Local variable selection and coefficient estimation for spatially-varying-coefficients models in the context of geographically-weighted regression using regularization

Wesley Brooks

1. Introduction

Varying-coefficients regression (VCR) (Hastie and Tibshirani, 1993) is a technique used to model non-stationary regression processes. Whereas the coefficients in an ordinary least squares (OLS) regression equation (e.g. $Y = X\beta + \varepsilon$) are scalar constants, the coefficients in a VCR regression equation (e.g. $Y = X\beta(s) + \varepsilon$) are functions - often *smooth* functions - of some effect-modifying parameter s. Methods for estimating the coefficient functions of VCR are typically divided between spline-based (e.g. Wood (2006)) and kernel-based (e.g. Fan and Zhang (1999); Loader (1999)) methods.

We can divide spatial data into three types: areal data, in which observations are made on (or aggregated over) areas like counties or grid cells; geostatistical data, in which one measures the value of a continuous process at some discrete locations; and point-process data, where the locations of observations are not selected but are observed to arise from a process (e.g. data consisting of the location of each tree in a forest plot). A key concept in the analysis of spatial data is that nearby locations are more alike than distant locations. Spatial data may be viewed as Clustering is a typical form of non-stationarity in spatial data. One way to analyze clustered data is to assume a constant underlying mean with random deviations that are clustered in space. This is the structure of an autoregressive model (). If the underlying mean is not constant but is given by a regression function with constant coefficients, then a conditionally autoregressive (CAR) model () is used to simultaneously estimate regression coefficients and clustering of the residuals.

Regression models for spatial data may incorporate either or both of endogenous and exogenous variables. Endogenous modeling means using a variable as both an input and an output of the model (e.g. smoothing), while exogenous modeling uses covariates other than the output variable to build a regression model. These two are commonly combined in spatial models - for instance, by combining a systematic effect of exogenous variables with an endogenous spatial random effect. Endogenous regression models for spatial data are tuned by autoregression, while for geospatial data

Spatially-clustered residuals may indicate that the random error component arises in a spatially-clustered way but it may also indicate oversmoothing, where the small net bias of a global estimate masks larger but offsetting local bias. In this case, smoothing methods that model the mean response as a function of the covariates can reduce or eliminate clustering of the residuals. Where the underlying mean is a regression function but the coefficients are not constants, a spatial VCR model is appropriate. Both spline-based (Wood, 2006) and kernel-based (Fotheringham et al., 2002) methods are available for estimating the coefficient functions.

Geographically weighted regression (GWR) (Fotheringham et al., 2002) is a kernel-based method of estimating the coefficients of a VCR model for spatial data. GWR uses kernel-weighted regression with weights based on the distance between observation locations. The presentation of GWR in Fotheringham et al. (2002) follows the development of local likelihood in Loader (1999).

GWR can be thought of as a kernel smoother for regression coefficients, and hence GWR coefficient estimates are likely to exhibit bias near the boundary of the region being modeled (Hastie and Loader, 1993). Modeling the coefficient surface as locally linear rather than locally constant (by including coefficient-by-location interactions) can reduce this boundary-effect bias (Hastie and Loader, 1993). Adding these interactions to the GWR model is analogous to a transition from kernel smoothing to local regression, and was introduced in Wang et al. (2008).

There is interest among practitioners of GWR not only in estimating the spatially-varying coefficient surface, but in doing variable selection to estimate which covariates are important predictors of the output variable, and in what regions they are important (citations?).

Some recent research has focused on variable selection in varying-coefficients models. In the context of varying-coefficients regression models, global variable selection (in which one compares the hypothesis that the coefficient on a given variable is zero everywhere against the hypothesis that the coefficient is nonzero somewhere) is distinguished from local variable selection (in which one compares the hypothesis that the coefficient on a given variable is zero at a given location against the hypothesis that the coefficient at that location is nonzero). Global variable selection for models where the varying coefficients are estimated using splines is addressed in Fan and Zhang (1999) for response variables that belong to an exponential-family distribution (as in the generalized linear model), and in Wang et al. (2008) for models with repeated measurements. Antoniadas et al. (2012) estimates the coefficient functions with P-splines, and then uses the nonnegative garrote of Breiman (1995) to do local variable selection by selecting P-spline bases.

Here we discuss a method of local variable selection in GWR models using the adaptive lasso of Zou (2006). The idea first appears in the literature as the geographically-weighted lasso (GWL) of Wheeler (2009), which uses 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 (specifically the AIC, but in principle one could use the BIC, *et cetera*). The AIC allows us to easily adapt our method to the setting of a generalized linear model. The local AIC presented here 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 $\boldsymbol{s}_1, \dots, \boldsymbol{s}_n$ in a spatial domain $D \subset \mathbb{R}^2$. For $i = 1, \dots, n$, 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 $\boldsymbol{\beta}_i(\boldsymbol{s}_i)$ may be spatially-varying and $\varepsilon(\boldsymbol{s}_i)$ is random noise at location $\boldsymbol{s}_i$.

$$y(\boldsymbol{s}_i) = \boldsymbol{x}'(\boldsymbol{s}_i)\boldsymbol{\beta}(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i) \tag{1}$$

Further assume that the error term $\varepsilon(\boldsymbol{s}_i)$ is normally distributed with zero mean and a possibly spatially-varying variance $\sigma^2(\boldsymbol{s}_i)$

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

In order to simplify the notation, let subscripts denote the values of data or parameters at the locations where data is observed. Thus, $\boldsymbol{x}(\boldsymbol{s}_i) \equiv \boldsymbol{x}_i \equiv (1, x_{i1}, \ldots, x_{ip})'$, $\boldsymbol{\beta}(\boldsymbol{s}_i) \equiv \boldsymbol{\beta}_i \equiv (\beta_{i0}, \beta_{i1}, \ldots, \beta_{ip})'$, $y(\boldsymbol{s}_i) \equiv y_i$, and $\sigma^2(\boldsymbol{s}_i) \equiv \sigma_i^2$. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ and $\boldsymbol{Y} = (y_1, \ldots, y_n)'$. Equations (1) and (2) can now be rewritten as

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta}_i + \varepsilon_i \text{ and } \varepsilon_i \sim \mathcal{N}\left(0, \sigma_i^2\right) \tag{3}$$

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

$$\ell\left(\boldsymbol{\beta}\right) = -1/2 \sum_{i=1}^{n} \left\{ \log\left(2\pi\sigma_i^2\right) + \sigma_i^{-2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_i\right)^2 \right\} \tag{4}$$

With n observations and $n \times (p+1)$ free parameters, 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 spatially-varying coefficients $\boldsymbol{\beta}(\boldsymbol{s})$ are smoothly varying, 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 ordinary GWR.

