Local Variable Selection and Parameter Estimation of Spatially Varying Coefficient Regression Models

Wesley Brooks

Abstract

Researchers who analyze spatial data often wish to discern how a certain response variable is related to a set of covariates. When it is believed that the effect of a given covariate may be different at different locations, a spatially varying coefficient regression model, in which the effects of the covariates are allowed to vary across the spatial domain, may be appropriate. In this case, it may be the case that the covariate has a meaningful association with the response in some parts of the spatial domain but not in others. Identifying the covariates that are associated with the response at a given location is called local model selection. Geographically weighted regression, a kernel-based method for estimating the local regression coefficients in a spatially varying coefficient regression model, is considered here. A new method is introduced for local model selection and coefficient estimation in spatially varying coefficient regression models. The idea is to apply a penalty of the elastic net type to a local likelihood function, with a local elastic net tuning parameter and a global bandwidth parameter selected via information criteria. Simulations are used to evaluate the performance of the new method in model selection and coefficient estimation, and the method is applied to a real data example in spatial demography.

Preprint *March 21, 2014*

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 varying coefficient regression (SVCR) model wherein that the regression coefficients vary over space. Statistical inference for the coefficients as functions of location in an SVCR model is more complicated than estimating the coefficients in a traditional linear regression model where the coefficients are constant across the spatial domain. My research concerns the development of new methodology for the analysis of spatial data using SVCR.

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

For areal data, the spatial domain $\mathcal{D}$ is partitioned into n regions $\{D_1, \ldots, D_n\}$ such that $\mathcal{D} = \bigcup_{i=1}^{n} D_i$. In the case of areal data, the random variables $\{Y(D_i), \boldsymbol{X}(D_i)\}$ are defined for regions instead of for point locations; population and spatial mean temperature are examples of areal data. The analytical method described herein can be applied to areal data if they are recast as geostatistical data by assuming that the data are point-referenced to a representative location of each region,

such as the centroid. That is, $\{\boldsymbol{X}(\boldsymbol{s}_i), Y(\boldsymbol{s}_i)\}$ where $\boldsymbol{s}_i$ is the centroid of D_i for $i = 1, \ldots, n$.

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 a vector of covariates $\boldsymbol{X}(\boldsymbol{s})$, and a vector of regression coefficients $\boldsymbol{\beta}$. The random error $\varepsilon(\boldsymbol{s})$ denotes white noise 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 pre-specify 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

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

where σ^2 is a variance parameter, ϕ is a range parameter, and $\delta(\boldsymbol{s}, \boldsymbol{t})$ is the Euclidean distance between locations $\boldsymbol{s}$ and $\boldsymbol{t}$. The general form of a covariance function in the Matérn class is

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

where ν denotes the degree of smoothness, K_ν denotes the modified Bessel equation of the second

kind, and as before ϕ denotes a range parameter and $\delta(\boldsymbol{s}, \boldsymbol{t})$ 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 random field is said to be stationary if the joint distribution of a the response at a finite set of locations does not change when the set of locations are all shifted in space by a fixed spatial lag. That is, letting $\{T(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ 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 $T(\boldsymbol{s})$ and $\boldsymbol{h}$ is a fixed spatial lag. The random field $\{T(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ is second-order stationary if the following are satisfied:

$$
\begin{aligned}
E\left\{T(\boldsymbol{s})\right\} &= \mu \text{ for all } \boldsymbol{s} \in \mathcal{D} \\
\operatorname{var}\left\{T(\boldsymbol{s})\right\} &= \sigma^2 < \infty \text{ for all } \boldsymbol{s} \in \mathcal{D} \\
\operatorname{cov}\left\{T(\boldsymbol{s}), T(\boldsymbol{s} + \boldsymbol{h})\right\} &= C(\boldsymbol{h}) \qquad (4)
\end{aligned}
$$

where the function $C(\cdot)$ depends only on the spatial lag $\boldsymbol{h}$ and not on the location $\boldsymbol{s}$.

The coefficient vector $\boldsymbol{\beta}$ in (1) is a fixed constant. The model can be made more flexible if the coefficients are described by a stationary random field. Such a model is written

$$
Y(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})'\boldsymbol{\beta}(\boldsymbol{s}) + \varepsilon(\boldsymbol{s}) \qquad (5)
$$

where $\boldsymbol{\beta}(\boldsymbol{s})$ is a random coefficient field with a Matérn-class covariance function and the spatial random effect $W(\boldsymbol{s})$ included in the intercept $\beta_0(\boldsymbol{s})$. The random coefficient field $\boldsymbol{\beta}(\boldsymbol{s})$ can be estimated by Markov Chain Monte Carlo (MCMC) methods under the assumption that $\boldsymbol{\beta}(\boldsymbol{s})$ is stationary (Gelfand et al., 2003).

Alternatively, kernel-based and spline-based methods can be considered for fitting VCR models without assuming the coefficients are described by a stationary random field.

Coefficients for a spline-based VCR model are estimated by maximizing a penalized global likelihood, with the penalty calculated from the wiggliness of the coefficient surface (Wood, 2006). This contrasts to kernel-based estimates of the coefficients in a VCR model, which maximize a local likelihood to estimate the local coefficients at each sampling location (Loader, 1999). 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 spatial domain, while local selection means including or excluding variables at individual locations within the spatial domain. For global model selection in spline-based VCR models, Wang et al. (2008) proposed 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.

Wavelet methods for fitting SVCR models were explored by Shang (2011) and Zhang and Clayton (2011). Sparsity in the wavelet coefficients is achieved either by ℓ_1-penalization (also known as the Lasso (Tibshirani, 1996)) (Shang, 2011) or by Bayesian variable selection (Zhang and Clayton, 2011). Sparsity in the wavelet domain does not imply sparsity in the covariates, though, so neither method is suitable for local variable selection.

Geographically weighted regression (GWR) is a kernel-based method for estimating the coefficients of an SVCR model where the kernel weights are based on the distance between sampling locations

(Brundson et al., 1998; Fotheringham et al., 2002). At each sampling location, traditional GWR estimates the local regression coefficients by the local likelihood (Loader, 1999). As a kernel-based smoother for regression coefficients, traditional GWR tends to exhibit bias near the boundary of the region being modeled (Hastie and Loader, 1993). One way to reduce the boundary-effect bias is to model the coefficient surface as locally linear rather than locally constant by including coefficient-by-location interactions (Wang et al., 2008).

Traditional GWR relies on *a priori* global model selection to decide which variables should be included in the model. The idea of using Lasso regularization for local variable selection in a GWR model appears in the literature as the geographically weighted Lasso (GWL) (Wheeler, 2009). The GWL applies the Lasso for local variable selection and uses a jackknife criterion for selection of the Lasso tuning parameters. Because the jackknife criterion can only be computed at sampling locations where the response variable is observed, the GWL cannot be used to impute missing values of the response variable nor to interpolate the coefficient surface and/or the response variable between sampling locations.

