Local Variable Selection and Parameter Estimation of Spatially Varying Coefficient Models

Wesley Brooks

1. Introduction

Whereas the coefficients in traditional linear regression are scalar constants, the coefficients in a varying coefficient regression (VCR) model are functions - often *smooth* functions - of some effect modifying variable (Hastie and Tibshirani, 1993). When the effect modifying variable represents location in a spatial domain, a VCR model implies a spatially local regression model such that the regression coefficients vary over space and will be referred to as a spatially varying coefficient model (SVCR). Statistical inference for the coefficients as functions of location in an SVCR model is more complicated than estimating the coefficients in a global linear regression model where the coefficients are constant across the spatial domain. This document concerns the development of new methodologies for the analysis of spatial data using SVCR.

The methodology described herein is directly applicable only to geostatistical data, which is spatial data observed at discrete locations. Let $\mathcal{D}$ be a spatial domain on which data is collected, and let $\boldsymbol{s}$ denote a location variable that indexes the domain $\mathcal{D}$. Let univariate $\{Y(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ and possibly multivariate $\{\boldsymbol{X}(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ denote random fields called the response and the covariates, respectively. For $i = 1, \ldots, n$, let $\boldsymbol{s}_i$ denote the location in $\mathcal{D}$ of the ith observation of the response and the covariates. Then the data are a realization of the random variables $\{Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n), \boldsymbol{X}(\boldsymbol{s}_1), \ldots, \boldsymbol{X}(\boldsymbol{s}_n)\}$. Let the observed data be denoted $\{y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n), \boldsymbol{x}(\boldsymbol{s}_1), \ldots, \boldsymbol{x}(\boldsymbol{s}_n)\}$.

Areal data is a different kind of spatial data in which the spatial domain $\mathcal{D}$ consists of n regions $\{r_1, \ldots, r_n\}$. In the case of areal data, the random variables $\{Y(r_1), \ldots, Y(r_n), \boldsymbol{X}(r_1), \ldots, \boldsymbol{X}(r_n)\}$ are defined for regions instead of for points; population and spatial mean temperature are examples of areal data. The analytical method described herein can be applied to areal data if it is recast as geostatistical data by assuming that the data are point-referenced to the centroid of each region, i.e. $\{\boldsymbol{X}(\boldsymbol{s}_i), Y(\boldsymbol{s}_i)\} = \{\boldsymbol{X}(r_i), Y(r_i)\}$ where $\boldsymbol{s}_i$ is the centroid of r_i for $i = i, \ldots, n$. The data example in section 5 uses areal data relating to county-level demographics in this way.

Common practice in the analysis of geostatistical and areal data is to model the response variable with a spatial linear regression model consisting of the sum of a fixed mean function, a spatial random effect, and random error all on domain $\mathcal{D}$, as in:

$$Y(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})'\boldsymbol{\beta} + W(\boldsymbol{s}) + \varepsilon(\boldsymbol{s}) \tag{1}$$

where $\boldsymbol{X}(\boldsymbol{s})'\boldsymbol{\beta}$ is the mean function consisting of $\boldsymbol{X}(\boldsymbol{s})$, a possibly multivariate spatial random field of covariates, and $\boldsymbol{\beta}$, a vector of regression coefficients. The random error $\varepsilon(\boldsymbol{s})$ denotes a white noise field such that the errors are independent and identically distributed with mean zero and variance σ^2, while the random component $W(\boldsymbol{s})$ denotes a mean-zero, second-order stationary random field that is independent of the random error. The mean function captures the large-scale systematic trend of the response, the spatial random field $W(\boldsymbol{s})$ can be thought of as a small-scale spatial random effect, and the error term $\varepsilon(\boldsymbol{s})$ captures micro scale variation (Cressie, 1993). It is common to prespecify the form of a covariance function for the spatial random effect $W(\boldsymbol{s})$ (Diggle and Ribeiro, 2007). For example, the exponential covariance function (a special case of the Matérn class of covariance functions) has the form

$$\mathrm{Cov}(W(\boldsymbol{s}), W(\boldsymbol{t})) = \exp\left\{-\phi^{-1}\delta(\boldsymbol{s}, \boldsymbol{t})\right\} \tag{2}$$

where ϕ denotes a range parameter and $\delta(\boldsymbol{s}, \boldsymbol{t})$ denotes the Euclidean distance between locations $\boldsymbol{s}$ and $\boldsymbol{t}$. The general form of a Matérn class covariance function is

$$\mathrm{Cov}(W(\boldsymbol{s}), W(\boldsymbol{t})) = \left\{\Gamma(\nu)2^{\nu-1}\right\}^{-1}\left\{\delta(\boldsymbol{s}, \boldsymbol{t})\phi^{-1}\sqrt{2\nu}\right\}^{\nu} K_{\nu}\left(\delta(\boldsymbol{s}, \boldsymbol{t})\phi^{-1}\sqrt{2\nu}\right) \tag{3}$$

where ϕ denotes a range parameter, ν denotes the degree of smoothness, K_{ν} denotes the modified Bessel equation of the second kind, and $\delta(\boldsymbol{s}, \boldsymbol{t})$ denotes the Euclidean distance between locations $\boldsymbol{s}$ and $\boldsymbol{t}$. The exponential covariance function corresponds to a Matérn class covariance function with $\nu = 1/2$.

A spatial field is said to be stationary if the joint distribution of a sample from the field does not change when the sample locations are all shifted in space by the same amount. Let $\{T(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ be a random field on spatial domain $\mathcal{D}$ that takes value $T(\boldsymbol{s}_i)$ at location $\boldsymbol{s}_i \in \mathcal{D}$ for $i = 1, \ldots, n$. The random field $T(\boldsymbol{s})$ is stationary if $F_n\left(T(\boldsymbol{s}_1), \ldots, T(\boldsymbol{s}_n)\right) = F_n\left(T(\boldsymbol{s}_1 + \boldsymbol{h}), \ldots, T(\boldsymbol{s}_n + \boldsymbol{h})\right)$ where $F_n(\cdot)$ is the joint distribution of a length n sample from $T(\boldsymbol{s})$. A spatial random field is second-order stationary if the joint distribution of any two observations from a sample does not change when the sample locations are shifted by the same amount.

