Published in final edited form as:

Comput Stat Data Anal. 2009 June 15; 53(8): 3057–3069. doi:10.1016/j.csda.2008.07.037.

Spatial Latent Class Analysis Model for Spatially Distributed Multivariate Binary Data

Melanie M. Wall¹ and Xuan Liu

University of Minnesota, Division of Biostatistics

Abstract

A spatial latent class analysis model that extends the classic latent class analysis model by adding spatial structure to the latent class distribution through the use of the multinomial probit model is introduced. Linear combinations of independent Gaussian spatial processes are used to develop multivariate spatial processes that are underlying the categorical latent classes. This allows the latent class membership to be correlated across spatially distributed sites and it allows correlation between the probabilities of particular types of classes at any one site. The number of latent classes is assumed fixed but is chosen by model comparison via cross-validation. An application of the spatial latent class analysis model is shown using soil pollution samples where 8 heavy metals were measured to be above or below government pollution limits across a 25 square kilometer region. Estimation is performed within a Bayesian framework using MCMC and is implemented using the OpenBUGS software.

Keywords

Mixture model; multinomial probit; latent variables

1 Introduction

Multivariate binary measurements are very common in many research areas, such as sociology, psychology, and medical diagnostic tests where multiple questions or tests are given to subjects (or patients) each with a simple 0-1 response. Two basic but commonly employed methods for dealing with the multivariate nature of the data are to either analyze all the measurements separately or instead to create some overall summary index of the indicators and then analyze the index as the single observed measure. Both methods fail to take account of the fact that different indicators might be correlated. On the other hand, the latent class model is a multivariate model which has been widely used for this type of data when it is reasonable to hypothesize the measurements as representations of some underlying clusters in the study population. The basic idea of the latent class analysis (LCA) model is that the indicators are correlated because there are unobserved clusters (i.e. latent classes) in the population and responses to the indicators differ for different clusters thus creating correlation between the indicators when the clusters are unknown. The model assumes that within each latent class, the responses to the indicators are, in fact, independent.

¹Corresponding Author: Melanie M. Wall, A460 Mayo Building MMC 303, 420 Delaware Street SE, Minneapolis MN, 55455. melanie@biostat.umn.edu, 612-625-2138(P), 612-626-0660(F).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Commonly the LCA models applied in the literature make the assumption that the indicators are correlated within the same subject, but are independent for different subjects. This is often a reasonable assumption in problems where measurements are taken on independent individuals, but may not be appropriate in cases such as multiple indicators collected repeatedly over time, or over different spatial locations. There are examples of the latent class model being used to model multivariate binary data collected across time. For example, Reboussin et al. (1999) considered a latent transition model for data consisting of repeated measures of multiple indicators of health. In contrast, the current paper considers the latent class model for multivariate binary data collected across geographic regions.

The motivating data of this paper (see Figure 1) includes binary indicators for eight different heavy metals found in the soil which indicate whether the level of the heavy metal is over the legal pollution threshold or not at 97 spatially referenced locations throughout a 25 square kilometer region. It is of interest to identify contaminated areas and specifically to examine whether there are parts of the region that tend to share high levels of certain combinations of heavy metals, which may then help identify pollution sources. An LCA type model can be useful for this data since there are multiple indicators for each location and the indicators are likely to be correlated due to some common unknown and unobserved pollution sources. The challenge of this data is that observations of each of the eight variables are likely correlated over different locations and different variables observed at different locations may also be cross-correlated.

In this paper we introduce a spatial latent class analysis model. In the first part of this hierarchical model we define the relationship between the observed indicators and the latent classes the same way as in a classic LCA model, and in the second part of the model we use a multinomial probit to model the latent classes by introducing continuous random variables underlying them. The benefit of the multinomial probit model is that spatial covariance structures can be introduced that take account of the spatial correlations between the latent class variables at different locations.

In Section 2 we review the classic LCA model, and introduce the spatial latent class analysis (SLCA) model. In Section 3 we describe inference for the parameters in a fully Bayesian framework and propose to use cross-validation methods for model comparison and examining fit. In Section 4 we show the results of the SLCA model applied to the soil pollution example. It is found in the soil pollution data example that a 3-class model does not fit well in favor of the smaller 2-class model. In order to demonstrate that the SLCA can indeed fit a model with more than 2 classes, the SLCA model is also fit to a simulated data example in Section 4 where the truth is 3 underlying spatially distributed latent classes. Section 5 provides a summary discussion.

2 The Model

2.1 Classic LCA

The latent class model was designed to explain relationships between variables in a multidimensional contingency table (Clogg and Goodman, 1984; McCutcheon, 1987). Consider a sample of size N where each individual has P binary indicator variables. Let $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iP})$ be the $P \times 1$ vector of observed binary indicators for individual i . Each individual is in one of the K underlying classes and we let C_i represent his latent class. It is assumed that the classes form a complete partition of the population and that $\sum_k P(C_i = k) = 1$ for all i . The number of distinct latent classes K is considered fixed and known, but will be examined through model comparison. LCA has two important assumptions. First, given a certain latent class, individuals have common response probabilities π_{jk} :

$$P(Y_{ij}=1|C_i=k)=\pi_{jk}, \quad i=1, \dots, N; j=1, \dots, P; k=1, \dots, K. \quad (1)$$

Second, given the class membership, the individual's responses are independent:

$$P(Y_{i1}=y_{i1}, \dots, Y_{ip}=y_{ip}|C_i=k)=\prod_{j=1}^P P(Y_{ij}=y_{ij}|C_i=k). \quad (2)$$

Then (1) and (2) imply the marginal distribution of $\mathbf{Y}_i$:

$$P(Y_{i1}=y_{i1}, \dots, Y_{ip}=y_{ip})=\sum_{k=1}^K P(C_i=k) \prod_{j=1}^P \pi_{jk}^{y_{ij}} (1 - \pi_{jk})^{1-y_{ij}}. \quad (3)$$

In the classic LCA, individuals are considered iid with $P(C_i = k) = \eta_k$ which does not depend on individual i . Bandeen-Roche et al. (1997) extended the classic LCA model to a latent class regression model to incorporate subject-specific observed covariates $\mathbf{X}_i$ where $P(C_i = k)$ is replaced with $P(C_i = k | \mathbf{X}_i) = \eta_{ki}$. The probability of a particular class membership for individual i depends on the individual's covariates $\mathbf{X}_i$ through some function, e.g. $\eta_{ki} = g(\boldsymbol{\beta}, \mathbf{X}_i)$. Thus, for N independent individuals i , the joint distribution for $\mathbf{Y}_1, \dots, \mathbf{Y}_N$ given possible covariates $\mathbf{X}_1, \dots, \mathbf{X}_N$ based on the latent class analysis model is

$$P(\mathbf{Y}_1, \dots, \mathbf{Y}_N | \mathbf{X}_1, \dots, \mathbf{X}_N) = \prod_{i=1}^N \left[\sum_{k=1}^K \eta_{ki} \prod_{j=1}^P \pi_{jk}^{y_{ij}} (1 - \pi_{jk})^{1-y_{ij}} \right].$$

Statistical inference and applications of this model have been well studied and are straightforward to implement. For example the Mplus software (Muthen and Muthen, 1998) can be used for inference using maximum likelihood implementing an EM algorithm and Winbugs can be used for a fully Bayesian approach, see e.g. the "Biops" example in the Winbugs example manual vol. 2 (Speigalhalter, et al., 2004).

