University of Massachusetts Amherst

From the Selected Works of C. Marjorie Aelion

2009

Variable Selection for Spatial Random Field Predictors Under a Bayesian Mixed Hierarchical Spatial Model

Ji-in Kim

Andrew B. Lawson

Suzanne McDermott

C. Marjorie Aelion, *University of Massachusetts - Amherst*

Available at: https://works.bepress.com/cmarjorie_aelion/1/

Published in final edited form as:

Spat Spatiotemporal Epidemiol. 2009 ; 1(1): 95–102. doi:10.1016/j.sste.2009.07.003.

Variable selection for spatial random field predictors under a Bayesian mixed hierarchical spatial model

Ji-in Kim¹, Andrew B. Lawson^{2,*}, Suzanne McDermott³, and C. Marjorie Aelion⁴

¹Department of Epidemiology and Biostatistics, University of South Carolina, USA

²Department of Biostatistics, Bioinformatics & Epidemiology, Medical University of South Carolina

³Department of Family and Preventive Medicine, University of South Carolina, USA

⁴Department of Environmental Health Sciences, University of South Carolina, USA

Abstract

A health outcome can be observed at a spatial location and we wish to relate this to a set of environmental measurements made on a sampling grid. The environmental measurements are covariates in the model but due to the interpolation associated with the grid there is an error inherent in the covariate value used at the outcome location. Since there may be multiple measurements made on different covariates there could be considerable uncertainty in the covariate values to be used. In this paper we examine a Bayesian approach to the interpolation problem and also a Bayesian solution to the variable selection issue. We present a series of simulations which outline the problem of recovering the true relationships, and also provide an empirical example.

Background

Health outcomes can be observed at spatial locations (such as residential addresses) and it is desirable to relate these outcomes to a set of environmental measurements made on a sampling grid. The environmental measurements can be covariates in the model but due to the interpolation associated with the grid there is an error inherent in the covariate value used at the outcome location. This uncertainty is in addition to measurement error which is typically present when multiple measurements are made on different covariates. Finally, it is often desirable to want carry out variable selection on the environmental covariates to find the best model for the health outcome. This scenario occurs with some frequency in studies of air pollution and health (e.g. PM 2.5, PM 10 interpolation and small area mortality: see e.g. Dominici et al, 1999, 2007; Reich et al, 2008). It is also a problem for other multivariate environment predictors that are used to ‘explain’ health outcomes: for example, soil chemical sampling. In what follows we examine a particular sampling design where a health outcome (mental retardation or developmental delay) is observed at residential addresses and we seek to relate this outcome to environmental chemicals found in soil samples taken from a regular grid covering the study area. More specifically, within a larger study of mental retardation and developmental delay (MRDD) in a Medicaid population in South

© 2009 Elsevier Ltd. All rights reserved.

*Correspondence to: Department of Biostatistics, Bioinformatics & Epidemiology, 135 Cannon St, Medical University of South Carolina, Charleston, SC 29425, USA, Phone: +1-843-876-1865, Fax: +1-843-876-1126, lawsonab@musc.edu.

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.

Carolina we seek to relate the MRDD to sampled measures of soil chemistry, and to do this we propose a Bayesian hierarchical spatial model. There are 16 predictors to consider including 9 spatial random fields (soil chemicals) which are interpolated from soil samples. The other variables in the problem are individual level mother or child characteristics.

Variable Selection Issues

When we fit a statistical model with explanatory covariates, we usually face the problem of variable selection. Traditionally, likelihood ratio tests under different strategies (backward, forward and stepwise selection) have been used. However, the use of these likelihood-based methods does not always guarantee finding the best submodel. These strategies tend to select a highly complicated model for large sample size. An alternative strategy is to use the Akaike information criterion (AIC) which corrects the log-likelihood ratio for bias by penalizing for extra parameters used. Hoeting et al (2006) discuss the application of AIC and related measures for spatial model selection.