The coefficient vector $\boldsymbol{\beta}$ in (1) is a specific example of the case where $\{\boldsymbol{\beta}(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ is a spatial random field. Specifically, the coefficient vector in (1) represents the case of where $\boldsymbol{\beta}(\boldsymbol{s}) \equiv \boldsymbol{\beta}$, $\forall \boldsymbol{s} \in \mathcal{D}$, i.e. the random field is constant. Clearly such random field is stationary. It is also possible to specify a non-constant random coefficient field that is nevertheless stationary. Gelfand et al. (2003) suggests a Bayesian hierarchical model where

Both kernel-based and spline-based methods are available for fitting varying coefficient models. For example, Wood (2006) demonstrated that it is straightforward to modify a thin plate regression spline model into a VCR model. Loader (1999) developed the local likelihood as a tool for fitting generalized linear models with varying coefficients using kernel smoothing, while Fan and Zhang (1999) demonstrated that the optimal kernel bandwidth estimate for a VCR model can be found via a two-step technique.

Model selection in VCR models may be local or global. Global selection means including or excluding variables everywhere in the model domain, while local selection means including or excluding variables at each observation location. Two methods have been proposed for global model selection in spline-based VCR models. Wang et al. (2008) applied a SCAD penalty (Fan and Li, 2001) for variable selection in spline-based VCR models with a univariate effect-modifying variable. Antoniadas et al. (2012) used the nonnegative Garrote penalty (Breiman, 1995) in P-spline-based VCR models having a univariate effect-modifying variable.

This document focuses on GWR, which is a kernel-based method of estimating the coefficients of a VCR model in the context of spatial data (Brundson et al., 1998; Fotheringham et al., 2002). GWR uses kernel-weighted regression with weights based on the distance between observation locations. The presentation of GWR in Fotheringham et al. (2002) followed the development of local likelihood in Loader (1999). GWR can be thought of as a kernel smoother for regression coefficients, which tends to exhibit bias near the boundary of the region being modeled (Hastie and Loader, 1993). One way to reduce the boundery-effect bias is to model the coefficient surface as locally linear rather than locally constant by including coefficient-by-location interactions (Hastie and Loader, 1993). Adding these interactions to the GWR model is analogous to a transition from kernel smoothing to local regression, which was introduced in Wang et al. (2008).

Three regularization methods were used in this work. The adaptive Lasso was implemented in two ways - once via the lars algorithm (Efron et al., 2004) which uses least squares, and once via coordinate descent using the R package glmnet (Friedman et al., 2010). The third regularization method implemented here uses the adaptive elastic net penalty (Zou and Zhang, 2009), also via coordinate descent using the glmnet package.

This document aims to develop a new method of local variable selection for Geographically Weighted Regression (GWR) models, which are a class of kernel-based VCR models for geostatistical data. Traditional GWR relies on *a priori* model selection to decide which variables should be included in the model. In the context of ordinary least squares regression, regularization methods such as the adaptive Lasso (Zou, 2006) have been shown to have appealing properties for automating variable selection, sometimes including the "oracle" property of asymptotically selecting exactly the correct variables for inclusion in a regression model.

The idea of using regularization for local variable selection in a GWR model first appeared in the literature as the geographically-weighted Lasso (GWL) of Wheeler (2009), which used a jackknife criterion for selection of the Lasso tuning parameters. Because the jackknife criterion can only be computed at locations where the response variable is observed, the GWL cannot be used for imputation of missing data nor for interpolation between observation locations. We avoid this limitation of the GWL by using a penalized-likelihood criterion to select the Lasso tuning parameters. Here we use a version of the AIC, but in principle one could use another information criterion like the BIC. The local AIC presented here is based on the local likelihood (Loader, 1999) and the total AIC is based on an *ad hoc* calculation of the sample size and degrees of freedom for estimating the spatially-varying coefficient surfaces.

2. Geographically Weighted Regression

2.1. Model

Consider n data observations, made at sampling locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ in a spatial domain $D \subset \mathbb{R}^2$. For $i = 1, \ldots, n$, let $y(\boldsymbol{s}_i)$ and $\boldsymbol{x}(\boldsymbol{s}_i)$ denote the univariate response variable, and a $(p+1)$-variate vector of covariates measured at location $\boldsymbol{s}_i$, respectively. At each location $\boldsymbol{s}_i$, assume that the outcome is related to the covariates by a linear model where the coefficients $\boldsymbol{\beta}(\boldsymbol{s}_i)$ may be spatially-varying and $\varepsilon(\boldsymbol{s}_i)$ is random noise at location $\boldsymbol{s}_i$. That is,

$$y(\boldsymbol{s}_i) = \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta}(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i) \tag{4}$$

Further assume that the error term $\varepsilon(\boldsymbol{s}_i)$ is normally distributed with zero mean and a possibly spatially-varying variance $\sigma^2(\boldsymbol{s}_i)$, and that $\varepsilon(\boldsymbol{s}_i)$, $i = 1, \ldots, n$ are independent.

$$\varepsilon(\boldsymbol{s}_i) \overset{indep}{\sim} \mathcal{N}\left(0, \sigma^2(\boldsymbol{s}_i)\right) \tag{5}$$

In order to simplify the notation, let $\boldsymbol{x}(\boldsymbol{s}_i) \equiv \boldsymbol{x}_i \equiv (1, x_{i1}, \ldots, x_{ip})'$, $\boldsymbol{\beta}(\boldsymbol{s}_i) \equiv \boldsymbol{\beta}_i \equiv (\beta_{i0}, \beta_{i1}, \ldots, \beta_{ip})'$, $y(\boldsymbol{s}_i) \equiv y_i$, and $\sigma^2(\boldsymbol{s}_i) \equiv \sigma_i^2$. Further, let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ and $\boldsymbol{y} = (y_1, \ldots, y_n)'$. Equations (4) and (5) can now be rewritten as

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_i + \varepsilon_i \text{ and } \varepsilon_i \overset{indep}{\sim} \mathcal{N}\left(0, \sigma_i^2\right) \tag{6}$$

Thus, given the design matrix $\boldsymbol{X}$, observations of the response variable at different locations are independent of each other. Then the total log-likelihood of the observed data is the sum of the log-likelihood of each individual observation.

$$\ell\left(\boldsymbol{\beta}\right) = -\left(1/2\right)\sum_{i=1}^{n}\left\{\log\left(2\pi\sigma_i^2\right) + \left(\sigma_i^2\right)^{-1}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_i\right)^2\right\} \tag{7}$$

Since there are a total of $n \times (p+1)$ free parameters for n observations the model is not identifiable, so it is not possible to directly maximize the total likelihood. One way to effectively reduce the number of parameters is to assume that the coefficients $\boldsymbol{\beta}(\boldsymbol{s})$ are smoothly varying over space, and use a kernel smoother to make pointwise estimates of the coefficients by maximizing the local likelihood. In the setting of spatial data and with the kernel smoother based on the physical distance between observation locations, this is the traditional GWR.

