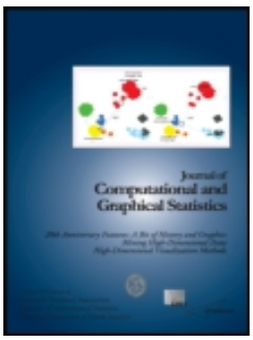


Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://www.tandfonline.com/loi/ucgs20

Spatial Lasso With Applications to GIS Model Selection

Hsin-Cheng Huang, Nan-Jung Hsu, David M. Theobald & F. Jay Breidt

To cite this article: Hsin-Cheng Huang, Nan-Jung Hsu, David M. Theobald & F. Jay Breidt (2010) Spatial Lasso With Applications to GIS Model Selection, Journal of Computational and Graphical Statistics, 19:4, 963-983, DOI: 10.1198/jcgs.2010.07102

To link to this article: https://doi.org/10.1198/jcgs.2010.07102

Supplementary materials for this article are available online.
Please click the JCGS link at http://pubs.amstat.org.

Spatial Lasso With Applications to GIS Model Selection

Hsin-Cheng HUANG, Nan-Jung HSU, David M. THEOBALD, and F. Jay BREIDT

Geographic information systems (GIS) organize spatial data in multiple two-dimensional arrays called layers. In many applications, a response of interest is observed on a set of sites in the landscape, and it is of interest to build a regression model from the GIS layers to predict the response at unsampled sites. Model selection in this context then consists not only of selecting appropriate layers, but also of choosing appropriate neighborhoods within those layers. We formalize this problem as a linear model and propose the use of Lasso to simultaneously select variables, choose neighborhoods, and estimate parameters. Spatially dependent errors are accounted for using generalized least squares and spatial smoothness in selected coefficients is incorporated through use of a priori spatial covariance structure. This leads to a modification of the Lasso procedure, called spatial Lasso. The spatial Lasso can be implemented by a fast algorithm and it performs well in numerical examples, including an application to prediction of soil moisture. The methodology is also extended to generalized linear models. Supplemental materials including R computer code and data analyzed in this article are available online.

Key Words: Cross validation; Generalized least squares; Least angle regression; Spatial regression; Variable selection.

1. INTRODUCTION

A raster-based geographic information system (GIS) organizes spatial data in multiple two-dimensional arrays called layers. A vector-based GIS describes spatial data as features: points, lines, and polygons. We focus on raster-based GIS, noting that it is possible to convert vector to raster and back. Layers in the raster-based GIS may include infor-

Hsin-Cheng Huang is Research Fellow, Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan (E-mail: *hchuang@stat.sinica.edu.tw*). Nan-Jung Hsu is Professor, Institute of Statistics, National Tsing-Hua University, Hsin-Chu 30043, Taiwan (E-mail: *njhsu@stat.nthu.edu.tw*). David M. Theobald is Associate Professor, Department of Human Dimensions of Natural Resources and Natural Resource Ecology Lab, Warner College of Natural Resources, Colorado State University, Fort Collins, CO 80523-1480 (E-mail: *davet@warnercnr.colostate.edu*). F. Jay Breidt is Professor and Chair of Statistics, Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877 (E-mail: *jbreidt@stat.colostate.edu*).

Journal of Computational and Graphical Statistics, Volume 19, Number 4, Pages 963–983
DOI: 10.1198/jcgs.2010.07102

mation on geology, topography, weather and climate, ownership, political administration, land cover, land use, and so on. In many applications, a response of interest is observed on a set of sites in the landscape, and it is of interest to build a regression model from the GIS layers to predict the response at unsampled sites. Some layers may have no explanatory power for the response of interest. A layer that is a useful predictor may influence the response at a particular site through its values at the same site or a set of neighboring sites. This neighborhood of influence may vary from layer to layer. Model selection in this context then consists not only of selecting appropriate layers, but also of choosing appropriate neighborhoods within those layers. Allowance for choice of neighborhoods leads to an enormous expansion in the number of models considered, and hence layers and neighborhoods are usually selected in a subjective way. To account for the combinatorially large set of possible covariates available in a GIS, we formulate this spatial regression as a linear model with a large number of unknown regression coefficients. We assume that regression coefficients are spatially homogeneous from neighborhood to neighborhood across the landscape, after correcting for known sources of spatial inhomogeneity. Even with this spatial homogeneity assumption, there is still a very large number of regression coefficients, for which direct least squares estimation would be inefficient. We propose the use of Lasso (Tibshirani 1996) to simultaneously select variables, choose neighborhoods, and estimate parameters. Lasso can be applied "off the shelf" to this problem, but this standard implementation does not exploit the possibility that in a given layer, regression coefficients in the same neighborhood may be similar. We further allow for spatially dependent errors and the possibility of spatial smoothness in selected coefficients through use of a priori spatial covariance structure. This leads to a modification of the Lasso procedure that we call *spatial Lasso*. The model and the implementation of standard Lasso and spatial Lasso are described in Sections 2.1–2.3. The least angle regression (LARS) algorithm of Efron et al. (2004) can be used in a fast implementation of Lasso, but does not apply directly to a fast implementation of spatial Lasso. Like the standard Lasso, the spatial Lasso also has a piecewise linear solution path. We briefly describe a closely related algorithm for computing spatial Lasso in Section 2.4. In Sections 3.1 and 3.2, we illustrate the Lasso methods in two numerical experiments. In Section 3.3 we consider an application to prediction of soil moisture, and in Section 3.4 we consider an application to an ecological dataset based on a Poisson regression model. Spatial Lasso performs well in all of these examples. Section 4 provides a summary and directions for future research.

2. MODEL AND METHODS

2.1 Spatial Regression With GIS Layers

Suppose that we have available p known GIS layers $\{\mathbf{x}_k(\mathbf{s}) : \mathbf{s} \in D\}$; $k = 1, \ldots, p$, mapped to a common spatial resolution described by a regular grid $D \subset \mathbb{Z}^2$ of m_1 by m_2 points, where we assume without loss of generality that $\mathbf{0} \in D$. Responses $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))'$ are observed at spatial locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subset D$. Our goal is to build up a regression model for the response $\mathbf{Y}$ from the p GIS layers by selecting not only layers, but also neighborhoods within those layers.

For each $k = 1, \ldots, p$, define the neighborhood set $\mathcal{N}_k \subset \mathbb{Z}^2$, where $\mathbf{0} \in \mathcal{N}_k$ if $\mathcal{N}_k$ is nonempty. The proposed spatial regression model is

$$Y(\mathbf{s}) = \sum_{j=1}^{J} a_j \phi_j(\mathbf{s}) + \sum_{k=1}^{p} \sum_{\mathbf{u} \in \mathcal{N}_k} b_k(\mathbf{s}, \mathbf{u}) x_k(\mathbf{s} + \mathbf{u}) + \varepsilon(\mathbf{s}); \qquad \mathbf{s} \in D, \tag{2.1}$$

where $\phi_1(\cdot), \ldots, \phi_J(\cdot)$ are known functions of spatial location only, $b_k(\mathbf{s}, \mathbf{u})$ is the coefficient corresponding to the neighbor $\mathbf{u}$ of $\mathbf{s}$ at the kth layer, and $\varepsilon(\cdot)$ is a (spatial dependence) error process with $\operatorname{var}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{\Sigma}$, where $\boldsymbol{\epsilon} \equiv (\varepsilon(\mathbf{s}_1), \ldots, \varepsilon(\mathbf{s}_n))'$. Note that $Y(\mathbf{s})$ does not depend on layer k if $\mathcal{N}_k$ is empty; depends on layer k only through $x_k(\mathbf{s})$ if $\mathcal{N}_k = \{\mathbf{0}\}$; and in general depends on layer k through the set of neighboring values $\{x_k(\mathbf{s} + \mathbf{u}) : \mathbf{u} \in \mathcal{N}_k\}$.

The coefficients $\{b_k(\mathbf{s}, \mathbf{u})\}$ depend on layer, neighborhood within layer, and site within neighborhood. To reduce the number of coefficients that we need to estimate, we make the assumption that neighborhoods within layers are spatially homogeneous across the landscape, after accounting for known sources of spatial heterogeneity. Specifically, we assume that

$$b_k(\cdot, \mathbf{u}) = \sum_{l=1}^{L_k} c_{k,l}(\mathbf{u}) \psi_{k,l}(\cdot); \qquad \mathbf{u} \in \mathcal{N}_k, \tag{2.2}$$

where $c_{k,l}(\mathbf{u})$ are unknown, spatially homogeneous regression coefficients (i.e., no dependence on $\mathbf{s}$), and $\psi_{k,l}(\cdot)$ are known basis functions (possibly depending on additional GIS layers) that adjust for known sources of spatial heterogeneity. The special case of spatial homogeneity is obtained with $L_k = 1$ and $\psi_{k,1}(\cdot) \equiv 1$, since $b_k(\cdot, \mathbf{u})$ then reduces to a constant function, for each $\mathbf{u}$. An interaction effect between $x_k(\mathbf{s} + \mathbf{u})$ and $x_{k'}(\mathbf{s})$ can also be introduced in (2.1) by taking $L_k = 1$ and $\psi_{k,1}(\cdot) = x_{k'}(\cdot)$. A common type of spatial heterogeneity arises when the domain D of interest is composed of M different subregions $D_1, \ldots, D_M$ (e.g., watersheds), in which case $\psi_{k,l}(\cdot)$ can be specified to account for different spatial features in different subregions as

$$\psi_{k,l}(\mathbf{s}) = \sum_{r=1}^{M} I(\mathbf{s} \in D_r) f_{k,l,r}(\mathbf{s}); \qquad l = 1, \ldots, L_k, k = 1, \ldots, p,$$

where $\{f_{k,l,1}(\cdot), \ldots, f_{k,l,M}(\cdot)\}$ are known functions. For example, different flow directions in watersheds can be reflected by choosing $f_{k,l,r}(\mathbf{s}) = f_{k,l}(\mathbf{B}_r \mathbf{s})$ $(\mathbf{s} \in \mathbb{R}^2, r = 1, \ldots, M)$ for some known function $f_{k,l}(\cdot)$, where $\mathbf{B}_r$ is a 2×2 rotation matrix corresponding to the rth watershed.

