterior estimates, including posterior means. The code developed for the space-time mixture of Poisson regression models with variable selection is attached as a supplementary file with the name 'WinBUGS code for mixture model with variable selection.pdf'. WinBUGS offers greater flexibility for modeling spatial data since a wide variety of spatial distributions (e.g., univariate or multivariate conditional autoregressive (CAR) models (Besag et al., 1991; Gelfand and Vounatsou, 2003)), are readily available for use. Another software, R-INLA (Rue et al., 2009) can also be used for fitting these models. This software uses Gaussian approximations to densities and may provide a similar level of flexibility and ensure quick convergences of model parameters.

Prior distribution specification and the identifiability condition

The prior distributions and the values for all the hyperpriors are chosen based on the recommendation provided in this area of research. For example, the prior distribution for the spatial dependence between the β_{iq}^* 's ($i=1, \dots, n$) will be specified by a proper CAR model, and similarly for the random effects, κ_{ij} 's, an improper CAR model will be used. Since the

regression coefficients are spatially varying, a condition $\sum_{i=1}^n \beta_{iq}^* = 0$ as mentioned in Assunção (2003) has to be imposed to ensure identifiability and the CAR models defined in WinBUGS will ensure this condition is satisfied while estimating the effects. The prior distribution of the precision parameter of CAR model is specified as *Gamma*(0.5, 0.0005) (Kelsall and Wakefield, 1999). Prior specifications of other parameters are provided in Appendix A.

4 Relabeling Algorithm

The problem of label-switching is defined as the condition where the likelihood function is invariant to the permutation of component labels. Since the inference in a Bayesian framework relies on the generated MCMC samples from the posterior distributions, the problem of label-switching is observed when the generated MCMC samples from the posterior distribution for a single observation visits more than one mode. This may happen when more than one cluster is present in the observed data and the likelihood values do not change even when the label to each observation is switched to a different value in MCMC

samples, i.e. the likelihood of a mixture model is invariant to the permutation of component labels. One approach to this problem is to use the posterior similarity matrix, and do the relabeling based on minimizing the expected loss function. The posterior similarity matrix is also known as a similarity matrix of posterior expected clustering. For our comparison to model selection criteria, we use the relabeling algorithms based on: Binder's loss function; Lau and Green criteria; and PEAR. Fritsch and Ickstadt (2009) mentioned the limitation of using PEAR as an optimality criterion for relabeling algorithm, it could fail if there is only one cluster or all are forming singleton clusters. Brief descriptions of these relabeling algorithms are provided in the next subsection.

Posterior similarity matrix

Let $\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(G)}$ be the MCMC samples of the allocation matrix $\mathbf{Z}$ where each matrix has an order of $n \times T$ for n geographical locations and T years, and G is the posterior sample size. These posterior samples are considered after ensuring the convergence of each model parameter. The elements of the matrix $\mathbf{Z}^{(g)}$ are given by $Z_{it}^{(g)} = l$, implying that in the g^{th} iteration of the MCMC, the (i, t) th observation is placed in cluster l , where $g = 1, \dots, G$ and $l = 1, \dots, L$. The element of the posterior similarity matrix for the locations i and j , and for an arbitrary year t is defined as

$$\pi_{ijt} = \Pr(Z_{it} = Z_{jt} | \mathbf{o}) \approx \frac{1}{G} \sum_{g=1}^G I(Z_{it}^{(g)} = Z_{jt}^{(g)}), \quad (2)$$

where $I(\cdot)$ is an indicator function takes value 1 or 0 depending on whether the equality condition within the parenthesis is satisfied or not, $\mathbf{o}$ is a matrix for observed space-time outcome variable of order $n \times T$, and $i, j = 1, \dots, n$, $t = 1, \dots, T$. From (2), we can derive that $E[(Z_{it} = Z_{jt}) | \mathbf{o}] = \pi_{ijt} = E[I(Z_{it} = Z_{jt}) | \mathbf{o}]$. This indicates that the posterior similarity matrix can also be viewed as a similarity matrix of posterior expected clustering. That is why the label switching does not affect the posterior similarity matrix and can be regarded as a similarity matrix for unknown true clustering.

Binder's loss function

Binder (1978) proposed a loss function based on pairwise similarities of observations. Here, the observations are the MCMC samples of allocation matrix $\mathbf{Z}$ with G realizations. The loss function for any arbitrary year t , is given by

$$L(\mathbf{Z}'_t \mathbf{Z}_t) = \sum_{i < j} l_1 \cdot I(Z'_{it} \neq Z'_{jt}) I(Z_{it} = Z_{jt}) + l_2 \cdot I(Z'_{it} = Z'_{jt}) I(Z_{it} \neq Z_{jt}),$$

where l_m ($m = 1, 2$) is a non-negative constant, and $\mathbf{Z}'_t$ is a proposed cluster estimate. The choice of l 's is critical in a sense that when $l_1 > l_2$, the expectation is that the observations are clustered into few clusters and the loss is minimized to reflect this expectation, and in the extreme all observations are in one cluster. Whereas, when $l_2 > l_1$, the loss is minimized to meet the expectation that the observations are clustered into many clusters, and in the extreme each observation forms its own cluster. When there is no information available to

determine this expectation, a reasonable assumption is to set $l_1 = l_2 = 1$ (Fritsch and Ickstadt, 2009) to allow the loss to be minimized for both ends equally.