2.2. Estimation

In geographically weighted regression, the coefficient surface $\boldsymbol{\beta}(\boldsymbol{s})$ is estimated at each sampling location $\boldsymbol{s}_i$. First calculate the Euclidean distance $\delta_{ii'} \equiv \delta\left(\boldsymbol{s}_i, \boldsymbol{s}_{i'}\right) \equiv \|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|_2$ between locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_{i'}$ for all i, i'. The bi-square kernel can be used to generate spatial weights based on the Euclidean distances and a bandwidth ϕ. The bisquare kernel assigns the maximum weight of one where $\boldsymbol{s}_i = \boldsymbol{s}_{i'}$ so $\delta_{ii'} = 0$, discontinuously differentiable, and assigns zero weight to observations at distances greater than one bandwidth from $\boldsymbol{s}_i$:

$$w_{ii'} = \begin{cases} \left[1 - \left(\phi^{-1}\delta_{ii'}\right)^2\right]^2 & \text{if } \delta_{ii'} < \phi \\ 0 & \text{if } \delta_{ii'} \geqslant \phi \end{cases} \tag{8}$$

For the purpose of estimation, define the local likelihood at each location (Fotheringham et al., 2002):

$$\mathcal{L}_i\left(\boldsymbol{\beta}_i\right)=\prod_{i'=1}^{n}\left[\left(2\pi\sigma_i^2\right)^{-1/2}\exp\left\{-\left(2\sigma_i^2\right)^{-1}\left(y_{i'}-\boldsymbol{x}_{i'}'\boldsymbol{\beta}_i\right)^2\right\}\right]^{w_{ii'}} \tag{9}$$

Thus, the local log-likelihood function is:

$$\ell_i\left(\boldsymbol{\beta}_i\right)\propto-(1/2)\sum_{i'=1}^{n}w_{ii'}\left\{\log\sigma_i^2+\left(\sigma_i^2\right)^{-1}\left(y_{i'}-\boldsymbol{x}_{i'}'\boldsymbol{\beta}_i\right)^2\right\} \tag{10}$$

The GWR coefficient estimates $\hat{\boldsymbol{\beta}}_{i,\mathrm{GWR}}$ maximize the local likelihood at location $\boldsymbol{s}_i$. From (9) and (10), it is apparent that $\hat{\boldsymbol{\beta}}_{i,\mathrm{GWR}}$ can be calculated using weighted least squares. Let $\boldsymbol{W}_i$ denote a diagonal weight matrix with

$$\boldsymbol{W}_i=\operatorname{diag}\left\{w_{ii'}\right\}_{i'=1}^{n} \tag{11}$$

Thus, it follows that

$$\hat{\boldsymbol{\beta}}_{i,\mathrm{GWR}}=\left(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{y} \tag{12}$$

The estimate of σ_i^2 is attained by maximizing (10). Thus,

$$\begin{aligned}\hat{\sigma}_i^2&=\left(\boldsymbol{1}_n'\boldsymbol{w}_i\right)^{-1}\left(\boldsymbol{y}-\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{y}\right)'\boldsymbol{W}_i\left(\boldsymbol{y}-\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{y}\right)\\&=\left(\boldsymbol{1}_n'\boldsymbol{w}_i\right)^{-1}\left(\boldsymbol{y}-\hat{\boldsymbol{y}}\right)'\boldsymbol{W}_i\left(\boldsymbol{y}-\hat{\boldsymbol{y}}\right)\end{aligned} \tag{13}$$

3. Model Selection

3.1. Variable Selection

3.1.1. Adaptive Lasso

The adaptive Lasso is an ℓ_1 regularization method for variable selection in regression models (Zou, 2006). Unlike the traditional Lasso (Tibshirani, 1996), which applies an equal penalty λ_i^* to each covariate in the local model at $\boldsymbol{s}_i$, the adaptive Lasso adjusts the penalty of each covariate based on the covariate's unpenalized local coefficient. Letting the vector of unpenalized local coefficients be $\boldsymbol{\gamma}_i$, the adaptive Lasso penalty for covariate j at location $\boldsymbol{s}_i$ is $|\lambda_i/\gamma_{ij}|$, where λ_i is a the local penalty that applies to all coefficients at location $\boldsymbol{s}_i$ and $\boldsymbol{\gamma}_i = \{\gamma_{ij}\}$ is the vector of adaptive weights at location $\boldsymbol{s}_i$. Thus, objective minimized by GWR to fit the local model at $\boldsymbol{s}_i$ using the adaptive Lasso is

$$\mathcal{S}_i = \sum_{i'=1}^{n} w_{ii'} \left(y_{i'} - \boldsymbol{x}'_{i'}\boldsymbol{\beta}_i \right)^2 + \lambda_i \sum_{j=1}^{p} |\beta_{ij}/\gamma_{ij}| \tag{14}$$

where $\sum_{i'=1}^{n} w_{ii'} \left(y_{i'} - \boldsymbol{x}'_{i'}\boldsymbol{\beta}_i \right)^2$ is the weighted least squares objective minimized by traditional GWR, and $\lambda_i \sum_{j=1}^{p} |\beta_{ij}/\gamma_{ij}|$ is the adaptive Lasso penalty.

To apply an adaptive Lasso to GWR, the design matrix $\boldsymbol{X}$ is first multiplied by $\boldsymbol{W}_i^{1/2}$, the diagonal matrix of geographic weights at $\boldsymbol{s}_i$. Since some of the weights $w_{ii'}$ may be zero, the matrix $\boldsymbol{W}_i^{1/2}\boldsymbol{X}$ is not of full rank. The matrices $\boldsymbol{Y}_i^*$, $\boldsymbol{X}_i^*$, and $\boldsymbol{W}_i^*$ are formed by dropping the rows of $\boldsymbol{X}$ and $\boldsymbol{W}_i$ that correspond to observations with zero weight in the regression model at location $\boldsymbol{s}_i$. Now, letting $\boldsymbol{U}_i^* = \boldsymbol{W}_i^{*1/2}\boldsymbol{X}_i^*$ and $\boldsymbol{V}_i^* = \boldsymbol{W}_i^{*1/2}\boldsymbol{Y}_i^*$, we seek to estimate the coefficients $\boldsymbol{\beta}_i$ of the regression model:

$$\boldsymbol{V}_i^* = \boldsymbol{U}_i^*\boldsymbol{\beta}_i + \boldsymbol{\varepsilon} \tag{15}$$

Each column of $\boldsymbol{U}_i^*$ is centered around zero and rescaled to have an ℓ_2-norm of one. Let $\widetilde{\boldsymbol{U}}_i^*$ denote the centered-and-scaled version of $\boldsymbol{U}_i^*$. Now the adaptive weights $\boldsymbol{\gamma}_i^*$ are calculated via least squares:

$$\boldsymbol{\gamma}_i = \left(\widetilde{\boldsymbol{U}}_i^{*\prime} \widetilde{\boldsymbol{U}}_i^* \right)^{-1} \widetilde{\boldsymbol{U}}_i^{*\prime} \boldsymbol{V}_i^* \tag{16}$$

For $j = 1, \ldots, p$, the jth column of $\tilde{\boldsymbol{U}}_i^*$ is multiplied by γ_{ij}, the j^{th} element of $\boldsymbol{\gamma}_i$. Call this rescaled matrix $\breve{\boldsymbol{U}}_i^*$.

