Fast Estimation and Selection of Autologistic Regression Models via Penalized Pseudolikelihood

Rao Fu[a], Andrew L. Thurman[b], Michelle M. Steen-Adams[c], Jun Zhu[d]

[a]*Department of Statistics, University of Wisconsin, Madison, WI 53706*
[b]*Department of Statistics, University of Wisconsin, Madison, WI 53706*
[c]*Department of Environmental Studies, University of New England, Biddeford, ME 04005*
[d]*Department of Statistics and Department of Entomology, University of Wisconsin, Madison, WI 53706*

Abstract

Autologistic regression models are suitable for relating spatial binary responses in ecology to covariates such as environmental factors. Model parameters can be estimated via maximum likelihood or Bayesian methods, although they often involve Monte Carlo simulation and tend to be computationally intensive. For big ecological data, pseudolikelihood estimation is appealing due to its ease of computation, but at least two challenges remain. Although an important issue, it is unclear how model selection may be carried out under pseudolikelihood. In addition, for assessing the variation of pseudolikelihood estimates, parametric bootstrap using Monte Carlo simulation is often used but may be infeasible for very large data sizes. Here both these issues are addressed by developing a penalized pseudolikelihood estimation method and an approximation of the variance of the parameter estimates. A simulation study is conducted to evaluate the performance of the proposed method, followed by a data example in a study of land cover in relation to land ownership characteristics. These models and methods are further extended to deal with spatial-temporal binary data.

Key words: Binary data; Markov random field; model selection; spatial statistics.

Email addresses: `rfu7@wisc.edu` (Rao Fu), `athurman@wisc.edu` (Andrew L. Thurman), `msteenadams@une.edu` (Michelle M. Steen-Adams), `jzhu@stat.wisc.edu` (Jun Zhu)

1. Introduction

Autologistic regression is suitable for modeling the relationship between a spatial binary response and covariates in environmental and ecological studies (see, e.g., Gumpertz *et al.* (1997); Huffer and Wu (1998); He *et al.* (2003)). For big ecological data, however, there are several challenging issues in terms of both computation and methodology for statistical inference. This paper aims to address some of these issues and in particular, we develop a new, computationally efficient method for selection of covariates and estimation of model parameters.

Several approaches are possible for modeling spatial binary response in relation to covariates. One possibility is spatial generalized linear mixed models (GLMM), where the response variable is categorical and is linked to a regression on covariates and a spatial random effect as two additive components in a link function. Bayesian hierarchical modeling provides a general framework for drawing statistical inference about the regression coefficients and the spatial random effects, although maximum likelihood estimation is also possible (Diggle and Ribeiro, 2007; Banerjee *et al.*, 2003). Bayesian or maximum likelihood estimation for spatial GLMM tends to be computationally intensive, as the computation often involves Markov chain Monte Carlo (MCMC) algorithms.

Autologistic regression offers an attractive alternative. For an autologistic model (without regression), maximum pseudolikelihood (MPL) estimation is easy to implement and fast to compute (Besag, 1972, 1974). Maximum likelihood estimation is a challenging problem because of a normalizing constant that is analytically intractable in the likelihood function. The idea of Monte Carlo maximum likelihood (MCML) can be adopted to deal with the normalizing constant problem (Geyer, 1994). Various researchers have considered Bayesian estimation using innovative ideas such as auxiliary variables (Moeller *et al.*, 2006), Monte Carlo approximations (Sun and Clayton, 2008), and reduced dependence approximations (Friel *et al.*, 2009). These methods tend to focus on autoregression but not as much on regression.

For an autologistic model with regression, the idea of MPL can be applied (Gumpertz *et*

al., 1997) as well as MCML (Huffer and Wu, 1998). More recently, Caragea and Kaiser (2009) suggested to center the autocovariates in the model, which allows more meaningful interpretation of the coefficients in the autologistic model. Here we will consider the centering suggested by Caragea and Kaiser (2009) and focus on MPL. Compared with MCML, MPL is much faster to compute, which can be a great advantage for the analysis of big data (Hughes *et al.*, 2011; Wang and Zheng, 2012; Zheng and Zhu, 2008). There are, however, at least two issues that appear to be unresolved. First, there appear to be no existing methods for the selection of covariates using MPL. For autologistic regression, Zhu *et al.* (2008) applied MCML for regression and used an approximate information criterion for model comparison. When the number of candidate models is large, however, the usefulness of information criteria is fairly limited due to a high computational cost. Second, the estimates of the variance of the MPL estimates are largely based on parametric bootstrap where the data in the bootstrap are simulated using MCMC (Gumpertz *et al.*, 1997; Zhu *et al.*, 2005). There clearly is a need to improve these aspects of the MPL method, if it is to be used for analyzing big ecological data.

Different approaches can be taken for variable selection in a standard linear regression assuming independent response variables. Regularization (or, penalized) methods for standard linear regression are becoming popular using, for example, least absolute shrinkage and selection operator (Lasso) (Tibshirani, 1996), adaptive Lasso (Zou, 2006), and least angle regression (LARS) algorithm (Efron *et al.*, 2004). Penalized methods have also been developed for spatial data with continuous response variables (see, e.g., Huang *et al.* (2010), Zhu and Liu (2009), Zhu *et al.* (2010)) but not much with discrete responses. In this paper, we utilize the idea of adaptive Lasso and develop a penalized pseudolikelihood estimation method for selection of covariates. In addition, unlike the existing approach to estimate the variance of the parameter estimates by bootstrapping, we propose an analytical form based on asymptotic results.

In the remainder of the paper, we will present the autologistic regression models in Section 2 and describe our method in Section 3. A simulation study is conducted in Section 4

to evaluate the performance of the proposed method, followed by a data example in a study of land cover in relation to land ownership characteristics in Section 5. Conclusions and a discussion about extending our method to deal with spatial-temporal binary data are in Section 6.

2. Autologistic Regression Models

2.1. Autologistic Regression

For $i = 1, \ldots, n$, let Z_i denote the response variable at the ith site on a spatial lattice, such that $Z_i = 0$ or 1. Let $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ denote the binary response variables at all n sites on this lattice and $\mathbf{Z}_{-i} = (Z_1, \ldots, Z_{i-1}, Z_i, \ldots, Z_n)'$ denote the vector that has all the response variables of $\mathbf{Z}$ except for Z_i. Further, consider a pre-specified spatial neighborhood structure. For example, the first-order neighborhood consists of the four nearest neighbors on a regular grid. Let $\mathcal{N}_i$ denote the set of indices of the neighbors of site i and let $i' \sim i$ denote that site i' is a neighbor of site i such that $i' \in \mathcal{N}_i$.

To model the response variables $\mathbf{Z}$, we assume a Markov random field in the sense that the probability of the ith response, Z_i, conditional on $\mathbf{Z}_{-i}$ depends on only the responses in the neighborhood, $Z_{i'}$, where $i' \in \mathcal{N}_i$. That is,

$$p(Z_i | \mathbf{Z}_{-i}) \;\; = \;\; p(Z_i | Z_{i'} : i' \sim i). \tag{1}$$

Further, we assume that the conditional distribution $p(Z_i | Z_{i'} : i' \sim i)$ is Bernoulli with success probability π_i such that

$$\pi_i \;\; = \;\; p(Z_i = 1 | Z_{i'} : i' \sim i) \tag{2}$$

where π_i depends on $Z_{i'}$ for $i' \in \mathcal{N}_i$ via a logit link function

$$\text{logit}(\pi_i) \;\; = \;\; \log \frac{\pi_i}{1 - \pi_i} = \boldsymbol{x}_i' \boldsymbol{\beta} + \sum_{i' \sim i} \eta_{ii'} Z_{i'}. \tag{3}$$

In (3), $\boldsymbol{x}_i$ denotes a $(p+1)$-dimensional vector of intercept 1 and p covariates at site i and $\boldsymbol{\beta}$ denotes the corresponding vector of the regression coefficients. Also, $\eta_{ii'}$ for $i' \sim i$ denotes

the autoregression coefficient between sites i and i' and the term $Z_{i'}$ can be thought of as an autocovariate. In the special case that all the autoregression coefficients are zero ($\eta_{ii'} = 0$), the model (1)–(3) reduces to a traditional logistic regression with independent responses.

We will restrict our attention to a constant autoregression coefficient $\eta_{ii'} = \eta$ for all $i' \in \mathcal{N}_i$ and the logit link becomes

$$\text{logit}(\pi_i) \;\; = \;\; \log \frac{\pi_i}{1 - \pi_i} = \boldsymbol{x}_i'\boldsymbol{\beta} + \sum_{i'\sim i} \eta Z_{i'}. \tag{4}$$

However, this assumption may be relaxed to include different orders of neighborhood (Zhu *et al.*, 2010).

2.2. Centered Model

Caragea and Kaiser (2009) proposed to center the autocovariate around its expected value to achieve more meaningful interpretations for regression purposes. Here we consider this centered autologistic regression model by modifying (4) to

$$\text{logit}(\pi_i) \;\; = \;\; \log \frac{\pi_i}{1 - \pi_i} = \boldsymbol{x}_i'\boldsymbol{\beta} + \sum_{i'\sim i} \eta(Z_{i'} - \mu_{i'}), \tag{5}$$

where $\mu_{i'}$ is the expectation of $Z_{i'}$ assuming independence (i.e., $\eta = 0$) and thus

$$\mu_{i'} \;\; = \;\; \frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})}. \tag{6}$$

The terms $\{\mu_i\}_{i=1}^n$ may be interpreted as the large-scale structure of the random field $\boldsymbol{Z}$ and with regression, relate the response variables to covariates. The difference between the autocovariate and μ_i is $Z_i - \mu_i$ and represents the small-scale structure after adjusting for the large-scale structure. The effects of covariates are therefore captured in the regression coefficients $\boldsymbol{\beta}$ and the local spatial dependence in the autoregression coefficient η. In contrast, the uncentered autologistic regression model (3) does not permit such a clear interpretation (Caragea and Kaiser, 2009).