Lasso regularization for model selection, while popular, can leave relevant covariates out of the model when they are correlated with other covariates, and the predictive performance of the Lasso may be dominated in such a case by ridge regression, which does not allow for local model selection (Tibshirani, 1996). The elastic net is a regularization method that combines a ℓ_1 (Lasso) and a ℓ_2 (ridge) penalty on the estimated coefficients, overcoming these drawbacks of the Lasso (Zou and Hastie, 2005).

Additionally, Lasso regularization does not generally produce consistent estimates of the relevant covariates (Leng et al., 2006). The adaptive Lasso (AL) (Zou, 2006) is an improvement to the Lasso

that does produce consistent estimates of the coefficients and has been shown to have appealing properties for automating variable selection, which under suitable conditions include the "oracle" property of asymptotically selecting exactly the correct set of covariates for inclusion in a regression model.

Combining these improvements to the Lasso, the adaptive elastic net (AEN) achieves an oracle property and performs better than other oracle-like methods when there is collinearity in the covariates (Zou and Zhang, 2009).

This document introduces a new regularization method, called the geographically weighted elastic net (GWEN), for local variable selection in GWR models. Model selection under the GWEN uses the AEN. A penalized-likelihood criterion is used to select the local GWEN tuning parameters, which means that a GWEN can be fit at any location within the domain, whether or not data were observed at that location. The particular information criterion used to select the GWEN tuning parameters is a type of local BIC, but in principle another information criterion like the AIC is also possible. The local BIC presented here is based on the local likelihood (Loader, 1999).

The remainder of this document is organized as follows. The traditional GWR is presented in Section 2. The new GWEN is introduced in section 3. In Section 4, a simulation study is conducted to assess the performance of the GWEN in variable selection and coefficient estimation. An application of the GWEN to real data is presented in Section 5. Planned future improvements to the GWEN are discussed in Section 6.

2. Geographically weighted regression

2.1. Model

Consider n data observations, taken 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 error at location $\boldsymbol{s}_i$. That is,

$$y(\boldsymbol{s}_i) = \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta}(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i). \tag{6}$$

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

$$\varepsilon(\boldsymbol{s}_i) \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right). \tag{7}$$

In order to simplify the notation, let $\boldsymbol{x}(\boldsymbol{s}_i) \equiv \boldsymbol{x}_i \equiv (1, x_{i1}, \ldots, x_{ip})'$, $\boldsymbol{\beta}(\boldsymbol{s}_i) \equiv \boldsymbol{\beta}_i \equiv (\beta_{i0}, \beta_{i1}, \ldots, \beta_{ip})'$, and $y(\boldsymbol{s}_i) \equiv y_i$. Equations (6) and (7) can now be rewritten as

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_i + \varepsilon_i \text{ and } \varepsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2\right). \tag{8}$$

Further, let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ and $\boldsymbol{y} = (y_1, \ldots, y_n)'$. Thus, conditional on the design matrix $\boldsymbol{X}$, observations of the response variable at different locations are independent of each other.

Traditional GWR estimates the regression coefficients as locally constant, as in the class of Nadaraya-Watson kernel smoothers (Härdle, 1990). As such, GWR suffers the problem of biased estimation that is common to that class of models, particularly where there is a gradient to the coefficient

surface at the boundary of the domain (Hastie and Loader, 1993).

In the context of nonparametric regression, the boundary-effect bias can be addressed by local polynomial modeling, usually in the form of a locally linear model (Fan and Gijbels, 1996). Locally linear coefficient estimation was proposed for the traditional GWR to counter the boundary effect by Wang et al. (2008).

Here, locally linear coefficients are estimated by augmenting the local design matrix with covariate-by-location interactions in two dimensions:

where $Z_i = (z_1 \cdots z_n)'$ is the augmented local design design matrix at location $\boldsymbol{s}_i$, consisting of the original design matrix X, augmented with covariate-by-location interactions. The augmented local design matrix is

$$\boldsymbol{Z}_i = (X_i \;\; L_i X_i \;\; M_i X_i) \tag{9}$$

where $L_i = \text{diag}\{s_{i',x} - s_{i,x}\}$ and $M_i = \text{diag}\{s_{i',y} - s_{i,y}\}$ for $i' = 1, \ldots, n$.

2.2. Estimation

The total log-likelihood of the observed data is the sum of the log-likelihood of each individual observation:

$$\ell(\boldsymbol{\beta}_i) = -(1/2) \sum_{i'=1}^{n} \left\{ \log \sigma_i^2 + \sigma_i^{-2} \left(y_{i'} - \boldsymbol{z}'_{i'} \boldsymbol{\beta}_i \right)^2 \right\}. \tag{10}$$

Since there are a total of $n \times 3(p+1)$ free parameters for n observations, the model is not identifiable and it is not possible to directly maximize the total likelihood.

The values of the local coefficients $\boldsymbol{\beta}_i$ are estimated at $\boldsymbol{s}_i$ by the weighted likelihood

$$\mathcal{L}_i\left(\boldsymbol{\beta}_i\right) = \prod_{i'=1}^{n}\left[\left(2\pi\sigma_i^2\right)^{-1/2}\exp\left\{-1/2\sigma_i^{-2}\left(y_{i'}-\boldsymbol{z}_{i'}'\boldsymbol{\beta}_i\right)^2\right\}\right]^{w_{ii'}}, \tag{11}$$

where the weights are calculated by a kernel function $K_h(\cdot)$ such as the Epanechnikov kernel:

$$\begin{gathered} w_{ii'} = K_h(\delta_{ii'}) = h^{-2}K\left(h^{-1}\delta_{ii'}\right) \\ K(x) = \begin{cases} (3/4)(1-x^2) & \text{if } \delta_{ii'} < h, \\ 0 & \text{if } \delta_{ii'} \geqslant h. \end{cases} \end{gathered} \tag{12}$$

Thus, the local log-likelihood function is, up to an additive constant:

$$\ell_i\left(\boldsymbol{\beta}_i\right) = -(1/2)\sum_{i'=1}^{n} w_{ii'}\left\{\log\sigma_i^2 + \sigma_i^{-2}\left(y_{i'}-\boldsymbol{z}_{i'}'\boldsymbol{\beta}_i\right)^2\right\}. \tag{13}$$

From (13), the maximum local likelihood estimate $\hat{\sigma}_i^2$ is:

$$\hat{\sigma}_i^2 = \left(\sum_{i'=1}^{n} w_{ii'}\right)^{-1}\sum_{i'=1}^{n} w_{ii'}\left(y_{i'}-\boldsymbol{z}_{i'}'\hat{\boldsymbol{\beta}}_i\right)^2 \tag{14}$$

3. Local variable selection and parameter estimation

3.1. Local variable selection

The AGL is explored as a penalty function for local variable selection in SVCR models.