Finally, the adaptive Lasso coefficient estimates minimizing (14) at location $\boldsymbol{s}_i$ are found, either by using the `lars` algorithm (Efron et al., 2004) to model $\boldsymbol{V}_i^*$ as a function of $\breve{\boldsymbol{U}}_i^*$ or by using the `glmnet` package to implement coordinate descent.

3.1.2. Adaptive Elastic Net

The adaptive elastic net combines the adaptive Lasso penalty with the ridge penalty (Zou and Zhang, 2009). Ridge regression is an ℓ_2 regularization technique that differs from the Lasso in that the ridge penalty $\lambda_i^\dagger$ is applied to the sum of the squared local regression coefficients (Hoerl and Kennard, 1970). The ridge penalty is used to estimate coefficients in regression models with correlated covariates because it stabilizes the inversion of the covariance matrix, which robustifies the coefficient estimates (Hastie et al., 2009). To implement the adaptive elastic net, the adaptive weights $\boldsymbol{\gamma}_i$ are calculated as for the adaptive Lasso, but there is an additional elastic net parameter α that controls the balance between the ℓ_1 and ℓ_2 penalties, so that the objective to be minimized is:

$$\begin{aligned}
&\sum_{i'=1}^{n} w_{ii'} \left(y_{i'} - \boldsymbol{x}_{i'}' \boldsymbol{\beta}_i\right)^2 + \alpha\lambda_i \sum_{j=1}^{p} |\beta_{ij}/\gamma_{ij}| + (1-\alpha)\lambda_i \sum_{j=1}^{p} \left(\beta_{ij}/\gamma_{ij}\right)^2 \\
= &\sum_{i'=1}^{n} w_{ii'} \left(y_{i'} - \boldsymbol{x}_{i'}' \boldsymbol{\beta}_i\right)^2 + \lambda_i \left(\alpha \sum_{j=1}^{p} |\beta_{ij}/\gamma_{ij}| + (1-\alpha) \sum_{j=1}^{p} \left[\beta_{ij}/\gamma_{ij}\right]^2 \right) \qquad (17)
\end{aligned}$$

In the simulation study (Section 4), α is calculated from the maximum global (i.e. for all data without weighting) Pearson correlation between any two covariates, $\rho_{\max}$: $\alpha = 1 - \rho_{\max}$.

3.2. Tuning Parameter Selection

At each location $\boldsymbol{s}_i$, it is necessary to select the Lasso tuning parameter λ_i. To compare different values of λ_i, we propose a locally-weighted version of the Akaike information criterion (AIC) (Akaike, 1974) which we call the local AIC, or $\mathrm{AIC}_{\mathrm{loc}}$. The local AIC is calculated by adding a penalty to the local likelihood, with the sum of the weights around $\boldsymbol{s}_i$, $\sum_{i'=1}^{n} w_{ii'}$, playing the role of the sample size and the "degrees of freedom" (df_i) at $\boldsymbol{s}_i$ given by the number of nonzero coefficients in $\boldsymbol{\beta}_i$ (Zou et al., 2007).

$$
\begin{aligned}
\mathrm{AIC}_{\mathrm{loc},i} &= -2 \sum_{i'=1}^{n} \ell_{ii'} + 2\mathrm{df}_i \\
&= -2 \times \sum_{i'=1}^{n} \log \left\{ \left(2\pi\hat{\sigma}_i^2\right)^{-1/2} \exp\left[-\frac{1}{2}\hat{\sigma}_i^{-2} \left(y_{i'} - \boldsymbol{x}'_{i'}\hat{\boldsymbol{\beta}}_{i'} \right)^2 \right] \right\}^{w_{ii'}} + 2\mathrm{df}_i \\
&= \sum_{i'=1}^{n} w_{ii'} \left\{ \log\left(2\pi\right) + \log\hat{\sigma}_i^2 + \hat{\sigma}_i^{-2} \left(y_{i'} - \boldsymbol{x}'_{i'}\hat{\boldsymbol{\beta}}_{i'} \right)^2 \right\} + 2\mathrm{df}_i \\
&= \hat{\sigma}_i^{-2} \sum_{i'=1}^{n} w_{ii'} \left(y_{i'} - \boldsymbol{x}'_{i'}\hat{\boldsymbol{\beta}}_{i} \right)^2 + 2\mathrm{df}_i + C_i \qquad (18)
\end{aligned}
$$

Since the estimated local variance $\hat{\sigma}_i^2$ is the variance estimate from the unpenalized local model, C_i does not depend on the choice of tuning parameter and can be ignored (Zou et al., 2007).

Wheeler (2009) proposed selecting the tuning parameter for the Lasso at location $\boldsymbol{s}_i$ to minimize the jackknife prediction error $|y_i - \hat{y}_i^{(i)}|$. Because the jackknife prediction error is undefined everywhere except for at observation locations, this choice restricts coefficient estimation to occur at the locations where data has been observed. By contrast, the local AIC can be calculated at any location where we can calculate the local likelihood. As a practical matter this allows for variable selection and coefficient surface estimation to be done at locations where no data was observed (interpolation) and for imputation of missing values of the response variable.

3.3. Bandwidth Selection

The bandwidth parameter ϕ in (8) is global and so a global statistic is needed, by which prospective bandwidths can be compared. We propose the following statistic, called the total AIC (AIC_{tot}):

$$\text{AIC}_{\text{tot}} = \sum_{i=1}^{n} \left\{ \log \hat{\sigma}_i^2 + \hat{\sigma}_i^{-2} \left(y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_i \right)^2 + 2 \left(\sum_{i'=1}^{n} w_{ii'} \right)^{-1} \text{df}_i \right\} \tag{19}$$

which is different than the AIC for traditional GWR (Fotheringham et al., 2002). The AIC for traditional GWR is based on the trace of the projection matrix of the GWR model. But using an ℓ_1 penalty (as in the adaptive Lasso or the adaptive elastic net) results in a non-linear smoother, as can be seen by equating the derivatives of (14) with respect to $\boldsymbol{\beta}$ to zero to obtain the maximum likelihood estimates $\hat{\boldsymbol{\beta}}_i$ (Zou et al., 2007).