2.2 Spatial LCA

The nature of the multivariate binary data in our pollution example and the appeal of the LCA model for modeling multiple binary indicators motivate us to consider the LCA model. But since the heavy metal samples are collected over certain spatial locations, it is necessary to further develop the LCA model to take account of the possible spatial correlations in the data.

We propose a Spatial Latent Class Analysis (SLCA) Model where the relationship between the binary indicators and latent classes is defined in the same way as (1) and (2), but the model will take account of the spatial correlation between observations at different sites by putting a spatial model on the underlying latent classes. This means that when forming the joint distribution of the $\mathbf{Y}$ from (3) we will model jointly the $P(C_1, C_2, \dots, C_n)$ via a spatial process so that $P(C_i = k)$ will depend on the values of the latent classes at other spatial locations.

Because in general the latent classes represent K unordered categories, it is necessary to build a spatial model that can handle multinomial responses. Spatial models for continuous responses

are much more developed than those for categorical outcomes though there is a relatively large literature focused on spatial models for binary responses. The autologistic model (Besag, 1974) was developed for binary data and is commonly used but it is not easily extended to multinomial responses. Diggle et al. (1998) considered a binomial logit link model with a spatially correlated normal random effect. A straightforward extension of their approach to multinomial spatial data is to generalize the binomial logit link to the multinomial logit link. However, a limitation of the multinomial logit model is that it assumes the alternative categories are mutually independent which may not be reasonable (Alvarez and Nagler 2001; Lacy and Burden 1999; McFadden et al. 1981; Cheng and Long (2007)). In the spatial econometrics literature, the probit model has been used extensively for modeling spatially distributed binary data (McMillen 1992; Beron and Vijverberg 2004; Holloway et al. 2002; Smith and LeSage 2004). Moreover, the probit model is easily extended to the multinomial probit (Daganzo 1979) which allows for interdependencies between the different categories through correlations of an error term specified within the model (described below). The multinomial probit model is appealing for modeling spatially referenced multinomial data since continuous latent variables are introduced allowing common spatial covariance structures for continuous outcomes to be included (Bolduc 1992, 1999, Bolduc et al. 1997; Schmidheiny 2003, Beron and Vijverberg 2004, Mohammadian et al. 2005). In this paper we use a multinomial probit model for modeling the spatial latent classes and incorporate spatial structure through a multivariate spatial process.

Now we define the multinomial probit model. Let C_i be the latent class variable with K classes at location i . For each C_i , an underlying normal vector $\mathbf{z}_i^T = (z_{i1}, \dots, z_{iK})$ is considered, where class k is “observed” if the k th component of $\mathbf{z}_i$ is larger than others, i.e.,

$$C_i = k \quad \text{if } z_{ik} = \max(\mathbf{z}_i) \quad (4)$$

and

$$\mathbf{z}_i = \boldsymbol{\mu}_z + \mathbf{u}_i,$$

where $\boldsymbol{\mu}_z$ is a $K \times 1$ vector of parameters, which can be replaced with regression covariates in general, and $\mathbf{u}_i$ is a $K \times 1$ normal random vector of errors with mean vector $\mathbf{0}$ and some unknown covariance matrix. Without further restrictions, the model for $P(C_i | \boldsymbol{\mu}_z)$ is not identified and can be shown to be a function of the differences between the elements of $\mathbf{z}_i$ (Daganzo 1979, Bunch 1991). Hence the last category K is chosen as the reference category and a new vector

$\mathbf{w}_i^T = (w_{i1}, \dots, w_{i,K-1})$ with $w_{ik} = z_{ik} - z_{iK}$ is used to identify the model. The new vector $\mathbf{w}_i$ is a $(K-1)$ vector of normal random variables, and we can write

$$\mathbf{w}_i = \boldsymbol{\mu}_w + \boldsymbol{\delta}_i, \quad (5)$$

where $\boldsymbol{\delta}_i$ is a $(K-1)$ vector of normal errors with mean vector $\mathbf{0}$ and some unknown covariance matrix. We also define a new class variable $d_i = 0, \dots, K-1$, such that

$$d_i = \begin{cases} 0 & \text{if } \max(\mathbf{w}_i) < 0 \\ k & \text{if } \max(\mathbf{w}_i) = w_{ik}. \end{cases} \quad (6)$$

Note d_i is a simple reparametrization of C_i by switching the class K to class 0. The new class variable d_i is completely determined by the values of the normal random vector $\mathbf{w}_i$, so we can take account of the spatial dependence of the multinomial class variable d_i by imposing a multivariate spatial process on $\mathbf{w}_i$ or δ_i .

A particularly flexible class of models for multivariate spatial data is developed by considering the elements of the multivariate spatial process as linear combinations of a set of independent univariate spatial processes. This type of model is often referred to as the linear model of coregionalization (LMC) in geostatistics (Grzebyk and Wackernagel 1994; Wackernagel 2003 Chapter 26). For lattice or areal data, the multivariate conditional autoregressive (MCAR) model (Gelfand and Vounatsou 2003) can similarly be described as a linear combination of independent univariate CAR models.

We consider a relatively simple example to illustrate this constructive modelling strategy. Suppose $\mathbf{Y}(\mathbf{s}_i)$ is a $p \times 1$ vector, a realization of a p -variate spatial process at location $\mathbf{s}_i (i = 1, \dots, n)$. Define

$$\mathbf{Y}(\mathbf{s}_i) = \mathbf{A}\mathbf{v}(\mathbf{s}_i),$$

where $\mathbf{A}$ is a $p \times p$ full rank matrix and the components of $\mathbf{v}(\mathbf{s}_i)$, $v_j(\mathbf{s}_i)$ ($j = 1, \dots, p$), are spatial processes independent across j . Let $\mathbf{C}$ be the $n \times n$ covariance matrix of each spatial process $v_j (j = 1, \dots, p)$. Let $\mathbf{Y}^T = (\mathbf{Y}^T(\mathbf{s}_1), \dots, \mathbf{Y}^T(\mathbf{s}_n))$ and Σ be the covariance matrix of $\mathbf{Y}$, then

