Local variable selection for varying-coefficients models in the context of geographically-weighted regression

Wesley Brooks

1. Introduction

Varying-coefficients regression (Hastie and Tibshirani, 1993) is a technique used in spatial statistics to model a non-stationary process. Geographically weighted regression (GWR) (Fotheringham et al., 2002) is a method of fitting varying-coefficients regression models for spatial data that 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 degrees of freedom used to estimate the spatially-varying coefficient surfaces.

2. Geographically-weighted regression models

2.1. Model

Consider n data observations, made at locations $s_1, \ldots, s_n$. For $i = 1, \ldots, n$, let $y(s_i)$ and $\boldsymbol{x}(s_i)$ be the univariate outcome of interest, and a $(p+1)$-variate vector of covariates measured at location s_i, respectively. At each location s_i, assume that the outcome is related to the covariates by a linear model with coefficients $\boldsymbol{\beta}_i(s_i)$ that may be spatially-varying.

$$y(s_i) = \boldsymbol{x}'(s_i)\boldsymbol{\beta}(s_i) + \epsilon(s_i) \tag{1}$$

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

$$\epsilon(s_i) \sim \mathcal{N}\left(0, \sigma^2(s_i)\right) \tag{2}$$

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

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_i + \epsilon_i \tag{3}$$

$$\epsilon_i \sim \mathcal{N}\left(0, \sigma_i^2\right) \tag{4}$$

Assume that, given the covariates $\boldsymbol{X}$, observations of the output 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.

$$\ell\left(\boldsymbol{\beta}\right) = -\frac{1}{2}\sum_{i=1}^{n}\left\{\log\left(2\pi\sigma_i^2\right) + \sigma_i^{-2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}_i\right)^2\right\} \tag{5}$$

With n observations and $n \times (p+1)$ free parameters, the model is overdetermined so it is not possible to directly maximize the total likelihood. To effectively reduce the number of parameters, assume that the spatially-varying coefficients $\boldsymbol{\beta}(s)$ 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 method is called geographically-weighted regression (GWR).

2.2. Geographically-weighted regression

Geographically-weighted regression estimates the value of the coefficient surface $\boldsymbol{\beta}(s)$ at each location s_i. Assume for now that there are known weights $w_{ii'}$ based on the distance $\|s_i - s_{i'}\|$ between locations s_i and $s_{i'}$ for all i, i'.

Coefficient estimation is done by maximizing the local likelihood at each location (Fotheringham et al., 2002).

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

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

The first and second derivatives of the local log-likelihood are

$$\left\{\frac{\partial \ell_i}{\partial \boldsymbol{\beta}_i}\right\}_j = \sum_{i'=1}^{n} \left\{ x_{i'j} w_{ii'} \sigma_i^{-2} \left(y_{i'} - \boldsymbol{x}_{i'}'\boldsymbol{\beta}_i\right) \right\} \tag{8}$$

$$\left\{\frac{\partial^2 \ell_i}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_i'}\right\}_{j,k} = -\sum_{i'=1}^{n} \left\{ x_{i'j} x_{i'k} w_{ii'} \sigma_i^{-2} \right\} \tag{9}$$

So the observed Fisher information in the locally weighted sample is

$$\boldsymbol{\mathcal{J}}_i = \sigma_i^{-2} \begin{pmatrix} \sum_{i'=1}^{n} w_{ii'} x_{i'1}^2 & \dots & \sum_{i'=1}^{n} w_{ii'} x_{i'1} x_{i'p} \\ \vdots & \ddots & \vdots \\ \sum_{i'=1}^{n} w_{ii'} x_{i'p} x_{i'1} & \dots & \sum_{i'=1}^{n} w_{ii'} x_{i'p}^2 \end{pmatrix} \tag{10}$$

$$= \sigma_i^{-2} \sum_{i'=1}^{n} w_{ii'} \begin{pmatrix} x_{i'1}^2 & \dots & x_{i'1} x_{i'p} \\ \vdots & \ddots & \vdots \\ x_{i'p} x_{i'1} & \dots & x_{i'p}^2 \end{pmatrix} \tag{11}$$

$$= \sigma_i^{-2} \sum_{i'=1}^{n} w_{ii'} \boldsymbol{x}_{i'} \boldsymbol{x}_{i'}' \tag{12}$$

The form of the observed Fisher information suggests that the information in the data $\boldsymbol{x}_{i'}$ about the coefficients at location s_i is proportional to the weight $w_{ii'}$.