For Bayesian models, the deviance information criterion (DIC) or Bayesian information criterion (BIC) are often employed to compare different competing models. However, these criteria cannot be used if the number of competing models is large due to computational burden. In this case, Bayesian model selection, a more automatic data-driven tool to identify a parsimonious model, can play an important role. Several Bayesian model selection methods have been developed in the literature. Bayesian models employ a likelihood for the observed data and combine this with prior distributions for the parameters in the model to form what is known as a posterior distribution (see e.g. Lawson and Banerjee (2009) for a recent survey of Bayesian approaches). These often employ Markov chain Monte Carlo (MCMC) techniques to provide samples of parameters from the posterior distribution. A range of different variable selection approaches have been proposed in the literature. Carlin and Chib (1995) proposed a variable selection method which uses a Gibbs sampler, while George and McCulloch (1993) propose a 'stochastic search variable selection' strategy (see also Cai and Dunson, 2006). Kuo and Mallick (1998) proposed a simpler approach which embeds indicator variables in the regression equation and Dellaportas et al. (2002) introduced a modification of Carlin and Chib's Gibbs sampler. A general review of these different Bayesian model selection methods can be found in Dellaportas et al. (2002).

For a normal regression model, the performance of different Bayesian variable selection methods has been presented previously using simulated data and it was found that several methods found the correct linear regression model (Dellaportas et al, 2002). For a generalized linear model, Kuo and Mallick (1998) presented a simulated example for a log-linear model but not for a logistic regression model. Using a simulated example for a logistic regression, Chen et al. (1999) found that their proposed Bayesian variable selection methods identified the correct model whereas likelihood-based variable selection methods such as AIC and BIC could not. When Bayesian variable selection was applied to a logistic regression model for real data examples, Dellaportas et al. (2002) obtained similar results using Bayesian methods and a likelihood-based method (BIC). Their example featured a simple logistic model with only two categorical variables. When a Bayesian variable selection method was applied to a more complicated real data example with about 50 covariates, different sets of variables were selected by a stepwise selection method compared to the Bayesian method, with some overlap of variables (Gerlach et al, 2002). Hence the performance of these methods is not clearly defined for actual large datasets, and indeed has never been applied to spatial problems where multiple spatially-referenced covariates are found.

In this paper we investigate the performance of a Bayesian variable selection method for a spatial logistic regression model in a simulation study. We also report a brief example where real MRDD data are examined. We consider several simulation set-ups for a logistic regression model and we investigate whether the prior distribution of the logistic regression parameters affect the variable selection results. We repeated our simulations 200 times for each scenario rather than obtaining the result from one simulated example. We followed the approach of Kuo and Mallick (1998) due to its simplicity and ease of implementation in standard software (WinBUGS). We will also used a variation of Kuo and Mallick's approach employing hyperpriors for indicator variables which follow a Bernoulli distribution. Finally, we applied this Bayesian variable selection to our real data example where we considered several latent variables as independent variables.

The remainder of this paper describes the development of a Bayesian variable selection method followed by a simulation study to verify the performance of Bayesian variable selection under different set-ups. Finally, we present the application to a real data example.

Problem Definition

In our application we considered a binary outcome (MRDD or not) which is spatially-referenced. We defined the outcome as y_i for a given subject and assumed that residential addresses of MRDD cases are the locations with binary mark $y_i = 1$ and non-MRDD (controls) as $y_i = 0$. The controls are the residential locations of pregnant women who had a normal child and the MRDD cases are the residential locations of pregnant women who had a child who received a diagnosis of MR or DD. The case group is a subset of the local birth population. Thus we use a logistic regression to model this spatially-referenced outcome.

In our real example, we examined the spatial distribution of cases of mental retardation and developmental delay (MRDD) in a Medicaid population in South Carolina who were born to mothers who were pregnant and resided at known locations between 1996 and 2001. Medicaid is a health insurance for individuals with low incomes in the United States, covering low-income parents, children, seniors, and people with disabilities. During each month of pregnancy, residential addresses for mothers are recorded. The resulting data consists of residential location of the mothers during pregnancy and a binary indicator denoting MRDD/normal status (case/control) of the child. Other available mother and child covariates in this study include: mother's age at birth, mother's ethnicity, mother's alcohol consumption during pregnancy, parity, birth weight, baby sex and follow-up time. We term these covariates as 'mother-child' or 'individual' covariates. Strips of land were identified within South Carolina so concentrations of inorganic chemicals could be measured using a grid for sampling. Figure 1 displays the distribution of cases and controls in one study strip. In this area soil samples were collected from a network of sample sites which has a regular grid pattern covering the whole study area. These soil sample sites are spatially misaligned to the residential addresses of pregnant women. Thus, interpolation of chemical fields is required to link MRDD cases and other births to environmental exposure risk. The concentrations of nine chemicals were measured in soil samples: arsenic (As), barium (Ba), beryllium (Be), chromium (Cr), copper (Cu), lead (Pb), manganese (Mn), nickel (Ni) and mercury (Hg). We denote these observed soil chemical concentrations at sample sites as 'environmental' covariates.