2.2. Estimation

Geographically-weighted regression estimates the value of the coefficient surface $\boldsymbol{\beta}(\boldsymbol{s})$ at each 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'. A bisquare kernel is used to generate spatial weights based on the euclidean distances and a bandwidth ϕ:

$$w_{ii'} = \begin{cases} \left[1 - \left(\phi^{-1}\delta_{ii'}\right)^2\right]^2 & \text{if } \delta_{ii'} < \phi \\ 0 & \text{if } \delta_{ii'} \geqslant \phi \end{cases} \tag{5}$$

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

$$\mathcal{L}_i\left(\boldsymbol{\beta}_i\right) = \prod_{i'=1}^{n} \left\{ \left(2\pi\sigma_i^2\right)^{-1/2} \exp\left[-\frac{1}{2}\sigma_i^{-2}\left(y_{i'} - \boldsymbol{x}_{i'}'\boldsymbol{\beta}_i\right)^2\right]\right\}^{w_{ii'}} \tag{6}$$

Thus, the local log-likelihood function is:

$$\ell_i\left(\boldsymbol{\beta}_i\right) \propto -1/2 \sum_{i'=1}^{n} w_{ii'} \left\{\log\sigma_i^2 + \sigma_i^{-2}\left(y_{i'} - \boldsymbol{x}_{i'}'\boldsymbol{\beta}_i\right)^2\right\} \tag{7}$$

From which it is apparent that the GWR coefficient estimates $\hat{\boldsymbol{\beta}}_{i,\mathrm{GWR}}$, which maximize the local likelihood at location $\boldsymbol{s}_i$, can be calculated using weighted least squares. Letting the diagonal weight matrix $\boldsymbol{W}_i$ be:

$$\boldsymbol{W}_i = \operatorname{diag}\left\{w_{ii'}\right\}_{i'=1}^{n} \tag{8}$$

We have:

$$\hat{\boldsymbol{\beta}}_{i,\mathrm{GWR}} = \left(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y} \tag{9}$$

And $\hat{\sigma}_i$, which maximizes (7), is:

$$\hat{\sigma}_i = \left(\mathbf{1}_n'\boldsymbol{w}_i\right)^{-1}\boldsymbol{w}_i'\left(\boldsymbol{Y} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y}\right) \tag{10}$$

3. Model selection

3.1. Variable selection

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.

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 doordinate 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.

3.1.1. Adaptive lasso

The adaptive lasso is applied to GWR by first multiplying the design matrix $\boldsymbol{X}$ by $\boldsymbol{W}_i^{1/2}$, the diagonal matrix of geographic weights centered at 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 the coefficients $\boldsymbol{\beta}_i$ of the regression model:

$$\boldsymbol{V}_i^* = \boldsymbol{U}_i^*\boldsymbol{\beta}_i + \boldsymbol{\varepsilon} \tag{11}$$

To apply the adaptive lasso for estimating these regression coefficients, each column of $\boldsymbol{U}_i^*$ is centered around zero and rescaled to have an L$_2$-norm of one. Let $\tilde{\boldsymbol{U}}_i^*$ be the centered-and-scaled version of $\boldsymbol{U}_i^*$. Adaptive weights are calculated using the OLS regression coefficients $\boldsymbol{\gamma}_i^*$ via ordinary least squares (OLS):

$$\gamma_i^* = \left(\tilde{\boldsymbol{U}}_i^{*\prime}\tilde{\boldsymbol{U}}_i^*\right)^{-1}\tilde{\boldsymbol{U}}_i^{*\prime}\boldsymbol{V}_i^* \tag{12}$$

Now a final scaling step is done: for $j = 1, \ldots, p$, the jth column of $\tilde{\boldsymbol{U}}_i^*$ is multiplied by $(\gamma_i^*)_j$, the corresponding coefficient from (12). Call this rescaled matrix $\check{\boldsymbol{U}}_i^*$.

Finally, the adaptive lasso coefficient estimates 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. Either way, the objective being minimized is the same:

$$\sum_{i'=1}^{n} w_{ii'} \left(y_{i'} - \boldsymbol{x}_{i'}' \boldsymbol{\beta}_i \right)^2 \quad + \quad \lambda_i \sum_{j=1}^{p} |\beta_{ij} / \gamma_{ij}^*| \tag{13}$$

$$\tag{14}$$

3.1.2. Adaptive elastic net

To implement the adaptive elastic net (Zou and Zhang, 2009), 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:

$$\sum_{i'=1}^{n} w_{ii'} \left(y_{i'} - \boldsymbol{x}_{i'}' \boldsymbol{\beta}_i \right)^2 + \alpha \lambda_i \sum_{j=1}^{p} |\beta_{ij} / \gamma_{ij}^*| + (1 - \alpha) \lambda_i \sum_{j=1}^{p} \left(\beta_{ij} / \gamma_{ij}^* \right)^2 \tag{15}$$

$$= \sum_{i'=1}^{n} w_{ii'} \left(y_{i'} - \boldsymbol{x}_{i'}' \boldsymbol{\beta}_i \right)^2 + \lambda_i \left(\alpha \sum_{j=1}^{p} |\beta_{ij} / \gamma_{ij}^*| + (1 - \alpha) \sum_{j=1}^{p} \left[\beta_{ij} / \gamma_{ij}^* \right]^2 \right) \tag{16}$$

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 AIC_{loc}. The local AIC is calculated by adding a penalty to the local likelihood, with the sum of the weights around 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).

$$
\begin{aligned}
\mathrm{AIC}_{\mathrm{loc},i} &= -2\sum_{i'=1}^{n} \ell_{ii'} + 2\mathrm{df}_i && (17)\\
&= -2 \times \sum_{i'=1}^{n} \log\left\{ \left(2\pi\hat{\sigma}_i^2\right)^{-1/2} \exp\left[-\frac{1}{2}\hat{\sigma}_i^{-2}\left(y_{i'} - \boldsymbol{x}'_{i'}\hat{\boldsymbol{\beta}}_{i'}\right)^2\right]\right\}^{w_{ii'}} + 2\mathrm{df}_i && (18)\\
&= \sum_{i'=1}^{n} w_{ii'}\left\{\log(2\pi) + \log\hat{\sigma}_i^2 + \hat{\sigma}_i^{-2}\left(y_{i'} - \boldsymbol{x}'_{i'}\hat{\boldsymbol{\beta}}_{i'}\right)^2\right\} + 2\mathrm{df}_i && (19)\\
&= \hat{\sigma}_i^{-2}\sum_{i'=1}^{n} w_{ii'}\left(y_{i'} - \boldsymbol{x}'_{i'}\hat{\boldsymbol{\beta}}_i\right)^2 + 2\mathrm{df}_i + C_i && (20)
\end{aligned}
$$

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

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 is global and so we need a global statistic for comparing prospective bandwidths. The objective minimized by GWL is:

$$
\sum_{i'=1}^{n} w_{ii'}\left(y_{i'} - \boldsymbol{x}'_{i'}\boldsymbol{\beta}_i\right)^2 + \sum_{j=1}^{p}\lambda_{ij}|\beta_{ij}| \qquad (21)
$$

Where $\lambda_{ij}, j = 1, \ldots, p$ are penalties from the adaptive lasso (Zou, 2006). Taking the derivatives with respect to β and setting to zero, we see that

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{i,\mathrm{GWL}} &= \left(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y} - \frac{1}{2}\left(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\right)^{-1}\boldsymbol{\lambda}_i && (22)\\
\hat{y}_i = \boldsymbol{x}'_i\hat{\boldsymbol{\beta}}_{i,\mathrm{GWL}} &= \boldsymbol{x}'_i\left(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{Y} - \frac{1}{2}\boldsymbol{x}'_i\left(\boldsymbol{X}'\boldsymbol{W}_i\boldsymbol{X}\right)^{-1}\boldsymbol{\lambda}_i && (23)
\end{aligned}
$$