With this choice of l_1 and l_2 being equal to one, the posterior expectation of Binder's loss is given by

$$E \left[L \left(\mathbf{Z}'_t, \mathbf{Z}_t \right) | \mathbf{o} \right] = \sum_{i < j} \left| I \left(Z'_{it} = Z_{jt} \right) - \pi_{ijt} \right|, \quad (3)$$

where $|\cdot|$ denotes the absolute value. The posterior expected loss for each MCMC sample, g , can be calculated from (3). The estimated cluster $\hat{\mathbf{Z}}_t$ for year t will be assigned to the proposed cluster $\mathbf{Z}'_t$ for which the posterior expected loss is minimum, i.e.

$$\hat{\mathbf{Z}}_t = \arg \min_g E \left[L \left(\mathbf{Z}'_t, \mathbf{Z}_t \right) | \mathbf{o} \right], \quad (4)$$

and an optimum expected loss value for this relabeling algorithm can be computed by $\sum_t \hat{\mathbf{Z}}_t$.

There is a potential problem with using (4) as a minimization rule since more than one of the MCMC realizations could attain the minimum posterior expected loss. When this happens, we randomly select a $\mathbf{Z}'_t$ among the MCMC samples with the minimum posterior expected loss as an estimate of $\hat{\mathbf{Z}}_t$. Instead of (3), Dahl (2006) proposed to use the quadratic functional

for the expectation of Binder's loss as $E \left[L \left(\mathbf{Z}'_t, \mathbf{Z}_t \right) | \mathbf{o} \right] = \sum_{i < j} \left(I \left(Z'_{it} = Z_{jt} \right) - \pi_{ijt} \right)^2$. Fritsch and Ickstadt (2009) showed that minimizing the Dahl's criteria is equivalent to minimizing the Binder's loss given by (3).

Lau and Green criteria

Lau and Green (2007) formulated the clustering algorithm as a binary integer programming problem and recommended the use of the linear function of the form

$$E \left[L \left(\mathbf{Z}'_t, \mathbf{Z}_t \right) | \mathbf{o} \right] = \sum_{i < j} I \left(Z'_{it} = Z_{jt} \right) (1 - \pi_{ijt}), \quad (5)$$

for the expectation of Binder's loss. The estimated cluster $\hat{\mathbf{Z}}_t$ for year t is obtained by (4).

Posterior Expected Adjusted Rand (PEAR)

Fritsch and Ickstadt (2009) argued that instead of minimizing the Binder's loss, a utility function given by the adjusted Rand index of estimated and true clustering $AR(\mathbf{Z}'_t, \mathbf{Z}_t)$ should be maximized. The estimated cluster $\hat{\mathbf{Z}}_t$ for year t will be assigned to the proposed cluster $\mathbf{Z}'_t$ for which the PEAR is maximum, i.e.

$$\hat{\mathbf{Z}}_t = \arg \max_g AR \left[\mathbf{Z}'_t, E \left(\mathbf{Z}_t | \mathbf{o} \right) \right], \quad (6)$$

$$AR[\mathbf{Z}'_t, E(\mathbf{Z}_t|\mathbf{o})] = \frac{\sum_{i < j} I(Z'_{it} = Z'_{jt}) \pi_{ijt} - \sum_{i < j} I(Z'_{it} = Z'_{jt}) \sum_{i < j} \pi_{ijt} / \binom{n}{2}}{\frac{1}{2} \left[\sum_{i < j} I(Z'_{it} = Z'_{jt}) + \sum_{i < j} \pi_{ijt} \right] - \sum_{i < j} I(Z'_{it} = Z'_{jt}) \sum_{i < j} \pi_{ijt} / \binom{n}{2}},$$

where,

and $AR[\mathbf{Z}'_t, E(\mathbf{Z}_t|\mathbf{o})]$ and $E[AR(\mathbf{Z}'_t, \mathbf{Z}_t)|\mathbf{o}]$ are equivalent.

5 Simulation Data Example

Our purpose in conducting the following simulation study is to evaluate how well the model selection criteria and relabeling algorithm perform for the mixture model when the covariate effects are assumed to be homogenous and have various level of spatial autocorrelation. More specifically, we address the question of whether the best model selected by the model selection criteria produces the optimum relabeling or optimum expected loss values. We anticipated that the number of clusters that existed in the simulated dataset may influence this observation. This motivated us to design two simulation studies with a small and a relatively large number of clusters, named as Simulation-1 and Simulation-2, respectively. Simulation-1 is designed to contain two clusters while Simulation-2 contains five clusters. In both simulation experiments, no specific temporal trend is assumed. Three covariates with an intercept are used to derive the homogenous covariate effect. Among the three covariate effects, one is assumed to be spatially random and the other two have medium and strong spatial autocorrelations. We anticipate that the number of clusters existing in the response variable may interact differently with the levels of spatial autocorrelation in covariates; for a smaller number of clusters the significance of spatial patterns in covariates could be higher than for a large number of clusters. In Figure 3, the maps for the β 's are given. The darker colors show the regions with higher coefficient values. The thematic maps for β_0 (top-left), β_2 (bottom-left), and β_3 (bottom-right) are depicting the presence of spatial autocorrelation, whereas the map for β_1 (top-right) has a random spatial pattern. The steps involved in generating both of the simulated datasets are given in Appendix B. We used the state of Ohio in the United States for the geographical base, with data on county- and year-specific expected lung cancer mortality for the period 1968–88 (source: www.stat.unimuenchen.de/service/datenarchiv/ohio/ohio_e.html) as a base for our simulation. The details of this dataset can be found in Waller et al., (1997). The number of replications in each simulation experiment was set at $S=100$ and was chosen as a balance between computation time and accuracy with the assumption that this size is reasonable to capture all sources of variations.