Notational Definitions

We observed covariates that related to the individual outcomes as 'individual' covariates and 'environmental' covariates which were described in the next section. We term this model 'logistic spatial'.

Data are in the form $(\mathbf{x}_i, y_i)$, $1 \leq i \leq n$, where the y_i is a binary outcome for $\mathbf{x}_i$ and $\mathbf{x}_i \in \mathcal{R}^2$ represents geographical locations. To allow for the variable selection we need to extend the usual idea of a linear model to include extra parameters. These entry parameters, as they are known, will allow us to discover which variables will be optimally included in the model. We define an entry parameter γ_i which takes the value 0 or 1. Then define a linear predictor as follows:

$$\eta = \sum_{i=1}^p \gamma_i \beta_i X_i \quad (1)$$

where X_i is the design matrix of covariates and β_i is the regression parameter associated with the i th term. Here $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)$ is the entry parameter vector representing the presence or absence of given variables.

The outcome y_i can be regarded as a Bernoulli random variable and we model the probability of having MRDD using a logit link function:

$$y_i \sim \text{Bernoulli}(p_i) \\ \text{logit}(p_i | \mathbf{u}_i, \mathbf{z}^*(\mathbf{x}_i)) = \beta_0 + \sum_{j=1}^p \gamma_j \beta_j \mathbf{z}_j^*(\mathbf{x}_i) + \sum_{k=1}^q \gamma_k \beta_k \mathbf{u}_{ik} + \mathbf{R}_i \quad (2)$$

where $\mathbf{z}^*(\mathbf{x}_i) = \{z_1^*(\mathbf{x}_i), \dots, z_p^*(\mathbf{x}_i)\}$ is a vector of latent (true) soil chemicals at $\mathbf{x}_i$; $\mathbf{u}_i = (u_{1i}, \dots, u_{qi})$ is a vector of individual covariates; $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p, \beta_{p+1}, \dots, \beta_{p+q})$ is a vector of corresponding regression parameters; $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p, \gamma_{p+1}, \dots, \gamma_{p+q})$ is a vector of corresponding entry parameters and R_i is a random effect term at the individual level. Here, the star symbol (*) denotes the unobserved 'true' covariates. Note that in this problem we have two important issues. First, we need to perform variable selection on both soil chemicals and individual covariates, when there are a large number of these available. Second the true (latent) values of the soil chemicals are not known at the outcome locations and must be estimated. In Bayesian inference, it is usual to assume a joint model for the soil chemicals and the health outcome, and so a model relating the observed soil chemical values to the true values must be considered. This involves interpolation and hence error will be propagated to the outcome sites. In a sophisticated model both the chemical interpolation model and the outcome model would be fitted jointly. This means that all the parameters in the interpolation model (geostatistical parameters such as sill, nugget, covariance parameters) as well as logistic regression parameters will be estimated. This has major methodological implications, as when posterior sampling is employed the chemical covariate fields will be varying at every iteration of the sampling algorithm and hence are not fixed. This leads to considerable difficulty in identification of the correct model.

In our real example we simply examine the first issue of variable selection with entry parameters, using separately estimated (Bayesian Kriged) soil chemical values. We do this to highlight the variable selection procedure. To make allowance for the error in this interpolation, and also to allow for extra noise in the design due to unobserved confounding we have included a random effect term in the model. This effect is assumed a priori to be independent and have a zero mean Gaussian prior distribution with precision parameter $(\tau_v = 1/\sigma^2, \sigma \sim U(0, c))$ with a uniform prior distribution on the standard deviation. We assume $c=5$ in our studies. This is a relatively non-informative prior specification. We did not assume a spatially correlated prior distribution for the random effect as it has been found that estimation of geo-referenced covariates can be aliased with such correlation and hence

masking can occur (Reich et al, 2006; Ma et al, 2007). In the next section we explore via simulation the ability to recover different spatial fields within a spatial logistic regression problem.

Simulation Study for Variable Selection

A simulation study was carried out to determine how well Bayesian variable selection methods perform when multiple spatial fields are used as covariates in a logistic model. The simulation data were set-up to represent our real data example.