Unlike in the case of ordinary geographically-weighted regression, the fitted values $\hat{\boldsymbol{Y}}$ are not a linear combination of the observations $\boldsymbol{Y}$. Because GWL is not a linear smoother it is not possible to calculate the AIC as in Fotheringham et al. (2002) (Zou, 2006). We propose a statistic called the total AIC (AIC_{tot}) for the purpose of selecting the bandwidth parameter. Because of the kernel weights and the application of the adaptive lasso, the sample size and the degrees of freedom are different at each location. The total AIC is found by taking the sum over all of the observed data:

$$\begin{aligned} \text{AIC}_{\text{tot}} &= -2 \times \sum_{i=1}^{n} \ell_i + 2\text{df} && (24) \\ &= \sum_{i=1}^{n} \left\{ \log \hat{\sigma}_i^2 + \hat{\sigma}_i^{-2} \left(y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_i \right)^2 \right\} + 2\text{df} && (25) \end{aligned}$$

What remains is to calculate df, the number of degrees of freedom used by the model. Ordinary GWR, as developed in Loader (1999) and Fotheringham et al. (2002) calculates df using the trace of the "hat" matrix, but because the GWL is not a linear smoother, there is no "hat" matrix associated with GWL. Instead, notice that df can be pulled into the summation in (26):

$$\text{df} = \sum_{i=1}^{n} \left(n^{-1} \text{df} \right) \tag{26}$$

Now, because we are considering the sum of local weights to be the sample size for the local models, we estimate df by $\sum_{i=1}^{n} \left\{ \left(\sum_{i'=1}^{n} w_{ii'} \right)^{-1} \text{df}_i \right\}$, and the total AIC is then:

$$\text{AIC}_{\text{tot}} = \sum_{i=1}^{n} \left\{ \log \hat{\sigma}_i^2 + \hat{\sigma}_i^{-2} \left(y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_i \right)^2 + 2 \left(\sum_{i'=1}^{n} w_{ii'} \right)^{-1} \text{df}_i \right\} \tag{27}$$

The bandwidth that minimizes (28) is found by a line search.

3.4. Confidence interval construction

Confidence intervals for the GWL's coefficient estimates can be calculated either by the bootstrap (Efron and Tibshirani, 1986) or by exploiting an assumption of normally-distributed residuals.

Then the, e.g., 95% confidence interval for each regression coefficient is defined by the (2.5, 97.5) percentiles of the coefficient estimates from the bootstrap replicates.

To compute coefficient confidence intervals via the bootstrap, the observations with non-zero geographic weights are resampled uniformly with replacement for each of n_B bootstrap replicates. For each bootstrap replicate, the GWL is used to estimate regression coefficients. The local likelihood of the bootstrap replicates may be different from that of the original sample, so the adaptive lasso tuning parameter may differ for each bootstrap replicate. Since the GWL is applied independently to each bootstrap replicate, the variables selected by GWL may be different for each replicate.

Unshrunk coefficient estimates are found by using the GWL at each location for variable selection only and then estimating the coefficients for the selected variables by weighted least squares. An unshrunk bootstrap confidence interval is found by estimating the unshrunk coefficients for each of the n_B bootstrap replicates and then calculating the percentiles as above.

A third way to estimate the coefficient confidence intervals is to use the GWL for variable selection only and then to use weighted least squares for both coefficient estimation and confidence interval construction:

$$\hat{\text{se}}_{\beta_i} = \left(\tilde{\boldsymbol{X}}_i' \boldsymbol{W}_i \tilde{\boldsymbol{X}}_i\right)^{-1} \tilde{\boldsymbol{X}}_i' \boldsymbol{W}_i \boldsymbol{Y} \tag{28}$$

where $\tilde{\boldsymbol{X}}_i$ is the model matrix including only those variables that are selected by GWL at location i.

4. Simulation

4.1. Simulation setup

A simulation study was conducted to assess the performance of the method described in Sections 2-3. There were twelve simulation settings, each of which was simulated 100 times. For each of the twelve settings, $\boldsymbol{\beta}_1$, the true coefficient surface for $\boldsymbol{Z}_1$, 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.

Data was simulated on $[0,1] \times [0,1]$, which was divided into a 30×30 grid. Each of $p = 5$ covariates $Z_1, \ldots, Z_p$ was simulated by a Gaussian random field (GRF) with mean zero and exponential spatial covariance $Cov\left(Z_{ji}, Z_{ji'}\right) = \sigma_z^2 \exp\left(-\tau_z^{-1}\delta_{ii'}\right)$ where $\sigma_z^2 = 1$ is the variance, τ_z 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 $\boldsymbol{Z}$ 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 $y_i = \boldsymbol{z}_i'\boldsymbol{\beta}_i + \varepsilon_i$ for $i = 1, \ldots, 900$ where for simplicity the vector of additive errors $\boldsymbol{\varepsilon}$ were iid Gaussian.

The simulated data include the output y and five covariates $Z_1, \ldots, Z_5$. The true data-generating model uses only Z_1, so $Z_2, \ldots, Z_5$ are included to test the variable-selection properties of GWL.

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 β_1 (step, gradient, and parabola - see Figure **??**); data was simulated with ($\rho = 0.5$) and without ($\rho = 0$) correlation between the covariates; and simulations were made with low ($\sigma^2 = 0.25$) and high ($\sigma^2 = 1$) variance for the random noise term.

The performance of GWL was compared to 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.

[Table 1 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

4.2. Results

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 (Tables 2 - 6) and coefficient estimation (Tables **??** - **??**) are presented in the tables below.

4.3. Discussion

4.4. Tables

4.4.1. Selection

[Table 2 about here.]

[Table 3 about here.]

[Table 4 about here.]

[Table 5 about here.]

[Table 6 about here.]

4.4.2. Estimation

[Table 7 about here.]

[Table 8 about here.]

[Table 9 about here.]

[Table 10 about here.]

[Table 11 about here.]

5. Data analysis

5.1. Census poverty data

An example analysis is presented to demonstrate application of the geographically-weighted lasso. The analysis focuses on creating a varying-coefficients regression model to describe 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 1990. This analysis looks specifically at the upper midwest states of Minnesota, Iowa, Wisconsin, Illinois, Indiana, and Michigan. All data is aggregated at the county level.

Table 12 lists the variables that were considered as potential predictors of county-level poverty rate. The outcome of interest (poverty rate) is a proportion and so takes values on $[0, 1]$, but to demonstrate the geographically-weighted lasso in a linear regression context, we model the logit-transformed poverty rate. The predictor variables were not transformed - county-level proportions were used.

[Table 12 about here.]

5.2. Modeling

Variable selection and coefficient estimation using the geographically-weighted lasso were done for each census separately (no attempt is made here to borrow strength across years). Therefore the bandwidth was chosen independently for the model of each census. Table 13 lists the bandwidth that were selected for each census.

[Table 13 about here.]

5.3. Figures

The coefficient estimates are plotted on maps of the upper midwest in Figures 3 - 8. It is immediately apparent that the estimated coefficient surfaces are non-constant for most variables.

[Figure 3 about here.]

[Figure 4 about here.]

[Figure 5 about here.]

[Figure 6 about here.]

[Figure 7 about here.]

[Figure 8 about here.]