2.3. Alternative Coding

Although it is common to use 0 or 1 to code failure or success in the Bernoulli trial, in the uncentered model, the positive-valued autocovariates would artificially inflate the chance

of success (Hughes *et al.*, 2011). An alternative coding is -1 for failure and $+1$ for success (Gaetan and Guyon, 2010). We believe that the issue with the 0-1 coding can be mitigated by the ± 1 coding, however. If there are more $+1$ autocovariates than -1 in a neighborhood, then the sum of autocovariates is positive and thus the success probability increases (or decreases) with a positive (or negative) autoregression coefficient η; and the opposite holds when there are more -1 autocovariates than $+1$ in the neighborhood.

Here, we let $\tilde{Z}_i = 2Z_i - 1$ denote the binary response variable at site i under this alternative coding. Define $\tilde{\boldsymbol{Z}}$ and $\tilde{\boldsymbol{Z}}_{-i}$ as the counterparts of $\boldsymbol{Z}$ and $\boldsymbol{Z}_{-i}$. Let $\tilde{\pi}_i = P(\tilde{Z}_i = +1 | \tilde{Z}_{i'} : i' \sim i)$ denote the success probability. The logit link function in the uncentered model is

$$\text{logit}(\tilde{\pi}_i) \;=\; \log \frac{\tilde{\pi}_i}{1 - \tilde{\pi}_i} = \boldsymbol{x}_i' \tilde{\boldsymbol{\beta}} + \sum_{i' \sim i} \tilde{\eta} \tilde{Z}_{i'}, \tag{7}$$

but in the centered model is

$$\text{logit}(\tilde{\pi}_i) \;=\; \log \frac{\tilde{\pi}_i}{1 - \tilde{\pi}_i} = \boldsymbol{x}_i' \tilde{\boldsymbol{\beta}} + \sum_{i' \sim i} \tilde{\eta} (\tilde{Z}_{i'} - \tilde{\mu}_{i'}), \tag{8}$$

where it can be shown that the expectation of $\tilde{Z}_i$ assuming independence is

$$\tilde{\mu}_{i'} \;=\; \frac{\sinh(\boldsymbol{x}_{i'}' \tilde{\boldsymbol{\beta}})}{\cosh(\boldsymbol{x}_{i'}' \tilde{\boldsymbol{\beta}})}. \tag{9}$$

In fact, there is a one-to-one correspondence between the coefficients under the two different codings. For uncentered models, we have $\tilde{\eta} = \eta/4$ and $\boldsymbol{x}_i' \tilde{\boldsymbol{\beta}} = \boldsymbol{x}_i' \boldsymbol{\beta}/2 + |\mathcal{N}_i| \eta/4$, where $|\mathcal{N}_i|$ denotes the cardinality of $\mathcal{N}_i$ (i.e., the number of neighbors of the ith site). A similar connection can be made for centered models. See Appendix for more details.

3. Estimation and Selection of Autologistic Models

3.1. Maximum Pseudolikelihood

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \eta)'$ denote the vector of all the parameters for either centered or uncentered models using the 0-1 coding of the response variable. Let $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}', \tilde{\eta})'$ be analogous to $\boldsymbol{\theta}$ but for the models using ± 1 coding. We adopt maximum pseudolikelihood here for estimating

the model parameters such that the pseudolikelihood function is the product of the full conditional probabilities at all sites (Cressie, 1993).

In an uncentered model, the full conditional distribution of Z_i is

$$p(Z_i = 1|\boldsymbol{Z}_{-i}; \boldsymbol{\theta}) = p(Z_i = 1|Z_{i'} : i' \sim i; \boldsymbol{\theta}) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i} Z_{i'})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i} Z_{i'})} \tag{10}$$

and $p(Z_i = 0|\boldsymbol{Z}_{-i}; \boldsymbol{\theta}) = 1/\{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i} Z_{i'})\}$. Thus, the log-pseudolikelihood function for the uncentered model is

$$\ell_{\mathrm{p}}^{\mathrm{u}}(\boldsymbol{\theta}|\boldsymbol{Z}) = \sum_{i=1}^{n} \log \frac{\exp\{Z_i(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i} Z_{i'})\}}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i} Z_{i'})}. \tag{11}$$

Similarly, in a centered model, the full conditional distribution of Z_i is

$$p(Z_i = 1|\boldsymbol{Z}_{-i}; \boldsymbol{\theta}) = \frac{\exp\{\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i}(Z_{i'} - \mu_{i'})\}}{1 + \exp\{\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i}(Z_{i'} - \mu_{i'})\}} \tag{12}$$

and thus, the log-pseudolikelihood function is

$$\ell_{\mathrm{p}}^{\mathrm{c}}(\boldsymbol{\theta}|\boldsymbol{Z}) = \sum_{i=1}^{n} \log \frac{\exp[Z_i\{\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i}(Z_{i'} - \mu_{i'})\}]}{1 + \exp\{\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i' \sim i}(Z_{i'} - \mu_{i'})\}}. \tag{13}$$

Maximizing the log-pseudolikelihood function gives the maximum pseudolikelihood estimate (MPLE) of $\boldsymbol{\theta}$ and we denote them as $\widehat{\boldsymbol{\theta}}_{\mathrm{p}} = \mathrm{argmax}\{\ell_{\mathrm{p}}(\boldsymbol{\theta}|\boldsymbol{Z})\}$ for either $\ell_{\mathrm{p}}^{\mathrm{u}}(\boldsymbol{\theta}|\boldsymbol{Z})$ in (11) or $\ell_{\mathrm{p}}^{\mathrm{c}}(\boldsymbol{\theta}|\boldsymbol{Z})$ in (13).

Under the alternative ± 1 coding of responses, the full conditional distribution of $\tilde{Z}_i$ in an uncentered model is

$$p(\tilde{Z}_i = 1|\tilde{\boldsymbol{Z}}_{-i}; \tilde{\boldsymbol{\theta}}) = \frac{\exp(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \sim i} \tilde{Z}_{i'})}{2 \cosh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \sim i} \tilde{Z}_{i'})} \tag{14}$$

and in a centered model is

$$p(\tilde{Z}_i = 1|\tilde{\boldsymbol{Z}}_{-i}; \tilde{\boldsymbol{\theta}}) = \frac{\exp\{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \sim i}(\tilde{Z}_{i'} - \tilde{\mu}_{i'})\}}{2 \cosh\{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \sim i}(\tilde{Z}_{i'} - \tilde{\mu}_{i'})\}}. \tag{15}$$

The log-pseudolikelihood function for the uncentered model is

$$\ell_{\mathrm{p}}^{\mathrm{u}}(\tilde{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \log \frac{\exp\{\tilde{Z}_i(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \sim i} \tilde{Z}_{i'})\}}{2 \cosh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta} \sum_{i' \sim i} \tilde{Z}_{i'})} \tag{16}$$

7

and for the centered model is

$$\ell_{\mathrm{p}}^{\mathrm{c}}(\tilde{\boldsymbol{\theta}}) \;=\; \sum_{i=1}^{n} \log \frac{\exp[\tilde{Z}_i\{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i'\sim i}(\tilde{Z}_{i'} - \tilde{\mu}_{i'})\}]}{2\cosh\{\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i'\sim i}(\tilde{Z}_{i'} - \tilde{\mu}_{i'})\}}. \tag{17}$$

We denote the MPLE of $\tilde{\boldsymbol{\theta}}$ by $\widehat{\tilde{\boldsymbol{\theta}}}_{\mathrm{p}} = \mathrm{argmax}\{\ell_{\mathrm{p}}(\tilde{\boldsymbol{\theta}}|\tilde{\boldsymbol{Z}})\}$ for either $\ell_{\mathrm{p}}^{\mathrm{u}}(\tilde{\boldsymbol{\theta}}|\tilde{\boldsymbol{Z}})$ in (16) or $\ell_{\mathrm{p}}^{\mathrm{c}}(\tilde{\boldsymbol{\theta}}|\tilde{\boldsymbol{Z}})$ in (17).

3.2. Variable Selection

For an uncentered or centered model using the 0-1 coding of the response, we propose a penalized log-pseudolikelihood function

$$\ell_{\mathrm{pp}}(\boldsymbol{\theta}) = \ell_{\mathrm{p}}(\boldsymbol{\theta}) - n\sum_{j=1}^{p}\lambda_j|\beta_j| \tag{18}$$

where in the second term of (18), the regression coefficients are subject to an ℓ_1-penalty function and λ_j is a regularization parameter for the jth regression coefficient β_j, where $j = 1,\ldots,p$. That is, no penalty is applied to the intercept β_0 nor the autoregression coefficient η. Maximizing (18) would enable variable selection and parameter estimation simultaneously, because the regression coefficients of the "unimportant" covariates will be shrunk to zero under the ℓ_1-penalty and nonzero estimates of the other parameters will be given. Because the regularization parameter λ_j varies by j, the penalization is akin to adaptive Lasso (Zou, 2006).