The model given by (2.1) and (2.2) can be written as a spatial regression model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = (\mathbf{X}_1'\boldsymbol{\beta} + \varepsilon(\mathbf{s}_1), \ldots, \mathbf{X}_n'\boldsymbol{\beta} + \varepsilon(\mathbf{s}_n))', \tag{2.3}$$

where $\mathbf{X}_i$ is a vector of GIS variables corresponding to $Y(\mathbf{s}_i)$, $\mathbf{X} \equiv (\mathbf{X}_1, \ldots, \mathbf{X}_n)'$ and $\boldsymbol{\beta}$ consists of all of the unknown coefficients $\{a_j\} \cup \{c_{k,l}(\mathbf{u}) : \mathbf{u} \in \mathcal{N}_k\}$. Because the total number of regression parameters $m = J + \sum_{k=1}^{p} L_k |\mathcal{N}_k|$ is typically very large, ordinary least squares (OLS) estimates or generalized least squares (GLS) estimates would be inefficient due to large estimation errors. On the other hand, traditional variable selection methods,

such as stepwise search, sequential testing, or AIC (Akaike 1973), may be either inefficient or computationally too intensive. Instead of using OLS or GLS with variable selection, we propose to select variables and estimate the parameters simultaneously for this problem by adapting the least absolute shrinkage and selection operator (Lasso) technique introduced by Tibshirani (1996).

2.2 Standard Lasso for Spatial Regression

In this section, we consider the usual linear regression with independent errors (i.e., $\boldsymbol{\Sigma} = \mathbf{I}$). For simplicity, the intercept effect is assumed to be subtracted out from the response variable and the covariates throughout the article. The Lasso is a constrained version of OLS that trades off bias for lower variance. Specifically, let $\mathbf{X}^* \equiv \mathbf{XM}$ be the column-wise standardized version of $\mathbf{X}$ and let $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_m^*)' \equiv \mathbf{M}^{-1}\boldsymbol{\beta}$. Then (2.3) can be rewritten as

$$\mathbf{Y} = \mathbf{X}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}. \tag{2.4}$$

The Lasso estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is obtained by minimizing the residual sum of squares,

$$(\mathbf{Y} - \mathbf{X}^*\boldsymbol{\beta}^*)'(\mathbf{Y} - \mathbf{X}^*\boldsymbol{\beta}^*),$$

with respect to $\boldsymbol{\beta}^*$ subject to $\sum |\beta_j^*| \le t$, where t is a tuning parameter. Equivalently, $\hat{\boldsymbol{\beta}}$ can be obtained by minimizing

$$(\mathbf{Y} - \mathbf{X}^*\boldsymbol{\beta}^*)'(\mathbf{Y} - \mathbf{X}^*\boldsymbol{\beta}^*) + \lambda \sum_{j=1}^{m} |\beta_j^*|, \tag{2.5}$$

where λ is a tuning parameter depending on t.

As explained by Tibshirani (1996), Lasso is readily understood graphically in two dimensions. Picture the (β_1^*, β_2^*) plane with elliptical contours of the residual sum of squares $\mathrm{RSS}(\beta_1^*, \beta_2^*)$ superimposed. The constraint $\sum |\beta_j^*| \le t$ is a diamond-shaped region centered on $(0, 0)$, with corners on the β_1^* and β_2^* axes. Lasso finds the point within the diamond that is as close as possible to the minimum residual sum of squares. If t is large enough, then the diamond includes the OLS estimate and Lasso yields OLS. If t is small enough, then the diamond excludes OLS, and the Lasso estimate is shrunken toward $(0, 0)$. In particular, the Lasso point may fall on a corner of the diamond, corresponding to $\hat{\beta}_1^* = 0$ or $\hat{\beta}_2^* = 0$, so that a submodel may be selected. Lasso thus computes shrinkage estimators and performs model selection simultaneously.

Note that $\hat{\boldsymbol{\beta}}^*$ is also the Bayes posterior mode if the prior distributions of $\{\beta_1^*, \ldots, \beta_m^*\}$ are conditionally independent given σ^2, and have a common double-exponential density:

$$p(\beta_j^*) = \frac{\lambda}{4\sigma^2} \exp\left(-\frac{\lambda}{2\sigma^2}|\beta_j^*|\right), \tag{2.6}$$

where $\lambda > 0$ is a scaling parameter, and σ^2 is the variance of $\varepsilon(\mathbf{s})$ in (2.1).

Because the spatial regression model in (2.3) is a linear model, standard Lasso can be applied directly to simultaneously select variables, choose neighborhoods, and estimate parameters. The tuning parameter λ in (2.5) can be estimated using cross-validation (CV)

or generalized cross-validation as in the article by Tibshirani (1996), or using Mallows' C_p as in the article by Efron et al. (2004), where the corresponding degrees of freedom can be estimated using Stein unbiased risk estimation (Zou, Hastie, and Tibshirani 2007) or a data perturbation technique (Ye 1998). For implementation, the Lasso estimate in (2.5) can be solved efficiently using the homotopy algorithm (Osborne, Presnell, and Turlach 2000a, 2000b) or the least angle regression (LARS) algorithm (Efron et al. 2004).

Note that standard Lasso assumes no structure among regression coefficients in the same neighborhood for a given layer. In practice, it may be reasonable to suppose that these neighboring regression coefficients are similar, so that the response reacts in a similar way to neighboring covariate values. One method for incorporating such spatial smoothness is to suppose that neighboring coefficients are spatially correlated, as we do in the next subsection.

2.3 SPATIAL LASSO

To account for both spatially dependent errors and spatial smoothness among coefficients within the same neighborhood, we consider a generalized residual sum of squares and specify a spatial dependence prior for $\boldsymbol{\beta}$. For simplicity, the error correlation structure $\boldsymbol{\Sigma}$ is assumed known; see Section 3.3 for an application with unknown $\boldsymbol{\Sigma}$. We modify (2.5) by incorporating a prior for $\boldsymbol{\beta}$ in which the components of $\boldsymbol{\beta}^{**} = (\beta_1^{**}, \ldots, \beta_m^{**})' \equiv \boldsymbol{\Gamma}^{-1/2}\boldsymbol{\beta}^*$ are independent and follow a common double-exponential density (2.6), where $\boldsymbol{\Gamma}$ is the prior correlation matrix of $\boldsymbol{\beta}^*$ accounting for the spatial smoothness in each layer. Note that $\boldsymbol{\Gamma}^{1/2}$ is the square root of $\boldsymbol{\Gamma}$, which is symmetric and nonnegative definite. Then the posterior mode of $\boldsymbol{\beta}$ can be obtained by minimizing

$$\begin{aligned}&\left(\mathbf{Y} - \mathbf{X}^*\boldsymbol{\Gamma}^{1/2}\boldsymbol{\beta}^{**}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{Y} - \mathbf{X}^*\boldsymbol{\Gamma}^{1/2}\boldsymbol{\beta}^{**}\right) + \lambda\sum_{j=1}^{m}|\beta_j^{**}| \\ &\quad = (\mathbf{Y}^* - \mathbf{X}^{**}\boldsymbol{\beta}^{**})'(\mathbf{Y}^* - \mathbf{X}^{**}\boldsymbol{\beta}^{**}) + \lambda\sum_{j=1}^{m}|\beta_j^{**}| \qquad (2.7)\end{aligned}$$

over $\boldsymbol{\beta}^{**} \in \mathbb{R}^m$, where $\mathbf{X}^{**} \equiv \boldsymbol{\Sigma}^{-1/2}\mathbf{X}^*\boldsymbol{\Gamma}^{1/2}$ and $\mathbf{Y}^* \equiv \boldsymbol{\Sigma}^{-1/2}\mathbf{Y}$. We call the resulting estimate $\hat{\boldsymbol{\beta}}_{\mathrm{SL}}$ of $\boldsymbol{\beta}$ the *spatial Lasso* estimate. An example of $\boldsymbol{\Gamma}$ is

$$\operatorname{corr}(c_{k,l}(\mathbf{s}), c_{k',l'}(\mathbf{s}')) = \delta_{k,k'}\delta_{l,l'}\exp(-\|\mathbf{s} - \mathbf{s}'\|/\gamma), \qquad (2.8)$$

$$\operatorname{corr}(a_j, c_{k,l}(\mathbf{u})) = 0, \qquad (2.9)$$

$$\operatorname{corr}(a_j, a_{j'}) = \delta_{j,j'} \qquad (2.10)$$

for $j, j' = 1, \ldots, J$; $k, k' = 1, \ldots, p$; $l, l' = 1, \ldots, L_k$; and $\mathbf{s}, \mathbf{s}' \in \mathcal{N}_k$, where $\|\cdot\|$ is the L_2 norm, δ is the Kronecker delta function, and $\gamma > 0$ is a spatial dependence parameter. The resulting estimator of $\mu(\mathbf{s})$ is $\hat{\mu}_{\mathrm{SL}}(\mathbf{s}) \equiv (\mathbf{x}(\mathbf{s}))'\hat{\boldsymbol{\beta}}_{\mathrm{SL}}$, where $\mathbf{x}(\mathbf{s})$ is the corresponding covariate vector at $\mathbf{s} \in D$.