Set up of two grids

For the simulation study, two grids were set up. One grid represents soil sample sites and the other grid represents the outcome locations. First, a 10 by 10 regular grid is created within a unit square $\{0.1, \dots, 1\} \times \{0.1, \dots, 1\}$. This first grid is assumed to be soil sample sites. Let $\mathbf{s}_j \in \mathbb{R}^2$, $1 \leq j \leq 100$ denote the first grid. Imitating our MRDD outcome which is observed on a fine grid, the second grid is created from all combination of $x_2 = \{0.42, 0.44, 0.46, 0.48, 0.52, 0.54, 0.56, 0.58, 0.62, 0.64, 0.66, 0.68\}$ and $y_2 = \{0.22, 0.24, 0.26, 0.28, 0.32, 0.34, 0.36, 0.38, 0.42, 0.44, 0.46, 0.48\}$. Let $\mathbf{x}_i \in \mathbb{R}^2$, $1 \leq i \leq 144$ denote the second grid. The graphical representation of these two grids is shown in Figure 2. In what is reported here, due to space limitations, we do not make use of the double grid and we only report chemical concentrations at observation sites. Future reports will examine the interpolation issues.

Logistic model with observed spatial fields

In this simulation, we assumed that all the soil chemical fields are observed at the outcome location $\mathbf{x}_i \in \mathbb{R}^2$. Firstly, we generated true soil chemicals, $z_p^{true}(\mathbf{x}_i)$, $1 \leq p \leq 3$ using stationary spatial Gaussian random field models with a first-order trend and Matern covariance function. Matérn covariance function was chosen as recommended by Ruppert et al. (2003). This is performed in R using GaussRF package. The Gaussian random field can be represented as

$$\begin{aligned} z_p^{true}(\mathbf{x}) &\sim MVN(\mu_{z_p}, \Sigma) \\ \mu_{z_p} &= [1 \ \mathbf{x}] \alpha_p \text{ where } \alpha_p = (\alpha_{1p}, \alpha_{2p}, \alpha_{3p}) \\ \Sigma(\mathbf{x}, \mathbf{x}') &= \sigma_p^2 (1 + \|\mathbf{x} - \mathbf{x}'\| / \rho_p) e^{(-\|\mathbf{x} - \mathbf{x}'\| / \rho_p)} \end{aligned} \quad (3)$$

Here, we assumed no nugget and zero mean. We used $\rho_p = 0.064$ which is $1/20 \times \max$ distance between sample sites following Ruppert et al. (2003)'s suggestion. We also used $\alpha_1 = (0.01, 0.01, 0.01)$, $\alpha_2 = (0.05, 0.05, 0.05)$, $\alpha_3 = (0.1, 0.1, 0.1)$ and $\sigma_p^2 = 0.1$.

The outcome $y^{true}(\mathbf{x}_i)$ is generated as follows:

$$\begin{aligned} \text{logit}(p_i^{true}) &= \beta_1 z_1^{true}(\mathbf{x}_i) + \dots + \beta_p z_p^{true}(\mathbf{x}_i) \\ y^{true}(\mathbf{x}_i) &\sim \text{rbern}(p_i^{true}) \end{aligned} \quad (4)$$

In this simulation, we assumed that chemical concentrations at the outcome sites are observed without error. We fitted a logistic model with entry parameters which follow a Bernoulli distribution.

$$\begin{aligned}
 y(\mathbf{x}_i) &\sim \text{Bernoulli}(p_i) \\
 \text{logit}(p_i | z_1^{\text{true}}(\mathbf{x}_i), \dots, z_p^{\text{true}}(\mathbf{x}_i)) &= b_0 + \gamma_1 b_1 z_1^{\text{true}}(\mathbf{x}_i) + \dots + \gamma_p b_p z_p^{\text{true}}(\mathbf{x}_i) + R_i \\
 \text{where } \gamma_1, \dots, \gamma_p &\text{ are entry parameters}
 \end{aligned} \tag{5}$$

Different scenarios for simulation study

First, we considered different logistic regression parameters to generate $y^{\text{true}}(\mathbf{x}_i)$. The parameter for the intercept was assumed zero for all the scenarios. Each set of parameters was chosen so that 40% of $y^{\text{true}} = 1$, which corresponds approximately to the overall rate of MRDD found in the Medicaid population in high risk areas of South Carolina.