To maximize (18), we propose a one-step approximation. We first set the MPLE of $\boldsymbol{\theta}$ to be the initial value. That is, $\widehat{\boldsymbol{\theta}}^{(0)} = \widehat{\boldsymbol{\theta}}_{\mathrm{p}} \equiv (\widehat{\boldsymbol{\beta}}^{(0)\prime}, \widehat{\eta})'$. We approximate the penalized log-pseudolikelihood function (18) up to a constant by

$$\begin{aligned}
\ell_{\mathrm{pp}}(\boldsymbol{\beta}) \;=\; & (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(0)})'\frac{\partial \ell_{\mathrm{p}}(\widehat{\boldsymbol{\theta}}^{(0)})}{\partial \boldsymbol{\beta}} - (1/2)(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(0)})'\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}}^{(0)})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}^{(0)}) \\
& -n\sum_{j=1}^{p}\lambda_j|\beta_j|
\end{aligned} \tag{19}$$

where $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = -\frac{\partial^2 \ell_{\mathrm{p}}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}$ is the negative of the second-order derivative of $\ell_{\mathrm{p}}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\beta}$. We update $\widehat{\boldsymbol{\beta}}^{(0)}$ to be $\widehat{\boldsymbol{\beta}}^{(1)} = \underset{\boldsymbol{\beta}}{\arg\max}\{\ell_{\mathrm{pp}}(\boldsymbol{\beta})\}$. It can be shown that this solution can

8

be attained equivalently by

$$\widehat{\boldsymbol{\beta}}^{(1)} = \arg\min_{\boldsymbol{\beta}} \left\{ (1/2)(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*)'(\boldsymbol{y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*) + n\sum_{j=1}^{p} |\beta_j^*| \right\} \quad (20)$$

where

$$\boldsymbol{y}^* = (\boldsymbol{B}^{-1})' \left\{ \frac{\partial \ell_{\mathrm{p}}(\widehat{\boldsymbol{\theta}}^{(0)})}{\partial \boldsymbol{\beta}} + \boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}}^{(0)})\widehat{\boldsymbol{\beta}}^{(0)} \right\}, \boldsymbol{X}^* = \boldsymbol{B}\mathrm{diag}\{\lambda_j^{-1}\}_{j=1}^{p},$$

$$\boldsymbol{\beta}^* = \mathrm{diag}\{\lambda_j\}_{j=1}^{p}\boldsymbol{\beta}, \ \boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}}^{(0)}) = \boldsymbol{B}'\boldsymbol{B}.$$

The minimization in (20) can be solved by a LARS algorithm and the computation is generally fast. We approximate the maximum penalized pseudolikelihood estimate (MPPLE), or $\widehat{\boldsymbol{\beta}}_{\mathrm{pp}}$, by $\widehat{\boldsymbol{\beta}}^{(1)}$ and let $\widehat{\boldsymbol{\theta}}_{\mathrm{pp}} = (\widehat{\boldsymbol{\beta}}_{\mathrm{pp}}', \widehat{\eta})'$. In particular, some of the entries of $\widehat{\boldsymbol{\beta}}_{\mathrm{pp}}$ are zero (or nonzero), indicating that the corresponding covariates are selected out of (or kept in) the final selected model.

In addition, for estimating the regularization parameters $\{\lambda_j\}_{j=1}^{p}$, we let $\lambda_j = \lambda \log(n)/(n|\widehat{\beta}_j|)$ and compute a Bayesian information criterion (BIC) type criterion for determining λ. In particular, define

$$\mathrm{BIC}_{\mathrm{p}}(\lambda) = -2\ell_{\mathrm{p}}(\widehat{\boldsymbol{\theta}}_{\mathrm{pp}}; \lambda) + e(\lambda)\log(n) \quad (21)$$

where $e(\lambda)$ denotes the number of nonzero estimates in $\widehat{\boldsymbol{\beta}}_{\mathrm{pp}}$. We estimate λ by $\widehat{\lambda} = \mathrm{argmin}_{\lambda}\{\mathrm{BIC}_{\mathrm{p}}(\lambda)\}$. Even though $\mathrm{BIC}_{\mathrm{p}}(\lambda)$ is not the actual BIC under maximum likelihood, it is in the same spirit and our empirical study via simulation will demonstrate that the method works well as a variable selection technique.

For models using ± 1 as the response variables, the MPPLE of $\tilde{\boldsymbol{\theta}}$ can be obtained analogously by considering $\ell_{\mathrm{p}}(\tilde{\boldsymbol{\theta}})$ and applying the same penalty as in (18). The resulting maximum penalized pseudolikelihood estimates are denoted as $\widehat{\tilde{\boldsymbol{\theta}}}_{\mathrm{pp}} = (\widehat{\tilde{\boldsymbol{\beta}}}_{\mathrm{pp}}', \widehat{\tilde{\eta}})'$.

3.3. Variance Estimation

Let $\boldsymbol{\mathcal{J}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{i' \sim i, i'=i} \frac{\partial \ell_{\mathrm{p}i}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \{\frac{\partial \ell_{\mathrm{p}i'}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\}'$, where $\ell_{\mathrm{p}i}$ is the log-pseudolikelihood of site i. By arguments similar to Gaetan and Guyon (2010), the following central limit theorem

9

holds for the MPLE $\widehat{\boldsymbol{\beta}}_{\mathrm{p}}$:

$$\{\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\beta}}_{\mathrm{p}})\}^{-1/2}\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}}_{\mathrm{p}})\{\widehat{\boldsymbol{\beta}}_{\mathrm{p}} - \boldsymbol{\beta}\} \to_d \ \mathcal{N}_p(0, \boldsymbol{I}_{p+1}). \tag{22}$$

Therefore, an estimate of the variance of $\widehat{\boldsymbol{\beta}}_{\mathrm{p}}$ is

$$\widehat{Var}(\widehat{\boldsymbol{\beta}}_{\mathrm{p}}) \approx \boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}}_{\mathrm{p}})^{-1}\boldsymbol{\mathcal{J}}(\widehat{\boldsymbol{\beta}}_{\mathrm{p}})\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}}_{\mathrm{p}})^{-1}. \tag{23}$$

For the MPPLE, we replace the MPLE $\widehat{\boldsymbol{\beta}}_{\mathrm{p}}$ in the variance formula (23) with the vector of non-zero entries of the MPPLE $\widehat{\boldsymbol{\beta}}_{\mathrm{pp}}$. Thus, the operations involved in the variance estimation are of dimension up to $(p+1) \times (p+1)$ and generally manageable. We will show by a simulation study that these variance estimates perform reasonably well for finite samples.

More specifically, for the uncentered model and the 0-1 coding, the $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta})$ and $\boldsymbol{\mathcal{J}}(\boldsymbol{\beta})$ are

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i} Z_{i'})}{\{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i} Z_{i'})\}^2} \quad \text{and}$$

$$\boldsymbol{\mathcal{J}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{i'\sim i, i'=i} \left[\boldsymbol{x_i} \left\{ Z_i - \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i} Z_{i'})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i} Z_{i'})} \right\} \right]$$
$$\left[\boldsymbol{x}_{i'} \left\{ Z_{i'} - \frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta} + \eta \sum_{i''\sim i'} Z_{i''})}{1 + \exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta} + \eta \sum_{i''\sim i'} Z_{i''})} \right\} \right]'.$$

For the centered model, the $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta})$ and $\boldsymbol{\mathcal{J}}(\boldsymbol{\beta})$ are

$$\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\boldsymbol{x}_i - \eta \sum_{i'\sim i} \frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})}{\{1 + \exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})\}^2} \boldsymbol{x}_{i'} \right] \left[\boldsymbol{x}_i - \eta \sum_{i'\sim i} \frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})}{\{1 + \exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})\}^2} \boldsymbol{x}_{i'} \right]'$$
$$\frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i}(Z_{i'} - \mu_{i'}))}{\{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i}(Z_{i'} - \mu_{i'}))\}^2}$$
$$+ \sum_{i=1}^{n} \sum_{i'\sim i} \eta \left\{ Z_i - \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i''\sim i}(Z_{i''} - \mu_{i''}))}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i''\sim i}(Z_{i''} - \mu_{i''}))} \right\} \frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta}) - \{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})\}^2}{\{1 + \exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})\}^3} \boldsymbol{x}_{i'} \boldsymbol{x}_{i'}'$$

$$\boldsymbol{\mathcal{J}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{i'\sim i, i'=i}$$
$$\left[\left\{ \boldsymbol{x}_i - \eta \sum_{i'\sim i} \frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta})}{(1 + \exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta}))^2} \boldsymbol{x}_{i'} \right\} \left\{ Z_i - \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i}(Z_{i'} - \mu_{i'}))}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i}(Z_{i'} - \mu_{i'}))} \right\} \right]$$
$$\left[\left\{ \boldsymbol{x}_{i'} - \eta \sum_{i''\sim i'} \frac{\exp(\boldsymbol{x}_{i''}'\boldsymbol{\beta})}{(1 + \exp(\boldsymbol{x}_{i''}'\boldsymbol{\beta}))^2} \boldsymbol{x}_{i''} \right\} \left\{ Z_{i'} - \frac{\exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta} + \eta \sum_{i''\sim i'}(Z_{i''} - \mu_{i''}))}{1 + \exp(\boldsymbol{x}_{i'}'\boldsymbol{\beta} + \eta \sum_{i''\sim i'}(Z_{i''} - \mu_{i''}))} \right\} \right]'.$$