The spatial Lasso predictor of $Y(\mathbf{s})$ is given by

$$\hat{Y}(\mathbf{s}) = \hat{\mu}_{\mathrm{SL}}(\mathbf{s}) + \operatorname{corr}(\varepsilon(\mathbf{s}), \boldsymbol{\epsilon})\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{SL}}).$$

Note that $\hat{Y}(\mathbf{s}) = Y(\mathbf{s})$ if $\mathbf{s} \in \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$.

Both the spatial dependence parameter γ and the tuning parameter λ in (2.7) can be estimated using cross-validation (CV) or generalized cross-validation as in the article by Tibshirani (1996), or by other methods as discussed in Section 2.2.

2.4 COMPUTATIONAL ALGORITHM

For implementation, the Lasso estimate in (2.5) can be solved very efficiently using LARS (Efron et al. 2004) because the column vectors of $\mathbf{X}^*$ are standardized. However, the spatial Lasso estimate in (2.7) cannot be obtained directly using LARS, because the column vectors of $\mathbf{X}^{**}$ in (2.7) are not standardized and cannot be standardized without distorting either the error correlation structure or the a priori correlation structure for the regression coefficients. Nevertheless, the LARS algorithm can be modified for nonstandardized covariates by processing along an equal-length projection path rather than an equiangular path, resulting in what we will refer to as an *equal-length projection* (ELP) algorithm. Just as the standard Lasso of (2.5) can be solved by the LARS with a slight modification (Efron et al. 2004), the spatial Lasso of (2.7) can be solved by an ELP algorithm with a similar modification. As pointed out by a referee, this modified ELP algorithm is equivalent to the algorithm proposed by Turlach (2005). Although Turlach's article focuses mainly on Lasso with standardized covariates, the algorithm developed there applies to nonstandardized covariates, and thus can also be used to compute the spatial Lasso.

3. EXAMPLES

We examine the performance of the proposed method based on two numerical experiments and one application. The first experiment is designed to mimic variable-selection problems for GIS data, in which covariates are available at multiple spatial layers. The second experiment is designed to further examine the efficiency gain of using spatial Lasso when the regression coefficients in each layer are smoothly varied. Finally, we consider an application of real GIS data to predict an index of soil moisture.

3.1 NUMERICAL EXPERIMENT I: LAYER AND NEIGHBORHOOD SELECTION

In our first numerical experiment, we considered $p = 6$ spatial layers, $\{x_k(\mathbf{s}) : \mathbf{s} \in D\}$; $k = 1, \ldots, 6$, defined on a regular lattice $D \equiv \{(i_1, i_2) : i_1, i_2 = 1, \ldots, 110\}$ of size 110×110. For each k, $\{x_k(\mathbf{s}) : \mathbf{s} \in D\}$ were independently generated from a zero-mean stationary Gaussian spatial process with an exponential covariance function,

$$C(\mathbf{h}) \equiv \operatorname{cov}(x_k(\mathbf{s}), x_k(\mathbf{s}+\mathbf{h})) = \exp(-\|\mathbf{h}\|/\theta); \qquad k = 1, \ldots, 6. \tag{3.1}$$

We used the "GaussRF" function in the "RandomFields" package of R, available through the Comprehensive R Archive Network (CRAN) website (*http://cran.r-project.org*), and chose a turning-band method, called TBM3, to generate these processes. We considered two values of θ (2 and 0.5), corresponding to strong and weak spatial dependence.

The underlying process was generated from

$$\mu(\mathbf{s}) = 1 + \frac{1}{3}\left\{x_1(\mathbf{s}) + \sum_{j=-1}^{1} x_1(\mathbf{s}+(j,1)) + \sum_{j=-2}^{2} x_1(\mathbf{s}+(j,2))\right\} + \frac{1}{3}\sum_{j=-1}^{1}\sum_{k=-1}^{1} x_2(\mathbf{s}+(j,k)),$$

for $\mathbf{s} \in D$. The neighborhood for layer 1, $\mathcal{N}_1$, is an inverted 5-3-1 pyramid consisting of one site, the three sites above it, and the five sites above the three. The neighborhood for layer 2, $\mathcal{N}_2$, is a centered 3×3 block of sites. Neighborhoods $\mathcal{N}_3$, $\mathcal{N}_4$, $\mathcal{N}_5$, $\mathcal{N}_6$ are all empty.

The response variables $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))'$ were generated by adding noise to the above mean function:

$$Y(\mathbf{s}_i) = \mu(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i); \qquad i = 1, \ldots, n, \tag{3.2}$$

where $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ were sampled from $D^* \equiv \{(i_1, i_2) : i_1, i_2 = 6, \ldots, 105\}$ using simple random sampling with sample size $n = 100$, and $\varepsilon(\mathbf{s}_1), \ldots, \varepsilon(\mathbf{s}_n)$ are independent standard normal random variables. Note that indices $6, \ldots, 105$ were picked to guarantee that all the neighborhoods we consider in (3.6) will lie in D. We could also define D^* to be D by applying some boundary treatment (see an example in Section 3.4). Figure 1(a) shows a realization of $\{\mu(\mathbf{s}) : \mathbf{s} \in D^*\}$, where the underlying spatial layers $x_1(\cdot)$ and $x_2(\cdot)$ were generated from (3.1) with $\theta = 2$.

We applied OLS, standard Lasso, and spatial Lasso to the generated data. Specifically, we considered three OLS estimators ($\hat{\mu}_{\mathrm{OLS},1}(\mathbf{s})$, $\hat{\mu}_{\mathrm{OLS},3}(\mathbf{s})$, and $\hat{\mu}_{\mathrm{OLS},5}(\mathbf{s})$) of the mean

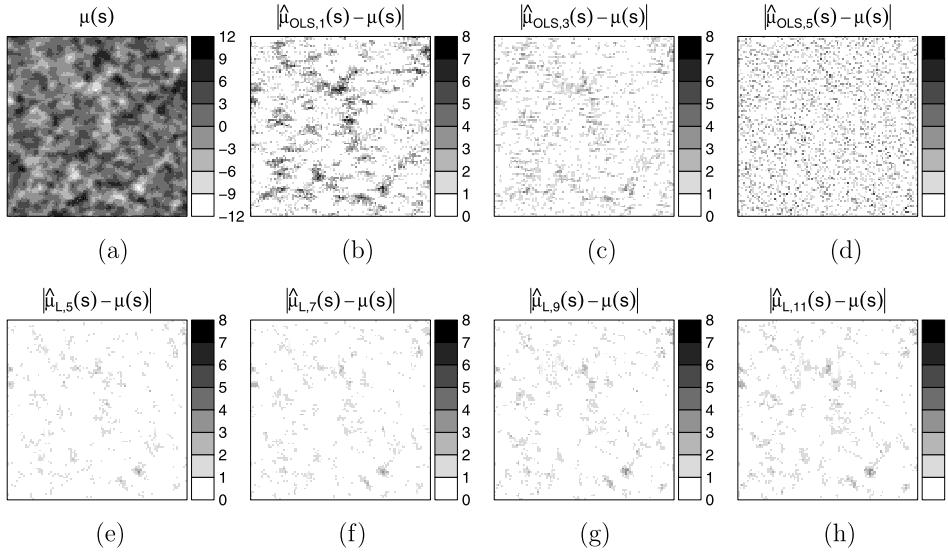


Figure 1. The underlying true mean function and the estimation errors based on various methods.

$\mu(\mathbf{s})$ based on the following three regression models:

$$Y(\mathbf{s}_i) = \beta_0 + \sum_{l=1}^{2} \beta_l x_l(\mathbf{s}_i) + \varepsilon(\mathbf{s}_i), \tag{3.3}$$

$$Y(\mathbf{s}_i) = \beta_0 + \sum_{l=1}^{2} \sum_{j=-1}^{1} \sum_{k=-1}^{1} \beta_{l,j,k} x_l(\mathbf{s}_i + (j,k)) + \varepsilon(\mathbf{s}_i), \tag{3.4}$$

$$Y(\mathbf{s}_i) = \beta_0 + \sum_{l=1}^{2} \sum_{j=-2}^{2} \sum_{k=-2}^{2} \beta_{l,j,k} x_l(\mathbf{s}_i + (j,k)) + \varepsilon(\mathbf{s}_i). \tag{3.5}$$

Note that in each case, the OLS estimator is based on the correct layers, but only the third OLS estimator has a neighborhood structure large enough to contain the true neighborhoods.

To apply standard Lasso, we set $J = L_1 = \cdots = L_6 \equiv 1$ with $\phi_1(\cdot) = \psi_{1,1}(\cdot) = \cdots = \psi_{6,1}(\cdot) \equiv 1$, corresponding to a model with an intercept and spatial homogeneity across neighborhoods. Unlike the OLS models, standard Lasso is not based on the correct layers. In every layer, we considered five different neighborhood sets,

$$\mathcal{N}^{(2q+1)} = \{0, \pm 1, \pm 2, \ldots, \pm q\}^2, \tag{3.6}$$

for $q = 1, \ldots, 5$, resulting in five regression models with $m = J + p(2q+1)^2 = 1 + 6(2q+1)^2 = 55, 151, 295, 487$, or 727 parameters to be selected and estimated. We used multiple neighborhood sets, rather than choosing only one large set and allowing Lasso to do all of the selection, for two reasons. First, since it is not clear how large a set is large enough, choice of a single set can lead to either bias or excessive estimation error in standard Lasso. Second, Lasso for different neighborhood sizes corresponds to different, nonnested models. In either case, selecting among different neighborhood sets using cross-validation leads to better performance empirically.