$$\Sigma = \mathbf{C} \otimes \mathbf{T},$$

where $\mathbf{T} = \mathbf{A}\mathbf{A}^T$. Note that this specification is equivalent to a separable covariance specification as in Mardia and Goodall (1993), where $\mathbf{T}$ is interpreted as the within-site covariance matrix between variables and $\mathbf{C}$ is interpreted as the across-site covariance matrix. In practice, $\mathbf{A}$ is commonly assumed to be lower or upper triangular without loss of generality, since for any positive definite matrix $\mathbf{T}$, there is a unique lower or upper triangular matrix $\mathbf{A}$ such that $\mathbf{T} = \mathbf{A}\mathbf{A}^T$. The model just introduced can be extended more generally if we allow the p independent processes $v_1, \dots, v_p$ to have different distributions with covariance matrices $\mathbf{C}_1, \dots, \mathbf{C}_p$, so

$$\Sigma = \sum_{j=1}^p \mathbf{C}_j \otimes \mathbf{T}_j,$$

where $\mathbf{T}_j = \mathbf{A}_j \mathbf{A}_j^T$ and $\mathbf{A}_j$ is the j^{th} column of $\mathbf{A}$.

Returning to the development of the SLCA model, we specify the δ_i in (5) as a multivariate spatial process using the method of linear combinations of independent univariate spatial processes, i.e.,

$$\delta_i \equiv \delta(\mathbf{s}_i) = \mathbf{A}\mathbf{v}(\mathbf{s}_i), \quad (i=1, \dots, n), \quad (7)$$

where $\mathbf{A}$ is a $(K-1) \times (K-1)$ lower triangular matrix and the $K-1$ elements of $\mathbf{v}(\mathbf{s}_i)$ are independent zero-mean Gaussian spatial processes. Let $\mathbf{C}(\alpha_k)$, $k = 1 \dots K-1$, denote the parametric covariance matrix of the k^{th} component of the process $\mathbf{v}(\mathbf{s}_i)$.

In the soil heavy metal pollution example, the data were collected on a regular grid, so the conditional autoregressive (CAR) model is reasonable to consider. Thus for each spatial component, $v_k(\mathbf{s}_i)$, the covariance matrix is

$$\mathbf{C}(\alpha_k) = \frac{1}{\tau_k} \mathbf{M}(\mathbf{I}_n - \rho_k \mathbf{H})^{-1} \quad (8)$$

where ρ_k is the “spatial association” parameter, τ_k is a parameter proportional to the conditional precision of $v_k(\mathbf{s}_i) \mid v_k(\mathbf{s}_h \in N(\mathbf{s}_i))$, where $N(\mathbf{s}_i)$ denotes the set of neighbors of $\mathbf{s}_i$, and $\mathbf{M}$ is a diagonal matrix with elements equal to $1/m_i$ where m_i is the number of neighbors of region $\mathbf{s}_i$. The neighbor adjacency matrix $\mathbf{H} = (h_{ii'})$ is a matrix containing the normalized neighboring information, such that $h_{ii'} = 1/m_i$ implies that region $\mathbf{s}_i$ is adjacent to region $\mathbf{s}_{i'}$ and otherwise $h_{ii'} = 0$ implies that these two regions are not neighbors. Note that $\mathbf{M}$ and the neighbor matrix $\mathbf{H}$ do not depend on k and are thus assumed to be the same for each independent $v_k(\mathbf{s}_i)$. We denote $\psi \sim \text{CAR}(\rho, \tau)$, if ψ has a CAR covariance structure with parameters ρ and τ .

The SLCA model is then made up of (1) and (2) relating the observed variables to the latent classes, and a multinomial probit model with spatial correlation for $P(C_1, C_2, \dots, C_n)$ with (6) defining the class membership where the $K-1$ multivariate Gaussian spatial process for δ_i in (7) is plugged in which allows both correlation across sites and between class probabilities within site. In the current paper, we treat the number of latent classes K as fixed and selection of it is done by comparing model fit across different K .

3 Bayesian Inference for the SLCA Model

3.1 Identifiability of the Model

In the relationship between the latent class C_i and latent vector $\mathbf{w}_i$ specified by (4) and (6), the mean vector μ_w and covariance matrix Σ_w of $\mathbf{w}_i$ are not identified without additional restrictions. This is because the value of the latent class variable C_i is determined by the comparison between the components of $\mathbf{w}_i$, so multiplying $\mathbf{w}_i$ by a positive constant s does not change the value of C_i . To demonstrate, let $P(C_i \mid \mu_w, \Sigma_w)$ be the probability function of C_i , then

$$P(C_i \mid \mu_w, \Sigma_w) = P(C_i \mid s\mu_w, s^2 \Sigma_w).$$

Because of this identifiability problem of scale, it is common to fix the the (1, 1) element of Σ_w to 1, to achieve identification of the model. This constraint is accomplished in the spatial version of the model by appropriately constraining elements of $\mathbf{A}$ in (7) and constraining variance parameters in $\mathbf{v}(\mathbf{s}_i)$. Specifically given a lower triangular matrix $\mathbf{A}$, we can solve the identifiability problem by fixing the (1, 1) element of $\mathbf{A}$ to 1 and by fixing unknown variance parameters in $\text{Var}(\mathbf{v}(\mathbf{s}_i))$ to be 1. When the CAR model in (8) is used, this amount to fixing $\tau_1 = \dots = \tau_{K-1} = 1$.

3.2 Prior Specifications

Let Φ be the vector that contains all the unknown parameters in the model,

$$\Phi = \left\{ \pi_{jk}, j=1, \dots, P, k=1, \dots, K, \mu_w, \mathbf{A}, \alpha_1, \dots, \alpha_{K-1} \right\}.$$

Let $P(\Phi)$ be the joint prior for these parameters. We define $P(\Phi)$ as the product of the prior distributions of each parameter in Φ , i.e., we use independent priors for all parameters. For the conditional response probabilities $\pi_{jk}, j=1, \dots, P, k=1, \dots, K$, we use Uniform(0,1) for each of them. The elements of μ_w are each assigned a diffuse normal prior with mean 0 and variance 1000. The (1, 1) element of $\mathbf{A}$ is fixed to 1 as discussed in Section 3.1. Other diagonal elements of $\mathbf{A}$ are each assigned a Uniform(0,1000) prior, and all off-diagonal elements are assigned a diffuse normal prior with mean 0 and variance 1000. In the data example, we use the CAR model for $\mathbf{C}(\alpha)$ with the τ_k parameters fixed to 1 for identifiability purposes and the $\rho_1, \dots, \rho_{K-1}$ each assigned Uniform(0,1) priors. Note that a priori we have put a positive prior on the spatial association parameter since we believe the presence of a particular latent class of heavy metals at one location means it is more likely to observe the same class of heavy metals at a nearby location. In general, our goal is to assign diffuse proper priors for all the unknown parameters.