With the alternative ± 1 coding, the variance estimation can be obtained in analogy to (23):

$$\widehat{Var}(\widehat{\tilde{\boldsymbol{\beta}}}_{\mathrm{p}}) \approx \boldsymbol{\mathcal{I}}(\widehat{\tilde{\boldsymbol{\beta}}}_{\mathrm{p}})^{-1}\boldsymbol{\mathcal{J}}(\widehat{\tilde{\boldsymbol{\beta}}}_{\mathrm{p}})\boldsymbol{\mathcal{I}}(\widehat{\tilde{\boldsymbol{\beta}}}_{\mathrm{p}})^{-1}. \tag{24}$$

For the uncentered model, the $\boldsymbol{\mathcal{I}}(\tilde{\boldsymbol{\beta}})$ and $\boldsymbol{\mathcal{J}}(\tilde{\boldsymbol{\beta}})$ are

$$
\begin{aligned}
\boldsymbol{\mathcal{I}}(\tilde{\boldsymbol{\beta}}) &= \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \left[1 - \left\{ \frac{\sinh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i'\sim i}\tilde{Z}_{i'})}{\cosh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i'\sim i}\tilde{Z}_{i'})} \right\}^2 \right] \\
\boldsymbol{\mathcal{J}}(\tilde{\boldsymbol{\beta}}) &= \sum_{i=1}^{n} \sum_{i'\sim i, i'=i} \left[\boldsymbol{x}_i \left\{ \tilde{Z}_i - \frac{\sinh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i'\sim i}\tilde{Z}_{i'})}{\cosh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i'\sim i}\tilde{Z}_{i'})} \right\} \right] \\
&\qquad \left[\boldsymbol{x}_{i'} \left\{ \tilde{Z}_{i'} - \frac{\sinh(\boldsymbol{x}_{i'}'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i''\sim i'}\tilde{Z}_{i''})}{\cosh(\boldsymbol{x}_{i'}'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i''\sim i'}\tilde{Z}_{i''})} \right\} \right]'.
\end{aligned}
$$

For the centered model, the $\boldsymbol{\mathcal{I}}(\tilde{\boldsymbol{\beta}})$ and $\boldsymbol{\mathcal{J}}(\tilde{\boldsymbol{\beta}})$ are

$$
\begin{aligned}
\boldsymbol{\mathcal{I}}(\tilde{\boldsymbol{\beta}}) &= \sum_{i=1}^{n} \left[\boldsymbol{x}_i - \tilde{\eta}\sum_{i'\sim i} \frac{1}{\{\cosh(\boldsymbol{x}_{i'}'\tilde{\boldsymbol{\beta}})\}^2}\boldsymbol{x}_{i'} \right] \left[\boldsymbol{x}_i - \tilde{\eta}\sum_{i'\sim i} \frac{1}{\{\cosh(\boldsymbol{x}_{i'}'\tilde{\boldsymbol{\beta}})\}^2}\boldsymbol{x}_{i'} \right]' \\
&\quad \cosh[\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i'\sim i}(\tilde{Z}_{i'} - \tilde{\mu}_{i'})]^{-2} \\
&\quad - \sum_{i=1}^{n} \sum_{i'\sim i} 2\tilde{\eta} \left\{ \tilde{Z}_i - \frac{\sinh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i''\sim i}(\tilde{Z}_{i''} - \tilde{\mu}_{i''}))}{\cosh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i''\sim i}(\tilde{Z}_{i''} - \tilde{\mu}_{i''}))} \right\} \frac{\sinh(\boldsymbol{x}_{i'}'\tilde{\boldsymbol{\beta}})}{\{\cosh(\boldsymbol{x}_{i'}'\tilde{\boldsymbol{\beta}}\}^3}\boldsymbol{x}_{i'}\boldsymbol{x}_{i'}' \\
\boldsymbol{\mathcal{J}}(\tilde{\boldsymbol{\beta}}) &= \\
&\sum_{i=1}^{n} \sum_{i'\sim i, i'=i} \left[\left\{ \boldsymbol{x}_i - \tilde{\eta}\sum_{i'\sim i}(\cosh(\boldsymbol{x}_{i'}'\tilde{\boldsymbol{\beta}}))^{-2}\boldsymbol{x}_{i'} \right\} \left\{ \tilde{Z}_i - \frac{\sinh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i'\sim i}(\tilde{Z}_{i'} - \tilde{\mu}_{i'}))}{\cosh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i'\sim i}(\tilde{Z}_{i'} - \tilde{\mu}_{i'}))} \right\} \right] \\
&\quad \left[\left\{ \boldsymbol{x}_{i'} - \tilde{\eta}\sum_{i''\sim i'}(\cosh(\boldsymbol{x}_{i''}'\tilde{\boldsymbol{\beta}}))^{-2}\boldsymbol{x}_{i''} \right\} \left\{ \tilde{Z}_{i'} - \frac{\sinh(\boldsymbol{x}_{i'}'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i''\sim i'}(\tilde{Z}_{i''} - \tilde{\mu}_{i''}))}{\cosh(\boldsymbol{x}_{i'}'\tilde{\boldsymbol{\beta}} + \tilde{\eta}\sum_{i''\sim i'}(\tilde{Z}_{i''} - \tilde{\mu}_{i''}))} \right\} \right]'.
\end{aligned}
$$

4. Simulation Study

4.1. Simulation Set-up

We conducted a simulation study to examine the finite-sample properties of the method developed in Sections 2–3. Consider an $m \times m$ square lattice, where $m = 15$ or 30, corresponding to sample sizes $n = 225$ or 900. For spatial dependence, the neighborhood structure

is of the first order and the autoregression coefficient η is either 0.3 or 0.7, corresponding to weaker or stronger spatial dependence.

Let $\boldsymbol{u}_j = (u_{j1}, \ldots, u_{jn})'$ denote the jth covariate vector such that $\{u_{ji} : i = 1, \ldots, n\}$ is a Gaussian random field with mean 0 and an exponential covariance function

$$Cov(u_{ji}, u_{ji'}) = \sigma^2 \exp(-|i - i'|/\tau), \tag{25}$$

where we let the variance parameter be $\sigma^2 = 1$ and the range parameter be $\tau = 0.1$. To obtain cross-covariate correlation, let $\boldsymbol{u}_i = (u_{1i}, \ldots, u_{pi})'$ and $\boldsymbol{x}_i = \boldsymbol{A}\boldsymbol{u}_i$ for site i, where $\boldsymbol{A}\boldsymbol{A}' = [\rho^{|j-j'|}]_{j,j'=1}^p$ and $\rho = 0.4$.

Let $p = 10$ be the number of covariates. The regression coefficients are set to be $\boldsymbol{\beta} = (1, \beta_1, 1, 1, \boldsymbol{0}_7')'$, that is, 3 out of 10 coefficients are non-zero and the remaining 7 coefficients are zero. For the 0-1 coding, we adopt the notion of an average large-scale structure as the average of μ_i over all sites and covariates (Caragea and Kaiser, 2009). Let

$$\bar{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mu_i = \frac{1}{n} \sum_{i=1}^{n} \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta})} \tag{26}$$

The large-scale structure is considered to be weak when $\bar{\mu}$ is around 0.5 and strong otherwise. For the ± 1 coding, we define the average large-scale structure analogously as

$$\bar{\tilde{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\mu}_i = \frac{1}{n} \sum_{i=1}^{n} \frac{\sinh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})}{\cosh(\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}})} \tag{27}$$

In this case, the large-scale structure is considered to be weak if $\bar{\tilde{\mu}}$ is close to 0 but strong otherwise. Here, we let $\beta_1 = 1$ or 5, which corresponds to a stronger or weaker large-scale structure, respectively.

In this simulation study, we restricted our attention to the 0-1 coding of the response variables. For each combination of m, β_1 and η, $S = 100$ samples were generated from the uncentered and centered model. In particular, the data were simulated by a Gibbs sampler using the full conditional probabilities in (10) and (12) (Wasserman, 2003). Since the initial value for the one-step approximation algorithm may not be in a strictly convex area of the likelihood function, the matrix $\boldsymbol{\mathcal{I}}(\cdot)$ is not always positive definite. This occurred only

occasionally to only the centered model, in which case we modified the $\mathcal{I}(\cdot)$ matrix by adding a positive diagonal matrix (Nocedal and Wright, 2000).

4.2. Simulation Results

Tables 1–2 provide the results of variable selection for sample size $n = 225$ and 900 in terms of the average numbers of correctly identified zero-valued and non-zero regression coefficients. The true number of non-zero and zero regression coefficients are 3 and 7, respectively. When the sample size n is larger, the number of correctly identified zero-valued (or the non-zero-valued) coefficients is closer to the truth. The number of correctly identified non-zero coefficients is closer to the truth compared with zero coefficients. When the sample size is smaller ($n = 225$), the average number of correctly identified zero (and non-zero) regression coefficients is largely closer to the truth when η is smaller corresponding to weaker spatial dependence. For the larger sample ($n = 900$), the effect of η is not as obvious. For either sample size, there is no apparent effect of β_1 (i.e., large-scale structure) on the average number of correctly identified zero or non-zero regression coefficients.

Figures 1–2 provide the results of parameter estimation for sample size $n = 225$ and 900 for the regression coefficients and the autoregression coefficient. In general, for both uncentered and centered models, the bias and variance of the MPPLE of $\{\beta_j\}$ and η are smaller as the sample size increases from 225 to 900, but are larger when the autoregression coefficient η is larger. Also, the bias and variance are larger when the large-scale structure is weaker. This is plausible, because when the large-scale structure effect is weaker, the small-scale structure induced by the spatial dependence is relatively stronger.