Secondly, several priors for logistic regression parameters were investigated. Kuo and Mallick (1998) used a relatively non-informative Gaussian distribution $\beta_p \sim N(0, 16)$. Following Gelman (2006)'s recommendation, we used a uniform distribution for hyperprior for standard deviation instead using a fixed variance as follows:

$$\begin{aligned}
 \beta_p &\sim N(0, \sigma_{\beta_p}^2) \\
 \sigma_{\beta_p} &\sim U(0, k)
 \end{aligned} \tag{6}$$

We have considered $k=1$, $k=10$ and $k=100$. In comparison, we also used $\beta_p \sim N(0, 16)$ following Kuo and Mallick (1998).

Finally, two different Bernoulli distributions were considered for entry parameters.

$$\begin{aligned}
 1) \gamma_1, \dots, \gamma_p &\sim \text{Bernoulli}(0.5) \\
 2) \gamma_1, \dots, \gamma_p &\sim \text{Bernoulli}(p_p), p_p \sim \text{Beta}(0.5, 0.5)
 \end{aligned} \tag{7}$$

Method for summarization of simulation results

For each dataset, simulations were performed 200 times. In each simulation, two MCMC chains were run 2000 times using WinBUGS. The first 500 iteration were used as burn-in and thinning=5 was used to reduce autocorrelation. Variable selection results were summarized by 1) marginal total inclusion probability and 2) frequency of variable subset selection. Marginal total inclusion probability is calculated as follows:

$$p(\gamma_p) = \frac{1}{200} \left(\sum_{i=1}^{200} \text{Posterior mean of } \gamma_p \text{ from 600 samples at } i\text{th simulation} \right)$$

Frequency of variable subset selection is calculated as follows:

$$\begin{aligned}
 p(m_j) &= \frac{1}{200} \left(\sum_{i=1}^{200} p(m_j) \text{ from 600 sample at } i\text{th simulation} \right) \\
 \text{where } j &= 0, \dots, 7
 \end{aligned}$$

The results of these calculations are displayed in Figure 3 and Table 2, Table 3 and Table 4. For 200 simulations, two MCMC chains were run 2000 times using WinBUGS. The first 500 iterations were used as burn-in and thinning=5 was used to reduce autocorrelation.

Summary statistics were calculated for 600 samples from both chains. Convergence was checked calculating Gelman-Rubin statistics and visual inspection of trace plots.

Results of simulation study

Table 1 displays the parameter settings for the regression parameters in the four different data sets. When the beta values were relatively large (dataset1), Bayesian variable selection found the right model most frequently (47%). This is also supported by the marginal distribution of entry parameters shown in Figure 3. In contrast, when the beta values were relatively small (dataset2), the null model usually selected (27%) followed by a model including z2 only (22%) (Table 2). When beta values were close to zero (dataset3), the null model was selected 63% of the time. When the beta is relatively large for one variable and relatively small for the other (dataset4), the model with z1 was selected most often (38%) followed by the true model (32%). Using the dataset1, different priors were used for a normal distribution for the logistic regression parameter as described in Equation (6). We compared the results from 4 different cases and the results are presented in Table 3. The first column shows the result from the dataset1 presented in Table 2. When different priors were used for the logistic regression parameters, the variable selection results were different. When $\sigma_{\beta_p} \sim U(0,100)$, $\sigma_{\beta_p} \sim U(0,10)$ and $\sigma_{\beta_p} = 4$ were used, the intercept only model was usually selected. The marginal distribution of each entry parameter also indicated that no variable is significant. This shows that Bayesian variable selection can be very sensitive to prior distribution of the logistic regression parameters.

In Table 4 we display the results of varying the entry parameter prior assumptions. The first column represents the result from dataset1 in Table 2. Instead of fixing $p_p = 0.5$, we used a hyperprior for p_p employing a beta distribution. When we used a hyperprior for p_p , the intercept only model appears more often. This also confirms the need for careful selection of prior parameters.