5.4. Discussion

If the model is to be believed, then it is onto uncommon for the same variable to have both positive and negative effects on poverty in different geographical areas - see, for instance, the coefficient surface for `pex` (mining employment) in the 1970 census. That surface indicates an interaction whereby the proportion of people working in mining in southern parts of the studied area is associated with an increase in the poverty rate, while in northern parts of the studied area it is associated with a decrease in the poverty rate. Often, a variable is found to be associated with an effect on poverty in some counties but not in others (see, for instance, the coefficient surface for `pserve` (services employment) in 1980 -

The trend from small bandwidth in the earlier censuses to large bandwidth in the more recent censuses is also interesting - perhaps the local nature

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.

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

Efron, B. and R. Tibshirani (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science 1*(1), 54–75.

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.

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.

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

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.

List of Figures

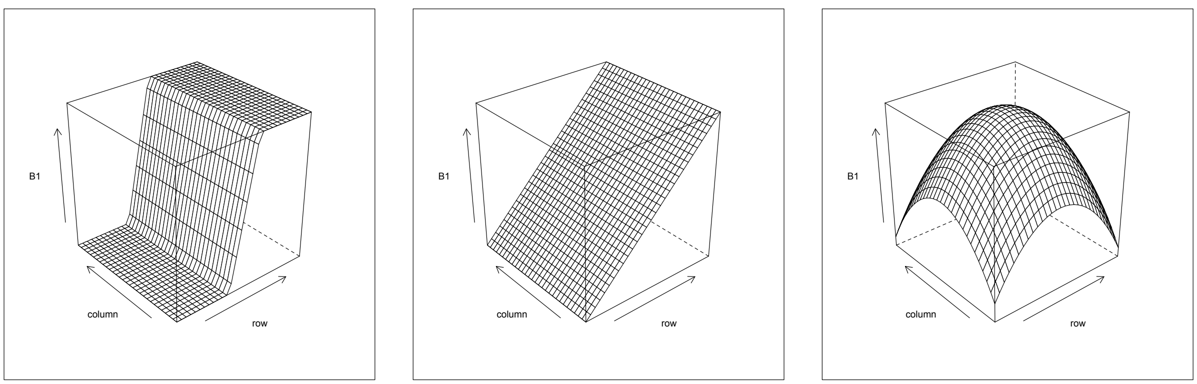

Figure 1: The actual β_1 coefficient surface used in the simulation.

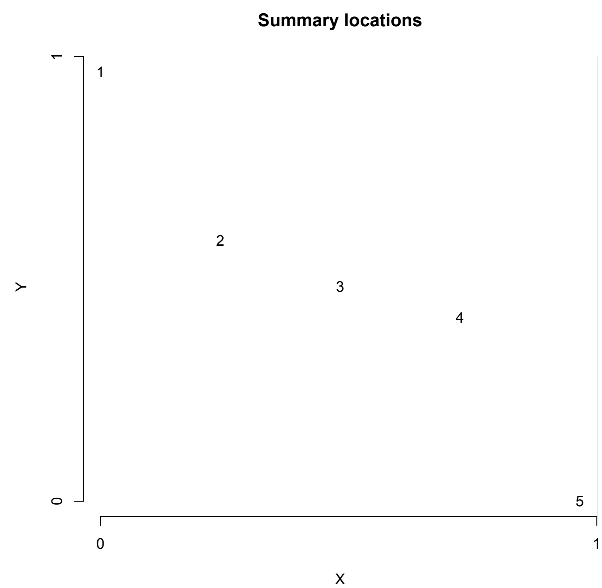

Figure 2: Locations where the variable selection and coefficient estimation of GWL were summarized.

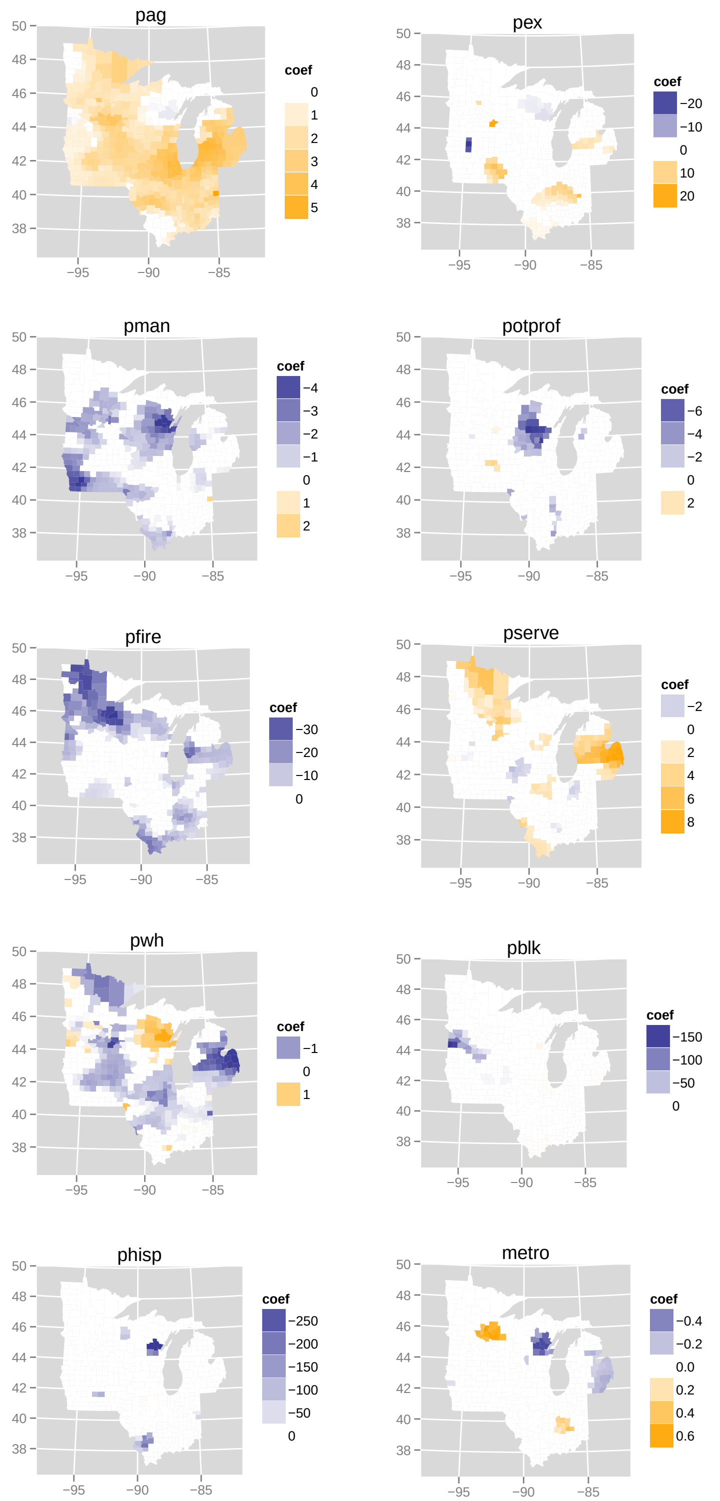

Figure 3: Estimated coefficient surfaces for the 1960 census.