At each location s_i, the ordinary geographically-weighted regression estimator minimizes the objective function:

$$\sum_{i'=1}^{n} w_{ii'} \left(y_{i'} - \boldsymbol{x}'_{i'} \boldsymbol{\beta}_i \right)^2 \tag{13}$$

Letting the weight matrix $\boldsymbol{W}_i$ be

$$\boldsymbol{W}_i = \begin{pmatrix} w_{i1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{in} \end{pmatrix} \tag{14}$$

estimation of the ordinary geographically-weighted regression coefficient surface is by weighted least squares:

$$\hat{\boldsymbol{\beta}}_{i,\text{GWR}} = \left(\boldsymbol{X}' \boldsymbol{W}_i \boldsymbol{X} \right)^{-1} \boldsymbol{X}' \boldsymbol{W}_i \boldsymbol{Y} \tag{15}$$

2.3. Smoothing kernel

The bisquare kernel function is used to generate geographic weights based on the distance between observation locations. For estimating the value of the coefficient surface at location s_i, the weight given to the observation at location $s_{i'}$ is

$$w_{ii'} = \begin{cases} \left[1 - \left(\text{bw}^{-1} \left\| s_i - s_{i'} \right\| \right)^2 \right]^2 & \text{if } \left\| s_i - s_{i'} \right\| < \text{bw} \\ 0 & \text{if } \left\| s_i - s_{i'} \right\| \geqslant \text{bw} \end{cases} \tag{16}$$

where bw is the kernel bandwidth.

3. Model selection and shrinkage

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 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 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:

$$\boldsymbol{V}_i^* = \boldsymbol{U}_i^*\boldsymbol{\beta}_i + \epsilon \tag{17}$$

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 $\widetilde{\boldsymbol{U}}_i^*$ 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):

$$\gamma_i^* = \left(\widetilde{\boldsymbol{U}}_i^{*\prime}\widetilde{\boldsymbol{U}}_i^*\right)^{-1}\widetilde{\boldsymbol{U}}_i^{*\prime}\boldsymbol{V}_i^* \tag{18}$$

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 (18). Call this rescaled matrix $\breve{\boldsymbol{U}}_i^*$.

Finally, the Adaptive LASSO coefficient estimates at location s_i are found by using the `lars` algorithm (Efron et al., 2004) to model $\boldsymbol{V}_i^*$ as a function of $\breve{\boldsymbol{U}}_i^*$.

3.1. Tuning parameter selection

The final task is to select the LASSO tuning parameter. Wheeler (2009) proposed selecting the tuning parameter for the LASSO at location s_i to minimize the jackknife prediction error $|y_i - \hat{y}_i^{(i)}|$, but this choice restricts coefficient estimation to occur at the locations where data has been observed. We instead propose to use a locally-weighted version of the Akaike Information Criterion (AIC (Akaike, 1974)) to select the tuning parameter. The local AIC allows coefficients to be estimated at any location where the local likelihood can be calculated. 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 number of nonzero coefficients in $\boldsymbol{\beta}_i$ playing the role of the "degrees of freedom" (df_i) (Zou et al., 2007).