Data Example

For demonstration purposes only, we have considered a partial analysis of a particular set of MRDD outcomes with associated soil chemical measurements. In this example we have used data obtained from a larger scale study of mental retardation and developmental delay (MRDD) in a Medicaid population in South Carolina. Detailed background to this study and its sampling methodology can be found elsewhere (Zhen et al (2008); Aelion et al (2008)). A variety of sampling sites was chosen within the state to satisfy various criteria. In particular, a primary criterion for selecting test sites was the need to examine areas where MRDD displayed a gradient of excess risk. This was assessed by Bayesian spatial cluster analysis using local likelihood methods (Hossain and Lawson, 2005; Lawson, 2006). During each month of pregnancy, residential addresses for mothers are recorded. The resulting data consists of residential location during a month of pregnancy and a binary indicator denoting MRDD state of the child. Other available mother and child covariates in this study include: mother's age, mother's ethnicity, mother's alcohol consumption during pregnancy, parity, birth weight, infant sex and follow-up time. We term these covariates as 'mother and child' or 'individual' covariates. In a chosen area, soil samples were collected from a network of sample sites which has a regular grid pattern covering the whole study area. These soil sample sites are spatially misaligned to the residential addresses of pregnant women. Thus, interpolation of chemical fields is required to link MRDD cases and normal controls to environmental exposure risk. Nine chemicals were measured in soil samples and include arsenic (As), barium (Ba), beryllium (Be), chromium (Cr), copper (Cu), lead (Pb), manganese (Mn), nickel (Ni) and mercury (Hg). We denote these observed soil chemical concentrations at sample sites as 'environmental' covariates.

In this analysis we examined all data for mothers pregnant and insured by Medicaid during 1996–2001 for strip2, which is anonymous for reporting purposes. This strip is mixed in terms of demographic characteristics and is primarily rural with a small town. The total number of mother child pairs in this strip is 144, with 38 cases of MRDD. Here, we simply exemplify the method by reporting a small set of mother child covariates and four chemicals (As, Cr, Hg, and Cu). The analysis is carried out using Bayesian Kriged soil chemical values as plug-in values for the variable selection. The soil chemicals were examined for the need for transformation and a Box-Cox transformation and a log transformation was employed. The Bayesian Kriged values (using `krige.bayes` in the R package `geoR`) were interpolated to the sites of the cases and controls based on an exponential covariance model and suitable non-informative priors for the variance and correlation parameters. All models were fitted with added random effect $R_i \sim N(0, \tau_v)$ with $\tau_v = 1/\sigma^2$, $\sigma \sim U(0, c)$. The four different modeling strategies (models 1–4) were as follows:

- **Model 1:** entry parameters ($\{\gamma_1, \dots, \gamma_7\}$) assumed for all 7 covariates (mom-age, birth-weight, parity, Cr, Cu, As, Hg) with $\{\gamma_k\} \sim \text{Bern}(p_p)$, and entry probability $p_p \sim \text{beta}(1,1)$ a uniform prior distribution for the entry probability. This is a uniform distribution similar to $\text{beta}(0.5,0.5)$ but more non-informative. An individual level uncorrelated random effect (v_i) was also included to allow for unobserved confounders in the data. This was assumed to have a zero mean Gaussian distribution with variance σ_v^2 where $\sigma_v \sim U(0,5)$
- **Model 2 :** as for model 1 but with $p_p = 0.5$, so that any variable is equally likely to be chosen apriori.
- **Model 3:** the three ‘mother child’ covariates are fixed in the model but the four soil chemicals have entry parameters with $p_p = 0.5$. Otherwise this was the same as model 2.
- **Model 4:** as for model 3 but with $p_p \sim \text{beta}(1,1)$.

For these models we ran samplers to convergence which took place within 20000 iterations. We have monitored the average deviance of these models, which measures the goodness of fit of the models and Table 5 displays the results. For model 3, where only the four chemicals have entry parameters and the ‘mother child’ covariates are forced into the model, it is clear that the deviance is considerably smaller and also the standard deviation is lower as well. From the simulation it was noted that fixing the Bernoulli parameter yielded a better recovery of the true model and it seems here that the lowest deviance model has that specified (comparison of model 3 and 4). Finally in Table 6 we present the parameter estimates found for model 3. The results are for a converged sample and a sample size of 10000. The posterior average parameter estimates and 95% credible intervals are reported as well as the posterior expected estimates for the inclusion probabilities. The latter are estimated as the average of the entry parameters over the converged sample for the 4 chemicals. Overall, it is noticeable that the main significant covariate is mother’s age which has a negative effect on the outcome, whereas the other ‘mother child’ covariates are not significant. Finally although the soil chemicals don’t show significance in this example, they do display considerably different inclusion probabilities. In this case, Cr has the highest probability of inclusion (0.708) closely followed by As (0.459), whereas Cu and Hg are both very low (0.132, 0.054). Using the criterion of Barbieri and Berger (2004) that supports all variables with inclusion probability > 0.5 , then Cr would be included in the model even when in the final sample it is not significant. As is marginally supported but also does not achieve significance.