Let $\hat{\mu}_{L,2q+1}(\mathbf{s})$ be the Lasso estimate of $\mu(\mathbf{s})$ associated with (2.5) based on the neighborhood set $\mathcal{N}^{(2q+1)}$ for $q = 1, \ldots, 5$, where the tuning parameter λ for each case was selected using tenfold CV. The final Lasso estimate, denoted as $\hat{\mu}_L^*$, is selected among $\{\hat{\mu}_{L,3}, \hat{\mu}_{L,5}, \ldots, \hat{\mu}_{L,11}\}$ having the smallest (tenfold) CV value.

For one particular realization of the response variables, Figures 1(b)–(d) show the absolute prediction errors of the three OLS estimates. Figures 1(e)–(h) show the absolute prediction errors of four standard Lasso estimates, $\{\hat{\mu}_{L,5}, \hat{\mu}_{L,7}, \hat{\mu}_{L,9}, \hat{\mu}_{L,11}\}$. For this particular realization, the first OLS model has large systematic errors that match features in the true mean function, due to incorrect specification of the neighborhoods. The second and third OLS estimates have smaller and less systematic errors than the first OLS, but larger absolute errors overall than any of the four Lasso estimates.

For each neighborhood set $\mathcal{N}^{(2q+1)}$, we used the proposed ELP algorithm with modifications to compute six spatial Lasso estimates (2.7) corresponding to $\mathbf{\Gamma}$ with $\gamma = 0, 1, \ldots, 5$ in (2.8), under $\mathbf{\Sigma} = \mathbf{I}$. Note that the spatial Lasso estimate with a larger γ corresponds to stronger spatial dependence of regression coefficients in a layer, and it reduces to the standard (nonspatial) Lasso estimate if $\gamma = 0$. For each $q \in \{1, \ldots, 5\}$, let $\hat{\mu}_{\mathrm{SL},2q+1}$ be

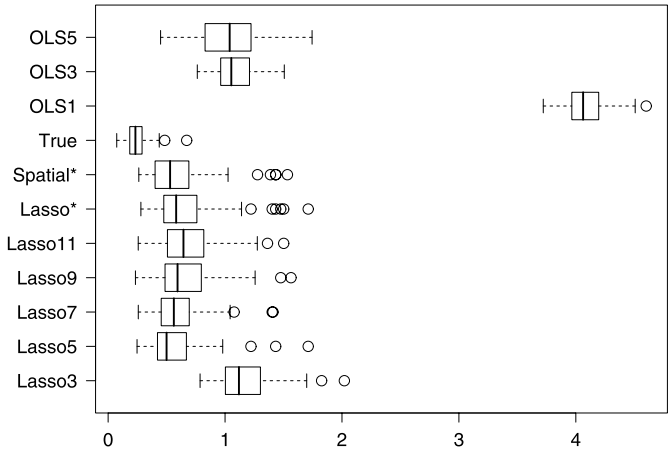


Figure 2. Boxplots of ASE performance for various estimation methods based on 100 simulation replicates for the first simulation experiment, where the underlying GIS layers are generated from (3.1) with $\theta = 2$.

the estimate that has the smallest (tenfold) CV value among the six spatial Lasso estimates corresponding to $\gamma = 0, 1, \ldots, 5$. Then the final spatial Lasso estimate $\hat{\mu}^*_{\mathrm{SL}}$ is defined to be the one among $\{\hat{\mu}_{\mathrm{SL},3}, \ldots, \hat{\mu}_{\mathrm{SL},11}\}$ having the smallest CV value.

To assess the performance of the various estimators across multiple realizations, we used the average squared error (ASE) of estimation across the entire spatial domain,

$$\mathrm{ASE} = \frac{1}{10{,}000} \sum_{\mathbf{s} \in D^*} (\hat{\mu}(\mathbf{s}) - \mu(\mathbf{s}))^2,$$

where $\hat{\mu}(\mathbf{s})$ is a generic estimate of $\mu(\mathbf{s})$. The ASE values for various methods based on 100 simulation replicates are summarized in Figure 2. The boxplot labeled "True" uses OLS based on exactly the correct neighborhood structure (usually infeasible in practice), so that model selection is unnecessary. As in the single simulated realization of Figure 1, the worst estimator corresponds to the first OLS model. The best estimators are Lasso and spatial Lasso with large enough neighborhood sizes to encompass the true neighborhoods. These estimators dominate the corresponding OLS estimators with sufficiently large neighborhoods, even though the OLS estimators have the added advantage of correct layer specification. Comparing between the best spatial Lasso and the best standard Lasso, the best spatial Lasso appears to perform better by having a slightly smaller median ASE value.

It is of interest to see how well the Lasso procedures pick up the correct neighborhood structures within the candidate neighborhood sets $\mathcal{N}^{(2q+1)}$; $q = 1, \ldots, 5$. Figure 3 shows the average proportions of $\{x_k(\mathbf{s}) : \mathbf{s} \in \mathcal{N}^{(2q+1)}\}$ being selected for each k ($k = 1, 2, 3$) and q ($q = 1, \ldots, 5$) using the standard Lasso under the spatial process with $\theta = 2$. (That is, the figure records the frequency out of 100 simulated realizations that the regression coefficient was estimated as nonzero.) Results for layers $k = 4, 5, 6$ are similar to those for $k = 3$ and are not shown. Clearly, the Lasso procedures are picking up the correct neighborhood structures: inverted pyramid for layer 1, centered block for layer 2, and empty set for layer 3. The results for $\theta = 0.5$ are similar and are not reported here.

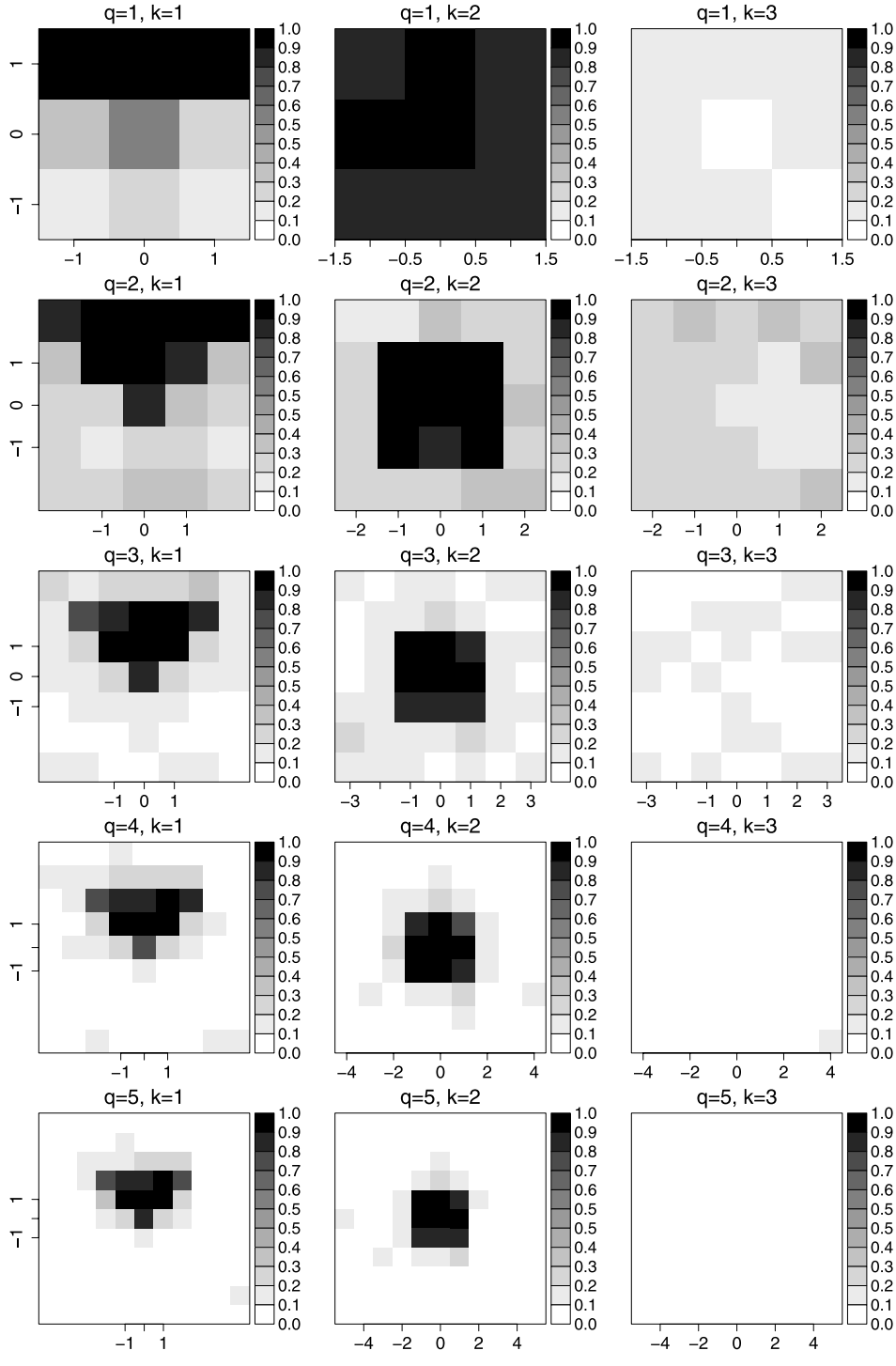


Figure 3. Average proportions of $\{x_k(\mathbf{s}) : \mathbf{s} \in \mathcal{N}^{(2q+1)}\}$ being selected for $\theta = 2$ under different neighborhoods and layers.