The proposed local variable selection with AGL penalty is an ℓ_1 regularization method for variable selection in regression models (Wang and Leng, 2008; Zou, 2006). The adaptive group lasso selects groups of covariates for inclusion or exclusion in the model.

3.1.1. Local variable selection and coefficient estimation with the adaptive Lasso

The objective function for the GWAL at $\boldsymbol{s}_i$ consists of the local log-likelihood and an additive penalty that is the weighted ℓ_1-norm of the coefficients, defined to be

$$\begin{aligned}\mathcal{S}(\boldsymbol{\beta}_i) &= -2\ell_i\left(\boldsymbol{\beta}_i\right) + \mathcal{J}_1(\boldsymbol{\beta}_i) \\ &= \sum_{i'=1}^{n} w_{ii'}\left\{\log\sigma_i^2 + \sigma_i^{-2}\left(y_{i'} - \boldsymbol{z}_{i'}'\boldsymbol{\beta}_i\right)^2\right\} + \lambda_i\sum_{j=1}^{p}\|\beta_{ij}\|/\gamma_{ij} \end{aligned} \tag{15}$$

where $\sum_{i'=1}^{n} w_{ii'}\left(y_{i'} - \boldsymbol{z}_{i'}'\boldsymbol{\beta}_i\right)^2$ is the weighted sum of squares minimized by traditional GWR, and $\mathcal{J}_1(\boldsymbol{\beta}_i) = \lambda_i\sum_{j=1}^{p}\|\beta_{ij}\|/\gamma_{ij}$ is the AGL penalty. With the vector of unpenalized local coefficients $\boldsymbol{\gamma}_i$, the AL penalty for the jth group of coefficients β_{ij} at location $\boldsymbol{s}_i$ is λ_i/γ_{ij}, where $\lambda_i > 0$ is a the local tuning parameter that applies to all coefficients at location $\boldsymbol{s}_i$ and $\boldsymbol{\gamma}_i = (\gamma_{i1}, \ldots, \gamma_{ip})'$ is the vector of adaptive weights at location $\boldsymbol{s}_i$.

3.2. Tuning parameter selection

A local tuning parameter λ_i is required for the variable selection step of fitting each local model by the GWAL method. The corrected AIC is used to select λ_i (Hurvich et al., 1998):

$$\begin{aligned}\mathrm{AIC}_{\mathrm{c},i} &= -2\sum_{i'=1}^{n}\ell_{ii'} + \log\left(\sum_{i'=1}^{n} w_{ii'}\right)\mathrm{df}_i \\ &= -2\times\sum_{i'=1}^{n}\log\left[\left(2\pi\hat{\sigma}_i^2\right)^{-1/2}\exp\left\{-\frac{1}{2}\hat{\sigma}_i^{-2}\left(y_{i'} - \boldsymbol{x}_{i'}'\hat{\boldsymbol{\beta}}_{i'}\right)^2\right\}\right]^{w_{ii'}} + \log\left(\sum_{i'=1}^{n} w_{ii'}\right)\mathrm{df}_i \\ &= \sum_{i'=1}^{n} w_{ii'}\left\{\log\left(2\pi\right) + \log\hat{\sigma}_i^2 + \hat{\sigma}_i^{-2}\left(y_{i'} - \boldsymbol{x}_{i'}'\hat{\boldsymbol{\beta}}_{i'}\right)^2\right\} + \log\left(\sum_{i'=1}^{n} w_{ii'}\right)\mathrm{df}_i \end{aligned} \tag{16}$$

The local BIC 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). Since the estimated variance

$\hat{\sigma}_i^2$ is the variance estimate from the unpenalized local model, its value does not depend on the choice of tuning parameter; it is constant in (16) (Zou et al., 2007).

For the geographically weighted Lasso (GWL), Wheeler (2009) proposed selecting the local Lasso tuning parameters for local selection in a SVCR model at location $\boldsymbol{s}_i$ to minimize the local jackknife prediction error $|y_i - \hat{y}_i^{(i)}|$. Because the jackknife prediction error is undefined everywhere except for at the sampling locations, this choice restricts coefficient estimation to occur at the locations where data has been observed. By contrast, the local BIC can be calculated at any location where the local log-likelihood can be obtained. As a practical matter this allows for variable selection and coefficient surface estimation to be done at locations where no data are observed and for imputation of missing values of the response variable.

3.3. Bandwidth parameter estimation

The bandwidth parameter is estimated to minimize an information criterion. It is common in nonparametric regression to select the bandwdth to minimize a corrected AIC where the degrees of freedom are given by the trace of a smoothing matrix (Hurvich et al., 1998). Because ℓ_1 penalization procedures like the AGL are not linear smoothers, there is no smoothing matrix for the AGL. There is need for a procedure to estimate the degrees of freedom of a nonparametric AGL model.

The degrees of freedom used to estimate the $\hat{y}_i$ are estimated by $\mathrm{df}_i w_{ii} / \sum_{j=1}^{n} w_{ij}$.

4. Simulation study

4.1. Simulation setup

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

Data were simulated on the domain $[0,1]^2$, 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 covariance function $\text{Cov}\left(X_{ji}, X_{ji'}\right) = \sigma_x^2 \exp\left(-\tau_x^{-1}\delta_{ii'}\right)$ where $\sigma_x^2 = 1$ is the variance, $\tau_x = 0.1$ 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 matrix $\boldsymbol{X} = (X_1 \cdots X_5)$ by $\boldsymbol{R}$, where $\boldsymbol{R}$ is the Cholesky decomposition of the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{R}'\boldsymbol{R}$. 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 was $y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_i + \varepsilon_i$ for $i = 1, \ldots, n$ where $n = 900$ and the ε_i's were iid Gaussian with mean zero and variance σ_ε^2. The simulated data included the response y and five covariates $x_1, \ldots, x_5$. The true data-generating model uses only x_1. The variables $x_2, \ldots, x_5$ are included to assess performance in model selection.

There were twelve simulation settings, each of which was simulated 100 times. For each of the twelve settings, $\beta_1(\boldsymbol{s})$, the true coefficient surface for x_1, was nonzero in at least part of the domain, with a minimum of zero and maximum of one. Three parameters were varied to produce the twelve settings: there were three functional forms for the coefficient surface $\beta_1(\boldsymbol{s})$, 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 error term. The twelve simulation settings are described in Table 1.

The three coefficient surfaces used to produce the response variable in the simulations are pictured in Figure 1. The first is a "step" function, which is equal to zero in 40% of the spatial domain, equal to one in a different 40% of the spatial domain, and increases linearly in the middle 20% of

the domain. The second is a gradient function, which increases linearly from zero at one end of the domain to one at the other. The final coefficient function is a parabola taking its maximum value of 1 at the center of the domain and falling to zero at each corner of the domain.

The performance of the GWEN, GWAL, GWEN-LLE, and GWAL-LLE were compared to that of the traditional GWR algorithm of Fotheringham et al. (2002) and that of "oracular" GWR, which is traditional GWR with oracular variable selection and locally linear fitting as described in Section **??**. Oracular selection means that exactly the correct set of covariates was used to fit the GWR model at each location in the simulation. The implementation of the AEN uses coordinate descent via the `R` package `glmnet` (Friedman et al., 2010).