For each simulated data set, we also computed the standard error of the MPPLE of the nonzero coefficients $\{\beta_j : j = 1, \ldots, 4\}$ by taking the square root of the diagonal elements of the estimated variance matrix (23). Figures 3–4 give the box plots of these standard errors in each of the $S = 100$ simulated datasets, along with the nominal standard error of the MPPLE from this 100 simulations. The bias and variance of standard error of the MPPLE of $\{\beta_j : j = 0, \ldots, 4\}$ are smaller when the sample size is larger, but are larger when η is larger. In particular, the bias and variance of the standard error of the MPPLE of β_1 are

13

larger when β_1 is larger corresponding to a weaker large-scale structure. Furthermore, the bias and variance of standard error of the intercept estimate $\widehat{\beta}_0$ tend to be larger than all other regression coefficients.

5. Data Example

In this section, we illustrate our method by a real data example in landscape ecology and environmental history. The study is aimed at assessing the influence of past land ownership characteristics on land cover structure in northern Wisconsin, USA. Landscape ecologists are interested in the effect of land ownership because it offers a spatially specific, quantifiable approach to assess the effect of important, yet often geographically amorphous social, economic, and historical factors on landscape structure (Turner *et al.*, 1996).

Data were derived from historical plat maps and the Wisconsin Land Economic Inventory, a land survey. The spatial unit of analysis is a quarter section (or, 1/36 township, 160 acres ≈ 65 ha) and there are 1429 units in the study area. We focus on a binary response variable indicating whether the land cover in a quarter section is aspen (*Populus* spp.)-paper birch (*Betula papryfera*) (APB), an early successional forest class, or not (Figure 5).

Forest composition can be associated with several land ownership characteristics, including ownership class, parcel size, and ownership size (Crow *et al.*, 1999; Stanfield *et al.*, 2002). The specific covariates of interest are: reservation, number of parcels, average size of parcels within a quarter section, proportion of the largest parcel, total area, and average size of all parcels in (but not necessarily all contained within) a quarter section. More specifically, reservation (`Reserv`) indicates whether the quarter section is on an Indian reservation or not. The number of parcels (`PolyNm`) in a given quarter section is a measure of parcel density. Average size (`PolyPr`) is the average size of parcel polygon, which is a measure of parcel size. Proportion of the largest parcel (`MxPolyPr`) reports the largest parcel polygon as a proportion of quarter section. Total area (`TotOwn`) shows the total property area (measured in acres) in this study area associated with owner of the largest parcel polygon in a given quarter section. Because the values of this covariate are skewed to the right, we use a log

14

transformation. Average size of all parcels (`AvParcel`) is the average size of all parcels that lay in whole or in part within a given quarter section (measured in acres). All the covariates except reservation are standardized to have mean 0 and variance 1 before being used in the regression procedure.

We considered the four combinations of uncentered or centered models and the 0-1 or ± 1 coding of APB. For autologistic regression, a first-order neighborhood structure is assumed as before. The MPPLE of the regression coefficients are given in Table 3, as well as the standard errors for those MPPLE that are non-zero. For the different models and codings, the variables selected largely agree. `Reserv` and `TotOwn` were selected for all four combinations, whereas the other covariates were not selected except `MxPolyPr` once in the uncentered model with 0-1 coding. The signs of all the nonzero regression coefficients also agree and for this data set, the signs are all positive. That is, there is a greater likelihood of APB on sites that are reservation and that contain a parcel held by a large owner in the study area. The autoregression coefficient η is consistently estimated to be positive, indicating a positive spatial dependence. The magnitude of the estimates is about four times as large for the 0-1 coding than the ± 1 coding. This is perhaps expected, as given in Appendix, the exact relation is $\tilde{\eta} = \eta/4$.

We see that reservation (`Reserv`) is a significant predictor of APB. That is, whether a site lies on the Indian reservation or not influences the likelihood that APB is the dominant land cover. This is plausible because sites on the reservation lay within a single administrative unit (the Indian Agency, later the Bureau of Indian Affairs), which historically has had authority over forest management decisions, regardless of ownership type. Indian Agency authority was especially pervasive in the early 20th century (for historical political reasons), which corresponds with the time period represented by the land cover data. Moreover, the autologistic regression model fitting finds that ownership size (`TotOwn`) is a significant predictor of APB presence vs. absence. In this particular ecoregion, the land cover class of interest, APB, is strongly associated with a specific forest management practice, i.e., short-rotation, even-aged management. Large landowners are most likely to implement this

practice: forest stands on some parcels could be harvested, while those on other parcels could be held for harvest. By distributing harvest rotations across a large land area, forest products could be continuously harvested. This management logic typically assumes large real estate holding, and functions less optimally on smaller ownerships.

Comparing the magnitude of the estimated nonzero regression coefficients in an uncentered model with the centered model, there are differences but no obvious pattern. However, an additional variable `MxPolyPr` was selected for the 0-1 coding in the uncentered model and the magnitudes of the autoregression coefficient estimates $\widehat{\eta}$ are comparable. We further estimated the average large-scale structure by plugging the parameter estimates from the logistic regression into (26) and (27). The estimated large-scale structure is 0.65 for the 0-1 coding and 0.14 for the ± 1 coding, which indicate a somewhat strong large-scale structure when compared with 0.5 and 0, respectively. Based on the simulation results in Section 4, therefore, we may conclude that the results using the centered model are more reliable in this data example.

6. Conclusions and Discussion

In this paper, we have proposed maximum penalized pseudolikelihood estimation for simultaneous selection of covariates and estimation of parameters. Our simulation study has shown desirable large-sample properties in the sense that the performance of variable selection and parameter estimation improves as the sample size increases. In addition, a variance estimation based on the limiting distribution of the parameter estimates appears to work reasonably well. We have also illustrated our method by a data example in the intersection of landscape ecology and environmental history. For future research, it would be interesting to adopt some of the innovations in computation such as Monte Carlo maximum likelihood and reduced dependence approximation.

We may further consider spatial-temporal autologistic models as an extension of the spatial-only model above. For site $i = 1, \ldots, n$ and time $t = 1, \ldots, T$, let $Z_{i,t} = 0$ or 1 denote

the response variable and define the set of indices of spatial-temporal neighbors as

$$\mathcal{N}_{i,t} = \{(i', t-l) : i' \sim i, \ l = 0, \ldots, L\} \cup \{(i, t-l) : l = 1, \ldots, L\} \tag{28}$$

where recall that $i' \sim i$ denotes that $i' \in \mathcal{N}_i$ by a pre-specified spatial neighborhood and L denotes the maximum time lag in the temporal neighborhood. That is, the spatial-temporal neighborhood contains the spatial neighbors of site i at times $t, t-1, \ldots, t-L$ and site i itself in the past $t-1, \ldots, t-L$. For a given time point t, let $\boldsymbol{Z}_t = (Z_{1,t}, \ldots, Z_{n,t})'$ denote the binary response variables at all n sites.

To model the response variables $\{Z_{i,t} : i = 1, \ldots, n, \ t = 1, \ldots, T\}$, we first assume that conditional on $\{\boldsymbol{Z}_{t'} : t' < t\}$, the probability of $\boldsymbol{Z}_t$ depends on only the most recent L time points:

$$p(\boldsymbol{Z}_t | \boldsymbol{Z}_{t'} : t' < t) = p(\boldsymbol{Z}_t | \boldsymbol{Z}_{t'} : t' = t-1, \ldots, t-L). \tag{29}$$

For time t, further assume a Markov random field such that

$$p(Z_{i,t} | Z_{i',t} : i' \neq i; \boldsymbol{Z}_{t'} : t' = t-1, \ldots, t-L) = p(Z_{i,t} | Z_{i',t'} : (i', t') \in \mathcal{N}_{i,t}). \tag{30}$$

The conditional distribution $p(Z_{i,t} | Z_{i',t'} : (i', t') \in \mathcal{N}_{i,t})$ is Bernoulli with success probability $\pi_{i,t} = p(Z_{i,t} = 1 | Z_{i',t'} : (i', t') \in \mathcal{N}_{i,t})$.

The logit function of the success probability can be defined analogously to the spatial-only autologistic regression model. For an uncentered model, let

$$\text{logit}(\pi_{i,t}) = \log \frac{\pi_{i,t}}{1 - \pi_{i,t}} = \boldsymbol{x}'_{i,t}\boldsymbol{\beta} + \sum_{(i',t') \in \mathcal{N}_{i,t}} \eta_{i,i',t,t'} Z_{i',t'} \tag{31}$$

for a centered model, let

$$\text{logit}(\pi_{i,t}) = \log \frac{\pi_{i,t}}{1 - \pi_{i,t}} = \boldsymbol{x}'_{i,t}\boldsymbol{\beta} + \sum_{(i',t') \in \mathcal{N}_{i,t}} \eta_{i,i',t,t'}(Z_{i',t'} - \mu_{i',t'}) \tag{32}$$

where $\boldsymbol{x}_{i,t}$ denotes the covariate vector at site i and time t and $\mu_{i,t} = \exp(\boldsymbol{x}'_{i,t}\boldsymbol{\beta})/\{1 + \exp(\boldsymbol{x}'_{i,t}\boldsymbol{\beta})\}$ is the expectation of $Z_{i,t}$ assuming independence.