3.2 NUMERICAL EXPERIMENT II: SPATIAL SMOOTHNESS

The setting of the second numerical experiment is designed to demonstrate the advantages of spatial Lasso when the regression coefficients vary smoothly within neighborhoods. The domain D is the same as that for the first experiment, but the underlying mean surface was generated from a single covariate layer,

$$\mu(\mathbf{s}) = 1 + 5 \sum_{i_1=-2}^{2} \sum_{i_2=-2}^{2} w_{i_1,i_2} x(\mathbf{s} + (i_1, i_2)); \qquad \mathbf{s} \in D,$$

where $\{x(\mathbf{s}) : \mathbf{s} \in D\}$ consists of independent standard normal random variables and the weights are given by a truncated Gaussian kernel,

$$w_{i_1,i_2} \equiv \frac{\exp(-(i_1^2 + i_2^2)/4)}{\sum_{i_1=-2}^{2} \sum_{i_2=-2}^{2} \exp(-(i_1^2 + i_2^2)/4)}; \qquad i_1, i_2 = 0, \pm 1, \pm 2$$

(see Figure 5(a) below). As in the first experiment, the response variables $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))'$ were generated according to (3.2), where $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ were sampled from D^* using simple random sampling of size $n = 100$. We considered $p = 1$, $J = L_1 = 1$, $\phi_1(\cdot) = \psi_{1,1}(\cdot) \equiv 1$, five different neighborhood sets $\{\mathcal{N}^{(2q+1)} : q = 1, \ldots, 5\}$ of (3.6), and $\mathbf{\Sigma} = \mathbf{I}$. For each neighborhood set $\mathcal{N}^{(2q+1)}$, six spatial Lasso estimates defined in (2.7) corresponding to $\mathbf{\Gamma}$ with $\gamma = 0, 1, \ldots, 5$ in (2.8) were computed using the proposed ELP algorithm. As in the first simulation experiment, for each $q \in \{1, \ldots, 5\}$, let $\hat{\mu}_{\mathrm{SL},2q+1}$ be the estimate that has the smallest (tenfold) CV value among the six spatial Lasso estimates corresponding to $\gamma = 0, 1, \ldots, 5$. Then the final spatial Lasso estimate $\hat{\mu}^*_{\mathrm{SL}}$ is defined to be the one among $\{\hat{\mu}_{\mathrm{SL},3}, \ldots, \hat{\mu}_{\mathrm{SL},11}\}$ having the smallest CV value.

The spatial Lasso estimate $\hat{\mu}^*_{\mathrm{SL}}$ is compared with the standard Lasso estimate $\hat{\mu}^*_L$, as well as with the three OLS estimates corresponding to models (3.3)–(3.5) with $\beta_2 = \beta_{2,j,k} = 0$ for all j and k. Note that $\hat{\mu}_{\mathrm{OLS},5}$ is the OLS estimate based on a true model. The ASE results for various estimation procedures based on 100 simulation replicates are summarized in Figure 4.

As in the first numerical experiment, $\hat{\mu}_{\mathrm{OLS},1}$ performs badly due to the incorrect neighborhood specification. The estimator $\hat{\mu}_{\mathrm{OLS},3}$ performs quite well relative to the estimator based on the true model, $\hat{\mu}_{\mathrm{OLS},5}$, because it imposes a hard shrinkage to zero on all of the small coefficients in the outer band of the neighborhood. Standard Lasso has no particular advantages in this numerical experiment. The OLS estimators based on larger neighborhoods have both the correct layer specification and correct or nearly correct neighborhood specification, so the model selection capability of standard Lasso is not useful in this example. Consequently, standard Lasso performs similarly to the OLS estimators. Further, standard Lasso does not take advantage of spatial smoothness of the regression coefficients. Spatial Lasso does use this information to advantage, and so $\hat{\mu}^*_{\mathrm{SL}}$ clearly outperforms all the other estimates by a large margin.

Finally, Figure 5 shows the images of the true regression coefficients and the average estimated coefficients from spatial Lasso under various neighborhood sets. Spatial Lasso is able to pick up the correct neighborhood structure and to capture the smooth structure of the regression coefficients.

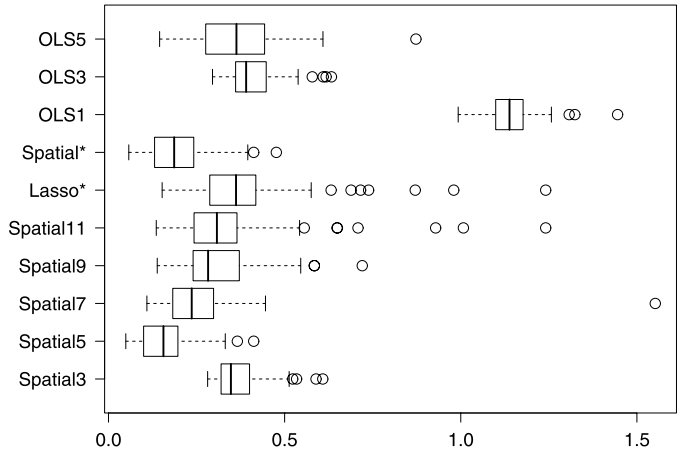


Figure 4. Boxplots of ASE performance for various estimation methods based on 100 simulation replicates for the second experiment.

3.3 APPLICATION TO SOIL MOISTURE PREDICTION

In this example, the response of interest $Y(\cdot)$ is a soil moisture index, known for all sites on a spatial lattice of 100×100 regular grid points $D \equiv \{(i_1, i_2) : i_1, i_2 = 1, \ldots, 100\}$. (The soil moisture index is a product derived from other GIS layers, not a field-measured value at every site.) To assess the prediction ability of the proposed spatial Lasso method, the response variable on a sample of sites was used for model fitting, and the responses

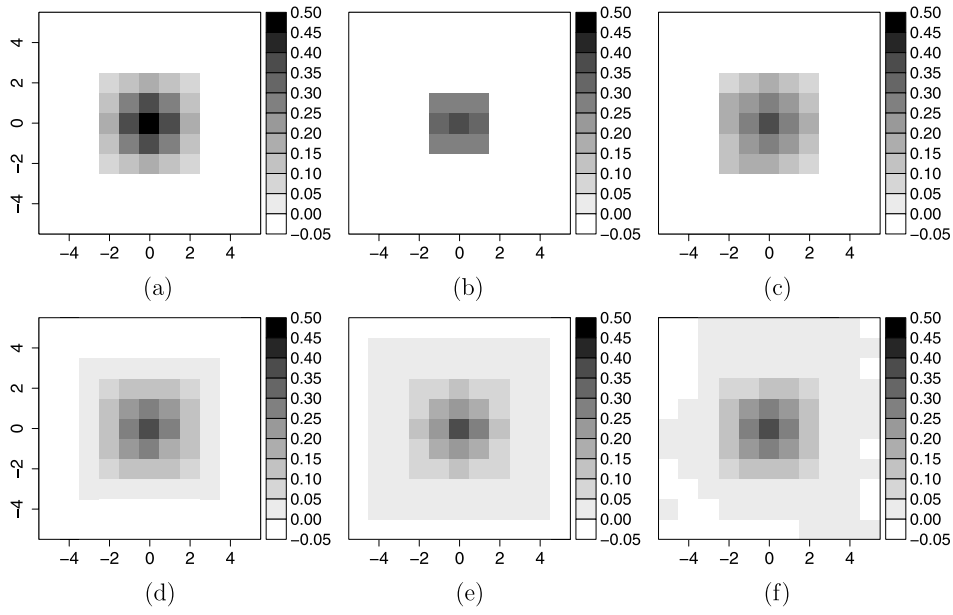


Figure 5. (a) Image of true regression coefficients. (b)–(f) Images of average estimated coefficients from spatial Lasso under $\mathcal{N}^{(3)}$, $\mathcal{N}^{(5)}$, $\mathcal{N}^{(7)}$, $\mathcal{N}^{(9)}$, $\mathcal{N}^{(11)}$, respectively.

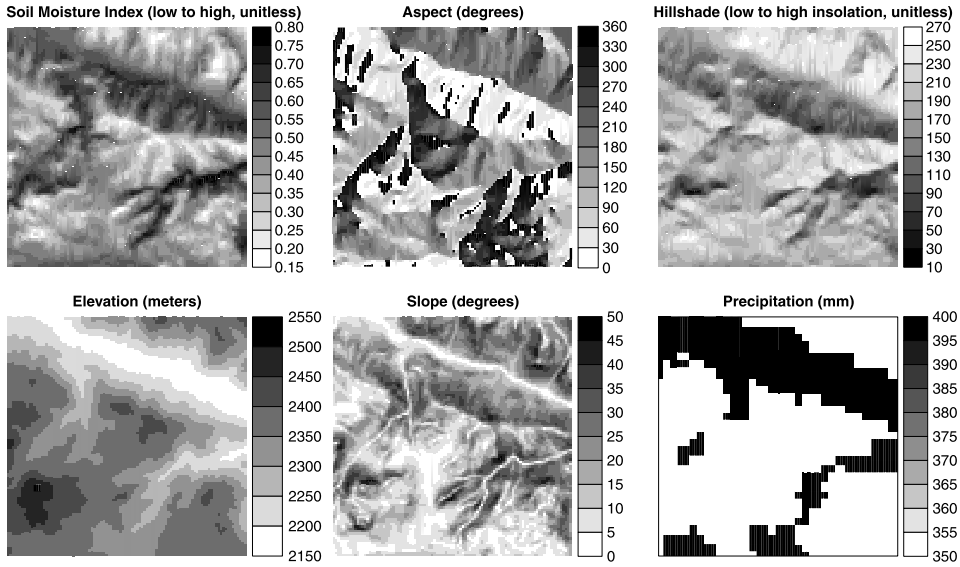


Figure 6. Six GIS layers.

on the remaining sites were put aside for testing. Specifically, we used $n = 100$ responses $\{Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)\}$ for model fitting, where $\{\mathbf{s}_1, \ldots, \mathbf{s}_{100}\}$ were sampled from $D^* \equiv \{(i_1, i_2) : i_1, i_2 = 6, \ldots, 95\}$ using simple random sampling.