Results of the simulation were summarized at five locations on the domain (Figure 2). Due to edge effects, we expect biased estimation at locations one and five (which are at opposite corners of the domain) from traditional GWR, particularly when the coefficient surface has nonzero gradient at the boundary (which is the case for the gradient and parabola functions). Because the GWEN-LLE, GWAL-LLE, and oracular GWR use locally linear fitting, they are expected to exhibit less bias at the boundaries.

Locations two and four are at the 'corners' of the step function. Because the step function is undifferentiable at these locations, locally linear fitting is not expected to be as effective at reducing bias here as at the boundaries of the gradient and parabola functions.

Local variable selection is expected to be ambiguous at locations where the underlying coefficient surface transitions from zero to nonzero. In the simulations, that occurs at location four of the step function, location five of the gradient, and locations one and five of the parabola.

Unlike the other two functions, the gradient is actually constant across the domain in terms of the

covariate-by-location interaction. As a result, the optimal kernel bandwidth ϕ is expected to be larger for estimating the gradient coefficient surface than for the step or the parabola. The result should be that the estimation is more accurate for the gradient function in terms of bias, variance, and MSE.

4.2. Simulation results

Variable selection. Table **??** lists the results of variable selection. The correct covariate was usually included in the local models, and the unimportant covariates were usually excluded. Ignore for now the ambiguous locations where the true β_1 surface transitions from zero to nonzero. Of the eighty simulated cases where $\beta_1(\boldsymbol{s})$ is unambiguously nonzero, more than half (52) saw no false negatives (over 100 simulations). The number with no false negatives and no false positives (i.e. exactly the correct model was recovered in all 100 simulations) was 26. Of the 120 total simulated cases, 72 had no false positives (i.e. no variable whose true coefficient is zero was included in the model during any of the 100 simulation runs).

Selection performance worsens at the increase of the noise variance from $\sigma_\varepsilon^2 = 0.25$ to $\sigma_\varepsilon^2 = 1$ and at the increase in colinearity from $\rho = 0$ to $\rho = 0.5$. Of the 44 cases where model selection recovered exactly the correct model in all 100 runs of the simulation, only five arose from cases where $\sigma_\varepsilon^2 = 1$, while 19 arose from cases where $\rho = 0.5$. The worst error rates that were observed in these unambiguous cases were a false positive rate of 6% (location one of the step function with $\sigma_\varepsilon^2 = 1$, $\rho = 0.5$, and selection via the elastic net) and a false negative rate of 16% (location three of the step function with $\sigma_\varepsilon^2 = 1$, $\rho = 0.5$, and selection via the lasso).

Model selection was ambiguous at locations where the true $\beta_1(\boldsymbol{s})$ transitions from zero to nonzero. At location four of the step function, the selection rate of β_1 ranged from 43% to 60% among the

different simulation settings. At location five of the gradient, the range of selection rates was 63% to 82%, and the selection rate across locations one and five of the parabola ranged from 27% to 66%.

There is no indication that the GWEN performed better in selection than the GWAL, even in cases where the covariates were moderately correlated ($\rho = 0.5$).

Coefficient estimation. The mean squared error, bias, and variance of $\hat{\beta}_1$ (MSE(β_1), bias($\hat{\beta}_1$), var($\hat{\beta}_1$)) are listed in Tables **??**, **??**, and **??**, respectively. The method of oracular selection led to the best MSE($\hat{\beta}_1$) in 41 of the 60 cases.

In terms of MSE($\hat{\beta}_1$), while oracular selection clearly was the most accurate estimation method in most cases, the difference in accuracy between the estimation methods was modest in most cases. There were a few cases when the difference in MSE($\hat{\beta}_1$) between estimation methods amounted to at least an order of magnitude. At locations one and five of the parabola, oracular selection produces much more accurate estimation than GWEN, GWAL, or GWR because locations one and five are on the domain boundary where the parabola has a strong gradient, and those methods don't use locally linear fits to account for the boundary effect. This can also be seen from the fact that the bias($\boldsymbol{\beta}_1$) of GWEN, GWAL, and GWR is large at locations one and five of the parabola, where it is nearly zero for GWEN-LLE, GWAL-LLE, and oracular GWR.

A similar boundary effect is apparent at location five of the gradient, where GWEN, GWAL, and GWR produce a bias($\hat{\beta}_1$) and MSE($\hat{\beta}_1$) that are an order of magnitude or more greater than those of GWEN-LLE, GWAL-LLE, and oracular GWR (the differences in var($\hat{\beta}_1$) are smaller).

At location one of the step function, the MSE($\hat{\beta}_1$) and var($\hat{\beta}_1$) for GWR are much smaller than

for the other estimation methods, including oracular GWR, while the bias($\hat{\beta}_1$) doesn't vary much between estimation methods. It is not clear why this is the case.

As was the case for selection, accuracy in coefficient estimation seemed to suffer at the increase of the noise variance or the increase of correlation in the covariates. Once again, this effect is probably most apparent at location three of the step function.

Fitted Values. The MSE of the $\hat{y}$, MSE($\hat{y}$), is listed in Table **??**. Nominally, MSE($\hat{y}$) should be equal to the noise variance, σ^2_ε, which is 1 for odd-numbered rows and 0.25 for even numbered rows. Of the 60 simulation cases, GWR's MSE($\hat{y}$) was nearest to the known noise variance for 16, compared to 15 for oracular GWR, 14 for the GWAL-LLE, nine for each of the GWEN and the GWAL, and seven for the GWAL-LLE.

5. Data example

An example data analysis is presented, demonstrating application of the GWEN-LLE for local model selection in a SVCR model of the poverty rate in the Upper Midwest states of the U.S. (Minnesota, Iowa, Wisconsin, Illinois, Indiana, and Michigan). The GWEN-LLE model is compared to a traditional GWR model for the same data.

The response variable for the models is the logit-transformed proportion of individuals in each county who were living below the poverty line in 1970. The covariates are related to the employment structure of each county, namely the proportion of each county's residents who were employed in the economic sectors of agriculture; finance, insurance, and real estate; manufacturing; mining; services; and other professions. While the response variable was logit transformed, raw proportions were used as the covariates.

The data used in this example are aggregated at the level of counties, which are areal units. Each county's sampling location is assumed to be its centroid. The data are from the U.S. Census Bureau's decennial census in the year 1970. The county-level poverty rate for the Upper Midwest from the 1970 census is plotted in Figure 3, and the covariates are plotted in Figure 4.

The coefficient estimates are plotted on maps of the upper midwest in Figure 5 (based on the GWEN-LLE) and Figure 6 (based on traditional GWR). The GWR coefficient estimates are compared to those from the GWEN-LLE in Figure 7, where a 1-1 line is included as a guide for where the points would lie if the two methods produced identical coefficient estimates.