$$\hat{\boldsymbol{\beta}}_i = \left(\boldsymbol{X}' \boldsymbol{W}_i \boldsymbol{X} \right)^{-1} \boldsymbol{X}' \boldsymbol{W}_i \boldsymbol{y} - (1/2) \left(\boldsymbol{X}' \boldsymbol{W}_i \boldsymbol{X} \right)^{-1} \boldsymbol{\lambda}_i$$

$$\hat{y}_i = \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_i = \boldsymbol{x}_i' \left(\boldsymbol{X}' \boldsymbol{W}_i \boldsymbol{X} \right)^{-1} \boldsymbol{X}' \boldsymbol{W}_i \boldsymbol{y} - (1/2) \boldsymbol{x}_i' \left(\boldsymbol{X}' \boldsymbol{W}_i \boldsymbol{X} \right)^{-1} \boldsymbol{\lambda}_i$$

where $\boldsymbol{\lambda}_i$ is a vector of adaptive Lasso penalties, the j^{th} component of which is $|\lambda_i / \gamma_{ij}|$.

Because of the kernel weights and the application of the adaptive Lasso, the sample size and the degrees of freedom are different for each local model. The total AIC is found by summing the local likelihood and local penalty over all of the locations $\boldsymbol{s}_i$. The local penalty requires that we calculate df_{tot}, the total degrees of freedom used by the local models.

The expression for df_{tot} is derived by analogy. It is certainly true that $\text{df}_{\text{tot}} = \sum_{i=1}^{n} \left(n^{-1} \text{df}_{\text{tot}} \right)$, which suggests that the total degrees of freedom is the sum of the local degrees of freedom. Rather than using the mean local degrees of freedom, $n^{-1}\text{df}_{\text{tot}}$ at each location, substitute df_i for df_{tot} and $\sum\limits_{i'=1}^{n} w_{ii'}$ for n to get

$$\text{df}_{\text{tot}} = \sum_{i=1}^{n} \text{df}_i \left(\sum_{i'=1}^{n} w_{ii'} \right)^{-1} \tag{20}$$

4. Simulation

4.1. Simulation setup

A simulation study was conducted to assess the performance of the method described in Sections 2–3.

Data was simulated on $[0,1] \times [0,1]$, which was divided into a 30×30 grid. Each of $p = 5$ covariates $X_1, \ldots, X_5$ was simulated by a Gaussian random field (GRF) with mean zero and exponential spatial covariance $Cov\left(X_{ji}, X_{ji'}\right) = \sigma_x^2 \exp\left(-\tau_x^{-1}\delta_{ii'}\right)$ where $\sigma_x^2 = 1$ is the variance, $\tau_x = 0$ is the range parameter, and $\delta_{ii'}$ is the Euclidean distance $\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|_2$. Correlation was induced between the covariates by multiplying the $\boldsymbol{X}$ matrix by $\boldsymbol{R}$, where $\boldsymbol{R}$ is the Cholesky decomposition of the covariance matrix $\Sigma = \boldsymbol{R}'\boldsymbol{R}$. The covariance matrix $\boldsymbol{\Sigma}$ is a 5×5 matrix that has ones on the diagonal and ρ for all off-diagonal entries, where ρ is the between-covariate correlation.

The simulated response is $y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_i + \varepsilon_i$ for $i = 1, \ldots, n$ where $n = 900$ and for simplicity the ε_i's were iid Gaussian with mean zero and variance σ_ε^2.

The simulated data include the output y and five covariates $X_1, \ldots, X_5$. The true data-generating model uses only X_1, so $X_2, \ldots, X_5$ are included to assess performance in variable-selection.

There were twelve simulation settings, each of which was simulated 100 times. For each of the twelve settings, $\boldsymbol{\beta}_1(\boldsymbol{s})$, the true coefficient surface for $\boldsymbol{X}_1$, was nonzero in at least part of the simulation domain. There were four other simulated covariates, but their true coefficient surfaces were zero across the area under simulation. The twelve simulation settings are described in Table 1. Three parameters were varied to produce the twelve settings: there were three functional forms for the coefficient surface $\beta_1(\boldsymbol{s})$ (step, gradient, and parabola - see Figure 1); data was simulated both with ($\rho = 0.5$) and without ($\rho = 0$) correlation between the covariates; and simulations were made with low ($\sigma_\varepsilon^2 = 0.25$) and high ($\sigma_\varepsilon^2 = 1$) variance for the random noise term.

The performance of the penalized GWR methods (adaptive Lasso via `lars` and via `glmnet`, and the adaptive elastic net (`enet`) was compared to that of oracular GWR (O-GWR), which is ordinary GWR with "oracular" variable selection, meaning that exactly the correct set of predictors was used to fit the GWR model at each location in the simulation. Also included in the comparison was the GWR algorithm of Fotheringham et al. (2002) without variable selection (`gwr`). Finally, there is a category of simulation results using the three penalized GWR methods for local variable selection and then ordinary GWR for coefficient estimation.

Results from the simulation were summarized at five locations on the simulated grid (see Figure 2). The five key locations were chosen because they represent interesting regions of the β_1 coefficient surfaces. The results of variable selection and coefficient estimation are presented in the tables below.

Selection: Tables 2

MSE of $\hat{\beta}_1(s_i)$ $(i = 1, \ldots, 5)$: Tables 4

Bias of $\hat{\beta}_1(s_i)$ $(i = 1, \ldots, 5)$: Tables 6

Variance of $\hat{\beta}_1(s_i)$ $(i = 1, \ldots, 5)$: Tables 7

MSE of $\hat{Y}(s_i)$ $(i = 1, \ldots, 5)$: Tables 8

4.2. Results

At locations where β_1 is nonzero, X_1 usually selected for inclusion in all or nearly all of the model runs. An exception is at location four for the step function, where X_1 was included in about half of the model runs. This is probably because location four is at the very point where β_1 transitions from zero to nonzero. Selection performance was relatively poor for the step function at location one, especially for data with $\sigma^2 = 1$. For those simulations, X_1 was correctly included in around 85% of the simulations. The bias, variance, and MSE of $\hat{\beta}_1$ under the same settings were also much larger than the baseline established by the standard `gwr` algorithm. The reason(s) for the poor performance under those particular conditions is currently unknown.

Otherwise, selection performance was good, with the rate of false positive selections for X_2–X_5 (and for X_1 where its true coefficient was zero) usually below 0.10. Selection (also bias, variance, and MSE of $\hat{\beta}_1(\boldsymbol{s})$) tended to suffer worse by the change from low to high error variance than by the change from low to high collinearity amongst the predictors.