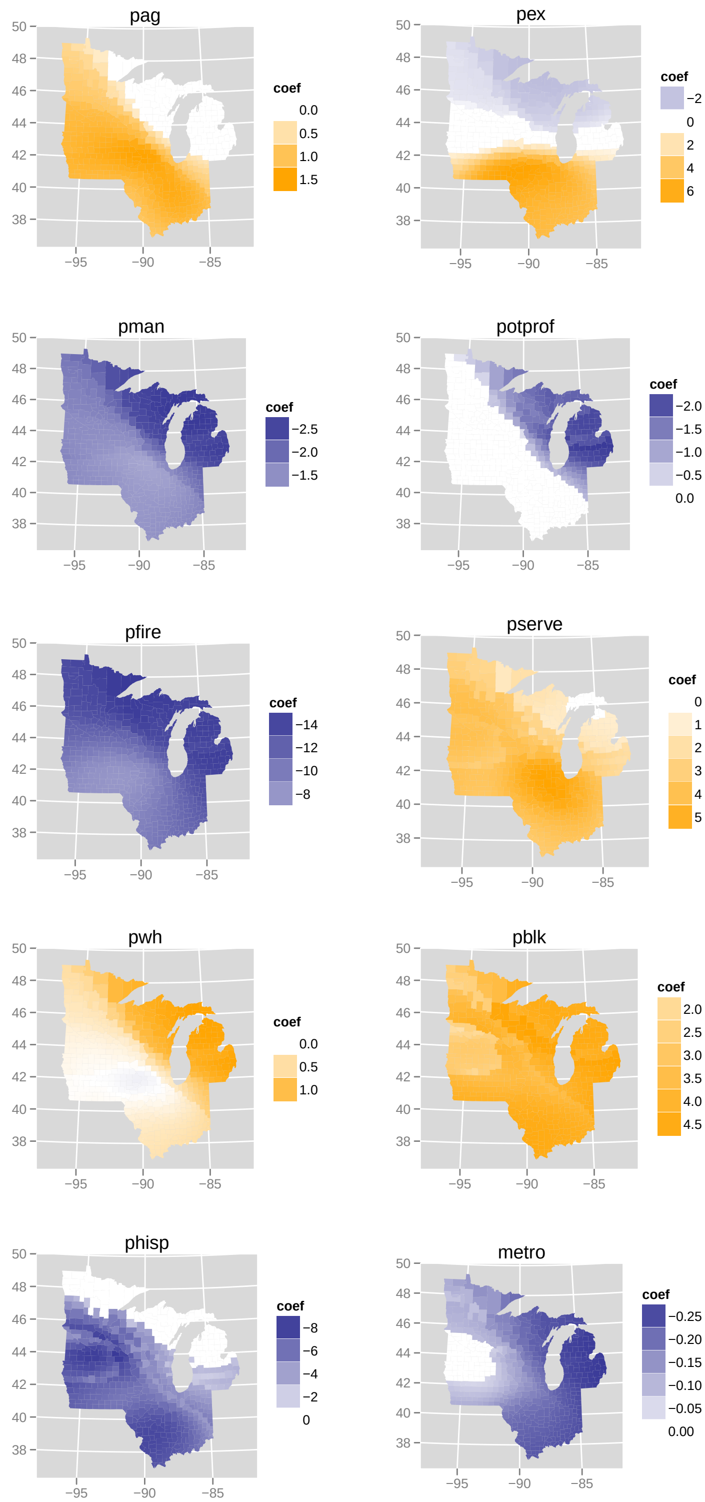

Figure 4: Estimated coefficient surfaces for the 1970 census.

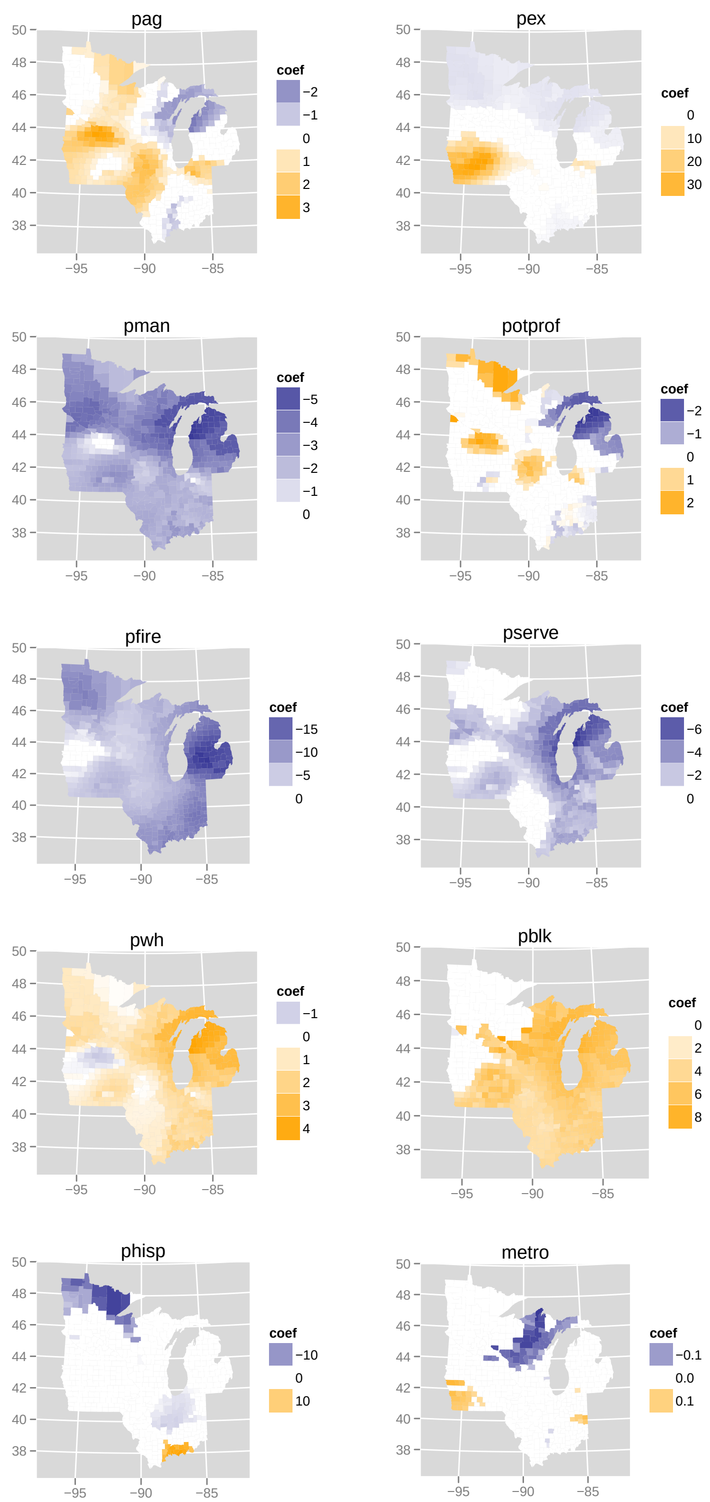

Figure 5: Estimated coefficient surfaces for the 1980 census.