The coefficient surfaces estimated by the GWEN-LLE are relatively constant as compared to the those estimated by GWR. The GWEN-LLE indicates that the proportion of residents employed in services or in the "other professions" category does not affect the poverty rate anywhere within the Upper Midwest states, while the proportion of residents employed in manufacturing or in finance, insurance, and real estate have negative coefficients (meaning a negative association with poverty rate) in all but one county of the Upper Midwest states (that one county is at the extreme northwest corner of Minnesota).

The coefficient of employment in finance, insurance, and real estate has a larger magnitude than the other covariates (minimum value near -20, as opposed to the next-largest-magnitude coefficient, that of manufacturing employment, with a minimum near -3), indicating that counties with many workers in this sector tended to have low poverty rates in 1970. The local coefficients for employment in finance, insurance and real estate as estimated by GWR are comparable to those estimated by the GWEN-LLE.

Manufacturing employment was also associated with reduced poverty rates across the entire Upper

Midwest in 1970. While the coefficient is smaller than for employment in the finance, insurance and real estate sector, there were many more people employed in manufacturing. In the eastern part of the Upper Midwest, many counties have about half of their workers employed in manufacturing, while there are just a few counties scattered through the Upper Midwest that approach the maximum of ten percent of workers employed in finance, insurance, and real estate (Figure 4). Like the GWEN-LLE, traditional GWR estimates that the coefficient of manufacturing employment was negative across most of the Upper Midwest, but traditional GWR estimates that the coefficient was very slightly greater than zero around Chicago and in northwestern Minnesota.

The proportion of residents employed in agriculture is estimated to have affected the poverty rate of 1970 only in the western half of Iowa. In the counties where agricultural employment was associated with the poverty rate, there was a north-south gradient to the coefficient surface, from 0.5 in the northern part of western Iowa to -1 in the southern part. In those counties, the proportion of residents employed in agriculture is between 0.25 and 0.5. In this part of western Iowa, GWR finds a very slight positive association between agricultural employment and the poverty rate. The coefficient estimated by GWR is of a greater magnitude almost everywhere else in the domain than in western Iowa.

There were relatively few people employed in mining in the Upper Midwest in 1970. In the extreme northern part of the domain, some sparsely populated counties had about 25% of residents working in mining, and in the extreme southern part of the domain, several counties had about 10% of residents employed in mining. According to the GWEN-LLE, the coefficient of mining employment is zero except in the far southern part of the domain. There, mining employment is associated with an increased poverty rate (the coefficient is about 3). Within the far southern part of the domain,

GWR and the GWEN-LLE produce similar estimates of the coefficient of mining employment, but in parts of the Upper Midwest where the GWEN-LLE says that mining employment is not associated with the poverty rate, GWR estimates large local coefficients for mining employment, ranging between 20 and -15.

6. Future work

The GWEN is presented here as a method of analysis for data where the response variable follows a Gaussian distribution with independent additive errors. A key feature of spatial data, though, is that the errors are typically autocorrelated. Additionally, it is common to encounter data that does not follow a Gaussian distribution but for which the GWEN would otherwise be a valuable tool for analysis.

Autocorrelated Errors. In order to get a sense of how the GWEN will perform when the assumption of independent errors is violated, the simulation study from Section 4 was repeated with the addition of a Matèrn-class spatial covariance structure in the noise (results not included here). Introducing autocorrelation in the errors causes a substantial degradation in the estimation accuracy of the GWAL and GWEN, accompanied by a tendency to prefer smaller kernel bandwidths. The likely explanation is that when the kernel bandwidth is small, autocorrelated errors are indistinguishable from a varying intercept. The residuals are reduced to the extent that the errors are incorporated in the intercept term. Since this effect is absent in the case of uncorrelated errors, greater autocorrelation in the error term will tend to mean a greater reduction in the errors - and therefore a greater increase in the log likelihood - as the kernel bandwidth decreases.

Since the optimal kernel bandwidth is balance between the log likelihood and the degrees of free-

dom consumed by the model, and because the effect of autocorrelated errors is an increase in log likelihood without an offsetting increase in the consumed degrees of freedom at a given kernel bandwidth, greater autocorrelation will tend to lead to a smaller optimal kernel bandwidth. One quick way to counter this effect is to increase the penalty that is added to the total log likelihood in (**??**). There is currently no rule to set the penalty based on the data, which would be necessary before using this adjustment to analyze real data.

Generalized linear models. To date, validation of the GWEN has been for Gaussian data. The extension to any exponential-family distribution involves generalizing the likelihood that is used to select AEN tuning parameters and the kernel bandwidth. Exploratory simulations of the generalized GWEN have been carried out with Poisson and binomial data.

7. References

References

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

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

Brundson, C., S. Fotheringham, and M. Charlton (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 (Revised Edition)*. Wiley, New York.

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

Fan, J. and I. Gijbels (1996). *Local Polynomial Modeling and its Applications*. Chapman and Hall, London.

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

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

Fotheringham, A., C. Brunsdon, and M. Charlton (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley, West Sussex, England.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*, 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*, 387–396.

Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Boston MA.

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

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

Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society Series B 60*, 271–293.

Leng, C., Y. Lin, and G. Wahba (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica 16*, 1273–1284.

Loader, C. (1999). *Local Regression and Likelihood*. Springer, New York.

Shang, Z. (2011). *Bayesian Variable Selection*. Ph. D. thesis, University of Wisconsin-Madison.

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

Wang, H. and C. Leng (2008). A note on adaptive group lasso. *Computational Statistics and Data Analysis 52*, 5277–5286.

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*, 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*. Chapman and Hall, Boca Raton, FL.

Zhang, J. and M. K. Clayton (2011). Functional concurrent linear regression model for images. *Journal of Agricultural, Biological, and Environmental Statistics 16*, 105–130.

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

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B 67*, 301–320.

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

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

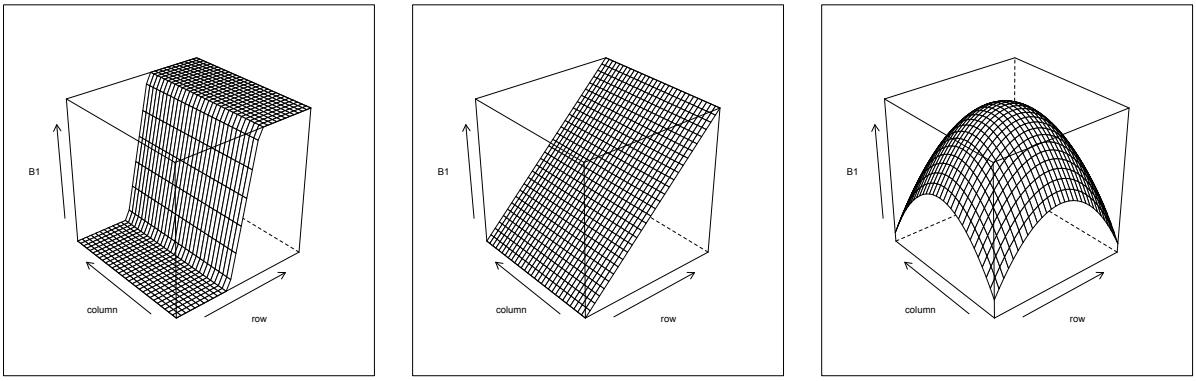

Figure 1: The actual β_1 coefficient surface used in the simulation.