As a comparison we have also fitted a simple mixed effect logistic regression with 3 mom-baby covariates and 4 chemicals as covariates i.e. model 3 but without entry parameters. The

posterior mean estimates of the parameters display similar behavior to those with entry parameters in that mother's age is significant but other variables are not. However, the posterior marginal density estimate for the Cr parameter (not shown) is skewed with most of its mass above 0, while the 95% credible interval crosses zero slightly. This is support for the finding that the inclusion probability is quite high but the estimated significance is low.

Discussion and Conclusions

We have highlighted important issues in the analysis of spatially-referenced health outcomes and explanatory covariates that have spatial expression. We have stressed the complex nature of the interpolation, measurement error and variable selection involved and the need for care in the choice of model formulation so that the best selection of covariates can be achieved. It is clear that these issues are very important in any application where environmental covariates are interpolated and selected in relation to a health outcome.

In particular, our simulation suggested that a fixed prior inclusion probability made the best recovery, while a particular choice of variance prior distribution for the regression parameters was also favored. This was confirmed in the MRDD example where we found that the lowest deviance model was one with fixed prior inclusion probability.

Note that the analysis of the real data example was completely carried out on WinBUGS which is freely available, and hence these methods could be adopted widely given such access. The specification of the entry parameter models for variable selection avoids the need to carry out time consuming stepwise procedures and also can be fitted within one model formulation. It is also simple to set up models with these parameters within the WinBUGS model. Nonetheless there should be considerable care taken to correctly identify the appropriate prior distributions used in the modeling both for the entry parameters and the regression parameters. The most successful priors based on the simulation and real example were $\sigma_\beta \sim U(0,10)$ for the regression parameters and fixed value of $0.5 p_p = 0.5$ for inclusion. In general, it is probably best to consider a range of prior specifications to assess sensitivity in any analysis.

Acknowledgments

We would like to acknowledge support for this work from the National Institutes of Health, National Institute of Environmental Health Sciences, R01 Grant No. ES012895-01A1.

References

- Aelion CM, Davis HT, McDermott S, Lawson AB. Metal concentrations in rural topsoil in South Carolina: Potential for human health impact. *Sci. Total Environment*. 2008; 402:149–156.
- Barbeieri M, Berger J. Optimal Predictive Model Selection. *Annals of Statistics*. 2004; 32(3):870–897.
- Cai B, Dunson D. Bayesian covariance selection in generalized linear mixed models. *Biometrics*. 2006; 62:446–457. [PubMed: 16918908]
- Carlin BP, Chib S. Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society. Series B (Methodological)*. 1995; 57(3):473–484.
- Chen M-H, Ibrahim JG, Yiannoutsos C. Prior elicitation, variable selection and Bayesian computation for logistic regression models. *Journal of the Royal Statistical Society (Series B): Statistical Methodology*. 1999; 61(1):223–242.
- Dellaportas P, Forster JJ, Ntzoufras I. On Bayesian model and variable selection using MCMC *Statistics and Computing*. 2002; 12:27–36.
- Dominici F, Zeger SL, Samet JM. Combining Evidence on Air Pollution and Daily Mortality from the Largest 20 US cities: A Hierarchical Modeling Strategy. *Royal Statistical Society, Series A, with discussion*. 1999; 163:263–302.