From the GIS, we also have five explanatory variables: `aspect`, `hillshade`, `elevation`, `slope`, and `precipitation`, denoted by $z_1(\mathbf{s}), \ldots, z_5(\mathbf{s})$. Figure 6 shows the images for the response and the five explanatory variables. We are interested in building a regression model to predict the soil moisture index that was produced using the method outlined by Iverson et al. (1996) that used layers constructed from the five explanatory variables. This method is based on a hydrologic process model, which incorporates the amount of upstream contributing area (a surrogate for precipitation input), soil permeability (the amount of water that permeates into the groundwater—the remainder is runoff), and solar insolation (a prime driver of evapotranspiration rate). The variable aspect is first transformed by taking the cosine of the angle from the north, so that 1 represents the north and -1 represents the south. Then the five explanatory variables were standardized to have zero mean and unit variance,

$$x_k = \frac{z_k - \texttt{mean}(z_k)}{\sqrt{\texttt{var}(z_k)}}; \qquad k = 1, \ldots, 5.$$

The model given by (2.1) and (2.2) was applied with $p = 10$ layers and $J = 1 = L_1 = \cdots = L_{10} = 1$, where $\phi_1(\cdot) = \psi_{1,1}(\cdot) = \cdots = \psi_{5,1}(\cdot) \equiv 1$ and the first five layers are $x_1, \ldots, x_5$. The sixth to eighth layers, $x_6 = x_3$, $x_7 = x_4$, and $x_8 = x_1$, were used to form interaction effects with respect to `precipitation` (x_5) through $\psi_{6,1} = \psi_{7,1} = \psi_{8,1} = x_5$. In addition, $x_9 = x_3$ and $x_{10} = x_4$ were used to form interaction effects with respect to `aspect` (x_1) through $\psi_{9,1} = \psi_{10,1} = x_1$.

We set $\mathcal{N}_k = \mathcal{N}^{(1)} = \{\mathbf{0}\}$ for layers $k = 5, 6, 7, 8$ involving `precipitation`, because `precipitation` takes on only two possible values, and allowing larger neighborhoods would lead to considerable redundancy. We chose $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3, \mathcal{N}_4$ separately among four different neighborhood sets $\{\mathcal{N}^{(1)}, \mathcal{N}^{(3)}, \mathcal{N}^{(7)}, \mathcal{N}^{(11)}\}$. In addition, $\mathcal{N}_9$ corresponding to the interaction effect between x_1 and x_3 was chosen to be the larger set between $\mathcal{N}_1$ and $\mathcal{N}_3$. Similarly, $\mathcal{N}_{10}$ corresponding to the interaction effect between x_1 and x_4 was chosen to be the larger set between $\mathcal{N}_1$ and $\mathcal{N}_4$. Note that the number of choices for layer selection in this example is 2^{10}, but the number of model choices for layer and neighborhood selection is $2^4(2^{121})^6$ when $\mathcal{N}_1 = \mathcal{N}_2 = \mathcal{N}_3 = \mathcal{N}_4 = \mathcal{N}^{(11)}$. Clearly, allowance for choice of neighborhoods leads to an enormous expansion in the number of models considered.

The spatial Lasso of (2.7) was applied with the prior covariance structure for the regression coefficients given by (2.8)–(2.10), where eight different values $\gamma = 0, 20, 40, \ldots, 140$, ranging from no spatial dependence to strong spatial dependence, were considered. The error vector $\boldsymbol{\epsilon}$ is considered to be spatially dependent with the (i, j)th element of $\boldsymbol{\Sigma}$ given by $\exp(-\|\mathbf{s}_i - \mathbf{s}_j\|/\phi)$, where seven different values $\phi = 0, 10, 20, \ldots, 60$, ranging from independent errors to strong spatially dependent errors, were used. For each neighborhood set $\mathcal{N}_k \in \{\mathcal{N}^{(1)}, \mathcal{N}^{(3)}, \mathcal{N}^{(7)}, \mathcal{N}^{(11)}\}$; $k = 1, 2, 3, 4$, each $\gamma \in \{0, 20, 40, \ldots, 140\}$, and each $\phi \in \{0, 10, 20, \ldots, 60\}$, we computed the (tenfold) CV value of the predicted soil moisture index and the corresponding average squared prediction error (ASPE) on the testing set:

$$\text{ASPE} = \frac{1}{90^2 - 100} \sum_{\mathbf{s} \in D^* \setminus \{\mathbf{s}_1, \ldots, \mathbf{s}_{100}\}} (\hat{Y}(\mathbf{s}) - Y(\mathbf{s}))^2.$$

Figure 7 shows the boxplots of the 50 smallest CV values corresponding to $\gamma = 0, 20, 40, \ldots, 140$ and $\phi = 0, 10, 20, \ldots, 60$. We found that the smallest CV value is achieved at $\mathcal{N}_1 = \mathcal{N}_2 = \mathcal{N}_4 = \mathcal{N}^{(1)}$, $\mathcal{N}_3 = \mathcal{N}^{(11)}$, $\gamma = 80$, and $\phi = 50$ among the $4^4 \times 8 \times 7 = 14{,}336$ possible combinations of $(\mathcal{N}_1, \ldots, \mathcal{N}_4, \gamma, \phi)$. Therefore, the corresponding model was selected as the final model. Across all models, the average squared prediction errors (based on the unobserved part of the population) and the corresponding CV values (based on the sample) have strong positive correlation (see Figure 8), indicating that the CV statistic is sensible. The improvement from selecting the spatially dependent

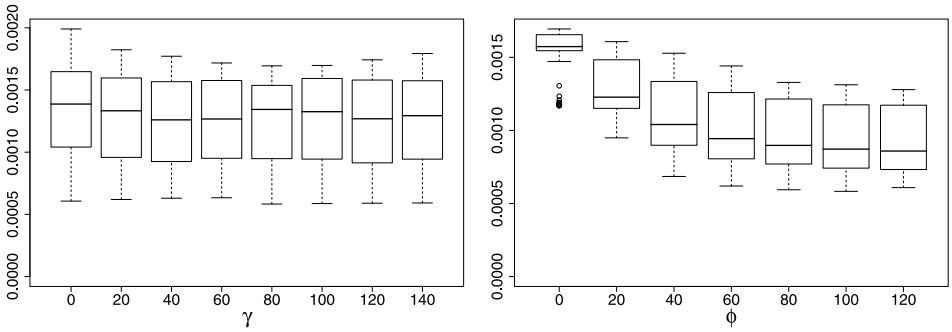


Figure 7. Boxplots of 50 smallest CV values of the predicted soil moisture index for various γ (left) and ϕ (right) values based on all 14,336 models considered for the soil moisture prediction example.

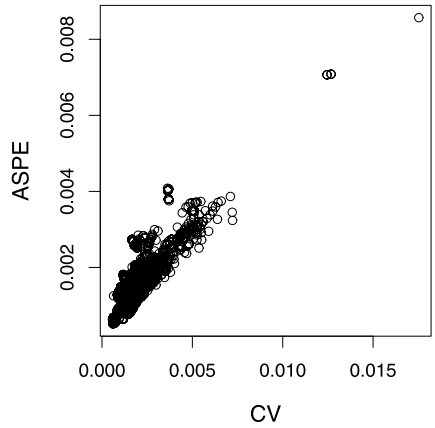


Figure 8. Average squared prediction errors of the soil moisture index (based on the unobserved part of the population) versus the corresponding CV values (based on the sample), for all 14,336 models considered for the soil moisture prediction example.

errors corresponding to $\phi = 50$ (CV $= 0.00058$ and ASPE $= 0.00064$) instead of independent errors corresponding to $\phi = 0$ (CV $= 0.00117$ and ASPE $= 0.00182$) is huge. The choice of γ turns out to have little impact on prediction, as suggested by Figure 7. The whole procedure is done using R, which takes about 4 hours on a Pentium-4 (3.4 GHz) PC in finding the smallest CV among the 4^4 combinations of neighborhood sets for each γ and each ϕ.

Figure 9(a) shows the predicted soil moisture index surface $\{\hat{Y}(\mathbf{s}) : \mathbf{s} \in D^*\}$ based on the final fitted model, where the crosses indicate the 100 locations sampled for model fitting. The corresponding prediction standard errors of $\{\hat{Y}(\mathbf{s}) - Y(\mathbf{s}) : \mathbf{s} \in D^*\}$ and the absolute standardized prediction errors are shown in Figures 9(b) and (c). Here the prediction standard errors were obtained using parametric bootstrap based on 100 replicates. Specifically, the kth bootstrap sample was generated from $Y^{*(k)}(\mathbf{s}) = \hat{\mu}_{\mathrm{SL}}(\mathbf{s}) + \varepsilon^{*(k)}(\mathbf{s})$; $\mathbf{s} \in D$, where $\{\varepsilon^{*(k)}(\mathbf{s}) : \mathbf{s} \in D\}$ were generated from a zero-mean stationary Gaussian process with spatial covariance function $\mathrm{cov}(\varepsilon^{*(k)}(\mathbf{s}), \varepsilon^{*(k)}(\mathbf{s}')) = \hat{\sigma}^2 \exp(-\|\mathbf{s}-\mathbf{s}'\|/50)$, and $\hat{\sigma}^2$ is the sample variance of $\{Y(\mathbf{s}_1) - \hat{\mu}_{\mathrm{SL}}(\mathbf{s}_1), \ldots, Y(\mathbf{s}_{100}) - \hat{\mu}_{\mathrm{SL}}(\mathbf{s}_{100})\}$. The spatial Lasso predictors