3.3 Posterior Inference

Let $\mathbf{Y}_i$ be a vector of P binary indicators at location i , and $\mathbf{Y}^T = (\mathbf{Y}_1^T, \dots, \mathbf{Y}_n^T)$ is the $PN \times 1$ vector containing all the data. Then based on the SLCA model, the joint posterior for all the unknown parameters Φ and all the latent class variables for each individual, i.e. $\mathbf{C} = (C_1, \dots, C_N)$ is

$$\begin{aligned} P(\Phi, \mathbf{C} | \mathbf{Y}) &\propto P(C_1, \dots, C_N) \prod_{i=1}^N \prod_{j=1}^P P(Y_{ij} | C_i) P(\Phi) \\ &\propto \prod_{k=1}^{K-1} P(\nu_k(\mathbf{s}_1) \dots \nu_k(\mathbf{s}_n) | \alpha_k) \prod_{i=1}^N \prod_{j=1}^P P(Y_{ij} | \{\pi_{jk}\}, \mu_w, \mathbf{A}, \mathbf{v}(\mathbf{s}_i)) P(\Phi) \end{aligned} \quad (9)$$

Note C_i is a deterministic function of $\mu_w, \mathbf{A}$, and $\mathbf{v}(\mathbf{s}_i)$ hence it does not appear (9). It is difficult to obtain individual posterior distributions by integrating (9) because of its analytical intractability, therefore we consider using the Markov Chain Monte Carlo (MCMC) method to obtain samples from the posterior distribution. MCMC gives a random sample from the marginal posterior distribution for each parameter and each latent class variable, and thus summaries such as the posterior means and quantiles of the parameters can be obtained. The 95% credible intervals (e.g., the interval formed by the posterior 0.025 and 0.975 quantiles) of the parameters can be used to assess the significance of the parameters. Computationally, we only need to calculate the full conditionals of each parameter given all other parameters, and we use the Metropolis-Hastings algorithm when closed-form full conditionals are not available. This model can be implemented using the WinBUGS software (Spiegelhalter, et al., 2004).

3.4 Model Fit and Comparison

A commonly used criteria for model comparison in hierarchical Bayesian models is the Deviance Information Criteria (DIC, Spiegelhalter, et al., 2002). But in the case of latent variable models, in particular mixture models such as the SLCA, the DIC is open to different variations depending on how the latent variables are considered. Celeux et al. (2006) present a careful examination of eight different variants of the DIC for mixture models, some which readily produce negative effective dimensions for no apparent reason. Because of uncertainty about the reliability of DIC as a model comparison method for mixture models, we considered cross-validation predictive distributions which have been advocated for mixture models (Dey et al. 1995)

In general, in cross-validation a validation set of observations $\mathbf{Y}_v$ is held to the side and the remaining data denoted $\mathbf{Y}_{(v)}$ is used as “training data” to fit the model. The conditional predictive distribution of the observed validation data $\mathbf{Y}_v$ given the training data and the model is then $P(\mathbf{Y}_v|\mathbf{Y}_{(v)}) = \int P(\mathbf{Y}_v|\mathbf{Y}_{(v)}, \Phi, \mathbf{C})P(\Phi, \mathbf{C}|\mathbf{Y}_{(v)})d\Phi d\mathbf{C}$. It is common to cycle through the data systematically taking different validation sets $\mathbf{Y}_v$ and computing the cross-validated conditional predictive distribution for each set. These values can then be summarized and used to assess and compare models. When leave-one-out cross-validation is used, the conditional predictive distribution value for each observation is often called the conditional predictive ordinate (CPO) (see e.g. Carlin and Louis 2008, Chapter 2.5).

With multivariate outcome data as we are modeling with the SLCA model, where the P-vector $\mathbf{Y}_i$ is modeled across locations $i = 1 \dots N$, there are two ways to interpret what is meant by leave-one-out cross validation. We can leave-out the entire vector of p-observations at location i , i.e. $\mathbf{Y}_v = \mathbf{Y}_i$, or we can leave out the j^{th} element of the p-vector at location i , i.e. $\mathbf{Y}_v = Y_{ij}$. When $\mathbf{Y}_v = \mathbf{Y}_i$, the CPOs can be empirically computed by first creating N different datasets each with one observation $\mathbf{Y}_i$ removed and replaced with missing values. Then the SLCA model is fit separately to each of the N “leave-one-out” datasets and $P(\mathbf{Y}_i|\mathbf{Y}_{(i)})$, i.e. the posterior of the currently missing observation given the rest of the data is computed and its posterior mean is taken as the i^{th} CPO. Note that when leave-one-out is taken to mean $\mathbf{Y}_v = Y_{ij}$, then $N \times P$ different CPOs are computed each requiring a separate run of the model. For our heavy metal soil pollution example this means $97 \times 8 = 776$ different runs of the model which is computationally very time consuming. Hence in lieu of deleting one element at a time separately for all the different locations, we propose a P-fold (here $P=8$) technique where the data is partitioned into 8 validation sets where at every location i , one random element (i.e. one randomly chosen heavy metal) is dropped. Each of the 8 validation sets has a different random element dropped for each location i as compared to the other validation sets. The conditional predictive ordinates can then be computed in the same way as described above. Generally, the log of the CPO is useful due to the small scale. In the case where $\mathbf{Y}_v = \mathbf{Y}_i$ the CPO values can be plotted to diagnose how well the model “predicts” observations at each location with larger values indicating better fit. Furthermore, the sum of the CPO across all validation sets in the partition can be used to compare different models again with larger values indicating better predictive performance. Though computationally intensive, cross-validation provides a very natural and interpretable way to assess the SLCA model. All cross-validation was performed using the **Brugs** R-package which allows calls to OpenBUGS from within R thus facilitating the looping and data gathering needed.

4 Pollution Data Example

4.1 Real Data Example

Thirty-six heavy metals were measured in 97 soil samples collected on a 500×500 meter grid in the urban area of Slovenia’s Capital, Ljubljana (Komac and Sajn, 2001). Among these thirty-six heavy metals, eight are considered most toxic, i.e., Cadmium-Cd, Cobalt-Co, Chromium-Cr, Copper-Cu, Nickel-Ni, Lead-Pb, Zinc-Zn, and Mercury-Hg, and are of interest in this example. The eight variables have been dichotomized based on whether the measurement of the particular heavy metal is over the limit defined by the Slovenian legislation. Let X_{ij} be the measurement of heavy metal j at location i , $i = 1, \dots, N$ and $j = 1, \dots, P$, then $Y_{ij} = I(X_{ij} \geq L_j)$, where $I()$ is an indicator function, and L_j is the legislative limit for heavy metal j .