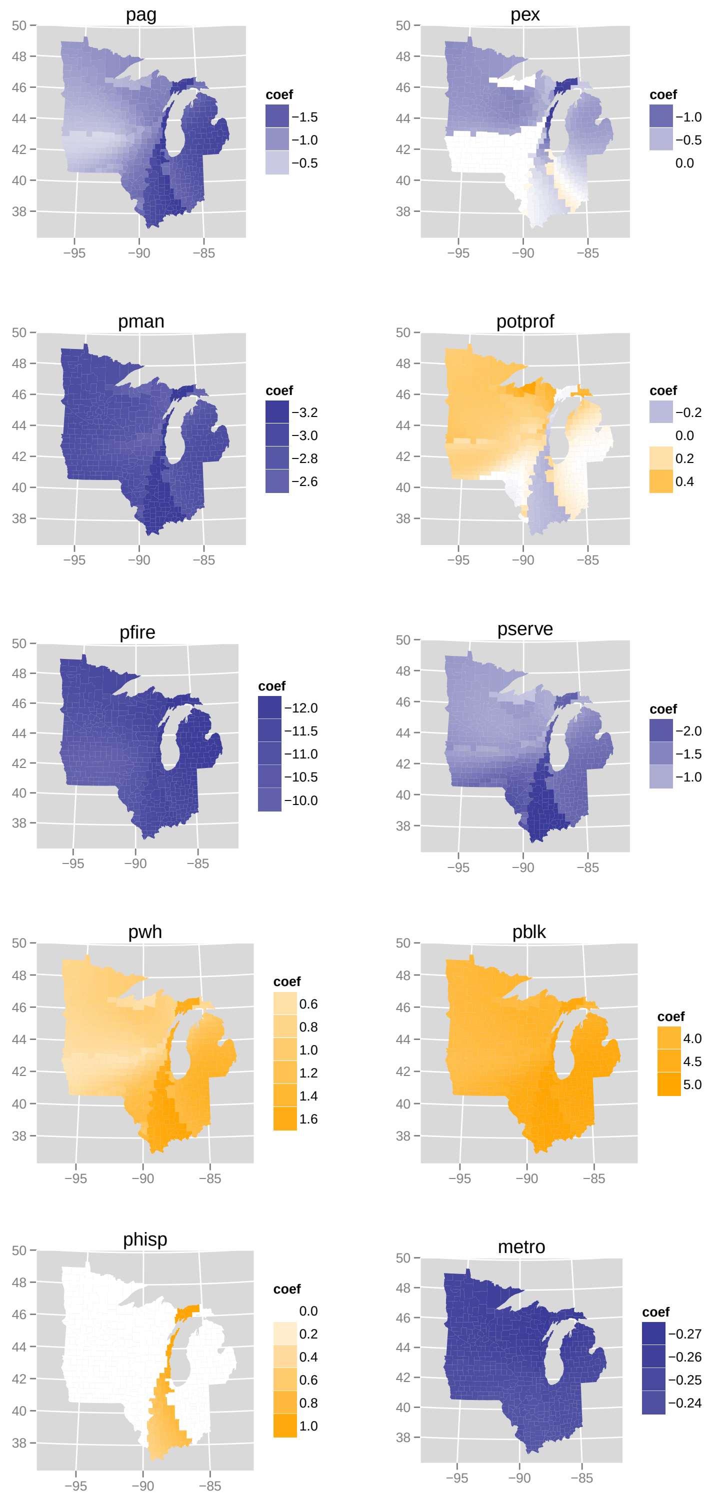

Figure 6: Estimated coefficient surfaces for the 1990 census.

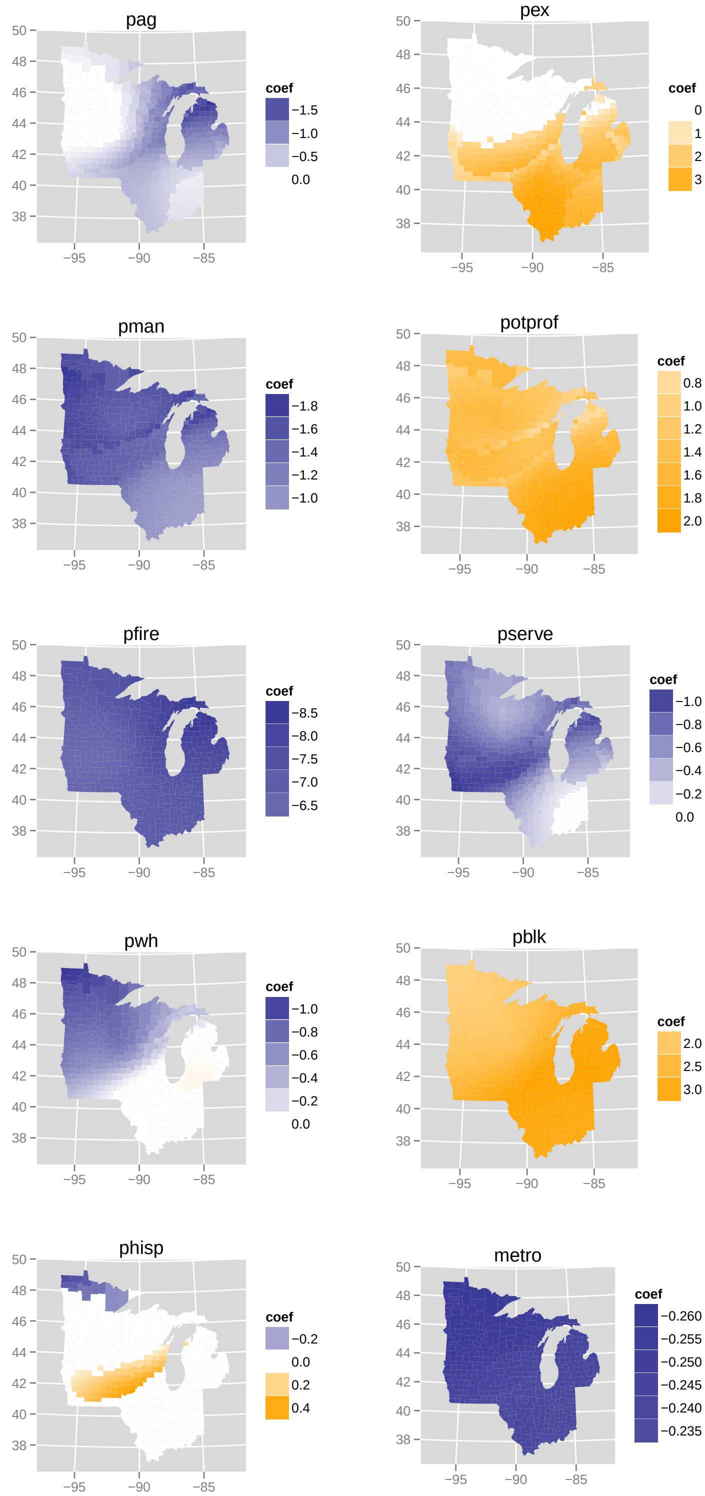

Figure 7: Estimated coefficient surfaces for the 2000 census.