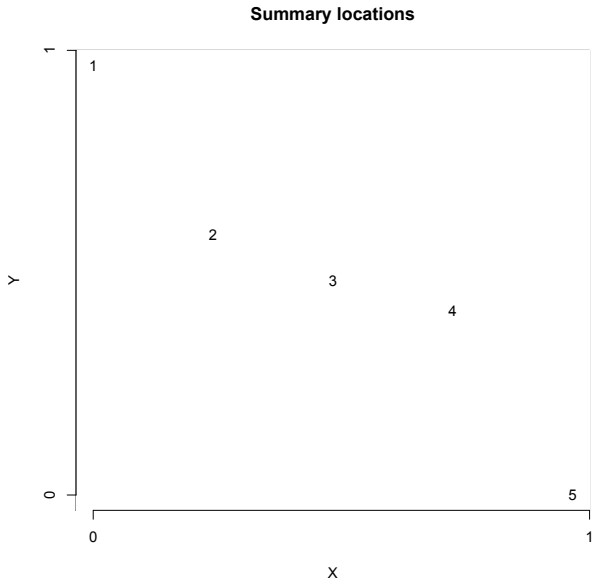

Figure 2: Locations where the variable selection and coefficient estimation of GWL were summarized.

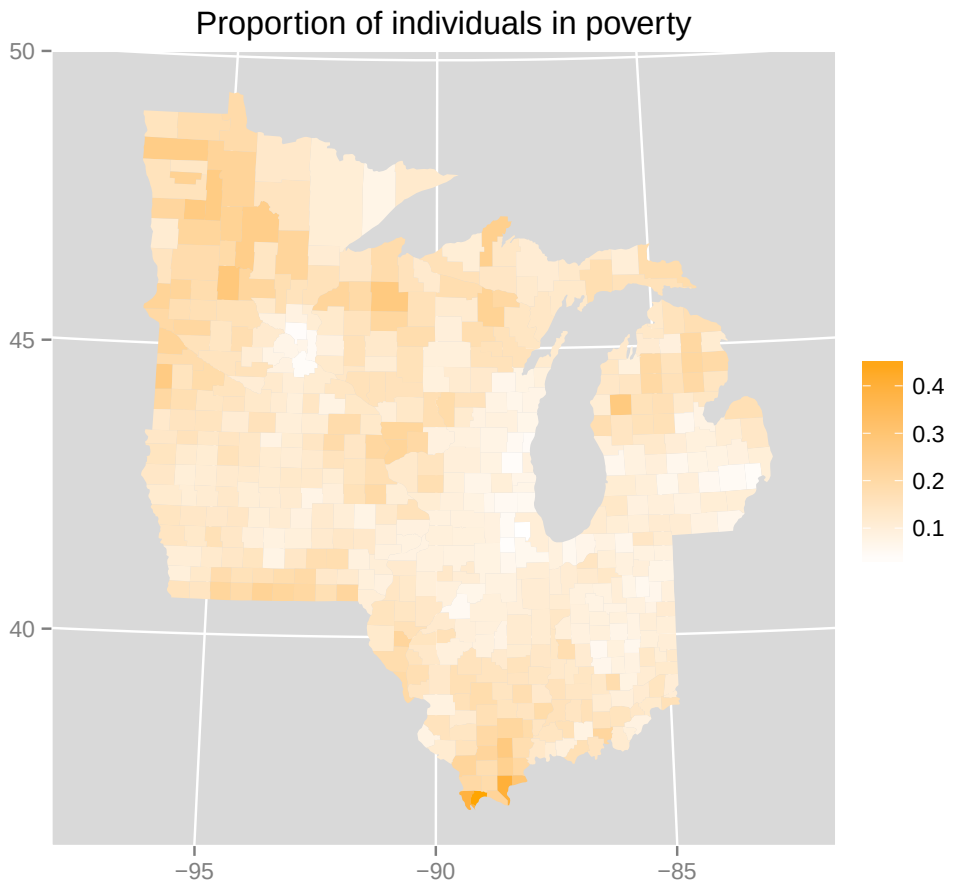

Figure 3: County-level poverty rate from the decennial U.S. census of 1970 for the Upper Midwest.