- Dominici F, Peng RD, Ebisu K, Zeger SL, Samet JM, Bell M. Does the Effect of PM10 on Mortality Depend on PM Nickel and Vanadium Content? A Reanalysis of the NMMAPS Data, Environmental Health Perspectives. 2007 (to appear).
- George EI, McCulloch RE. Variable Selection Via Gibbs Sampling. Journal of the American Statistical Association. 1993; 88(423):881–889.
- Gerlach R, Bird R, Hall A. Bayesian variable selection in logistic regression: predicting company earnings direction. Australian & New Zealand Journal of Statistics. 2002; 44(2):155–168.
- Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 1995; 82(4):711–732.
- Hoeting J, Davis R, Merton A, Thompson S. Model selection for geostatistical Models. Ecological Applications. 2006; 16:87–98. [PubMed: 16705963]
- Hossain MM, Lawson AB. Local Likelihood Disease Clustering: Development and Evaluation Environmental and Ecological Statistics. 2005; 12(3):259–273.
- 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. 2008 (to appear).
- Kuo L, Mallick B. Variable selection for regression models. Sankhya. 1998; B60:65–81.
- Lawson AB. Disease Cluster Detection: A critique and a Bayesian Proposal. Statistics in Medicine. 2006; 25:897–916. [PubMed: 16453377]
- Lawson, AB.; Banerjee, S. Bayesian Spatial Analysis. In: Fotheringham, S.; Rogerson, P., editors. The Handbook of Spatial Analysis. Sage; 2008. ch 9
- Ma B, Lawson AB, Liu Y. Evaluation of Bayesian Models for focused clustering in Health data. Environmetrics. 2007; 18:1–16.
- Reich B, Hodges J, Zadnik V. Effects of Residual Smoothing on the Posterior of the Fixed Effects in Disease-Mapping Models. Biometrics. 2006; 62:1197–1206. [PubMed: 17156295]
- Reich BJ, Fuentes M, Burke J. Analysis of the effects of ultrafine particulate matter while accounting for human exposure. Environmetrics. 2008 (to appear).
- Ruppert, D.; Wand, M.; Carroll, R. Semi-parametric Regression. New York: Cambridge University Press; 2003.
- Smith M, Fahrmeier L. Spatial Bayesian Variable Selection with application to Functional Magnetic Resonance Imaging. Journal of the American Statistical Association. 2007; 102:417–431.
- Zhen H, Lawson AB, McDermott S, Pande Lamichhane A, Aelion CM. Spatial analysis of mental retardation and maternal residence during pregnancy. Geospatial Health. 2008; 2(2):173–182. [PubMed: 18686266]

Figure 1.

Study region strip with the residential addresses of MRDD cases (1) and controls (o) marked. The axes are anonymous by shifting and the x axis covers a distance of approximately 9.4 kilometers.

Figure 2.

Two sets of grids for simulation: X represents the first grid and O represents the second grid

Figure 3.
Distribution of marginal total inclusion probability from 200 simulations for Dataset I

Table 1

Different logistic regression parameters for each scenario

	dataset 1	dataset 2	dataset 3	dataset 4
b1	−0.7	−0.3	−0.1	−0.8
b2	0.7	0.3	0.1	0.3
b3	0	0	0	0

Table 2

Mean of Posterior model probability from 200 simulations for different scenarios. The different models and datasets are described in the text.

	dataset1	dataset2	dataset3	dataset4
Intercept only	5	27	63	8
z1	16	18	13	38
z2	15	22	9	6
z1+z2	47	16	2	32
z3	1	6	9	1
z1+z3	3	3	2	6
z2+z3	4	4	2	2
z1+z2+z3	8	3	0	6

Table 3

Mean of Posterior model probability from 200 simulations with different logistic regression coefficient priors

model	$\sigma_{\beta} \sim U(0,10)$	$\sigma_{\beta} \sim U(0,100)$	$\sigma_{\beta} \sim U(0,1)$	$\sigma_{\beta} = 4$
Intercept only	5	73	20	38
z1	16	9	12	16
z2	15	10	14	17
z1+z2	47	1	14	8
z3	1	5	11	10
z1+z3	3	1	9	4
z2+z3	4	1	10	5
z1+z2+z3	8	0	10	2

Table 4

Mean of Posterior model probability from 200 simulations with different p

model	p=0.5	p~Beta(0.5, 0.5)
Intercept only	5	37
z1	16	14
z2	15	16
z1+z2	47	10
z3	1	11
z1+z3	3	5
z2+z3	4	5
z1+z2+z3	8	3

Table 1

Results of variable selection model fits for the study area

	Model 1	Model 2	Model 3	Model 4
Deviance	73.07	72.97	66.22	68.50
sd	20.95	24.01	19.16	22.90

Parameter estimates for the model 3: posterior average estimates with 95% credible intervals and standard deviation obtained from a posterior sample size 10,000 following convergence. Posterior expected inclusion probability for the four chemicals is also shown

Table 6

variable	estimate	2.5% level	97.5% level	sd
intercept	-0.825	-9.331	8.277	4.473
Mom age	-0.302	-0.618	-0.048	0.157
weight	-1.46E-5	-0.002	0.002	0.001
parity	0.407	-0.454	1.481	0.49
Cr	0.171	-3.563	3.691	1.547
Cu	-0.023	-6.113	5.938	2.697
As	0.333	-4.784	5.101	2.174
Hg	0.0547	-6.199	6.265	2.876
				0.054