17

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}', \eta_{i,i',t,t'})'$ denote the vector of the regression coefficients $\boldsymbol{\beta}$ and autoregression coefficients $\{\eta_{i,i',t,t'}\}$. The log-pseudolikelihood function for the uncentered model is

$$\ell_{\text{p(st)}}^{\text{u}}(\boldsymbol{\theta}|\boldsymbol{Z}) \;\; = \;\; \sum_{t=L+1}^{T} \sum_{i=1}^{n} \log \frac{\exp\{Z_{i,t}(\boldsymbol{x}_{i,t}'\boldsymbol{\beta} + \sum_{(i',t')\sim(i,t)} \eta_{i,i',t,t'} Z_{i',t'})\}}{1 + \exp(\boldsymbol{x}_{i,t}'\boldsymbol{\beta} + \sum_{(i',t')\sim(i,t)} \eta_{i,i',t,t'})} \tag{33}$$

and for the centered model is

$$\ell_{\text{p(st)}}^{\text{c}}(\boldsymbol{\theta}|\boldsymbol{Z}) \;\; = \;\; \sum_{t=L+1}^{T} \sum_{i=1}^{n} \log \frac{\exp[Z_{i,t}\{\boldsymbol{x}_{i,t}'\boldsymbol{\beta} + \sum_{(i',t')\sim(i,t)} \eta_{i,i',t,t'}(Z_{i',t'} - \mu_{i',t'})\}]}{1 + \exp\{\boldsymbol{x}_{i,t}'\boldsymbol{\beta} + \sum_{(i',t')\sim(i,t)} \eta_{i,i',t,t'}(Z_{i',t'} - \mu_{i',t'})\}} \tag{34}$$

For simultaneous variable selection and parameter estimation, a penalized log-pseudolikelihood function similar to (18) can be defined:

$$\ell_{\text{pp(st)}}(\boldsymbol{\theta}) = \ell_{\text{p(st)}}(\boldsymbol{\theta}) - N \sum_{j=1}^{p} \lambda_j |\beta_j| \tag{35}$$

where $N = n(T - L)$. Maximization of (35) can be attained following the computation algorithm in Section 3 and similarly, the variance estimation of the MPPLE can be derived. In both cases, fast computation is attainable given the computational efficiency of adaptive Lasso and relatively low dimension of the variance matrix.

Appendix: The relationship between 0-1 coding with ± 1 coding

For an uncentered model, we have

$$
\begin{aligned}
p(Z_i|Z_{i'} : i' \sim i) \;\; &= \;\; \frac{\exp\{Z_i(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i} Z_{i'})\}}{1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i} Z_{i'})} \\
&= \;\; \frac{\exp[\{(\tilde{Z}_i + 1)/2\}\{\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i}(\tilde{Z}_{i'} + 1)/2\}]}{1 + \exp\{\boldsymbol{x}_i'\boldsymbol{\beta} + \eta \sum_{i'\sim i}(\tilde{Z}_{i'} + 1)/2\}} \\[2mm]
&= \;\; \frac{\exp[\tilde{Z}_i\{(\boldsymbol{x}_i'\boldsymbol{\beta}/2 + |\mathcal{N}_i|\eta/4) + (\eta/4)\sum_{i'\sim i}\tilde{Z}_{i'}\}]}{2\cosh\{(\boldsymbol{x}_i'\boldsymbol{\beta}/2 + |\mathcal{N}_i|\eta/4) + (\eta/4)\sum_{i'\sim i}\tilde{Z}_{i'}\}}
\end{aligned}
$$

Thus, $\tilde{\eta} = \eta/4$ and $\boldsymbol{x}_i'\tilde{\boldsymbol{\beta}} = \boldsymbol{x}_i'\boldsymbol{\beta}/2 + |\mathcal{N}_i|\eta/4$.

Now, for a centered model, we have

$$
\begin{aligned}
p(Z_i | Z_{i'} : i' \sim i) &= \frac{\exp[Z_i \{ \boldsymbol{x}_i' \boldsymbol{\beta} + \eta \sum_{i' \sim i}(Z_{i'} - \mu_{i'})\}]}{1 + \exp\{\boldsymbol{x}_i' \boldsymbol{\beta} + \eta \sum_{i' \sim i}(Z_{i'} - \mu_{i'})\}} \\
&= \frac{\exp[\{(\tilde{Z}_i + 1)/2\}\{\boldsymbol{x}_i' \boldsymbol{\beta} + \eta \sum_{i' \sim i}(\tilde{Z}_{i'}/2 + 1/2 - \mu_{i'})\}]}{1 + \exp\{\boldsymbol{x}_i' \boldsymbol{\beta} + \eta \sum_{i' \sim i}(\tilde{Z}_{i'}/2 + 1/2 - \mu_{i'})\}} \\
&= \frac{\exp[\tilde{Z}_i \{\boldsymbol{x}_i' \boldsymbol{\beta}/2 + |\mathcal{N}_i|\eta/4 - \eta \sum_{i' \sim i}\mu_{i'}/2 + \eta \sum_{i' \sim i}\tilde{\mu}_{i'}/4 + (\eta/4)\sum_{i' \sim i}(\tilde{Z}_{i'} - \tilde{\mu}_{i'})\}]}{2\cosh\{\boldsymbol{x}_i' \boldsymbol{\beta}/2 + |\mathcal{N}_i|\eta/4 - \eta \sum_{i' \sim i}\mu_{i'}/2 + \eta \sum_{i' \sim i}\tilde{\mu}_{i'}/4 + (\eta/4)\sum_{i' \sim i}(\tilde{Z}_{i'} - \tilde{\mu}_{i'})\}}
\end{aligned}
$$

Thus, $\tilde{\eta} = \eta/4$ and $\boldsymbol{x}_i' \tilde{\boldsymbol{\beta}} = \boldsymbol{x}_i' \boldsymbol{\beta}/2 + |\mathcal{N}_i|\eta/4 - \eta \sum_{i' \sim i}\mu_{i'}/2 + \eta \sum_{i' \sim i}\tilde{\mu}_{i'}/4$

Acknowledgments

References

Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2003) *Hierarchical Modeling and Analysis for Spatial Data.* Chapman and Hall: Boca Raton.

Besag, J. (1972) Nearest-neighbour systems and the auto-logistic model for binary data. *Journal of the Royal Statistical Society Series B*, **34**, 75–83.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B*, **36**, 192–236.

Caragea, P.C. and Kaiser, M.S. (2009) Autologistic models with interpretable parameters. *Journal of Agricultural, Biological, and Environmental Statistics*, **14**, 281–300.

Cressie, N. (1993) *Statistics for Spatial Data, Rev Ed.* Wiley: New York.

Crow, T.R., Host, G.E., and Mladenoff, D.J. (1999) Ownership and ecosystem as sources of spatial heterogeneity in a forested landscape, Wisconsin USA. *Landscape Ecology*, **14**, 449–463.

Diggle, P.J. and Ribeiro, P.J. (2007) *Model-based Geostatistics*. New York: Springer.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004) Least angle regression (with discussion). *Annals of Statistics*, **32**, 407–499.

Friel, N., Pettitt, A.N., Reeves, R. and Wit, E. (2009) Bayesian inference in hidden Markov random fields for binary data defined on large lattices. *Journal of Computational and Graphical Statistics*, **18**, 243–261.

Gaetan, C. and Guyon, X. (2010) *Spatial Statistics and Modeling*. New York: Springer.

Geyer, C. J. (1994) On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Society of Statistics Series B*, **56**, 261–274.

Gumpertz, M.L., Graham, J.M., and Ristaino, J.B. (1997) Autologistic model of spatial pattern of *Phytophthora* epidemic in bell pepper: Effects of soil variables on disease presence. *Journal of Agricultural, Biological, and Environmental Statistics*, **2**, 131–156.

He, F. , Zhou, J., and Zhu, H. (2003) Autologistic regression model for the distribution of vegetation. *Journal of Agricultural, Biological and Environmental Statistics*, **8**, 205–222.

Huang, H.-C., Hsu, N.-J., Theobald, D.M., and Breidt, F.J. (2010) Spatial LASSO with applications to GIS model selection. *Journal of Computational and Graphical Statistics*, **19**, 963–983.

Huffer, F.W. and Wu, H. (1998) Markov chain Monte Carlo for autologistic regression models with application to the distribution of plant species. *Biometrics*, **54**, 509–524.

Hughes, J., Haran, M., and Caragea, P.C. (2011) Autologistic models for binary data on a lattice. *Environmetrics*, **22**, 857–871.

Moeller, J., Pettitt, A.N., Reeves, R., and Berthelsen, K.K. (2006) An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, **93**, 451–458.

Nocedal, J. and Wright, S.J. (2000) *Numerical Optimization, 2nd Ed.* New York: Springer.

Stanfield, B.J., Bliss, J.C. and Spies, T.A. (2002) Land ownership and landscape structure: A spatial analysis of sixty-six Oregon (USA) Coast Range watersheds. *Landscape Ecology*, 17, 685–697.

Sun, L. and Clayton, M.K. (2008) Bayesian analysis of cross-classified spatial data with autocorrelation. *Biometrics*, **64**, 74–84.

Tibshirani, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, **58**, 267–288.

Turner, M.G., Wear, D.N., and Flamm, R.O. (1996) Land ownership and land-over change in the Southern Appalachian Highlands and the Olympic Peninsula. *Ecological Applications*, **6**, 1150–1172.

Wasserman, L. (2003) *All of Statistics: A Concise Course in Statistical Inference.* New York: Springer.

Wang, Z. and Zheng, Y. (2012) Analysis of binary data via a centered spatial-temporal autologistic regression model. *Environmental and Ecological Statistics* (In press).

Zheng Y. and Zhu J. (2008) Markov chain Monte Carlo for spatial-temporal autologistic regression model. *Journal of Computational and Graphical Statistics*, **17**, 123–127.

Zhu, J., Huang, H.-C., and Reyes, P.E. (2010) On selection of spatial linear models for lattice data. *Journal of Royal Statistics Society Series B*, **72**, 389–402.