The objective minimized by the geographically-weighted adaptive lasso (GAL) is:

$$\sum_{i'=1}^{n} w_{ii'} \left(y_{i'} - \boldsymbol{x}_{i'}' \boldsymbol{\beta}_i \right)^2 + \sum_{j=1}^{p} \lambda_{ij} \beta_{ij} \tag{19}$$

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

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{i,\text{GAL}} &= \left(\boldsymbol{X}' \boldsymbol{W}_i \boldsymbol{X} \right)^{-1} \boldsymbol{X}' \boldsymbol{W}_i \boldsymbol{Y} - \frac{1}{2} \left(\boldsymbol{X}' \boldsymbol{W}_i \boldsymbol{X} \right)^{-1} \boldsymbol{\lambda}_i & (20) \\ \hat{y}_i = \boldsymbol{x}_i \hat{\boldsymbol{\beta}}_{i,\text{GAL}} &= \boldsymbol{x}_i \left(\boldsymbol{X}' \boldsymbol{W}_i \boldsymbol{X} \right)^{-1} \boldsymbol{X}' \boldsymbol{W}_i \boldsymbol{Y} - \frac{1}{2} \boldsymbol{x}_i \left(\boldsymbol{X}' \boldsymbol{W}_i \boldsymbol{X} \right)^{-1} \boldsymbol{\lambda}_i & (21) \end{aligned}$$

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 GAL is not a linear smoother the AIC and confidence intervals as calculated in Fotheringham et al. (2002) are not accurate for the GAL (Zou, 2006). The local AIC (AIC_{loc}) is minimized to select the adaptive lasso tuning parameter.

$$\begin{aligned}
\text{AIC}_{\text{loc},i} &= -2 \sum_{i'=1}^{n} \ell_{ii'} + 2\text{df}_i && (22) \\
&= -2 \times \sum_{i'=1}^{n} \log \left\{ \left(2\pi\hat{\sigma}_i^2\right)^{-1/2} \exp\left[-\frac{1}{2}\hat{\sigma}_i^{-2} \left(y_{i'} - \boldsymbol{x}'_{i'}\hat{\boldsymbol{\beta}}_{i'} \right)^2 \right] \right\}^{w_{ii'}} + 2\text{df}_i && (23) \\
&= \sum_{i'=1}^{n} w_{ii'} \left\{ \log\left(2\pi\right) + \log\hat{\sigma}_i^2 + \hat{\sigma}_i^{-2} \left(y_{i'} - \boldsymbol{x}'_{i'}\hat{\boldsymbol{\beta}}_{i'} \right)^2 \right\} + 2\text{df}_i && (24) \\
&= \hat{\sigma}_i^{-2} \sum_{i'=1}^{n} w_{ii'} \left(y_{i'} - \boldsymbol{x}'_{i'}\hat{\boldsymbol{\beta}}_{i} \right)^2 + 2\text{df}_i + C_i && (25)
\end{aligned}$$

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. The Maximum-Likelihood Estimate (MLE) of σ_i^2 is found by differentiating the local likelihood with respect to σ_i^2:

$$\begin{aligned}
\left.\frac{\partial \ell_i}{\partial \sigma_i^2}\right|_{\hat{\beta}_i} &= -\frac{1}{2} \sum_{i'=1}^{n} w_{ii'} \left\{ \left(\sigma_i^2\right)^{-1} - \left(\sigma_i^2\right)^{-2} \left(y_i - \boldsymbol{x}'_i\hat{\boldsymbol{\beta}}_i \right)^2 \right\} && (26) \\
\hat{\sigma}_i^2 &= \left(\sum_{i'=1}^{n} w_{ii'} \right)^{-1} \sum_{i'=1}^{n} w_{ii'} \left(y_i - \boldsymbol{x}'_i\hat{\boldsymbol{\beta}}_i \right) && (27)
\end{aligned}$$

3.2. Bandwidth selection

The bandwidth is selected to minimize the total AIC (AIC_{tot}). Because of the kernel weights and the application of the Adaptive LASSO, the sample size and degrees of freedom are different at each location. The total AIC is found by taking the sum over all of the observed data:

$$
\begin{aligned}
\text{AIC}_{\text{tot}} &= -2 \times \sum_{i=1}^{n} \ell_i + 2 \times \text{df} &(28)\\
&= \sum_{i=1}^{n} \left\{ \log \hat{\sigma}_i^2 + \hat{\sigma}_i^{-2} \left(y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_i \right)^2 \right\} + 2 \times \text{df} &(29)
\end{aligned}
$$