Figure 1 presents the observed data (1’s and 0’s) for the 8 different heavy metals at the 97 spatial locations. The color shading in the images was drawn using the **interp** function in R and simply provides a smooth linear interpolation of the observed data. Cobalt (Co) and Lead (Pb) are seen to be above the limit the most commonly while Nickel (Ni) is the least common to be observed above the limit. There appears to be some spatial clustering of high values within

the different maps and similarities across the different heavy metals in terms of where the high values are located. The SLCA model will be used to model both the relationship among the different heavy metals as well as their spatial similarity.

The plot on the lower right of Figure 1 shows the locations where the heavy metal samples were collected and provides numerical labels for those locations used in later Figures. The units of distance on the x and y axes are in kilometers so each square within the grid is 500×500 meters. As the samples were collected on a regular grid, we define two locations to be neighbors if they are adjacent by one step along the grid horizontally, vertically or diagonally. In addition, there were four sites (location 4, 22, 49, and 64) where a second observation was taken very nearby the regular grid locations and these sites are considered neighbors with the respective nearby grid location as well as with the adjacent grid sites.

The goal of this study is to examine whether the spatial similarity of the 8 different heavy metals can be explained by some shared underlying latent classes that are themselves spatially varying. The SLCA model can: identify the number of potential underlying classes, describe the relationship between the heavy metal indicators and the underlying classes (i.e. π_{jk}), and determine the probability that certain locations are of a particular latent class type $P(C_i|\mathbf{Y})$.

To determine the number of underlying classes and compare the performance of the classic LCA model and SLCA model, we consider four models. Model 1 and Model 2 have two latent classes, and Model 3 and Model 4 have three latent classes. Model 1 and Model 3 are classic LCA model, assuming the latent classes are independent across locations, and Model 2 and Model 4 are the SLCA model, assuming the latent classes at different locations are spatially correlated with a CAR covariance structure. More specifically:

- **Model 1:** In Model 1, (1), (2), and (3) are assumed for $K = 2$ to model the relationship between the binary indicator $\mathbf{Y}_i$ with the underlying classes. We also assume the multinomial probit model (4) and (6) for $K = 2$ to model the underlying classes. Note that when $K = 2$, $\mathbf{w}_i$ is univariate, i.e., $w_i = \mu_w + \delta_i$ where μ_w is a scalar parameter and δ_i is a normal random variable, and (6) becomes

$$d_i = \begin{cases} 0 & \text{if } w_i < 0 \\ 1 & \text{if } w_i > 0 \end{cases} \quad (10)$$

We further assume δ_i , $i = 1, \dots, N$ are i.i.d and normally distributed with mean 0 and precision τ_δ , and τ_δ is fixed to be 1 to preserve the identifiability of the model as discussed in the Section 3.1.

- **Model 2:** Model 2 has all the same assumptions as Model 1, except that we assume δ_i , $i = 1, \dots, N$ are spatially correlated over different locations and have a CAR distribution, i.e., $\delta \sim \text{CAR}(\rho, 1)$.
- **Model 3:** In Model 3, latent class model (1), (2) and (3), and multinomial logit model (4), and (6) are assumed for $K = 3$. Here $\mathbf{w}_i = \mu_w + \delta_i$ in (6) is a vector of length 2, and we assume δ_i is a bivariate normal random vector independent over locations $i = 1, \dots, N$, i.e., $\delta_i \stackrel{i.i.d.}{\sim} N(\mathbf{0}, \mathbf{\Omega})$, where $\mathbf{\Omega}$ is a positive definite matrix of size 2, with the (1,1) element fixed to 1.
- **Model 4:** In Model 4, we assume (1), (2), (3), (4), and (6) for $K = 3$ as in Model 3. Also $\mathbf{w}_i = \mu_w + \delta_i$ in (6) is a vector of length 2, different from Model 3, we assume δ_i is a bivariate normal random vector that is spatially correlated over locations $i = 1, \dots, N$ according to (7), i.e., $\delta_i = \mathbf{A}\mathbf{v}(\mathbf{s}_i)$, where $\mathbf{A}$ is a lower triangular matrix of size 2

with (1,1) element fixed to 1, $\mathbf{v}(\mathbf{s}_i)^T = (v_1(\mathbf{s}_i), v_2(\mathbf{s}_i))$ is a vector of two uncorrelated spatial process with $v_1 \sim \text{CAR}(\rho_1, 1)$ and $v_2 \sim \text{CAR}(\rho_2, 1)$.

Tables 1 and 2 show the posterior means and 95% credible intervals of the the probabilities of heavy meatal j being above the limit given the different class memberships k , i.e. the π_{jk} as defined in (1). In the last column for reference are the simple observed proportions of locations where the respective heavy metal is over the legal limit. Note from Table 1 that the two two-class models Model 1 and Model 2 have very similar posterior means and 95% credible intervals for π_{jk} hence incorporating the spatial correlation does not have much influence in the estimates for the π_{jk} .

Examining the 3-class models, we see that Classes 0 and 1 are very similar to the those in the 2-class model and in Class 2 most of the posteriors for π_{jk} have hardly moved from the priors which were uniform (0,1). In all cases the Class 2 95% credible intervals overlap the other two intervals for the other latent classes. Hence it is not advised to interpret the values of the posterior means π_{jk} for Class 2, but instead to conclude that there is not enough information in this data to support a third class. Furthermore, when we examine the cross-validation CPO summary where one whole location is left out at a time, and when we examine the 8-fold cross-validation which leaves one element out from each location at a time, we find that the 3-class models fit worse (lower scores) than their respective 2-class models.

Focusing on the 2-class models, it is of interest to examine the differences between Class 0 and Class 1 in terms of which heavy metals are more or less likely to be over the limit. Within Class 0, Co has the highest probability of being over the limit (0.84) and this is higher than when the location is of Class type 1 (0.64). On the other hand, all of the rest of the heavy metals have higher probability of being over the limit in Class type 1 than when the location is of Class type 0. Further we note that Zn is the element with the most distinct difference in the two classes, with it only being over the limit 3% of the time in Class 0 and 73% of the time in Class 1. Notice that for all three of Cu, Pb, and Zn that the 95% credible intervals do not overlap between the two classes suggesting that these are distinguished best by the two classes. On the other hand, Cr is the least distinguished between the classes with it having probability 0.19 in class 0 and only slightly higher at 0.25 in class 1 and its credible intervals greatly overlap.

In Figure 2, the upper left hand plot shows the CPO values for the spatial verses iid 2-class models. In sum, the CPOs are larger for the spatial model compared to the iid model (also see Table 1). Similary the 8-fold cross-validation summary (Table 1) supports the 2-class spatial model (SLCA) as having better cross-validation prediction than the 2-class iid model (LCA) for this data. In the 2-class SLCA model, ρ has a posterior mean of 0.72 with the 2.5% and 97.5% percentile 0.17 and 0.97 respectively. Though this posterior is quite variable, the addition of this vague spatial correlation has improved the predictability of the model which implies there is some information gained by incorporating neighboring points. We see in the upper right-hand plot of Figure 2 the trend that the CPO values get worse as the number of heavy metals over the limit at a particular site increases. This means that the SLCA model is not predicting as many heavy metals to be over the limit as there actually are. Recall though that it not the goal of the model to predict any particular heavy metal or the total number of heavy metals but to predict latent classes that result in regions with different distributions of heavy metals.