Results

The model for the simulated data with auxiliary indicator variables for the variable selection is defined by

$$\eta_{it} = \beta_{i0} + \tau_1 \beta_{i1}^* x_{it1} + \tau_2 \beta_{i2}^* x_{it2} + \tau_3 \beta_{i3}^* x_{it3}.$$

Table 1 and Table 3 give the average posterior mean estimate of the proportion of time each covariate is selected and the corresponding standard error of this estimate for Simulation-1 and Simulation-2, respectively. The averages are taken over 100 replicated datasets. As

expected in both simulations, the covariate x_3 has the highest posterior mean, whereas x_1 has the smallest posterior mean and x_2 is near the middle of covariates x_1 and x_3 . A relatively large number of clusters for Simulation-2 influences a higher average posterior mean estimate for x_1 than for Simulation-1. The standard error estimates for the covariate x_3 for both simulations remain close mainly because of strong spatial correlation in this covariate. The standard error estimates for the other two covariates increases for Simulation-2 because of the increase in cluster size.

Next, we constructed eight models with all possible combinations of covariates x_1 , x_2 and x_3 for each simulated dataset to check how the values for the model selection criteria and the optimum expected loss values change with the covariates included. The results after averaging over the replication from Simulation-1 and Simulation-2 are presented in Table 2 and Table 4, respectively. In both tables, the average values for the model selection criteria and the optimum expected loss are reported for the models with various combinations of covariates. The values for the best fitted model and the optimum expected loss are highlighted in bold. As expected, the model with covariate x_3 produces the minimum values for all the model selection criteria when the number of clusters is small (Simulation-1), indicating that the highly significant covariate x_3 produces the best model. Interestingly, the best model does not produce the optimum expected loss values. The model with covariates x_2 and x_3 produces the minimum Binder's and Lau & Green loss values, and the maximum PEAR value. The results for the presence of relatively large number of clusters (Simulation-2) are little different from those with the smaller number of clusters as indicated by the variable selection results in Table 3. The model with intercept only indicates the best model, and the model with covariate x_3 produces the optimum expected loss values. Overall, we observed for this simulated experiment with the scenarios of small and large numbers of clusters that the best model selected by the model selection criteria does not produce the optimum expected loss values.

6 Real Data Example

As with all real data, if any clusters exist, the number is always unknown. So, in addition to assessing the significance of each covariate, the number of clusters must also be estimated. We analyzed the county specific South Carolina low birth weight (LBW) incidence data for the years 1997 to 2007 to show how the values for the model selection criteria are related to the optimum expected loss values computed from the relabeling algorithm. The data are publicly available and can be acquired from the South Carolina Department of Health and Environmental Control (<http://www.scdhec.gov/>); and population, income, poverty and unemployment data were obtained from the US Census Bureau (<http://www.census.gov>). LBW is defined as a child weighting less than 2500 grams at live-birth. The county and year specific expected LBW incidence, E_{it} , is calculated by $n_{it}R$, where n_{it} is the total birth in county i and year t , and R is the South Carolina overall LBW incidence rate calculated by the ratio of total LBW to total birth over the entire spatial-temporal domain. The histogram and density plots of observed county specific standardized incidence ratios (SIRs) for one year (year: 1998) are presented in Figure 2. The figure shows the presence of clustering pattern in the observed SIRs and justify why we are using mixture models for this data set.

The number of breaks and bandwidth in this figure are estimated by using a robust method available through the `bw.SJ` function in R rather than fixing these to predetermined values.

The full model for the South Carolina LBW data is defined by

$$\eta_{it} = \beta_{i0} + \beta_{i1}PD_{it} + \beta_{i2}PAA_{it} + \beta_{i3}MHI_{it} + \beta_{i4}PP_{it} + \beta_{i5}UR_{it} + \beta_{i6}t$$

where the county and year specific covariates are: population density (PD), proportion of African-American (PAA), median household income (MHI), proportion in poverty (PP), and unemployment rate (UR). The last coefficient in the above equation, the parameter, β_{i6} ; measures the county-specific temporal effect and can be interpreted as a space-time interaction effect (Assunção, 2003).

When data are observed spatially, determining the number of clusters and selecting significant variables simultaneously is a complex issue. Two methods have recently been proposed for nonspatial data. One method combined the Dirichlet process mixture with the variable selection method (Kim et al., 2006), while the other is the reversible jump MCMC with the variable selection method (Tadesse et al., 2005). Since we are dealing with spatial data, we follow an approach that applies the variable selection method with a fixed components number, and determines the components number by following the model selection criteria. Following the variable selection terminology, the above full model can be written as

$$\eta_{it} = \beta_{i0} + \tau_1\beta_{i1}^*PD_{it} + \tau_2\beta_{i2}^*PAA_{it} + \tau_3\beta_{i3}^*MHI_{it} + \tau_4\beta_{i4}^*PP_{it} + \tau_5\beta_{i5}^*UR_{it} + \tau_6\beta_{i6}^*t.$$

All the models with the real data are implemented in WinBUGS with one chain and 100,000 iterations from which 80,000 is the burn-in period. The iteration size was determined as a compromise between computational time and relative precision. A relatively longer burn-in period was considered for removing the autocorrelation in MCMC samples since no thinning was done with the posterior samples. Table 5 gives the DIC, DIC₃, MSPE, and posterior mean estimates of each auxiliary indicator variable (with standard error estimates in parentheses) for each number of components from 2 to 7 for the full model. Model with $L = 6$ gives the minimum values for DIC, DIC₃ and MSPE (highlighted in bold). A supplementary file, with the name 'Posterior estimates - MCMC history - kernel density.pdf' attached with this submission, includes posterior estimates, history and density plots of all stochastic nodes. The convergence of each parameter was assessed by visual inspection of history and density plots of posterior samples. The history plots for the Θ 's show structured patterns which are also reflected by the multi-modal patterns in the density plots. Jasra et al. (2005) and Stephens (2000b) related this to the identifiability issues that are inherent to mixture models and proposed the use of a relabeling algorithm. The history plots for the precision parameters κ also show a similar pattern; we anticipate that changing the prior distribution, for example, to Christiansen and Morris (1997) type may improve the posterior estimates further.