What remains is to calculate df, the number of degrees of freedom used by the model. Classical GWR, as developed in Loader (1999) and Fotheringham et al. (2002) calculates df using the trace of the "hat" matrix, but because the GAL is not a linear smoother, there is no "hat" matrix associated with GWR. Instead, notice that df can be pulled into the summation in (29):

$$
\text{df} = \sum_{i=1}^{n} \left(n^{-1} \text{df} \right) \tag{30}
$$

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

$$
\text{AIC}_{\text{tot}} = \sum_{i=1}^{n} \left\{ \log \hat{\sigma}_i^2 + \hat{\sigma}_i^{-2} \left(y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}_i \right)^2 + 2 \times \left(\sum_{i'=1}^{n} w_{ii'} \right)^{-1} \text{df}_i \right\} \tag{31}
$$

The bandwidth that minimizes (31) is found by a line search.

3.3. Confidence interval construction

Confidence intervals for the GAL's coefficient estimates can be calculated either by the bootstrap (Efron and Tibshirani, 1986) or by exploiting an assumption of normally-distributed residuals. The, e.g., 95% confidence interval for each regression coefficient is then the (2.5, 97.5) percentiles of the coefficient estimates from the bootstrap replicates.

3.3.1. Bootstrap confidence interval

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 GAL 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 GAL is applied independently to each bootstrap replicate, the variables selected by GAL may be different for each replicate.

Unshrunk coefficient estimates are found by using the GAL at each location for variable selection only and then estimating the coefficients for the selected variables by GWR. 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.

3.3.2. Normal approximation-based confidence interval

A third way to estimate the coefficient confidence intervals is to use the GAL for variable selection only and then to use GWR to calculate a confidence interval based on the assumption of an independent, identically distributed, Gaussian error structure. In this case, the standard error of the regression coefficients is

$$\hat{\text{se}}_{\beta_i} = \left(\tilde{\boldsymbol{X}}_i' \boldsymbol{W}_i \tilde{\boldsymbol{X}}_i\right)^{-1} \tilde{\boldsymbol{X}}_i' \boldsymbol{W}_i \boldsymbol{Y} \tag{32}$$

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

4. Simulation

4.1. Simulation setup

A simulation study was conducted to assess the finite-sample properties of the method described in Sections **??**-3. 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 $Cov\left(Z_{ji}, Z_{ji'}\right) = \sigma_z^2 \exp\left(-\tau_z^{-1} \|s_i - s_{i'}\|\right)$ where $\sigma_z^2 = 1$ is the variance and τ_z is a range parameter. 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 $y_i = \boldsymbol{z}_i' \boldsymbol{\beta}_i + \epsilon_i$ for $i = 1, \ldots, 900$ where the vector of additive errors $\boldsymbol{\epsilon}$ is generated from a GRF with spatial covariance $Cov\left(\epsilon_i, \epsilon_{i'}\right) = \sigma_\epsilon^2 \exp\left(-\tau_\epsilon^{-1} \|s_i - s_{i'}\|\right)$ where $\sigma_\epsilon^2 = 1$.

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 GAL. The coefficient surface of β_1 is described by the "step" function:

$$\beta_1(s) = \begin{cases} 0 & \text{if } s_y < 0.4 \\ 5(s_y - 0.4) & \text{if } 0.4 \leqslant s_y < 0.6 \\ 1 & \text{o.w.} \end{cases} \tag{33}$$

.

In order to evaluate the performance of GAL under a range of conditions, the data was simulated under 18 different settings (Table 1): high (0.1) and low (0.03) levels of τ_z, the autoregression range parameter for the covariate GRFs $Z_1, \ldots, Z_5$; three levels (0, 0.5, 0.8) of between-covariate correlation ρ; and three levels (0, 0.03, 0.1) of the autoregression range parameter τ_ϵ for the error-term GRF $\boldsymbol{\epsilon}$. Each case was simulated 100 times.