An image plot of the posterior probabilities that each location is in latent Class 1 is shown in the bottom left of Figure 2. We see a distinctive contiguous line of locations with high probability of being in this class running from the bottom of the region up and to the right. There is also a region in the upper center with higher probability of being from latent Class 1. The parameter μ_w in model 2 has posterior mean of 0.12 which translates on the probit scale to a probability of 0.55, which represents the overall marginal probability of a location being

in class 0. This matches what is seen in the image plot where close to half of the region has a high probability of being in class 1. The bottom right plot in Figure 2 figure relates these latent class probabilities to a simple count of the number of heavy metals over the limit at each point. We see that as the total sum increases, the probability of that site being from Class 1 increases, specifically if the count is 5 or more out of 8 then the probability is effectively 1. On the other hand, if there are none or only 1 metal over the limit, the probability is that the site is from Class 0. The values of particular interest are those in the 2-4 range where depending on which heavy metals are present, the probability of class 1 is either high or low. We see for example that site 11 has 3 heavy metals over the limit (Ni, Cr, and Co, from Figure 1) but is predicted to be in the latent class 0. As was seen in Table 1, Ni and Cr provide little distinction in the two classes and Cr is found to be higher in Class 0 so it follows that this site would be from Class 0. If we examine site 89, it also has 3 heavy metals over the limit but they are Cu, Pb, and Zn the metals found to have much higher probability of being over the limit in Class 1 as compared to Class 0, and so this site has high probability of being from Class 1. The SLCA model does not just quantitatively describe the data in terms of the number of heavy metals over the limit but it provides some qualitative distinctions of the region.

4.2 Simulated Example

To demonstrate the SLCA model can be used to identify more than 2 underlying latent classes, we provide a generated data example where the truth is 3 underlying spatially varying latent classes. The set up of the simulated example is the same as the heavy metals example with $N=97$ observation on the same grid as in the example and 8 different binary measured variables at each location. The true values for the π_{jk} are either 0.1 or 0.8 and are shown as x's in the top 3 plots in Figure 3. The true values of the latent classes at each location are shown as numbers in the bottom 3 plots in Figure 3, and the other parameters governing the spatial process are shown in Table 3.

The four models considered in the previous subsection were fit to the simulated data and as expected the SLCA models had better cross-validation prediction than the non-spatial models and the 3-class SLCA model performed best in terms of the 8-fold CV summary. The resulting posterior estimates from the 3-class SLCA model are presented. In the top of Figure 3, the 95% credible intervals of the π_{jk} parameters are shown and in all but one case, (π_{41}), the true values are well covered.

Image plots of the posterior fitted probability that each region is from each of the 3 latent classes is shown in the bottom of Figure 3. There is very good agreement between the true value of the latent class (shown as numbers) and the posterior probability that the location is in the particular class (shown via the coloring with darker color indicating higher probability). In other words, with high probability, the model fits the the correct latent class at each location.

Table 3 shows the posterior estimates for the parameters governing the underlying spatial structure of the latent classes. Using vague priors as described in Section 3.2 for all of the parameters we find that the posterior for the parameters in **A** governing the covariance between the latent classes are quite wide. We replace $a_3 \sim \text{Uniform}(0, 1000)$ and $a_2 \sim N(0, 0.001)$ with $a_3 \sim \text{Uniform}(0, 10)$ and $a_2 \sim N(0, 0.1)$ and present those results also in Table 3 labeled as “informative priors for a_2 and a_3 ”. Note, the well-behaved posterior intervals for the π_{jk} and the $P(C_i|\mathbf{Y})$ described in Figure 3 are based on using all vague priors and they are unchanged when more informative priors are used for the underlying spatial parameters. This is important since π_{jk} and $P(C_i|\mathbf{Y})$ are the most interpretable parameters and they are found to be very robust to changing prior distributions. Overall, this simulated example demonstrates that the 3-class SLCA model can very closely replicate the data but that with a sample size of only $n = 97$ it has some trouble precisely estimating the parameters governing the underlying spatial process unless tighter priors are given.

5 Discussion

In this paper, we developed a spatial latent class analysis model that extends the classic LCA model by adding spatial structure to the latent class distribution through the use of the multinomial probit model. The multinomial probit model transfers a difficult problem of modeling correlated multinomial data to a relative straightforward multivariate Gaussian process modeling problem.

A simple extension to the proposed SLCA model taking advantage of the multinomial probit is to replace μ_w by a regression component including observed covariates. Although the data was not available to us, it would be a great next step in the heavy metal soil pollution example to include as covariates, for example, observed distances to known pollution sources.

In this paper we presented 2 cross-validation methods useful for choosing a best model and hence choosing the value for K . While this method of model choice appeared to behave reasonably, it is computationally intensive and it is of interest to further develop useful fit criterion for mixture models like the SLCA. Moreover, an interesting extension would be to treat K as an unknown parameter directly in the model. While intuitively nice, this extension is expected to be computationally quite difficult as it requires sampling from multiple parameter spaces as k changes using a technique such as reversible jump (Green 1995).

Finally, in our data example, we chose to use a CAR model to account for the spatial similarity but it may have been more natural given that the spatial support was a continuous region to use a geostatistical model. For example, a covariance structure based on an exponential variogram may have worked well. While this may be desirable, the practical matter is that computationally the geostatistical models are computationally very slow to run in Winbugs. The algorithm is of order N^3 . Hence for practical reasons we used the CAR model which has better computational stability. Furthermore, while we used linear combinations of univariate CAR models to build a multivariate spatial process, there are more general formulations of the multivariate conditional autoregressive (MCAR) model (Jin et al. 2005) which provide even more flexible modeling in cases where there is enough data to inform the parameter estimation.