For this model, it appears that space-time interaction is the most significant covariate and MHI has limited evidence of significance, although this evidence is very small. Figure 4 also provides additional supports to these results, where the county level p -values are plotted, obtained from univariate Poisson regression model for each covariate with expected count as an offset variable. The p -values for the regression coefficients are derived from testing of the null hypothesis. In each plot, along the x-axis the counties are placed in alphabetic order and a horizontal line is drawn at p -value=0.05 to help visualize how many counties are at, below or closer to this line. Although this county level univariate Poisson regression model does not adjust for all the covariate effects simultaneously and also for the spatial characteristics, the plots in Figure 4 reveal that the p -values for MHI and time are generally lower than for the covariates: PD, PAA, PP and UR. Also, the variable selection method that we applied for the results in Table 5 is insensitive to multicollinearity (O'Hara and Sillanpaa, 2009), although some of the covariates (e.g., PAA, MHI, PP and UR) are expected to be correlated.

After the number of components was determined and the significant covariates were selected, we compared the optimum expected loss values with the model selection criteria in determining how these two values changes with the change of models for various sets of covariates. We considered four models specified by: covariate MHI, space-time interaction, covariate MHI and space-time interaction, and a full model (i.e., a model with all the covariates and space-time interaction). Table 6 gives the values for the model selection criteria and the minimum expected loss values for the Binder's loss and the Lau & Green loss, and the maximum PEAR value for each model. The values for best fitted models and the optimum expected loss are highlighted in bold. As shown, all of the estimates are very close and the highest values for the model selection criteria are obtained for the space-time interaction model; whereas the model with the covariate MHI and space-time interaction, and the full model provide the lowest model selection criteria values. The lowest values for the Binder's loss and the Lau & Green loss, and the maximum PEAR value are obtained for models with space-time interaction.

7 Conclusions

Previously, Best et al. (2005) showed that the spatial model with convolution prior (e.g. Besag et al., 1991) overestimates the risk surface. Mixture modeling is a popular tool in health or environmental investigation to identify clusters since its Bayesian implementation introduces a label for each observed data point generated by an unobserved allocation variable. Fitting the model in a Bayesian framework allows generating a large number of posterior MCMC samples for the allocation variable which has the potential to resolve the label switching problem and find the appropriate label by using the relabeling algorithm. Before applying a relabeling algorithm, a model from which these posterior samples are generated must be selected. In this paper, we used a space-time mixture of Poisson regression model with homogeneous covariate effects to illustrate that the best model selected by using the model selection criteria does not always support the model if chosen by optimum relabeling algorithm.

We used both real and simulated datasets for this illustration. Two simulation studies were designed with a small and a relatively large numbers of clusters, and with common covariate

effects. The covariate effects were generated with weak to stronger levels of spatial autocorrelation. We observed that the data with a larger number of clusters tend to exclude the covariate effects in the best fitted model even though the covariates effects were significant according to the variable selection method. We suspect that the spatial autocorrelation induced in the simulation for the intercept, β_0 , may be the reason for this result. The model selected by the values for optimum expected loss always includes covariate effects, irrespective of the number of clusters. For both the simulated datasets and the real dataset, we observed that the model selected by the model selection criteria does not produce optimum expected loss values.

This research has a number of limitations. For the model selection criteria, we used only the criteria that are based on predictive performance and the results presented may be sensitive to this choice. There are other proposals to improve the DIC performance especially when the number of clusters is expected to be large. These include the use of the variational Bayes approach (McGrory and Titterton, 2007), viewing DIC as an approximate penalized loss function leading to a better estimate of effective number of parameters (Plummer, 2008), or using the model selection criteria from a predictive viewpoint, e.g., Bayesian predictive information criteria (Ando, 20012). Our choice of relabeling algorithm was limited to the methods related to the posterior similarity matrix. An alternative approach could be model-based where the label for each observation is chosen by maximizing the classification probability (Yao, 2012). We hope to address these in future research.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are gratefully acknowledging the NIH (grant # R21 HL088654-01A2) funding. Our sincere thanks go to the two reviewers and the Associate Editor for many constructive and insightful comments which contributed to the further improvement of this manuscript.

Appendix A

The likelihood function for the space-time mixture of Poisson model in Section 2 is defined as

$$L(\mathbf{o}|\mathbf{e}, \boldsymbol{\theta}, \boldsymbol{\eta}, L, \boldsymbol{\omega}) = \prod_{i=1}^n \prod_{t=1}^T \sum_{l=1}^L \omega_{itl} \text{Poisson}(\theta_l e_{it} \exp(\eta_{it})),$$