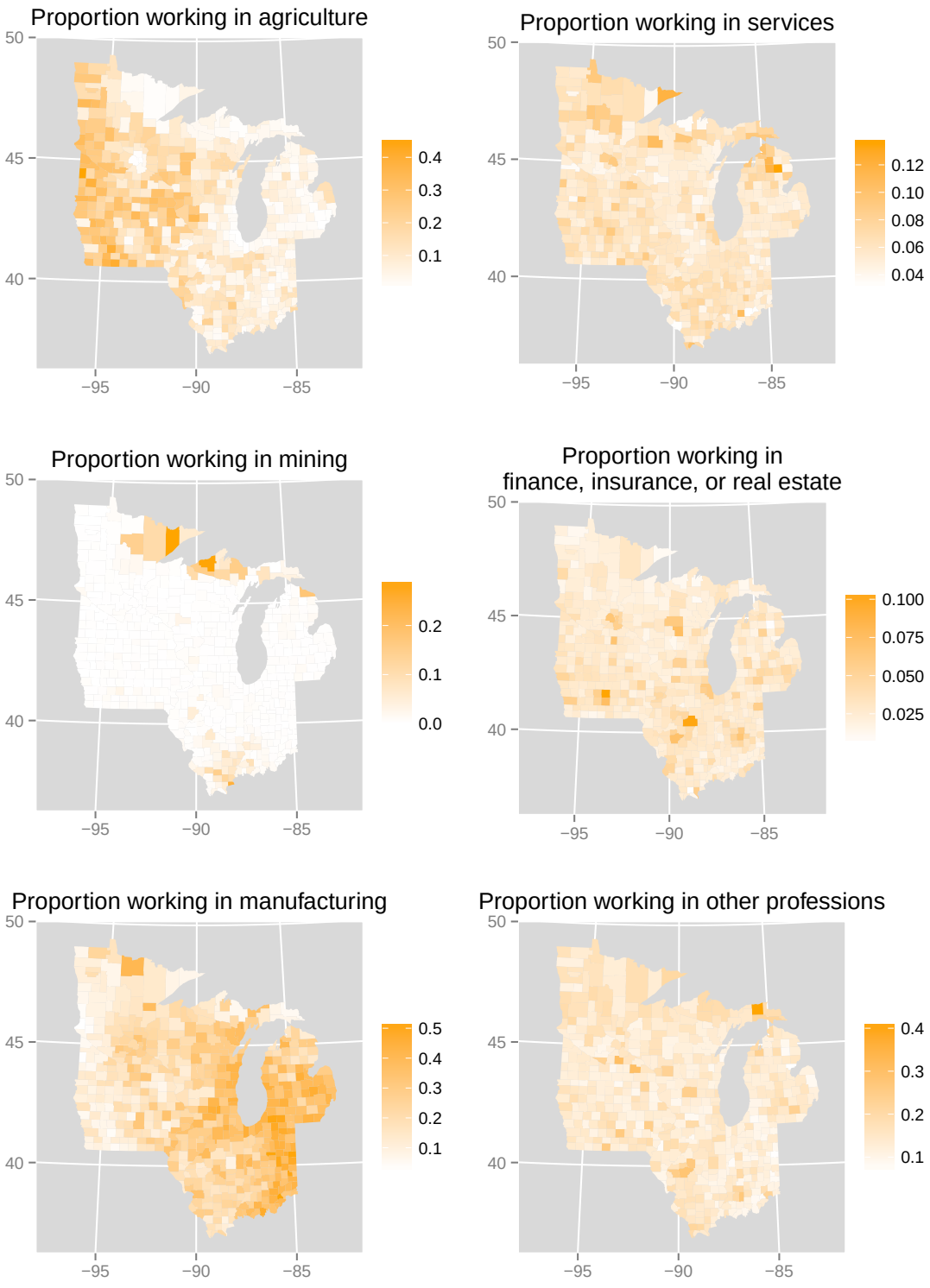

Figure 4: Covariates for the model of poverty for 1970.

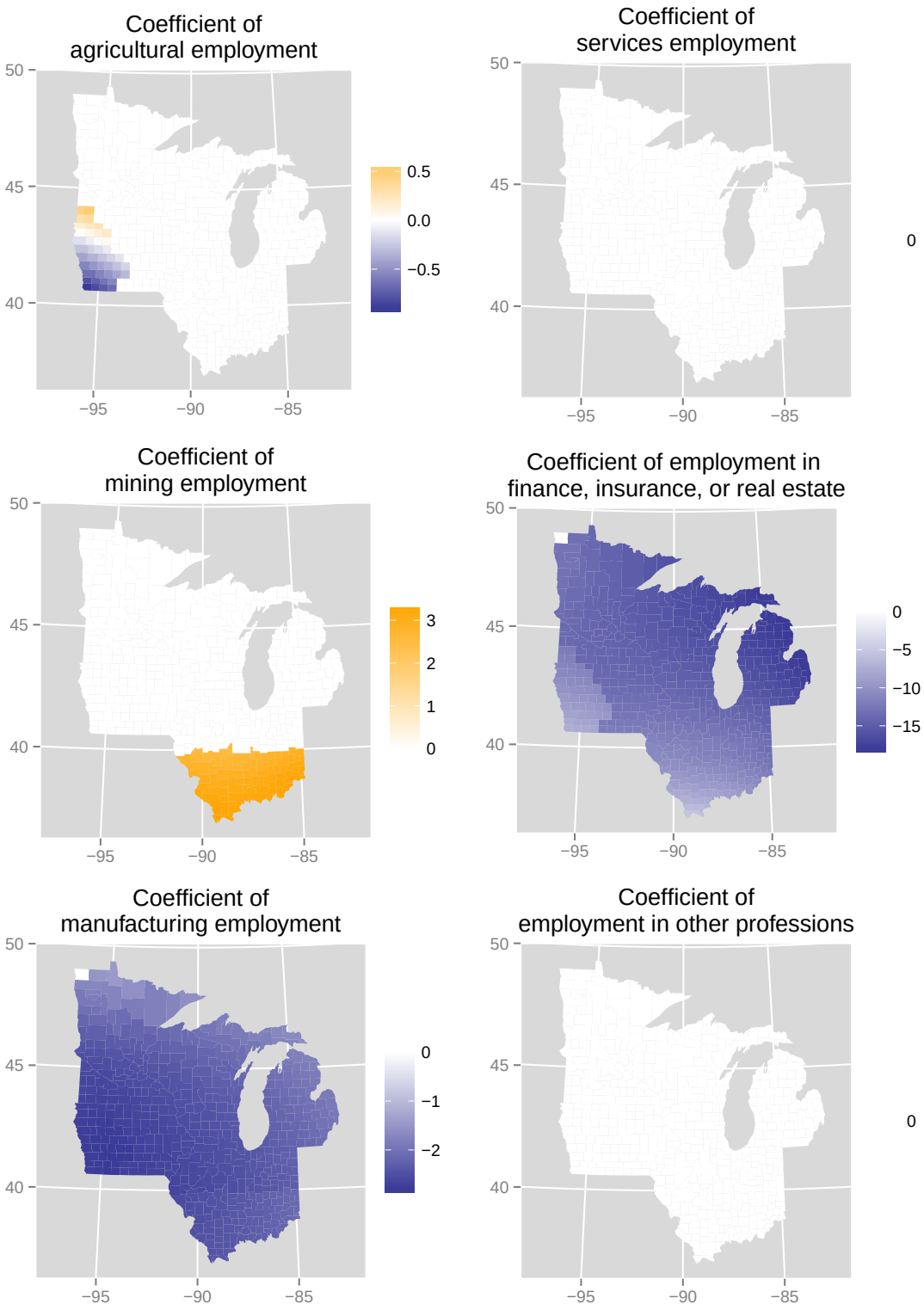

Figure 5: Local coefficient estimates for GWEN-LLE model of the logit of poverty rate, based on the 1970 census.