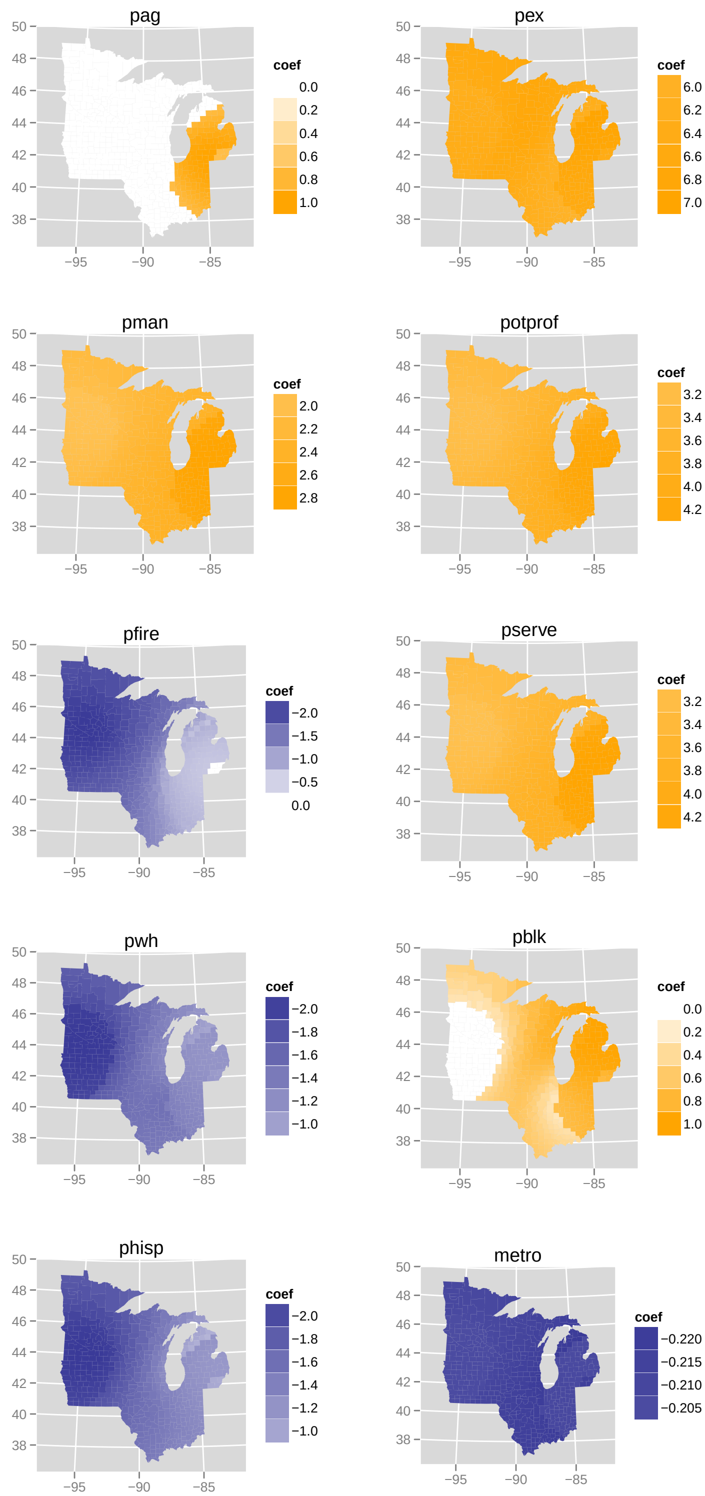

Figure 8: Estimated coefficient surfaces for the 2006 census.

List of Tables

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.

lars		enet		glmnet	
β_1	β_4 - β_5	β_1	β_4 - β_5	β_1	β_4 - β_5
0.98	0.04	1.00	0.04	1.00	0.05
0.89	0.09	0.86	0.09	0.82	0.07
0.96	0.07	0.99	0.10	0.96	0.09
0.84	0.04	0.84	0.07	0.88	0.05
1.00	0.04	1.00	0.03	1.00	0.03
0.99	0.08	0.97	0.07	0.97	0.07
1.00	0.07	1.00	0.06	1.00	0.04
0.90	0.08	0.92	0.08	0.92	0.08
0.94	0.06	0.95	0.06	0.94	0.06
0.80	0.06	0.81	0.07	0.80	0.06
0.95	0.06	0.94	0.09	0.95	0.04
0.78	0.12	0.79	0.12	0.80	0.12

Table 2: Selection frequency at location 1

lars		enet		glmnet	
β_1	β_4 - β_5	β_1	β_4 - β_5	β_1	β_4 - β_5
1.00	0.07	1.00	0.07	1.00	0.07
1.00	0.06	1.00	0.06	1.00	0.07
1.00	0.05	1.00	0.06	1.00	0.05
0.99	0.03	1.00	0.07	0.99	0.04
1.00	0.10	1.00	0.08	1.00	0.07
0.98	0.07	0.98	0.08	0.99	0.07
1.00	0.07	1.00	0.06	1.00	0.05
0.98	0.06	0.99	0.08	0.99	0.05
1.00	0.09	1.00	0.08	1.00	0.08
0.97	0.12	0.98	0.11	0.98	0.10
1.00	0.06	1.00	0.05	1.00	0.05
0.94	0.08	0.94	0.10	0.94	0.08

Table 3: Selection frequency at location 2

lars		enet		glmnet	
β_1	β_4 - β_5	β_1	β_4 - β_5	β_1	β_4 - β_5
0.99	0.05	0.99	0.06	0.99	0.06
0.84	0.08	0.84	0.08	0.82	0.07
0.96	0.05	0.97	0.08	0.92	0.04
0.78	0.08	0.81	0.11	0.80	0.08
1.00	0.09	1.00	0.08	1.00	0.07
0.98	0.08	0.95	0.08	0.96	0.07
1.00	0.07	1.00	0.06	1.00	0.04
0.93	0.09	0.95	0.09	0.94	0.09
1.00	0.09	1.00	0.09	1.00	0.09
0.96	0.10	0.97	0.09	0.97	0.10
1.00	0.08	1.00	0.07	1.00	0.07
0.93	0.10	0.94	0.10	0.96	0.10

Table 4: Selection frequency at location 3

lars		enet		glmnet	
β_1	β_4 - β_5	β_1	β_4 - β_5	β_1	β_4 - β_5
0.57	0.08	0.64	0.06	0.59	0.06
0.48	0.07	0.48	0.07	0.49	0.07
0.45	0.08	0.51	0.12	0.40	0.07
0.53	0.08	0.52	0.07	0.51	0.07
1.00	0.06	1.00	0.06	1.00	0.06
0.98	0.07	0.95	0.07	0.93	0.06
1.00	0.09	1.00	0.08	1.00	0.10
0.96	0.07	0.95	0.11	0.95	0.08
1.00	0.09	1.00	0.08	1.00	0.08
0.93	0.07	0.92	0.08	0.94	0.08
1.00	0.08	1.00	0.08	1.00	0.08
0.96	0.08	0.96	0.09	0.96	0.09

Table 5: Selection frequency at location 4

lars		enet		glmnet	
β_1	β_4 - β_5	β_1	β_4 - β_5	β_1	β_4 - β_5
0.04	0.03	0.03	0.03	0.03	0.03
0.07	0.05	0.06	0.04	0.04	0.05
0.02	0.04	0.02	0.03	0.03	0.05
0.05	0.04	0.04	0.03	0.06	0.06
0.92	0.05	0.93	0.05	0.94	0.04
0.71	0.08	0.70	0.07	0.70	0.07
0.93	0.10	0.95	0.14	0.95	0.10
0.60	0.07	0.63	0.13	0.64	0.06
0.93	0.10	0.93	0.09	0.92	0.10
0.80	0.05	0.81	0.05	0.79	0.05
0.93	0.07	0.94	0.12	0.94	0.07
0.81	0.09	0.81	0.11	0.83	0.08