There was not a clear and consistent difference in performance between the three selection methods. It might be expected that the adaptive elastic net would outperform the adaptive Lasso under greater covariate collinearity, but if such effect is real it is not apparent from this simulation. The unshrunk coefficient-estimation methods tended to exhibit more bias than the selection-plusshrinkage methods when the true coefficient value was near zero, and vice versa when the true coefficient was not near zero. The unshrunk methods were perhaps more consistent in their performance and for that reason they are probably preferable in practice.

Bias in coefficient estimation was greater and variance less for the standard `gwr` algorithm than for the methods described here. This is probably due to the fact that the methods described here show a preference for smaller bandwidths than those select by `gwr`. Accuracy (as measured by MSE) in fitting the true Y variables was comparable for all the methods.

4.3. Tables

4.3.1. Selection

4.3.2. Estimation

5. Data Analysis

5.1. Census Poverty Data

An example data analysis is presented to demonstrate application of penalized GWR. In this example we use penalized GWR to do local variable selection and coefficient estimation for a varying-coefficients model of how poverty is related to a list of demographic and social variables. The data is from the U.S. Census Bureau's decennial census from 1970. This analysis looks specifically at the upper midwestern states of Minnesota, Iowa, Wisconsin, Illinois, Indiana, and Michigan. This is areal data, aggregated at the county level.

Three kinds of variables were considered as potential predictors of county-level poverty rate.

- Variables that describe the county's employment structure (pag, the proportion of residents employed in agriculture, pex, the proportion of residents employed in mining, man, the proportion of residents employed in manufacturing, pfire, the proportion of residents employed in finance, insurance, and real estate, pserve, the proportion of residents employed in services, and potprof, the proportion of residents employed in other professions)

- Variables that describe the county's racial makeup (pwh, the proportion of residents who are white, pblk, the proportion of residents who are black, and phisp, the proportion of residents who are hispanic)

- pmetro: an indicator of whether the county is in a metropolitan area.

The outcome of interest (poverty rate) is a proportion, taking values in $[0, 1]$. To demonstrate the geographically-weighted Lasso in a linear regression context, we model the logit-transformed poverty rate. The predictor variables were not transformed - raw proportions were used.

5.2. Modeling

The adaptive elastic net was used for variable selection, and then coefficients for the selected variables were estimated by weighted least squares without shrinkage. The standard gwr algorithm was used to fit a model to the same data for the sake of comparison.

5.3. Figures

The coefficient estimates are plotted on maps of the upper midwest in Figure 3 (based on the adaptive elastic net) and Figure 4 (for standard GWR).

5.4. Discussion

It is immediately apparent that the estimated coefficient surfaces are non-constant for most variables. The same large-scale patterns appear in both figures, but with differences. First of all, the adaptive elastic net has selected a larger bandwidth than base GWR, so there is less variability in the coefficient estimates from the adaptive elastic net. This may be one reason that the adaptive elastic net coefficient estimates are less extreme than those for base GWR. In a model with a logit-transformed proportion as the output, the coefficients can be interpreted as log odds ratios, so, e.g., the estimate of -100 as the coefficient of `phisp` (albeit at the edge of the domain) seems unrealistic.

Assessing variable selection for this data is difficult, since the adaptive elastic net almost never removed any variables from the model. Indeed, some coefficients seem nearly constant across the domain. An exception is the coefficient surface for `pex` (mining employment). That surface indicates an interaction whereby the proportion of people working in mining in southern parts of the domain is associated with an increase in the poverty rate, while in northern parts of the domain it is associated with a decrease in the poverty rate.

6. References

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Antoniadas, A., I. Gijbels, and A. Verhasselt (2012). Variable selection in varying-coefficient models using p-splines. *Journal of Computational and Graphical Statistics 21*(3), 638–661.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics 51*, 373–384.

Brundson, C., S. Fotheringham, and M. Charltonn (1998). Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. *Environment and Planning A 30*, 1905–1927.

Cressie, N. (1993). *Statistics for spatial data*. Wiley.

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

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*(2), 407–499.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fan, J. and W. Zhang (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics 27*(5), 1491–1518.

Fotheringham, A., C. Brunsdon, and M. Charlton (2002). *Geographically weighted regression: the analysis of spatially varying relationships*. Wiley.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Gelfand, A. E., H.-J. Kim, C. F. Sirmans, and S. Banerjee (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association 98*(462), 387–396.

Hastie, T. and C. Loader (1993). Local regression: automatic kernel carpentry. *Statistical Science 8*(2), 120–143.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological) 55*(4), pp. 757–796.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer New York.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Loader, C. (1999). *Local regression and likelihood.* Springer New York.

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

Wang, L., H. Li, and J. Z. Huang (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association 103*(484), 1556–1569.

Wang, N., C.-L. Mei, and X.-D. Yan (2008). Local linear estimation of spatially varying coefficient models: an improvement on the geographically weighted regression technique. *Environment and Planning A 40*, 986–1005.

Wheeler, D. C. (2009). Simultaneous coefficient penalization and model selection in geographically weighted regression: the geographically weighted lasso. *Environment and Planning A 41*, 722–742.

Wood, S. (2006). *Generalized additive models: an introduction with R*. Texts in statistical science. Chapman & Hall/CRC.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.

Zou, H., T. Hastie, and R. Tibshirani (2007). On the "degrees of freedom" of the lasso. *Annals of Statistics 35*(5), 2173–2192.

Zou, H. and H. Zhang (2009). On the adaptive elastic net with a diverging number of parameters. *The Annals of Statistics 37*(4), 1733–1751.

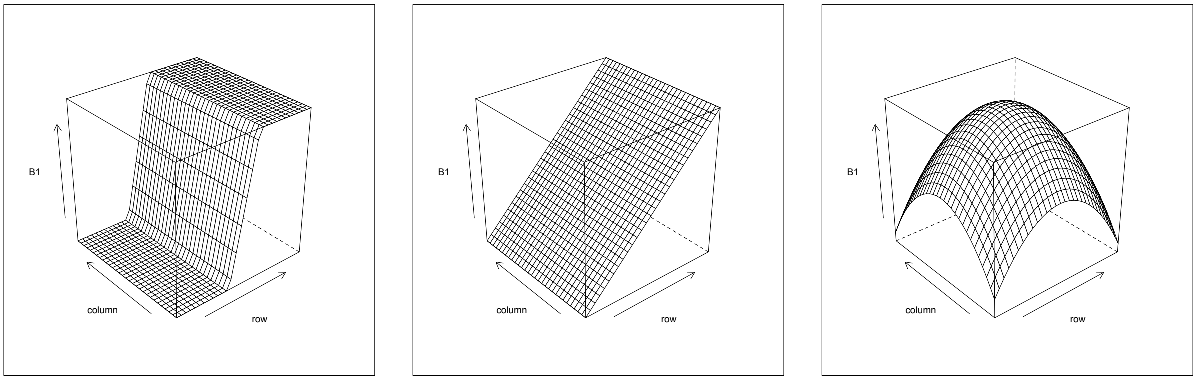

Figure 1: The actual β_1 coefficient surface used in the simulation.