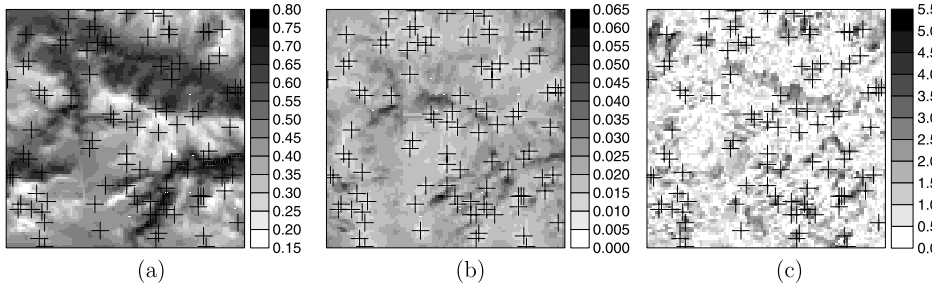


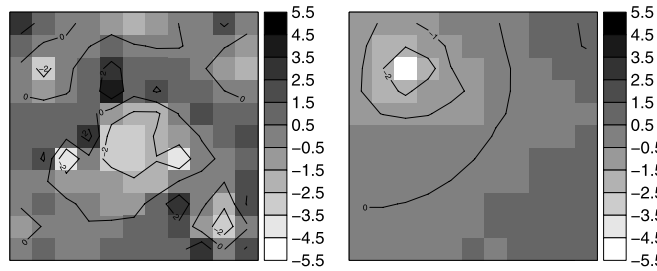
Figure 9. (a) Predicted soil moisture index. (b) Prediction standard errors. (c) Absolute standardized prediction errors.

Table 1. Standardized coefficient with the largest absolute value on the whole neighborhood set for each variable, where the standard errors of the estimated coefficients were obtained using a bootstrap method.

Variable	Effect	Standardized coefficient	Selected neighborhood set
x_1	`aspect`	−0.427	$\mathcal{N}^{(1)}$
x_2	`hillshade`	−1.827	$\mathcal{N}^{(1)}$
x_3	`elevation`	−3.985	$\mathcal{N}^{(11)}$
x_4	`slope`	−1.756	$\mathcal{N}^{(1)}$
x_5	`precipitation`	0.211	$\mathcal{N}^{(1)}$
x_6	`elevation*precipitation`	−0.257	$\mathcal{N}^{(1)}$
x_7	`slope*precipitation`	2.192	$\mathcal{N}^{(1)}$
x_8	`aspect*precipitation`	−0.435	$\mathcal{N}^{(1)}$
x_9	`aspect*elevation`	−5.209	$\mathcal{N}^{(11)}$
x_{10}	`aspect*slope`	2.076	$\mathcal{N}^{(1)}$

$\{\hat{Y}^{*(k)}(\mathbf{s}) : \mathbf{s} \in D^*\}$ were obtained by selecting among the 4^4 combinations of neighborhood sets under $\gamma = 80$ and $\phi = 50$ based on $\{Y^{*(k)}(\mathbf{s}_1), \ldots, Y^{*(k)}(\mathbf{s}_{100})\}$. Then the bootstrap prediction standard error of $\hat{Y}(\mathbf{s})$ is given by $(\frac{1}{100}\sum_{k=1}^{100}(\hat{Y}^{*(k)}(\mathbf{s}) - Y^{*(k)}(\mathbf{s}))^2)^{1/2}$ for $\mathbf{s} \in D^*$. The predicted soil moisture index can be seen to match the true index well.

For each covariate, the standardized coefficient with the largest absolute value on the whole neighborhood set is shown in Table 1, where the standard errors were obtained using the same bootstrap method described above. The variables `aspect` and `hillshade` are highly correlated (see Figure 6), and their effect depends on elevation and slope. As expected, the soil moisture index decreases at higher elevations (x_3) and on steeper slopes (x_4). The negative effect of elevation increases on north-facing aspects (x_9), and the negative effect of slope decreases on north-facing aspects (x_{10}). Naturally, the negative effect of slope on the soil moisture index decreases with increased precipitation (x_7). None of the other variables has a significant impact on the soil moisture index. The standardized coefficients of x_3 and x_9 are shown in Figure 10, where some spatial dependence features among the coefficients are apparent, as expected.



(a) (b)

Figure 10. Standardized coefficients of some variables: (a) `elevation`; (b) `aspect*elevation`, where the standard errors of the estimated coefficients were obtained using a bootstrap method.

3.4 APPLICATION TO GENERALIZED LINEAR REGRESSION MODELS

In this subsection, we extend the spatial regression model of (2.1), (2.2), and (2.3) to generalized linear regression. Let $\mathbf{X}_i$ be the vector of the GIS variables at the ith location defined in (2.3). Suppose that the joint density of responses $\mathbf{Y} \equiv (Y_1, \ldots, Y_n)'$ can be expressed as

$$L(\mathbf{Y}; \beta_0, \boldsymbol{\beta}) = \prod_{i=1}^{n} \exp(a(\mu_i)Y_i + b(\mu_i) + c(Y_i)), \tag{3.7}$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions, and $\mu_i = E(Y_i)$ is related to $\mathbf{X}_i$ through a known link function, $g(\mu_i) = \beta_0 + \mathbf{X}_i'\boldsymbol{\beta}$. Then, by (2.4) and (2.7) with $\boldsymbol{\Sigma} = \mathbf{I}$,

$$(g(\mu_1), \ldots, g(\mu_n))' = \beta_0\mathbf{1} + \mathbf{X}\boldsymbol{\beta} = \beta_0\mathbf{1} + \mathbf{X}^*\boldsymbol{\beta}^* = \beta_0\mathbf{1} + \mathbf{X}^{**}\boldsymbol{\beta}^{**},$$

recalling that $\boldsymbol{\beta}^{**}$ is a reparameterization of $\boldsymbol{\beta}$ designed to capture spatial smoothness of $\boldsymbol{\beta}$ in a lower-dimensional model. The spatial Lasso estimates $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ under model (3.7) are defined by minimizing

$$-2\ell(\beta_0, \boldsymbol{\beta}^{**}) + \lambda \sum_{j=1}^{m} |\beta_j^{**}|, \tag{3.8}$$

where $\ell(\beta_0, \boldsymbol{\beta}^{**}) \equiv \log(L^*(\mathbf{Y}; \beta_0, \boldsymbol{\beta}^{**})) \equiv \log(L(\mathbf{Y}; \beta_0, \boldsymbol{\beta}))$. Suppose that the log-likelihood function has the first two derivatives with respect to β_0 and $\boldsymbol{\beta}$. Then $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ can be obtained by iteratively applying the following penalized least squares algorithm:

$$\begin{aligned}
&\big(\hat{\beta}_0^{(N)}, \hat{\boldsymbol{\beta}}^{**(N)}\big) \\
&\quad = \arg\min_{(\beta_0, \boldsymbol{\beta}^{**})} \Big\{ -2\big(\beta_0 - \hat{\beta}_0^{(N-1)}, \big(\boldsymbol{\beta}^{**} - \hat{\boldsymbol{\beta}}^{**(N-1)}\big)'\big)\nabla\ell\big(\beta_0^{(N-1)}, \boldsymbol{\beta}^{**(N-1)}\big) \\
&\qquad - \big(\beta_0 - \hat{\beta}_0^{(N-1)}, \big(\boldsymbol{\beta}^{**} - \hat{\boldsymbol{\beta}}^{**(N-1)}\big)'\big) \\
&\qquad \times \nabla^2\ell\big(\beta_0^{(N-1)}, \boldsymbol{\beta}^{**(N-1)}\big) \begin{pmatrix} \beta_0 - \hat{\beta}_0^{(N-1)} \\ \boldsymbol{\beta}^{**} - \hat{\boldsymbol{\beta}}^{**(N-1)} \end{pmatrix} + \lambda \sum_{j=1}^{m} |\beta_j^{**}| \Big\}
\end{aligned} \tag{3.9}$$

for $N = 1, 2, \ldots$, where $\hat{\beta}_0^{(0)}$ and $\hat{\boldsymbol{\beta}}^{**(0)}$ are some given initial estimates. Note that (3.9) is equivalent to (2.7) if $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

For illustration, we applied the spatial Lasso of (3.8) to the Lansing Woods data (Gerrard 1969) in a Poisson regression model. The original data consist of the locations of 2251 trees classified into six categories: hickory, maple, red oak, white oak, black oak, and miscellaneous trees. We considered only hickory and maple trees, and investigated if the presence of hickory trees hinders the presence of maple trees. Similarly to the work of Fingleton (1986) and Lin (2008), we divided the region into 24×24 quadrats indexed by $\mathbf{s} \in D = \{(i_1, i_2) : i_1, i_2 = 1, \ldots, 24\}$. Let $Y(\mathbf{s})$ and $x(\mathbf{s})$ be the number of maple trees and the number of hickory trees in quadrat $\mathbf{s}$ (Figures 11(a) and (b)). We applied model (3.7) with response Y and a GIS layer x using Poisson regression with the log link function, where $p = J = L_1 = 1$, and $\phi_1(\cdot) = \psi_{1,1}(\cdot) \equiv 1$ for the mean structure in (2.1) and (2.2).