where $\eta_{it} = \beta_{i0} + \tau_1 \beta_{i1}^* x_{ij1} + \dots + \tau_p \beta_{ip}^* x_{ijp}$. The value for the mixing components number was determined by using the information criteria for which the value for information criteria is the minimum. Then, all other model parameters are estimated by following the profile likelihood approach where the mixing components number, L , is known. Using the conditional independence assumption, the joint posterior distribution of all the variables $(\boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\beta}^*, \mathbf{z}, \boldsymbol{\kappa}, \boldsymbol{\gamma}, \boldsymbol{\varphi}, \boldsymbol{\sigma}_\kappa, \boldsymbol{\sigma}_\gamma)$ is proportional to $L(\mathbf{o} | \mathbf{e}, L, \boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\beta}^*) p(\boldsymbol{\beta}^* | L, \boldsymbol{\omega}, \boldsymbol{\theta}) p(\boldsymbol{\tau} | L, \boldsymbol{\omega}, \boldsymbol{\theta}) p(\boldsymbol{\theta} | \alpha_0, \beta_0, L) p(\mathbf{z} | L, \boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{\tau}, \boldsymbol{\beta}) p(\boldsymbol{\kappa} | \boldsymbol{\sigma}_\kappa, L) p(\boldsymbol{\gamma} | \boldsymbol{\sigma}_\gamma, L) p(\boldsymbol{\sigma}_\gamma | L) p(\boldsymbol{\varphi})$. The posterior samples for each

unknown parameter are generated by the joint implementation of Gibbs and Metropolis-Hasting updating schemes. The parameter sampling schemes are given as: For each

$\beta_{q'}^*$'s $(=\beta_{iq'}^*, \dots, \beta_{nq'}^*)^T$, assuming a proper CAR prior distribution, $\beta_{iq'}^* | \{\beta_{-iq'}^*\}, \sigma_{\beta q'}^2, \rho_{\beta q'} \sim N(\rho_{\beta q'} \bar{\beta}_{iq'}^*, \sigma_{\beta q'}^2/n_i), q'=0, 1, \dots, p$ gives the posterior density as

$$\beta_{iq'}^* \left| \dots \propto N(\beta_{iq'}^* | \rho_{\beta q'} \bar{\beta}_{iq'}^*, \sigma_{\beta q'}^2/n_i) \prod_{t=1}^T \sum_{l=1}^L \omega_{itl} \text{Poisson}(o_{it} | e_{it}, \theta_l, \eta_{it}).$$

The prior distribution for the spatial autocorrelation parameter, $\rho_{\beta q'}$, is defined to have a uniform distribution with the limits specified by the minimum and maximum eigenvalues of a function of the neighborhood matrix (Stern and Cressie, 1999). Assuming an inverse gamma distribution for the variance parameter will give the posterior distribution as,

$$\sigma_{\beta q'}^2 \left| \dots \sim \text{Inv-Gamma} \left(\sigma_{\beta q'}^2 | a_{\beta} + n/2, b_{\beta} + \sum_{i=1}^n n_i (\beta_{iq'}^* - \bar{\beta}_{iq'}^*)^2 / 2 \right).$$

The posterior distribution of each $\tau_{q,q} = 1, \dots, p$ assuming a Bernoulli prior distribution with parameter $a_{\tau q}$, is

$$\tau_q | \dots \propto \text{Bernoulli}(\tau_q | a_{\tau q}) \prod_{t=1}^T \sum_{l=1}^L \omega_{itl} \text{Poisson}(o_{it} | e_{it}, \theta_l, \eta_{it}),$$

where the prior distribution for the $a_{\tau q}$, is defined to have a beta distribution with parameters 1, which is equivalent to an uniform distribution. For the structured spatial random effects $\kappa_l = (\kappa_{1l}, \dots, \kappa_{nl})$, assuming a CAR prior distribution,

$\kappa_{il} | \{\kappa_{-il}\}, \sigma_{\kappa l}^2 \sim N(\bar{\kappa}_{il}, \sigma_{\kappa l}^2/n_i), l=1, \dots, k, l=1, \dots, k$, the posterior distribution will be of the form,

$$\kappa_{il} | \dots \propto \prod_{t=1}^T \{\omega_{itl} \text{Poisson}(o_{it} | e_{it}, \theta_l, \eta_{it})\} N(\kappa_{il} | \bar{\kappa}_{il}, \sigma_{\kappa l}^2/n_i).$$

Assuming an inverse gamma distribution for the variance parameter gives the posterior distribution as,

$$\sigma_{\kappa l}^2 | \dots \sim \text{Inv-Gamma} \left(\sigma_{\kappa l}^2 | a_{\kappa} + n/2, b_{\kappa} + \sum_{i=1}^n n_i (\kappa_{il} - \bar{\kappa}_{il})^2 / 2 \right).$$

For the structured temporal random effects $\gamma_l = (\gamma_{1l}, \dots, \gamma_{Tl})$, assuming to have a first-order autoregressive prior specification, $\gamma_{tl} | \gamma_{t-1l}, \sigma_{\gamma l}^2 \sim N(\gamma_{t-1l}, \sigma_{\gamma l}^2), l=1, \dots, L$, the posterior distribution will be of the form,

$$\gamma_{it} | \dots \propto \prod_{i=1}^n \{\omega_{itl} \text{Poisson}(O_{it}|e_{it}, \theta_l, \eta_{it})\} N(\gamma_{it} | \gamma_{t-1l}, \sigma_{\gamma_l}^2).$$

The prior distribution for the inverse of $\sigma_{\gamma_l}^2$ is specified according to Christiansen and Morris (1997), in the form, $1/\sigma_{\gamma_l}^2 \sim \pi(1/\sigma_{\gamma_l}^2) = m_0(m_0 + 1/\sigma_{\gamma_l}^2)^{-2}$. This is equivalent to adopting a uniform (0, 1) prior for $B_0 = 1/\sigma_{\gamma_l}^2(m_0 + 1/\sigma_{\gamma_l}^2)^{-2}$. Setting $m_0 = 1$ is a standard choice. Assuming a categorical prior distribution with $\omega_{itl} = p(Z_{it} = l)$, the posterior distribution for the allocation variable has the form,

$$Z_{it} = l | \dots \sim \omega_{itl} \text{Poisson}(o_{it}|e_{it}, \theta_l, \eta_{it}) I(l \in \{1, \dots, L\}).$$