Zhu, J., Huang, H.-C., and Wu, J.-P. (2005) Modeling spatial-temporal binary data using Markov random fields. *Journal of Agricultural, Biological, and Environmental Statistics*, **10**, 212–225.

Zhu, J., Zheng, Y., Carroll, A.L., and Aukema, B.H. (2008) Autologistic regression analysis of spatial-temporal binary data via Monte Carlo maximum likelihood. *Journal of Agricultural, Biological, and Environmental Statistics*, **13**, 84–98.

Zhu, Z. and Liu, Y. (2009) Estimating spatial covariance using penalized likelihood with weighted L_1 penalty. *Journal of Nonparametric Statistics*, **21**, 925–942.

Zou, H. (2006) The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, **101**, 1418–1429.

Table 1: Average number of correctly identified non-zero and zero regression coefficients when $\beta_1 = 5$ (weak large-scale structure) for uncentered and centered model, sample size $n = 225$ or 900, and antoregression coefficient $\eta = 0.3$ or 0.7.

Model	$\{\beta_j\}$ n	Number of non-zero estimates $\eta = 0.3$	$\eta = 0.7$	Number of zero estimates $\eta = 0.3$	$\eta = 0.7$
Uncentered	225	2.77	2.69	6.12	6.14
	900	3.00	3.00	6.81	6.88
Centered	225	2.75	2.67	6.21	6.17
	900	3.00	3.00	6.87	6.82

Table 2: Average number of correctly identified non-zero and zero regression coefficients when $\beta_1 = 1$ (strong large-scale structure) for uncentered and centered model, sample size $n = 225$ or 900, and antoregression coefficient $\eta = 0.3$ or 0.7.

Model	$\{\beta_j\}$ n	Number of non-zero estimates $\eta = 0.3$	$\eta = 0.7$	Number of zero estimates $\eta = 0.3$	$\eta = 0.7$
Uncentered	225	2.94	2.52	6.09	6.12
	900	3.00	3.00	6.88	6.95
Centered	225	2.92	2.90	6.31	6.16
	900	3.00	3.00	6.87	6.92

Table 3: Maximum penalized pseudolikelihood estimates (MPPLE) of the regression coefficients and the autoregression coefficient η for uncentered or centered models and 0-1 or ±1 coding of the response variable. The standard errors of MPPLE of the regression coefficients are given in the parentheses.

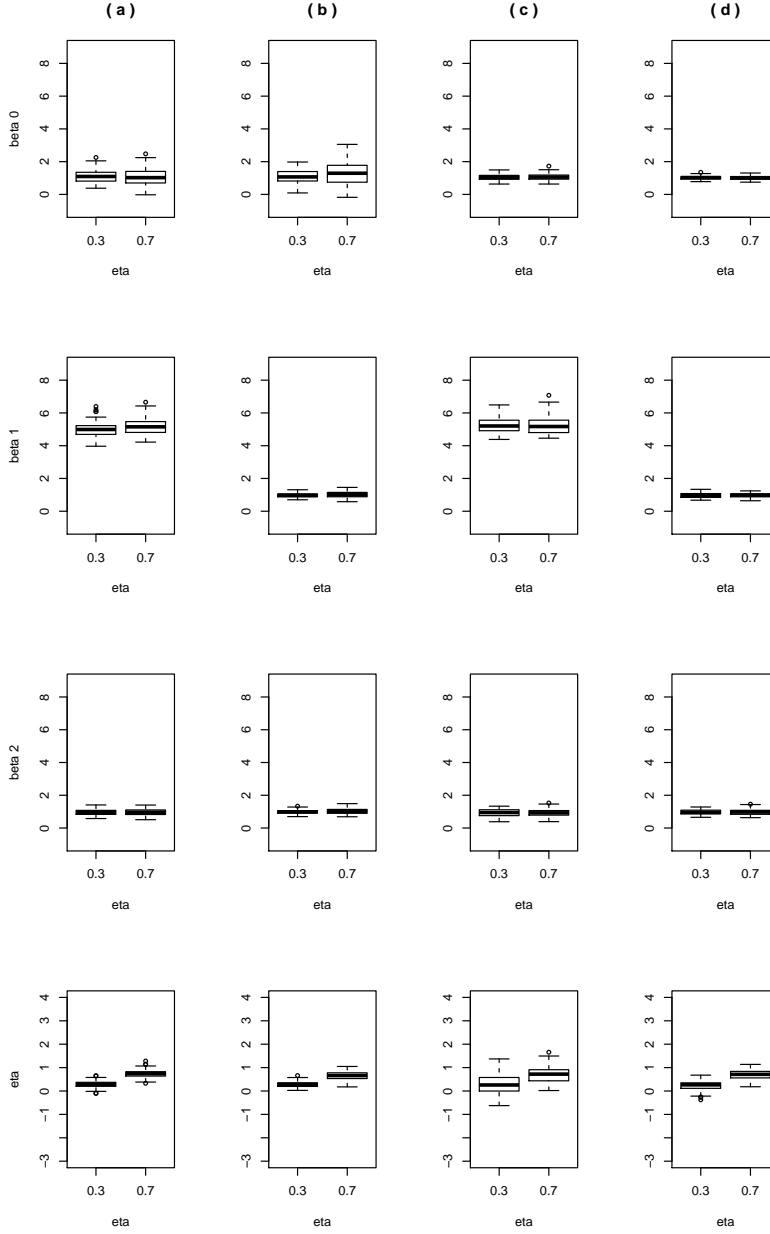| | 0-1 coding | | ±1 coding | |
Covariates	Uncentered	Centered	Uncentered	Centered
Reserv	0.728 (0.132)	1.889 (0.302)	0.230 (0.042)	0.944 (0.072)
PolyNm	–	–	–	–
PolyPr	–	–	–	–
MxPolyPr	0.114 (0.088)	–	–	–
log(TotOwn)	0.420 (0.095)	0.178 (0.068)	0.174 (0.023)	0.090 (0.025)
AvParcl	–	–	–	–
η	1.217	1.302	0.331	0.325
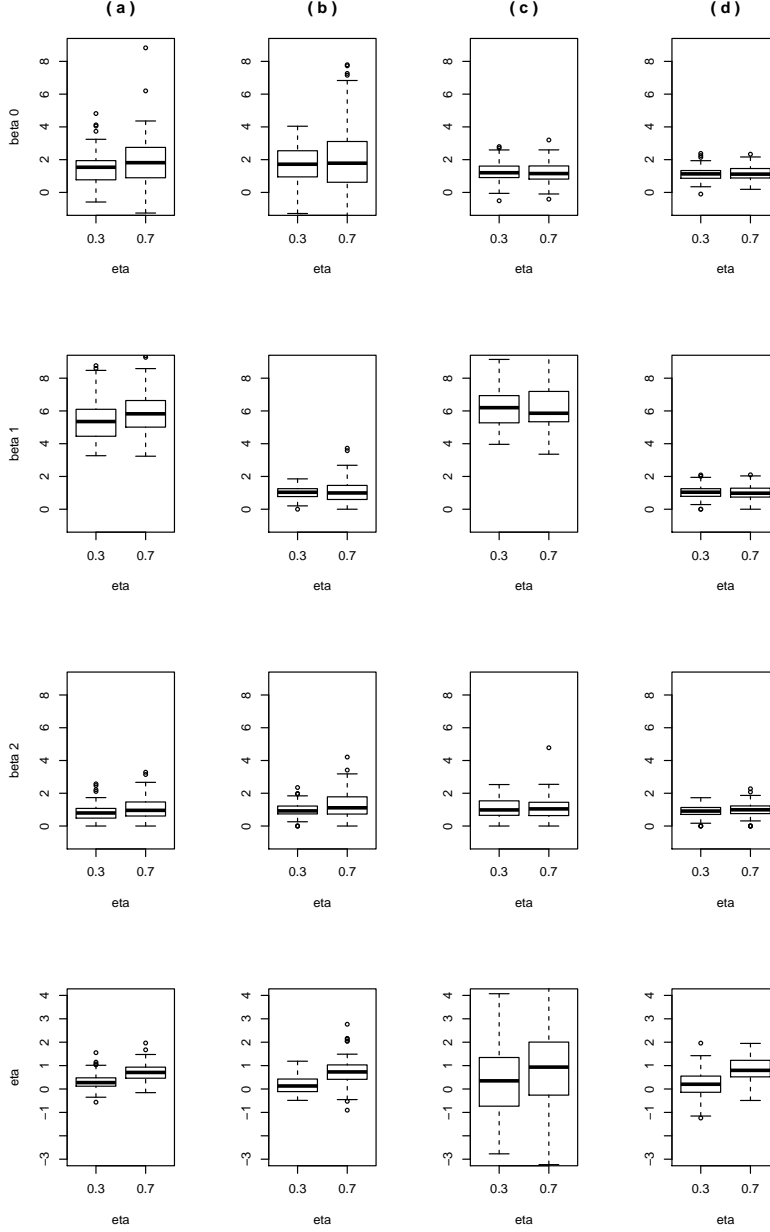
Figure 1: Box plot of the MPPLE $\widehat{\beta}_0$ (row 1), $\widehat{\beta}_1$ (row 2), $\widehat{\beta}_2$ (row 3) and $\widehat{\eta}$ (row 4) from the 100 simulations with sample size $n = 225$. Column (a): uncentered model with $\beta_1 = 5$ (weak large-scale structure); (b): uncentered model with $\beta_1 = 1$ (strong large-scale structure); (c): centered model with $\beta_1 = 5$ (weak large-scale structure); (d): centered model with $\beta_1 = 1$ (strong large-scale structure). The true values are $\beta_0 = 1$, $\beta_1 = 1$ or 5, $\beta_2 = 1$, $\beta_3 = 1$, and $\eta = 0.3$ or 0.7. The box plot of $\widehat{\beta}_3$ is similar to $\widehat{\beta}_2$ but omitted to save space.
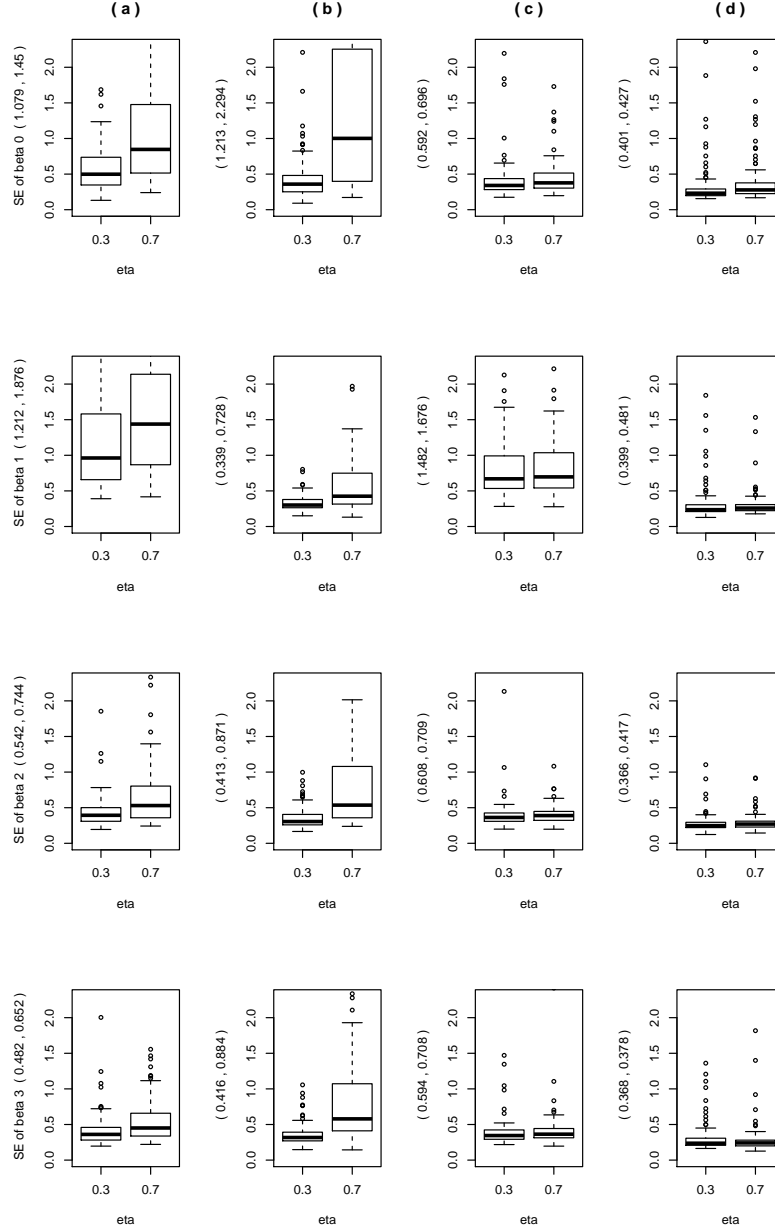
Figure 2: Box plot of the MPPLE $\widehat{\beta}_0$ (row 1), $\widehat{\beta}_1$ (row 2), $\widehat{\beta}_2$ (row 3) and $\widehat{\eta}$ (row 4) from the 100 simulations with sample size $n = 900$. Column (a): uncentered model with $\beta_1 = 5$ (weak large-scale structure); (b): uncentered model with $\beta_1 = 1$ (strong large-scale structure); (c): centered model with $\beta_1 = 5$ (weak large-scale structure); (d): centered model with $\beta_1 = 1$ (strong large-scale structure). The true values are $\beta_0 = 1$, $\beta_1 = 1$ or 5, $\beta_2 = 1$, $\beta_3 = 1$, and $\eta = 0.3$ or 0.7. The box plot of $\widehat{\beta}_3$ is similar to $\widehat{\beta}_2$ but omitted to save space.