Appendix

```
#####
### WinBugs code for fitting a 3-class SLCA model (Model 4) ###
#####
model { for (i in 1:N){          ## N=97
  for (j in 1:P){                ## P=8
    Y[i,j]~dbern(error[index[i],j])
  }
  index[i]<-class[i]+1

  ### Spz there are 3 latent variable z1_i,..., z3_i, underline the three classes
  ### define w1_i=z1_i-z3_i, and w2_i=z1_i-z2_i.
  ### the next statement defines class_i = 0 if w1_i <0 and w2_i < 0
  ###                                     1 if w1_i >0 and w1_i > w2_i
  ###                                     2 if w2_i >0 and w2_i > w1_i

  class[i] <- (abs(step(w1[i]) + step(w2[i])) +1)/2-abs(step(w1[i]) + step(w2[i])) -1)/2)
    * (2 - step(w1[i] - w2[i])) )
  class0[i]<-step(-w1[i])*step(-w2[i])
  class1[i]<-step(w1[i])*step(w1[i]-w2[i])
  class2[i]<-step(w2[i])*step(w2[i]-w1[i])
  w1[i] <-eta1[i]
  w2[i] <-eta2[i]
  m[i]<- 1/num[i]      ### standardized version of CAR model

##### coregionalization#
# The coregionalization formulas are :
# eta1[i] <-a[2]* v1[i] + a[3]* v2[i] +mu1
# eta2[i]<-a[1] * v1[i] +mu2
#Where v1 and v2 are two ind unit variance "CAR" process with
#correlation parameter alpha[1] and alpha[2]. To improve computing
#efficiency and convergence of the MCMC algorithm, we use an
#alternative equivalent formulation by integrating v1 and v2 out.

  muf2[i] <- (eta1[i]-mu1) * a[2]/a[1] + mu2      ## The conditional mean of eta2[i] given eta1[i]
  muf1[i] <- mu1
}

  a[1]<- 1/sqrt(a1.sq)
  a[2]<- a2      ### to make a[2] notation legal in the initial values
  a[3]<- 1/sqrt(a3.sq)
eta1[1:N] ~ car.proper(muf1[], C[], adj[], num[], m[],
a1.sq, alpha[1])
eta2[1:N] ~ car.proper(muf2[], C[], adj[], num[], m[], a3.sq,
```

```

alpha[2])

##Priors
mu1~dnorm(0,0.001) mu2~dnorm(0,0.001)
for(j in 1:P){
  error[1,j] ~dunif(0,1)
  error[2,j]~dunif(0,1)
  error[3,j]~dunif(0,1)
}
nclass0<-sum(class0[1:N])
nclass1<-sum(class1[1:N])
nclass2<-sum(class2[1:N])
alpha[1] ~ dunif(0,1)
alpha[2] ~ dunif(0,1)
a1.sq<- 1 ##fix a1
a3.sq ~ dunif(0,1000)
a2~ dnorm(0, 0.001)
## set up the weight matrix C for CAR model
cumsum[1] <- 0
for(i in 2:(N+1)) {
  cumsum[i] <- sum(num[1:(i-1)])
}
for(k in 1 : sumNumNeigh) {
  for(i in 1:N) {
    pick[k,i] <- step(k - cumsum[i] - epsilon) * step(cumsum[i+1] - k)
    # pick[k,i] = 1 if cumsum[i] < k <= cumsum[i+1]; otherwise, pick[k,i] = 0
  }
  C[k] <- 1 / inprod(num[], pick[k,]) # weight for each pair of neighbours
}
epsilon <- 0.0001
}

```

References

- Alvarez RM, Nagler J. Correlated disturbances in discrete choice models: a comparison of multinomial probit and logit models. Working Paper at California Institute of Technology, Division of the Humanities and Social Sciences. 2001
- Bandeén-Roche K, Miglioretti DL, Zeger SL, Rathouz PJ. Latent variable regression for multiple discrete outcomes. *Journal of the American Statistical Association* 1997;92:1375–1386.
- Beron, K.; Vijverberg, WPM. Probit in a spatial context: A Monte Carlo analysis. In: Anselin, L.; Florax, RJGM.; Rey, SJ., editors. *Advances in Spatial Econometrics: Methodology Tools and Applications*. Vol. Chapter 8. Springer; 2004. p. 169-192.
- Besag J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B* 1974;23:192–236.
- Bolduc D. Generalized autoregressive errors in the multinomial probit model. *Transportation Research B* 1992;26(2):155–170.
- Bolduc D. A practical technique to estimate multinomial probit models in transportation. *Transportation Research B* 1999;33:63–79.
- Bolduc D, Fortin B, Gordon S. Multinomial probit estimation of spatially interdependent choices: An empirical comparison of two new techniques. *International Regional Science Review* 1997;20:77–101.
- Bunch D. Estimability in the multinomial probit model. *Transportation Research B* 1991;25(1):1–12.
- Carlin, BP.; Louis, TA. *Bayesian Methods for Data Analysis*. 3. Chapman and Hall/CRC Press; Boca Raton, FL: 2008.
- Celeux G, Forbes F, Robert CP, Titterton DM. Deviance information criteria for missing data models. *Bayesian Analysis* 2006;4(1):651–674.
- Cheng S, Long JS. Testing for IIA in the multinomial logit model. *Sociological Methods and Research* 2007;35:583–600.
- Clogg CC, Goodman LA. Latent structure analysis of a set of multidimensional contingency tables. *Journal of the American Statistical Association* 1984;79:762–771.
- Daganzo, C. *Multinomial Probit: The theory and its application to demand forecasting*. Academic Press; New York: 1979.

- Dey DK, Kuo L, Sahu SK. A Bayesian predictive approach to determining the number of components in a mixture distribution. *Statistics and Computing* 1995;5:297–305.
- Diggle PJ, Tawn JA, Moyeed RA. Model-based geostatistics (with discussion). *Applied Statistics* 1998;47:299–350.
- Gelfand A, Vounatsou P. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* 2003;4(1):11–25. [PubMed: 12925327]
- Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 1995;82(4):711–732.
- Grzebyk, M.; Wackernagel, H. Multivariate analysis and spatial/temporal scales: real and complex models. *Proceedings of the XVIIth International Biometrics Conference; Hamilton, Ontario. 1994. p. 19-33.*
- Holloway G, Shankar B, Rahmanb S. Bayesian spatial probit estimation: A primer and an application to HYV rice adoption. *Agricultural Economics* 2002;27(3):383–402.
- Jin X, Carlin BP, Banerjee S. Generalized hierarchical multivariate CAR models for areal data. *Biometrics* 2005;61:950–961. [PubMed: 16401268]
- Komac M, Sajin R. Polluted or nonpolluted-A fuzzy approach determining soil pollution. *Proceeding of the Annual Conference of the International Association for Mathematical Geology. 2001*
- Lacy D, Burden BC. The vote-stealing and turnout effects of Ross Perot in the 1992 U.S. presidential election. *American Journal of Political Science* 1999;43(1):233–55.
- Mardia, KV.; Goodall, C. Spatial-temporal analysis of multivariate environmental monitoring data. In: Patil, GP.; Rao, CR., editors. *Multivariate Environmental Statistics*. Elsevier; Amsterdam: 1993. p. 347-386.
- McCutcheon, AL. *Latent Class Analysis*. Sage Publications; Beverly Hills and London: 1987. Sage University Paper Series on Quantitative Applications in the Social Sciences.
- McFadden D, Train K, Tye W. An application of diagnostic tests for the independence from irrelevant alternatives property of the multinomial logit model. *Transportation Research Board Record* 1981;637:39–46.
- McMillen DP. Probit with spatial autocorrelation. *Journal of Regional Science* 1992;32:335–48.
- Mohammadian, A.; Haider, M.; Kanaroglou, P. Incorporating observed spatial dependencies in random parameter discrete choice models. *Proceedings of the 84th Annual Meeting of the Transportation Research Board; Washington D.C. 2005.*
- Muthen, LK.; Muthen, BO. *Mplus User's Guide*. Muthen & Muthen; Los Angeles: 1998.
- Reboussin BA, Liang KY, Reboussin DM. Estimating equations for a latent transition model with multiple discrete indicators. *Biometrics* 1999;55(3):839–845. [PubMed: 11315015]
- Schmidheiny K. Income segregation and local progressive taxation: Empirical evidence from Switzerland. *Working Paper, Diskussionsschriften dp0311, Universitaet Bern, Departement Volkswirtschaft. 2003*
- Smith, TE.; LeSage, JP. A Bayesian probit model with spatial dependencies. In: LeSage, JP.; Pace, RK., editors. *Spatial and Spatio-Temporal Econometrics*. Elsevier; Amsterdam: 2004.
- Spiegelhalter DJ, Best N, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit (with discussion). *J Roy Statist Soc, Ser B* 2002;64:583–639.
- Spiegelhalter, DJ.; Thomas, A.; Best, N.; Lunn, D. *Winbugs version 1.4.1 manual*. 2004. <http://www.mrc-bsu.cam.ac.uk/bugs>
- Wackernagel, H. *Multivariate Geostatistics - An introduction with Applications*. 3. Springer-Verlag; New York: 2003.