Finally, assuming a gamma prior distribution for the θ_l , $l = 1, \dots, L$, the posterior distribution of $\theta = (\theta_1, \dots, \theta_L)$ has the form,

$$\theta \left| \dots \sim \prod_{l=1}^L \text{Gamma}(\theta_l | a_\theta + \sum_{i,t:Z_{it}=l} o_{it}, b_\theta + \sum_{i,t:Z_{it}=l} e_{it}).$$

Candidate values for $(\beta'_{i1}, \dots, \beta'_{iq}), (\kappa'_{i1}, \dots, \kappa'_{ik})$ and $(\gamma'_{t1}, \dots, \gamma'_{tk})$ are drawn by a Metropolis-Hasting updating scheme from the posterior distributions for $(\beta^*_{i1}, \dots, \beta^*_{iq}), (\kappa^*_{i1}, \dots, \kappa^*_{ik})$, and $(\gamma_{t1}, \dots, \gamma_{tk})$, respectively. For all other parameters, samples are drawn directly by Gibbs sampling.

Appendix B

We used the county and year specific expected lung cancer mortality for the period 1968–88 of state of Ohio as a base for our simulation, yielding $n = 88$ and $T = 21$. A brief description of the simulation steps is as follows.

Simulation-1 (The number of mixing component is set at 2):

- a. Set $\omega_1 = 0.6$ (and, $\omega_2 = 0.4$), and generate $n_1 \sim \text{Binomial}(n, \omega_1)$ and $n_2 = n - n_1$,
- b. Set $\theta_1 = 1$, and generate $\theta_2 \sim \text{Uniform}(\theta_1, 3.0)$ such that $\theta_1 < \theta_2$,
- c. The covariates are generated by

$$x_{itq} \sim \text{Uniform}(0, 1); i=1, \dots, n, t=1, \dots, T, \text{ and } q=1, \dots, 3$$

- d. Set values for $\beta_{i0}, \beta_{i1} (\sim \text{Uniform}(-0.5, 0.5))$, β_{i2} , and β_{i3} (Figure 3)
- e. Calculate $\xi_{it} = e_{it} \times \exp\left(\beta_{i0} + \sum_{q=1}^3 \beta_{iq} x_{itq}\right)$, and

- f. The county and year specific lung cancer mortality count was generated from a Poisson model with:

$$y_{it}^s | \theta_1, \xi_{it} \sim \text{Poisson}(\theta_1 \xi_{it}), i=1, \dots, n_1, \text{ and } y_{it}^s | \theta_2, \xi_{it} \sim \text{Poisson}(\theta_2 \xi_{it}), i=n_1+1, \dots, n,$$

- g. where $s = 1, \dots, S$ is the size of replication.

Simulation-2 (The number of mixing component is set at 5)

- a. Set $\omega = (0.30, 0.25, 0.20, 0.15, 0.10)$, and generate $(n_1, n_2, n_3, n_4, n_5) \sim \text{Multinomial}(n, \omega)$,
- b. Generate $\theta_1 \sim \text{Uniform}(0.0, 0.5)$, $\theta_2 \sim \text{Uniform}(\theta_1, \theta_1 + 3.0)$, $\theta_3 \sim \text{Uniform}(\theta_2, \theta_2 + 3.0)$,
- $\theta_4 \sim \text{Uniform}(\theta_3, \theta_3 + 3.0)$, and $\theta_5 \sim \text{Uniform}(\theta_4, \theta_4 + 3.0)$ such that
- $\theta_1 < \theta_2 < \theta_3 < \theta_4 < \theta_5$,

Steps: iii, iv, and v, are the same as Simulation-1, and

- a. The county and year specific lung cancer mortality count was generated from a Poisson model with:

$$\begin{aligned} y_{it}^s | \theta_1, \xi_{it} &\sim \text{Poisson}(\theta_1 \xi_{it}), i=1, \dots, n_1, \\ y_{it}^s | \theta_2, \xi_{it} &\sim \text{Poisson}(\theta_2 \xi_{it}), i=n_1+1, \dots, n_2, \\ y_{it}^s | \theta_3, \xi_{it} &\sim \text{Poisson}(\theta_3 \xi_{it}), i=n_2+1, \dots, n_3, \\ y_{it}^s | \theta_4, \xi_{it} &\sim \text{Poisson}(\theta_4 \xi_{it}), i=n_3+1, \dots, n_4, \text{ and} \\ y_{it}^s | \theta_5, \xi_{it} &\sim \text{Poisson}(\theta_5 \xi_{it}), i=n_4+1, \dots, n, \end{aligned}$$

where $s = 1, \dots, S$ is the size of replication.

References

- Ando T. Predictive Bayesian model selection. *American Journal of Mathematical and Management Sciences*. 2012; 31:13–38.
- Assunção RM. Space varying coefficient models for small area data. *Environmetrics*. 2003; 14:453–473.
- Banerjee S, Carlin BP. Semiparametric spatio-temporal frailty modeling. *Environmetrics*. 2003; 14:523–535.
- Besag J, York J, Mollie A. Bayesian image-resorption, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*. 1991; 43:1–20.
- Best N, Richardson S, Thomson A. A comparison of Bayesian spatial models for Disease mapping. *Statistical Methods in Medical Research*. 2005; 14:35–59. [PubMed: 15690999]
- Binder DA. Bayesian cluster analysis. *Biometrika*. 1978; 65:31–38.
- Binder DA. Approximations to Bayesian clustering rules. *The Annals of Statistics*. 1981; 68:275–285.