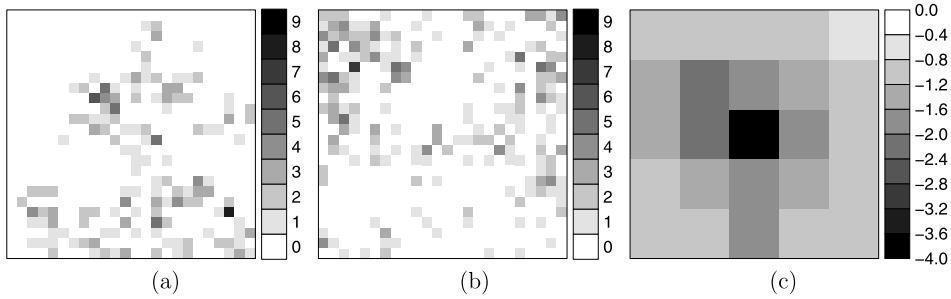


Figure 11. (a) The number of maple trees. (b) The number of hickory trees. (c) Estimated Poisson regression coefficients.

Our spatial Lasso estimates were obtained by applying (3.9), leading to the following algorithm:

$$
\begin{aligned}
&\bigl(\hat{\beta}_0^{(N)}, \hat{\boldsymbol{\beta}}^{**(N)}\bigr) \\
&\quad = \underset{(\beta_0, \boldsymbol{\beta}^{**})}{\arg\min} \Biggl\{ -2 \sum_{i=1}^{n} \bigl(Y_i - \hat{\mu}_i^{(N-1)}\bigr)\bigl(\beta_0 + (\mathbf{X}_i^{**})' \boldsymbol{\beta}^{**}\bigr) \\
&\qquad + \sum_{i=1}^{n} \hat{\mu}_i^{(N-1)} \bigl(\beta_0 + (\mathbf{X}_i^{**})' \boldsymbol{\beta}^{**} - \log\bigl(\hat{\mu}_i^{(N-1)}\bigr)\bigr)^2 + \lambda \sum_{j=1}^{m} |\beta_j^{**}| \Biggr\} \\
&\quad = \underset{(\beta_0, \boldsymbol{\beta}^{**})}{\arg\min} \Biggl\{ \sum_{i=1}^{n} \hat{\mu}_i^{(N-1)} \Biggl(\log\bigl(\hat{\mu}_i^{(N-1)}\bigr) + \frac{Y_i - \hat{\mu}_i^{(N-1)}}{\hat{\mu}_i^{(N-1)}} - \beta_0 - (\mathbf{X}_i^{**})' \boldsymbol{\beta}^{**} \Biggr)^2 \\
&\qquad + \lambda \sum_{j=1}^{m} |\beta_j^{**}| \Biggr\}, \qquad (3.10)
\end{aligned}
$$

where $\hat{\mu}_i^{(N-1)} = \exp(\hat{\beta}_0^{(N-1)} + (\mathbf{X}_i^{**})' \hat{\boldsymbol{\beta}}^{**(N-1)})$. We considered five different neighborhood sets, $\{\mathcal{N}^{(2q+1)} : q = 0, 1, \ldots, 4\}$, defined in (3.6). For $\mathcal{N}^{(2q+1)}$ with $q \geq 1$, the x variables are extended outside D by reflection at the boundaries. The maximum likelihood (ML) estimates of β_0 and $\boldsymbol{\beta}$ based on $\mathcal{N}^{(1)}$ (i.e., with the covariate only at the same quadrat) were used as the initial estimates, $\hat{\beta}_0^{(0)}$ and $\hat{\boldsymbol{\beta}}^{**(0)}$, in (3.10). Recall that the prior covariance structure for the regression coefficients are given by (2.8)–(2.10), where ten different values $\gamma = 0, 5, \ldots, 45$, ranging from no spatial dependence to strong spatial dependence, were considered. The final combination of λ, γ, and the neighborhood set is selected by minimizing the tenfold cross-validated deviance:

$$
2 \sum_{i=1}^{n} \bigl\{ \hat{\mu}_i^{(\mathrm{CV})} - Y_i \log\bigl(\hat{\mu}_i^{(\mathrm{CV})}\bigr) \bigr\},
$$

where $\hat{\mu}_i^{(\mathrm{CV})}$ is a generic estimate of μ_i obtained from the corresponding CV sample, for $i = 1, \ldots, n$.

The proposed method selects $\gamma = 5$ and $\mathcal{N}^{(5)}$ as the neighborhood set. The tenfold cross-validated deviance for the ML estimates based on $\mathcal{N}^{(1)}$ is 1045.7, whereas the tenfold

cross-validated deviance for our selected model is 942.1, showing some evidence that our selected method has better predictive ability. The estimated Poisson regression coefficients of $\{x(\cdot + \mathbf{u}) : \mathbf{u} \in \mathcal{N}^{(5)}\}$ are shown in Figure 11(c), which are all negative, indicating that presence of maple trees is inhibited by presence of hickory trees not only through the same quadrat but also through neighboring quadrats.

4. DISCUSSION

In this article, we have considered the problem of spatial regression using a sparse spatial sample and multiple GIS layers. This is an increasingly important problem as GIS data become more and more prevalent. Model selection in this context is complex because there are typically many layers from which to choose, and appropriate neighborhoods within layers must also be chosen. Our approach is to formalize this problem by constructing a large and flexible linear model, then using a Lasso method to simultaneously select variables, choose neighborhoods, and estimate parameters. Spatial Lasso is our extension of Lasso that allows for spatially dependent errors and the possibility of smoothness in selected coefficients, incorporated through use of a GLS formulation and a priori spatial covariance structure. Standard Lasso is a special case with an OLS formulation and a priori independence. The spatial Lasso performs well in the numerical examples we have considered, including an application to prediction of soil moisture using real GIS data. In addition, the methodology is extended to generalized linear regression, illustrated by a Poisson regression model for the Lansing Woods dataset.

Other Lasso related approaches, such as the fused Lasso (Tibshirani et al. 2005), the grouped Lasso (Yuan and Lin 2006), and the constraint given in the simultaneous variable selection approach (Turlach, Venables, and Wright 2005), may also be adapted to select GIS variables. However, these methods either can only be extended to account for the spatial smoothness in each GIS layer in a less natural way or require more computationally intensive algorithms for implementation.

The spatial Lasso has an advantage of performing variable selection and estimation simultaneously. Nevertheless, like the standard Lasso, a two-step procedure can also be considered by first applying the spatial Lasso to select the model but not to estimate the coefficients, and then refitting the model using GLS. The resulting GLS estimate has a smaller bias and a smaller generalized residual sum of squares at the cost of having a larger variance.

Although we focus on the Lasso method in this article, other variable selection methods based on different penalized least squares formulations, such as SCAD (Fan and Li 2001), the nonnegative garrote (Breiman 1995), and the elastic net (Zou and Hastie 2005), can also be considered. The proposed method can be easily extended under these settings.

SUPPLEMENTAL MATERIALS

Data Set: Soil moisture dataset for the soil moisture example in Section 3.3. (data.txt)
Spatial Lasso Code: R function for implementing spatial Lasso. (spatial_lasso.R)

Source Code: R script for reproducing soil moisture prediction in Section 3.3. (soil_moisture.R)

ACKNOWLEDGMENT

The work of all four authors was supported in part by STAR Research Assistance Agreement CR-829095 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University. This work has not been formally reviewed by EPA. The views expressed here are solely those of the authors. EPA does not endorse any products or commercial services mentioned in this report.

[Received August 2007. Revised January 2010.]

REFERENCES

Akaike, H. (1973), "Information Theory and the Maximum Likelihood Principle," in *International Symposium on Information Theory*, eds. V. Petrov and F. Csáki, Budapest: Akademiai Kiádo, pp. 267–281. [966]

Breiman, L. (1995), "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373–384. [981]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression" (with discussion), *The Annals of Statistics*, 32, 407–499. [964,967,968]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [981]

Fingleton, B. (1986), "Analyzing Cross-Classified Data With Inherent Spatial Dependence," *Geographic Analysis*, 18, 48–61. [979]

Gerrard, D. J. (1969), "Competition Quotient: A New Measure of the Competition Affecting Individual Forest Trees," Research Bulletin 20, Agricultural Experiment Station, Michigan State University. [979]

Iverson, L. R., Dale, M. E., Scott, C. T., and Prasad, A. (1996), "A GIS-Derived Integrated Moisture Index to Predict Forest Composition and Productivity of Ohio Forests (U.S.A.)," *Landscape Ecology*, 12, 331–348. [975]

Lin, P.-S. (2008), "Estimating Equations for Spatially Correlated Data in Multi-Dimensional Space," *Biometrika*, 95, 847–858. [979]

Osborne, M., Presnell, B., and Turlach, B. (2000a), "A New Approach to Variable Selection in Least Squares Problems," *IMA Journal of Numerical Analysis*, 20, 389–403. [967]

——— (2000b), "On the LASSO and Its Dual," *Journal of Computational and Graphical Statistics*, 9, 319–337. [967]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288. [964,966-968]

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), "Sparsity and Smoothness via the Fused Lasso," *Journal of the Royal Statistical Society, Ser. B*, 67, 91–108. [981]

Turlach, B. A. (2005), "On Algorithms for Solving Least Squares Problems Under an l_1 Penalty or an l_1 Constraint," in *2004 Proceedings of the Statistical Computing Section*, Alexandria, VA: American Statistical Association, pp. 2572–2577. [968]

Turlach, B. A., Venables, W. N., and Wright, S. J. (2005), "Simultaneous Variable Selection," *Technometrics*, 47, 349–363. [981]

Ye, J. (1998), "On Measuring and Correcting the Effects of Data Mining and Model Selection," *Journal of the American Statistical Association*, 93, 120–131. [967]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society, Ser. B*, 68, 49–67. [981]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society, Ser. B*, 67, 301–320. [981]

Zou, H., Hastie, T., and Tibshirani, R. (2007), "On the 'Degrees of Freedom' of the Lasso," *The Annals of Statistics*, 35, 2173–2192. [967]