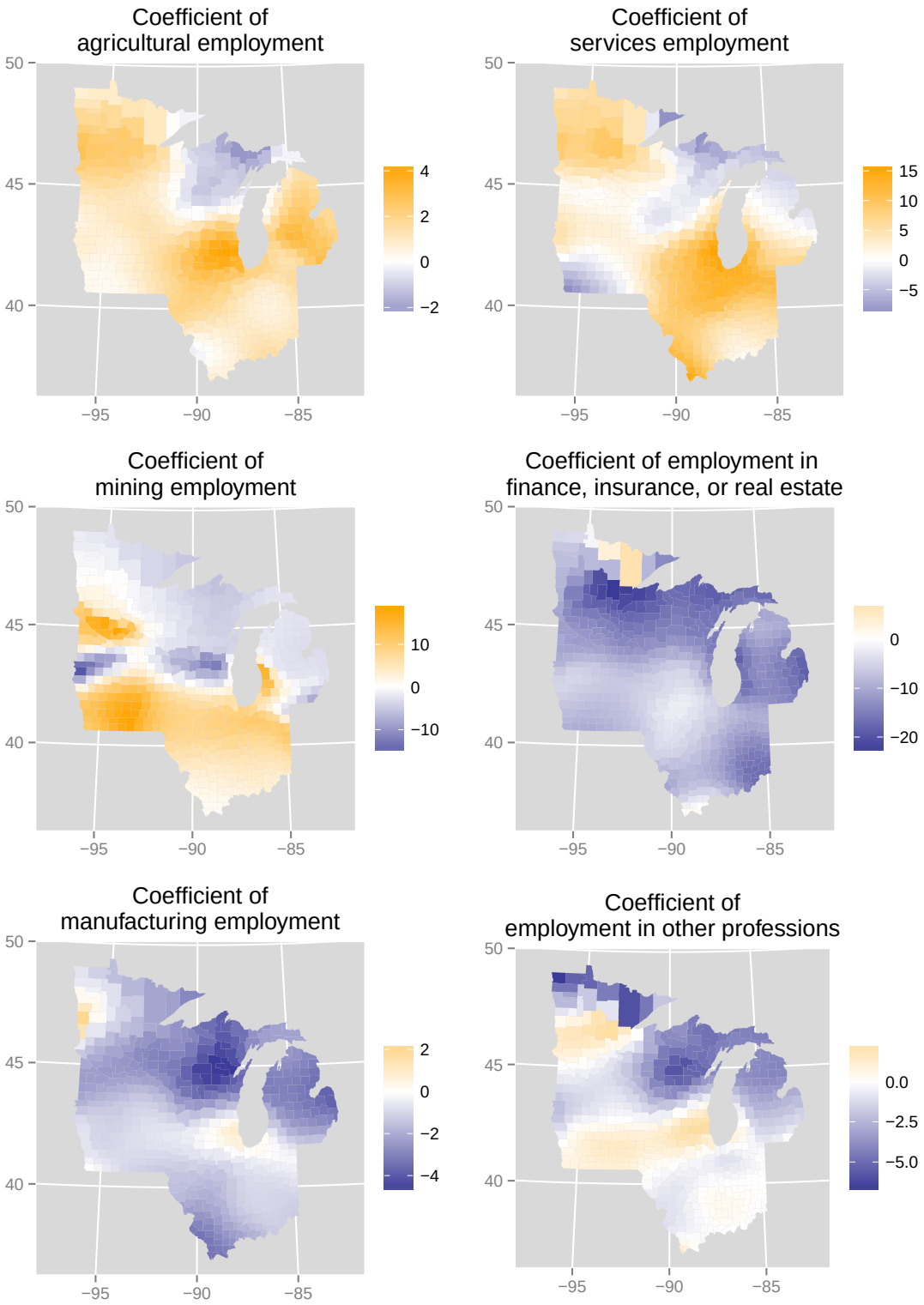

Figure 6: Local coefficient estimates for GWR model of the logit of poverty rate, based on the 1970 census.

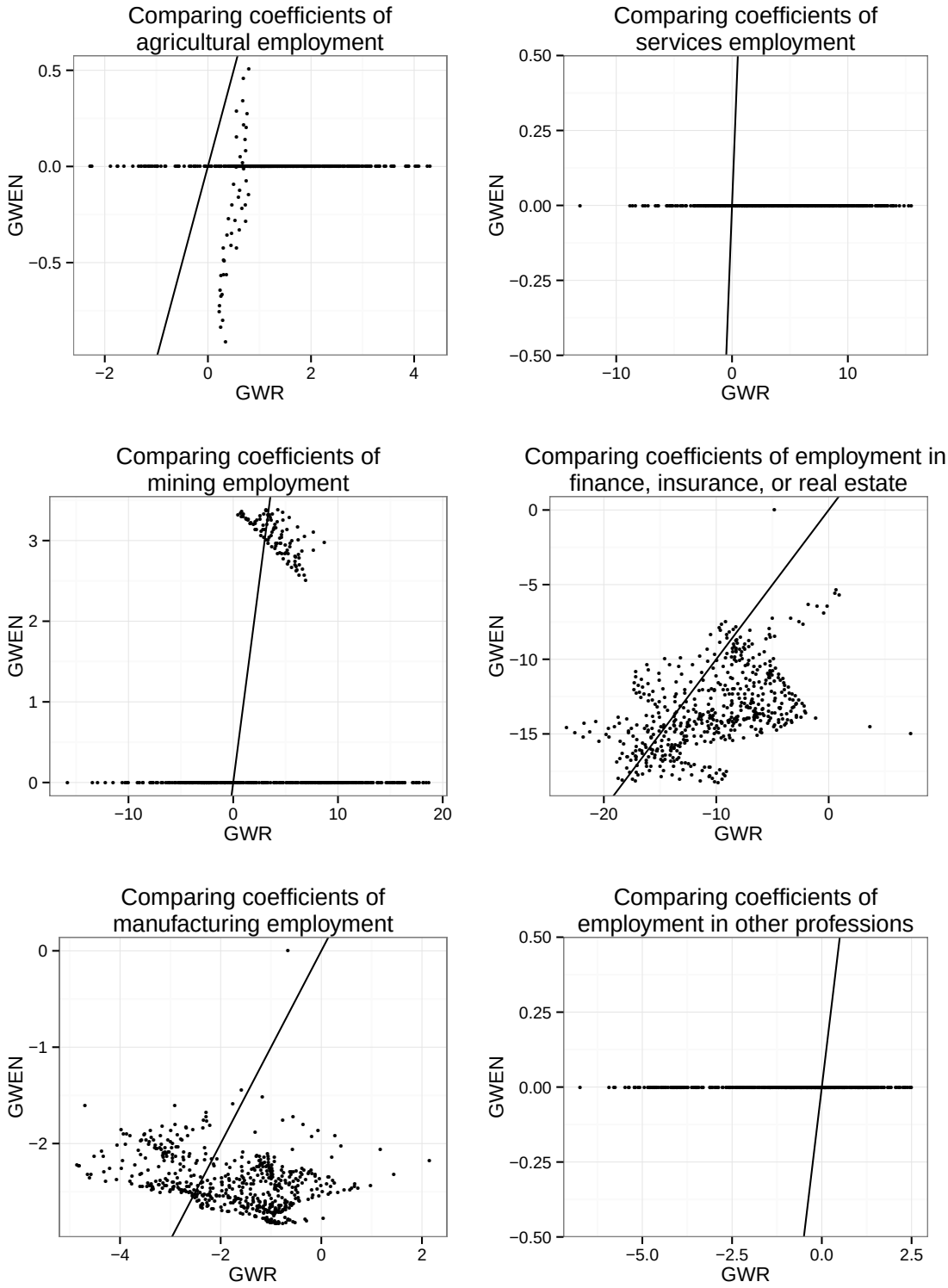

Figure 7: Comparison of the coefficients of employment structure variables for a SVCR model of poverty rate in the Upper Midwest states in 1970. Each plot includes a 1-1 line, along which the points would lie if GWR and the GWEN produced identical coefficient estimates.

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.