Figure 1. Sampled locations and values of heavy metals in the soil data. Values of 1 and 0 indicate soil sample has above and below regulatory limit for the respective heavy metal. Plot on bottom right provides location labels

2-class SLCA results

Figure 2.
Result of fitting the 2-class SLCA model to the heavy metal soil example

Figure 3.
Results of fitting the 3-class SLCA model to simulated example

Table 1

Posterior means and 95% credible intervals for π_{jk} from the heavy metal soil data - Model 1 and 2

π_{jk}	Model 1 iid 2-class		Model 2 spatial 2-class		Marginal proportion
	0	1	0	1	
Cd	0.06 (0.01,0.14)	0.22 (0.10,0.37)	0.06 (0.01,0.14)	0.22 (0.10,0.37)	0.11
Co	0.84 (0.72,0.93)	0.65 (0.49,0.78)	0.84 (0.72,0.93)	0.64 (0.47,0.78)	0.76
Cr	0.19 (0.10,0.31)	0.25 (0.13,0.40)	0.19 (0.10,0.31)	0.26 (0.13,0.40)	0.21
Cu	0.10 (0.02,0.22)	0.61 (0.45,0.77)	0.11 (0.02,0.23)	0.62 (0.46,0.78)	0.32
Ni	0.05 (0.01,0.13)	0.09 (0.03,0.19)	0.05 (0.01,0.12)	0.10 (0.03,0.21)	0.05
Pb	0.22 (0.05,0.38)	0.96 (0.85,0.99)	0.25 (0.09,0.40)	0.94 (0.84,0.99)	0.55
Zn	0.03 (0.01,0.11)	0.71 (0.52,0.91)	0.03 (0.00,0.10)	0.74 (0.54,0.93)	0.32
Hg	0.07 (0.01,0.16)	0.25 (0.13,0.40)	0.06 (0.01,0.15)	0.27 (0.14,0.42)	0.13
$\sum_{i=1}^{97} \log CPO_i$	-524.5		-508.4		
8-fold CV	-421.4		-416.4		

Table 2

Posterior means and 95% credible intervals for π_{jk} from the heavy metal soil data - Model 3 and 4

Model 3 iid 3-class				Model 4 spatial 3-class			Marginal proportion
π_{jk}	0	1	2	0	1	2	
Cd	0.05 (0.00 0.12)	0.22 (0.09 0.39)	0.40 (0.05 0.87)	0.06 (0.01 0.14)	0.22 (0.10 0.37)	0.43 (0.03 0.95)	0.11
Co	0.85 (0.72 0.94)	0.59 (0.38 0.77)	0.79 (0.28 0.99)	0.84 (0.73 0.92)	0.61 (0.40 0.77)	0.65 (0.05 0.99)	0.76
Cr	0.19 (0.09 0.31)	0.19 (0.03 0.36)	0.71 (0.21 0.99)	0.19 (0.10 0.30)	0.22 (0.06 0.39)	0.61 (0.05 0.99)	0.21
Cu	0.10 (0.03 0.21)	0.57 (0.38 0.75)	0.66 (0.07 0.99)	0.13 (0.04 0.24)	0.60 (0.42 0.76)	0.63 (0.05 0.99)	0.32
Ni	0.04 (0.01 0.12)	0.05 (0.00 0.15)	0.58 (0.11 0.98)	0.05 (0.01 0.12)	0.07 (0.00 0.19)	0.56 (0.04 0.98)	0.05
Pb	0.25 (0.12 0.40)	0.90 (0.77 0.99)	0.69 (0.08 0.99)	0.27 (0.14 0.41)	0.94 (0.82 0.99)	0.66 (0.06 0.99)	0.55
Zn	0.05 (0.00 0.15)	0.66 (0.48 0.84)	0.64 (0.06 0.98)	0.02 (0.00 0.08)	0.77 (0.57 0.93)	0.63 (0.05 0.98)	0.32
Hg	0.08 (0.02 0.17)	0.21 (0.08 0.36)	0.46 (0.05 0.91)	0.06 (0.01 0.14)	0.26 (0.11 0.42)	0.52 (0.04 0.96)	0.13
$\sum_{i=1}^{97} \log CPO_i$	-536.4			-518.4			
8-fold CV	-420.6			-420.3			

Table 3
Results for parameters of distribution of underlying latent classes for simulated 3-class SLCA model

	true value	all vague priors		informative priors for a_2 and a_3	
		posterior mean	95 % credible interval	posterior mean	95 % credible interval
a_1	.9	0.725	(0.112, 0.988)	0.726	(.136, .9887)
a_2	.4	0.277	(0.011, 0.735)	0.292	(.011, 0.749)
μ_{w1}	0	0.099	(-.335, 0.600)	0.040	(-0.357, 0.533)
μ_{w2}	0	-0.677	(-2.346, 0.883)	-.057	(-0.680, 0.554)
a_2	.1	1.88	(-2.18, 6.92)	0.65	(-.92, 1.9)
a_3	2	6.82	(2.15, 9.87)	2.4	(0.96, 4.09)