- Celeux G, Forbes F, Robert C, Titterton M. Deviance information criteria for missing data models. *Bayesian Analysis*. 2006; 1:651–674.
- Christiansen CL, Morris CN. Hierarchical Poisson regression modeling. *Journal of the American Statistical Association*. 1997; 92:618–632.
- Dahl, DB. Model-based clustering for expression data via a Dirichlet process mixture model. In: Do, KA.; Miller, P.; Vannucci, M., editors. *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press; 2006. p. 201–218.
- Fernandez C, Green PJ. Modelling spatially correlated data via mixtures: a Bayesian approach. *Journal of the Royal Statistical Society (Series B)*. 2002; 64:805–826.
- Fritsch A, Ickstadt K. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Analysis*. 2009; 4:367–392.
- Gelfand AE, Ghosh SK. Model choice: A minimum posterior predictive loss. *Biometrika*. 1998; 85:1–11.
- Gelfand A, Vounatsou P. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*. 2003; 4:11–25. [PubMed: 12925327]
- Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. *Bayesian data analysis*. Boca Raton: Chapman & hall/CRC; 2004.
- Green PJ, Richardson S. Modelling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*. 2001; 28:355–375.
- Hastie DJ, Green PJ. Model choice using reversible jump Markov chain Monte Carlo. *Statistica Neerlandica*. 2012; 66:309–338.
- Hossain MM, Lawson AB, Cai B, Choi J, Liu J, Kirby RS. Space-time stick-breaking processes for small area disease cluster estimation. *Environmental and Ecological Statistics*. 2012
- Hossain MM, Lawson AB. Space-time Bayesian small area disease risk models: Development and evaluation with a focus on cluster detection. *Environmental and Ecological Statistics*. 2010; 17:73–95. [PubMed: 20953293]
- Jasra A, Holmes CC, Stephens DA. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*. 2005; 20:50–67.
- Kelsall, JE.; Wakefield, JC. Discussion of “Bayesian models for spatially correlated disease and exposure data”. In: Best, N.; Waller, L.; Thomas, A.; Conlon, E.; Arnold, R.; Bernardo, JM.; Berger, JO.; Dawid, AP.; Smith, AFM., editors. *Sixth Valencia International Meeting on Bayesian Statistics*. London: Oxford University Press; 1999.
- Kim S, Tadesse MG, Vannucci M. Variable selection in clustering via Dirichlet process mixture models. *Biometrika*. 2006; 93:877–893.
- Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*. 2000; 19:2555–2567. [PubMed: 10960871]
- Kuo L, Mallick B. Variable selection for regression models. *Sankhya (Series B)*. 1998; 60:65–81.
- Lau JW, Green PJ. Bayesian model based clustering procedure. *Journal of Computational and Graphical Statistics*. 2007; 16:526–558.
- Lawson AB, Song HR, Cai B, Hossain MM, Huang K. Space-time latent component modeling of geo-referenced health data. *Statistics in Medicine*. 2010; 29:2012–2027. [PubMed: 20683893]
- Marin, J-M.; Mengersen, K.; Robert, C. Bayesian modelling and inference on mixtures of distributions. In: Rao, C.; Dey, D., editors. *Handbook of Statistics*. Vol. 25. New York: Springer-Verlag; 2005.
- McGrory CA, Titterton DM. Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics and Data Analysis*. 2007; 51:5352–5367.
- Meligkotsidou L. Bayesian multivariate Poisson mixtures with an unknown number of components. *Statistics and Computing*. 2007; 17:93–107.
- O’Hara RB, Sillanpaa MJ. A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis*. 2009; 4:85–118.
- Plummer M. Penalized loss functions for Bayesian model comparison. *Biostatistics*. 2008; 9:523–539. [PubMed: 18209015]

- Richardson S, Green PJ. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society (Series B)*. 1997; 59:731–792.
- Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society (Series B)*. 2009; 71:319–392.
- Spiegelhalter, DJ.; Thomas, A.; Best, NG. WinBUGS: Bayesian inference using Gibbs sampling, Manual version 1.4. Imperial College. Cambridge: London and Medical Research Council Biostatistics Unit; 2003.
- Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of complexity and fit (with discussion). *Journal of the Royal Statistical Society (Series B)*. 2002; 64:583–639.
- Stephens M. Bayesian analysis of mixture models with an unknown number of components – an alternative to reversible jump methods. *The Annals of Statistics*. 2000a; 28:40–74.
- Stephens M. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society (Series B)*. 2000b; 62:795–809.
- Stern, HS.; Cressie, NA. Inference for extremes in disease mapping. In: Lawson, AB.; Biggeri, A.; Bohning, D.; Lesaffre, E.; Viel, J-F.; Bertollini, R., editors. *Disease mapping and risk assessment for public health*. Chichester: Wiley; 1999. p. 63-84.
- Tadesse MG, Sha N, Vannucci M. Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*. 2005; 100:602–617.
- Viallefont V, Richardson S, Green PJ. Bayesian analysis of Poisson mixtures. *Journal of Nonparametric Statistics*. 2002; 14:181–202.
- Waller LA, Carlin BP, Xia H, Gelfand AE. Hierarchical spatio temporal-mapping of disease rates. *Journal of the American Statistical Association*. 1997; 92:607–617.
- Yao W. Model based labeling for mixture models. *Statistics and Computing*. 2012; 22:337–347.

Figure 1.
Density plot for the last 20,000 posterior samples of θ . The model is fitted with six components for the county specific South Carolina low birth weight data.