Figure 3: Box plot of the standard error of $\widehat{\beta}_0$ (row 1), $\widehat{\beta}_1$ (row 2), $\widehat{\beta}_2$ (row 3) and $\widehat{\beta}_3$ (row 4) from the 100 simulations with sample size $n = 225$. Column (a): uncentered model with $\beta_1 = 5$ (weak large-scale structure); (b): uncentered model with $\beta_1 = 1$ (strong large-scale structure); (c): centered model with $\beta_1 = 5$ (weak large-scale structure); (d): centered model with $\beta_1 = 1$ (strong large-scale structure). The nominal standard error of the MPPLE from the 100 simulations are given along the y-axis on the left.
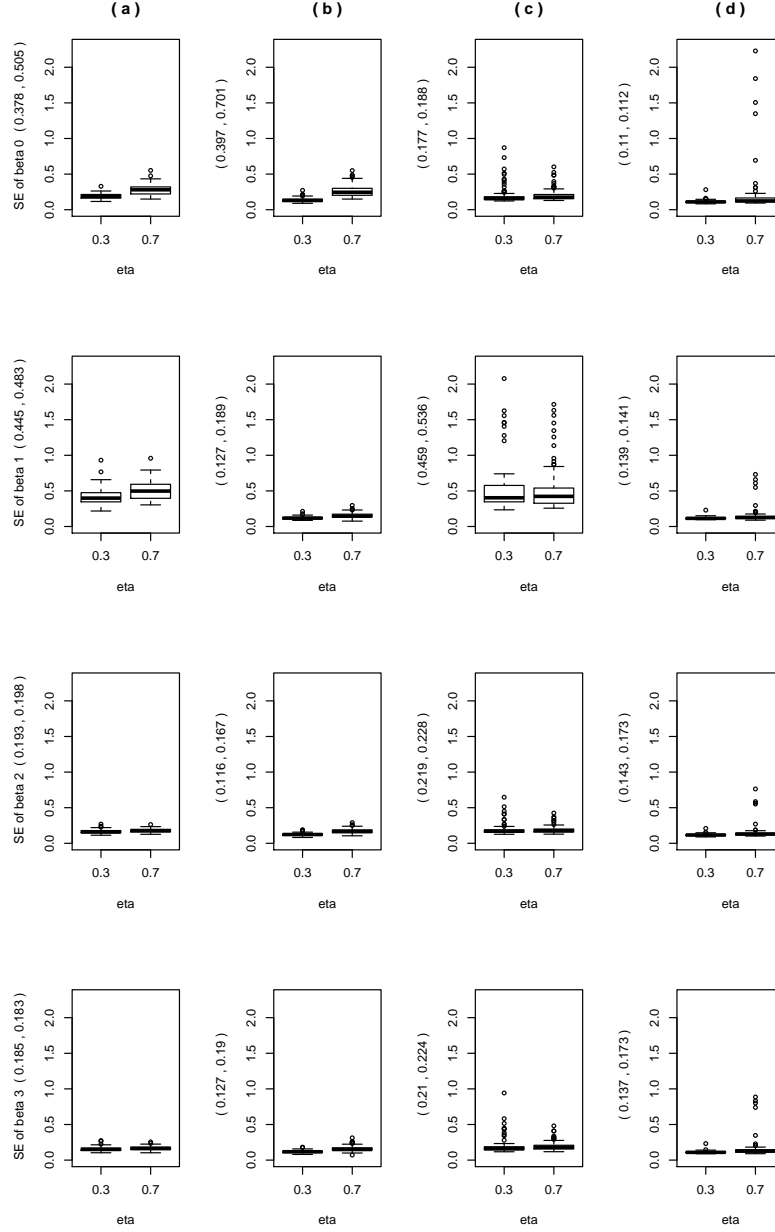
Figure 4: Box plot of the standard error of $\widehat{\beta}_0$ (row 1), $\widehat{\beta}_1$ (row 2), $\widehat{\beta}_2$ (row 3) and $\widehat{\beta}_3$ (row 4) from the 100 simulations with sample size $n = 900$. Column (a): uncentered model with $\beta_1 = 5$ (weak large-scale structure); (b): uncentered model with $\beta_1 = 1$ (strong large-scale structure); (c): centered model with $\beta_1 = 5$ (weak large-scale structure); (d): centered model with $\beta_1 = 1$ (strong large-scale structure). The nominal standard error of the MPPLE from the 100 simulations are given along the y-axis on the left.
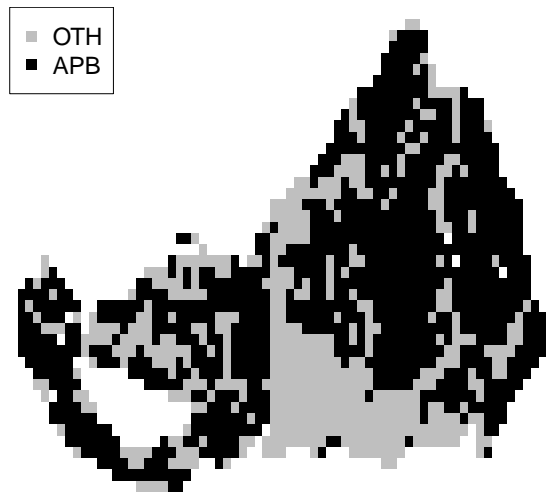
Figure 5: Map of the response variable in a study area of northern Wisconsin: Each of the 1,429 quarter sections is classified to be aspen-paper birch (APB) or other (OTH).