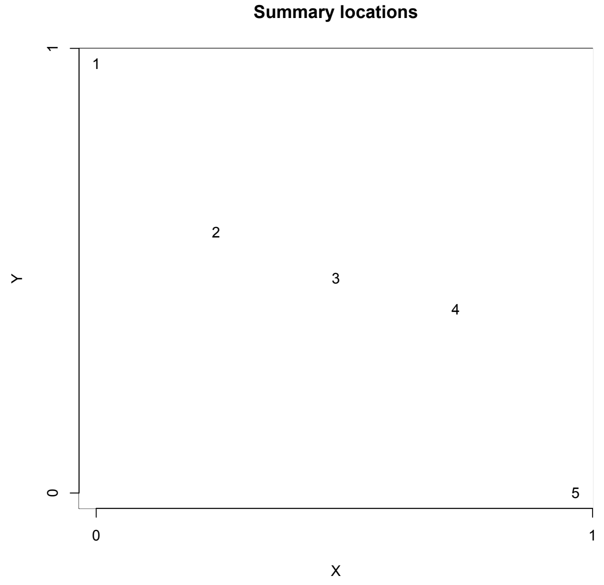

Figure 2: Locations where the variable selection and coefficient estimation of GWL were summarized.

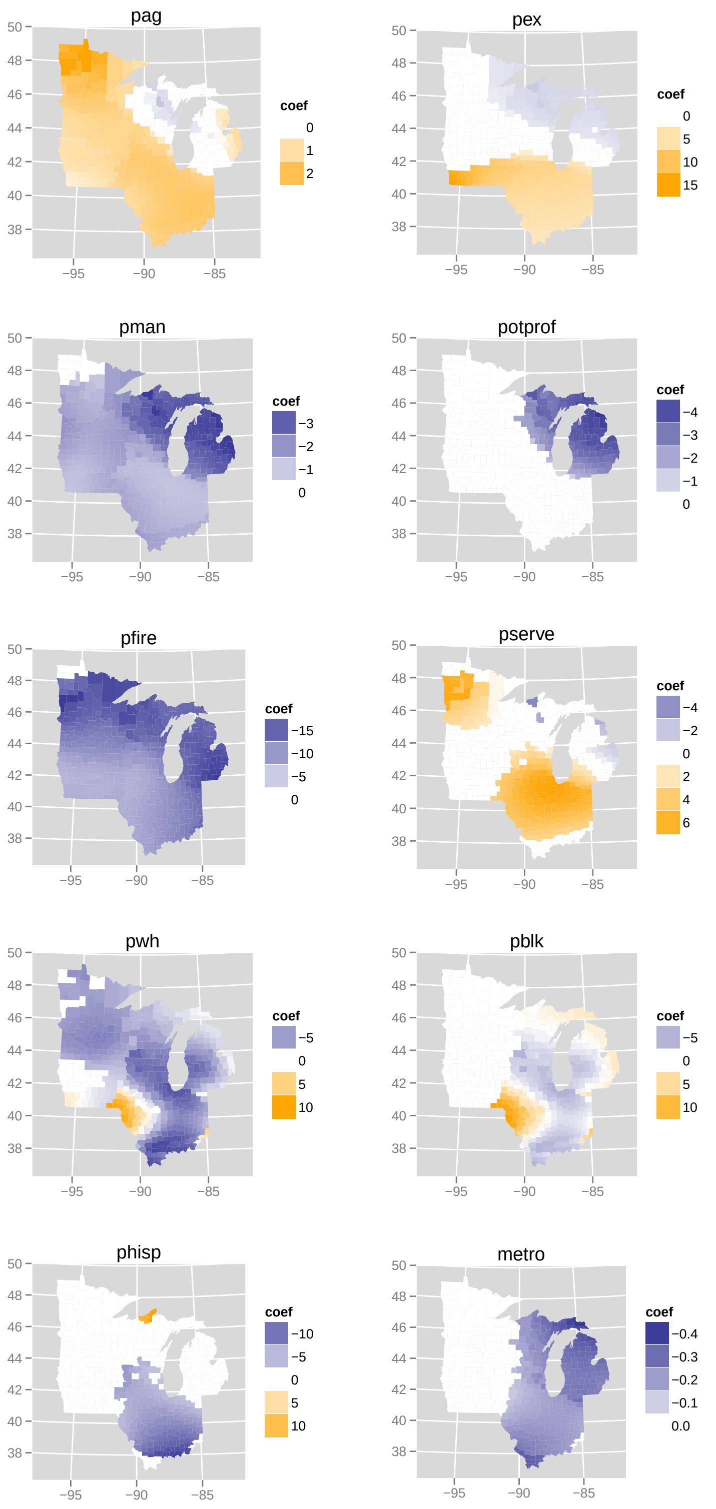

Figure 3: Coefficient surfaces for the logit of poverty rate, based on the 1970 census and estimated by the unshrunk adaptive elastic net.