Table 6: Selection frequency at location 5

lars	enet	glmnet	unshrunk.lars	unshrunk.enet	unshrunk.glmnet	oracular	gwr
0.046	0.025	*0.023*	0.151	0.127	0.124	0.082	**0.005**
0.146	0.186	0.216	0.290	0.376	0.375	*0.134*	**0.009**
0.072	*0.045*	0.073	0.172	0.134	0.205	0.101	**0.011**
0.214	0.218	0.179	0.441	0.425	0.369	*0.154*	**0.022**
0.066	0.069	0.070	0.007	**0.007**	*0.007*	0.010	0.016
0.084	0.094	0.096	0.161	0.078	0.085	*0.045*	**0.042**
0.065	0.070	0.069	0.009	**0.007**	*0.008*	0.009	0.019
0.161	0.149	0.144	0.149	0.123	0.121	**0.040**	*0.050*
0.074	0.075	0.074	0.020	*0.020*	**0.020**	0.022	0.105
0.079	0.078	0.077	0.041	*0.040*	**0.039**	0.063	0.106
0.077	0.069	0.076	0.024	**0.018**	0.023	*0.023*	0.099
0.083	0.072	0.083	*0.048*	**0.044**	0.050	0.067	0.110

Table 7: Mean squared error of estimates for β_1 at location 1 (**minimum**, *next best*).

lars	enet	glmnet	unshrunk.lars	unshrunk.enet	unshrunk.glmnet	oracular	gwr
0.024	0.024	0.024	**0.020**	0.021	*0.021*	0.021	0.042
0.061	0.063	0.068	*0.050*	0.054	0.056	**0.042**	0.070
0.022	0.027	0.021	**0.017**	0.021	*0.017*	0.018	0.044
0.069	0.071	0.071	0.057	*0.056*	0.061	**0.043**	0.075
0.003	0.003	0.003	*0.001*	0.001	0.001	**0.001**	0.001
0.014	0.013	0.008	0.015	0.013	0.009	*0.002*	**0.002**
0.003	0.003	0.003	0.001	*0.001*	**0.001**	0.001	0.002
0.015	0.012	0.012	0.014	0.011	0.011	**0.003**	*0.004*
0.005	0.005	0.005	0.007	0.007	0.007	**0.004**	0.007
0.018	0.016	0.016	0.026	0.021	0.022	*0.008*	**0.007**
0.007	0.007	*0.007*	0.011	0.010	0.009	**0.004**	0.008
0.020	0.022	0.020	0.022	0.022	0.023	**0.007**	*0.009*

Table 8: Mean squared error of estimates for β_1 at location 2 (**minimum**, *next best*).

lars	enet	glmnet	unshrunk.lars	unshrunk.enet	unshrunk.glmnet	oracular	gwr
0.011	0.011	0.010	0.007	0.007	0.007	**0.004**	*0.005*
0.043	0.043	0.047	0.049	0.049	0.054	*0.009*	**0.008**
0.016	0.014	0.022	0.013	0.011	0.021	**0.005**	*0.005*
0.048	0.047	0.045	0.049	0.045	0.044	*0.008*	**0.008**
0.001	0.001	0.001	**0.001**	0.001	0.001	*0.001*	0.001
0.007	0.017	0.015	0.007	0.019	0.017	*0.002*	**0.002**
0.001	0.001	0.000	0.000	*0.000*	**0.000**	0.001	0.002
0.022	0.017	0.019	0.023	0.017	0.021	**0.002**	*0.003*
0.015	0.015	0.015	0.015	0.015	*0.015*	**0.005**	0.022
0.032	0.029	0.029	0.030	0.027	0.027	**0.012**	*0.023*
0.019	0.018	0.019	0.018	*0.017*	0.018	**0.005**	0.024
0.037	0.037	0.030	0.037	0.034	0.029	**0.012**	*0.024*

Table 9: Mean squared error of estimates for β_1 at location 3 (**minimum**, *next best*).

lars	enet	glmnet	unshrunk.lars	unshrunk.enet	unshrunk.glmnet	oracular	gwr
0.014	0.014	**0.014**	0.017	0.019	0.018	0.021	0.042
0.037	**0.036**	0.039	0.039	0.042	0.046	0.047	0.074
0.010	0.012	*0.011*	0.013	0.016	0.014	0.020	0.044
0.038	**0.028**	0.038	0.048	0.047	0.048	0.043	0.082
0.003	0.003	0.003	0.002	0.001	*0.001*	**0.001**	0.001
0.009	0.014	0.016	0.007	0.012	0.015	**0.002**	*0.003*
0.003	0.002	0.003	0.002	0.001	*0.001*	**0.001**	0.002
0.013	0.015	0.014	0.013	0.014	0.014	**0.003**	*0.004*
0.006	0.006	*0.006*	0.009	0.009	0.009	**0.004**	0.008
0.025	0.027	0.023	0.027	0.029	0.024	*0.010*	**0.009**
0.008	0.008	0.008	0.011	0.011	0.011	**0.004**	0.010
0.018	0.020	0.017	0.022	0.022	0.022	**0.009**	*0.010*

Table 10: Mean squared error of estimates for β_1 at location 4 (**minimum**, *next best*).

lars	enet	glmnet	unshrunk.lars	unshrunk.enet	unshrunk.glmnet	oracular	gwr
0.002	*0.001*	0.002	0.006	0.004	0.004	**0.000**	0.007
0.003	0.006	*0.002*	0.016	0.024	0.009	**0.000**	0.011
0.002	0.002	0.003	0.009	0.009	0.009	**0.000**	0.010
0.017	*0.004*	0.022	0.046	0.038	0.043	**0.000**	0.015
0.067	0.068	0.069	0.004	0.004	*0.004*	**0.000**	0.016
0.054	0.051	0.052	*0.019*	0.019	0.019	**0.000**	0.044
0.062	0.060	0.064	0.009	0.010	*0.007*	**0.000**	0.021
0.050	0.047	0.053	*0.017*	0.020	0.017	**0.000**	0.051
0.074	0.075	0.075	**0.018**	0.020	0.021	*0.020*	0.104
0.075	0.074	0.073	0.024	**0.022**	*0.023*	0.055	0.104
0.077	0.069	0.076	*0.021*	0.023	**0.020**	0.025	0.099
0.081	0.075	0.081	0.037	*0.036*	**0.035**	0.042	0.113

Table 11: Mean squared error of estimates for β_1 at location 5 (**minimum**, *next best*).

Variable name	Description
pag	Proportion working in agriculture
pex	Proportion working in extraction (mining)
pman	Proportion working in manufacturing
pserve	Proportion working in services
pfire	Proportion working in finance, insurance, and real estate
potprof	Proportion working in other professions
pwh	Proportion who are white
pblk	Proportion who are black
phisp	Proportion who are hispanic
metro	Is the county in a metropolitan area?

Table 12: Description of the variables used in the census-data example

Year	Bandwidth
1960	0.033
1970	0.381
1980	0.122
1990	0.802
2000	0.854
2006	0.976

Table 13: Bandwidth selected for each year's model.