Figure 2.

Histogram and density plots of observed standardized incidence ratio (SIR) of county specific South Carolina low birth weight count data for the year 1998. The number of breaks and bandwidth are determined by using the `bw.SJ` function in R.

Figure 3. Thematic maps of β 's, the data used for simulating the county specific lung cancer deaths for the state of Ohio: β_0 (top-left), β_1 (top-right), β_2 (bottom-left), and β_3 (bottom-right).

Figure 4.

Plotting of p -values against county (county names are alphabetically ordered) obtained from county specific univariate Poisson regression model for each covariate with expected count as an offset variable for South Carolina low birth weight data. The covariates are: population density (PD), proportion of African-American (PAA), median household income (MHI), proportion in poverty (PP), unemployment rate (UR), and time.

Table 1

Average posterior mean and standard error of auxiliary indicator variable for each covariate effect for the Simulation-1 dataset

Indicator variable for covariate effect	Posterior mean	Standard error
τ_1	0.0576	0.2144
τ_2	0.7853	0.2186
τ_3	0.9974	0.0044

Table 2

Average values for the model selection criteria and the optimum expected loss for the Simulation-1 dataset for the models with various combinations of covariates x_1, x_2 and x_3

Covariate	DIC	DIC3	MSPE	Binder's loss	Lau & Green loss	Maximum PEAR
Intercept only	11618.51	11944.69	189.7469	1174.993	1173.514	0.37789
Intercept and x_1	11619.55	11953.45	185.6524	1093.262	1091.638	0.39438
Intercept and x_2	11986.08	12277.73	214.2977	337.1096	336.5138	0.57292
Intercept and x_3	11541.8	11873.84	173.6958	221.3878	220.0612	0.545087
Intercept, x_1 and x_2	11573.86	11915.37	180.5074	291.3908	291.0096	0.4914627
Intercept, x_1 and x_3	11615.46	11958.0	174.4494	199.1183	198.8899	0.5548457
Intercept, x_2 and x_3	11580.64	11920.02	175.5993	137.805	137.7711	0.6375331
Full (Intercept, x_1, x_2 and x_3)	11564.58	11906.14	175.4762	175.4762	167.6591	0.4951442

Table 3

Average posterior mean and standard error of auxiliary indicator variable for each covariate effect for the Simulation-2 dataset

Indicator variable for covariate effect	Posterior mean	Standard error
τ_1	0.67662	0.00937
τ_2	0.83869	0.01011
τ_3	0.95003	0.01030

Table 4

Average values for the model selection criteria and the optimum expected loss for the Simulation-2 dataset for the models with various combinations of covariates x_1 , x_2 and x_3

Covariate	DIC	DIC3	MSPE	Binder's loss	Lau & Green loss	Maximum PEAR
Intercept only	13089.64	13635.52	422.1604	1237.826	1084.311	0.2069
Intercept and x_1	14840.08	15964.27	1221.702	1030.294	991.5413	0.3707
Intercept and x_2	15512.65	16973.48	1487.344	873.979	851.8798	0.4744
Intercept and x_3	16293.81	17920.75	1981.996	873.1445	850.3804	0.4700
Intercept, x_1 and x_2	15574.62	16992.66	1523.662	901.3975	877.3205	0.4447
Intercept, x_1 and x_3	15287.31	16695.92	1358.119	877.1459	853.2936	0.4658
Intercept, x_2 and x_3	15235.93	16609.79	1388.221	878.995	854.6614	0.4707
Full (Intercept, x_1 , x_2 and x_3)	14131.80	15034.43	970.9007	1091.618	1037.593	0.3477

Table 5

Model selection criteria values and posterior mean estimates (standard error estimate in parenthesis) for the auxiliary indicator variables for the varying components number of spatially varying homogeneous covariate effect model. The results are for the county specific counts of South Carolina low birth weight incidences.

Number of components	DIC	DIC ₃	MSPE	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
2	3555.2	3635.3	220.9	0.27 (0.44)	0.16 (0.37)	0.16 (0.37)	0.13 (0.34)	0.15 (0.36)	1.00 (0.00)
3	3563.0	3640.9	222.2	0.17 (0.37)	0.20 (0.40)	0.06 (0.24)	0.17 (0.37)	0.13 (0.33)	1.00 (0.00)
4	3555.2	3635.3	220.9	0.20 (0.40)	0.22 (0.41)	0.10 (0.30)	0.13 (0.34)	0.19 (0.40)	1.00 (0.00)
5	3583.7	3680.5	229.2	0.27 (0.44)	0.10 (0.30)	0.37 (0.48)	0.40 (0.49)	0.14 (0.35)	0.86 (0.35)
6	3547.8	3633.4	218.8	0.24 (0.43)	0.19 (0.39)	0.29 (0.45)	0.14 (0.35)	0.15 (0.36)	0.88 (0.32)
7	3557.3	3638.0	220.7	0.10 (0.30)	0.16 (0.36)	0.34 (0.47)	0.20 (0.40)	0.12 (0.33)	0.62 (0.49)

Table 6
Values for the model selection criteria and minimum expected loss for the South Carolina low birth weight data

Covariate	DIC	DIC ₃	MSPE	Binder's loss	Lau & Green loss	Maximum PEAR
MHI	3556.4	3638.2	219.8	506.6	506.4	0.021
Time	3560.5	3644.5	220.1	460.5	458.4	0.145
MHI, time	3552.4	3633.8	219.7	507.4	507.2	0.020
Full (PD, PAA, MHI, PP, UR, time)	3552.3	3634.6	219.5	507.6	507.3	0.0191