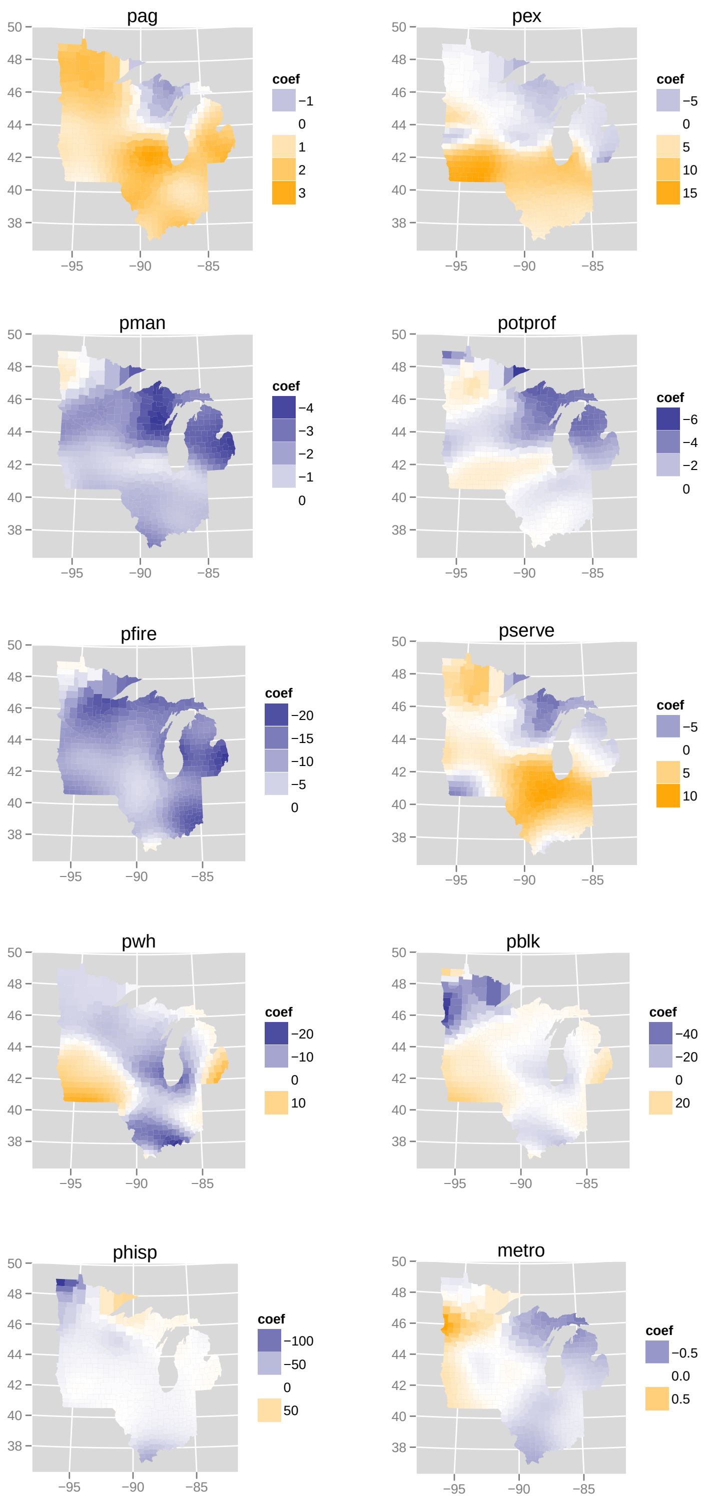

Figure 4: Coefficient surfaces for the logit of poverty rate based on the 1970 census and estimated by base GWR.

Setting	function	ρ	σ^2
1	step	0	0.25
2	step	0	1
3	step	0.5	0.25
4	step	0.5	1
5	gradient	0	0.25
6	gradient	0	1
7	gradient	0.5	0.25
8	gradient	0.5	1
9	parabola	0	0.25
10	parabola	0	1
11	parabola	0.5	0.25
12	parabola	0.5	1

Table 1: Simulation parameters for each setting.

	step						gradient						parabola					
location	`lars`		`enet`		`glmnet`		`lars`		`enet`		`glmnet`		`lars`		`enet`		`glmnet`	
	β_1	β_4 - β_5	β_1	β_4 - β_5	β_1	β_4 - β_5	β_1	β_4 - β_5	β_1	β_4 - β_5	β_1	β_4 - β_5	β_1	β_4 - β_5	β_1	β_4 - β_5	β_1	β_4 - β_5
1	0.98	0.04	1.00	0.04	1.00	0.05	1.00	0.04	1.00	0.03	1.00	0.03	0.94	0.06	0.95	0.06	0.94	0.06
	0.89	0.09	0.86	0.09	0.82	0.07	0.99	0.08	0.97	0.07	0.97	0.07	0.80	0.06	0.81	0.07	0.80	0.06
	0.96	0.07	0.99	0.10	0.96	0.09	1.00	0.07	1.00	0.06	1.00	0.04	0.95	0.06	0.94	0.09	0.95	0.04
	0.84	0.04	0.84	0.07	0.88	0.05	0.90	0.08	0.92	0.08	0.92	0.08	0.78	0.12	0.79	0.12	0.80	0.12
2	1.00	0.07	1.00	0.07	1.00	0.07	1.00	0.10	1.00	0.08	1.00	0.07	1.00	0.09	1.00	0.08	1.00	0.08
	1.00	0.06	1.00	0.06	1.00	0.07	0.98	0.07	0.98	0.08	0.99	0.07	0.97	0.12	0.98	0.11	0.98	0.10
	1.00	0.05	1.00	0.06	1.00	0.05	1.00	0.07	1.00	0.06	1.00	0.05	1.00	0.06	1.00	0.05	1.00	0.05
	0.99	0.03	1.00	0.07	0.99	0.04	0.98	0.06	0.99	0.08	0.99	0.05	0.94	0.08	0.94	0.10	0.94	0.08
3	0.99	0.05	0.99	0.06	0.99	0.06	1.00	0.09	1.00	0.08	1.00	0.07	1.00	0.09	1.00	0.09	1.00	0.09
	0.84	0.08	0.84	0.08	0.82	0.07	0.98	0.08	0.95	0.08	0.96	0.07	0.96	0.10	0.97	0.09	0.97	0.10
	0.96	0.05	0.97	0.08	0.92	0.04	1.00	0.07	1.00	0.06	1.00	0.04	1.00	0.08	1.00	0.07	1.00	0.07
	0.78	0.08	0.81	0.11	0.80	0.08	0.93	0.09	0.95	0.09	0.94	0.09	0.93	0.10	0.94	0.10	0.96	0.10
4	0.57	0.08	0.64	0.06	0.59	0.06	1.00	0.06	1.00	0.06	1.00	0.06	1.00	0.09	1.00	0.08	1.00	0.08
	0.48	0.07	0.48	0.07	0.49	0.07	0.98	0.07	0.95	0.07	0.93	0.06	0.93	0.07	0.92	0.08	0.94	0.08
	0.45	0.08	0.51	0.12	0.40	0.07	1.00	0.09	1.00	0.08	1.00	0.10	1.00	0.08	1.00	0.08	1.00	0.08
	0.53	0.08	0.52	0.07	0.51	0.07	0.96	0.07	0.95	0.11	0.95	0.08	0.96	0.08	0.96	0.09	0.96	0.09
5	0.04	0.03	0.03	0.03	0.03	0.03	0.92	0.05	0.93	0.05	0.94	0.04	0.93	0.10	0.93	0.09	0.92	0.10
	0.07	0.05	0.06	0.04	0.04	0.05	0.71	0.08	0.70	0.07	0.70	0.07	0.80	0.05	0.81	0.05	0.79	0.05
	0.02	0.04	0.02	0.03	0.03	0.05	0.93	0.10	0.95	0.14	0.95	0.10	0.93	0.07	0.94	0.12	0.94	0.07
	0.05	0.04	0.04	0.03	0.06	0.06	0.60	0.07	0.63	0.13	0.64	0.06	0.81	0.09	0.81	0.11	0.83	0.08

Table 2: Selection frequency at location 1