For measuring performance, we look at the pointwise selection frequency of $\beta_1, \ldots, \beta_5$ and the coverage frequency of β_1 (for a nominal 95% confidence interval).

As a baseline, ordinary GWR was used to estimate the coefficients using the same data but under an "oracle" setting (oracular GWR, or O-GWR), meaning that GWR was provided with the exactly correct set of predictors as used in the data-generating process. The ratio of the coverage frequency of the GAL to the coverage frequency of O-GWR is called the relative efficiency of the GAL.

	τ_z	ρ	τ_ϵ
1	0.03	0.00	0.00
2	0.03	0.00	0.03
3	0.03	0.00	0.10
4	0.03	0.50	0.00
5	0.03	0.50	0.03
6	0.03	0.50	0.10
7	0.03	0.80	0.00
8	0.03	0.80	0.03
9	0.03	0.80	0.10
10	0.10	0.00	0.00
11	0.10	0.00	0.03
12	0.10	0.00	0.10
13	0.10	0.50	0.00
14	0.10	0.50	0.03
15	0.10	0.50	0.10
16	0.10	0.80	0.00
17	0.10	0.80	0.03
18	0.10	0.80	0.10

Table 1: Simulation parameters for each setting.

4.2. Simulation results

Results of the simulation experiment were summarized to asses the consistency in selection and estimation, as well as the coverage properties of the confidence intervals. The confidence intervals based on the bootstrap (without shrinkage) were used for the GAL because they seemed to uniformly outperform the other options.

AL	AL-Unshrunk	AL-Precon	AL-Precon-Unshrunk	Oracle
0.023	0.017	0.024	*0.017*	**0.012**
0.029	*0.023*	0.030	**0.023**	0.031
0.027	*0.023*	0.030	**0.023**	0.056
0.028	0.023	0.030	*0.022*	**0.012**
0.038	0.034	0.040	*0.034*	**0.029**
0.032	*0.028*	0.038	**0.028**	0.057
0.041	0.039	0.042	*0.036*	**0.012**
0.077	0.081	0.081	*0.077*	**0.030**
0.062	0.060	0.076	*0.060*	**0.055**
0.026	0.023	0.026	*0.023*	**0.016**
0.040	**0.039**	0.041	*0.040*	0.061
0.057	**0.050**	0.067	*0.051*	0.125
0.029	0.026	0.030	*0.025*	**0.016**
0.055	0.056	0.057	**0.055**	0.059
0.078	**0.074**	0.094	*0.076*	0.130
0.046	0.046	0.045	*0.042*	**0.016**
0.119	0.134	0.122	0.132	**0.063**
0.167	*0.161*	0.210	0.166	**0.125**

4.3. Figures

Figures 1 - 18 show the frequency with which the true value of the parameter β_1 was covered by the 95% confidence intervals at each location under each simulation setting. The left column shows the coverage frequency of the 95% CI of the GAL using the unshrunk-bootstrap method of CI construction. The middle column is the coverage frequency of the 95% CI the O-GWR using the bootstrap to generate the CI. The right column is the relative efficiency of the GAL to O-GWR.

	AL	AL-Unshrunk	AL-Precon	AL-Precon-Unshrunk	Oracle
1	0.968	**0.934**	0.965	0.936	0.964
2	0.886	0.783	0.871	0.784	**0.714**
3	0.896	0.601	0.858	0.600	**0.167**
4	0.975	0.945	0.973	**0.944**	0.972
5	0.900	0.801	0.886	0.800	**0.726**
6	0.875	0.596	0.841	0.593	**0.161**
7	0.967	0.940	0.965	**0.938**	0.965
8	0.887	0.786	0.872	0.785	**0.724**
9	0.890	0.609	0.853	0.605	**0.165**
10	1.009	**0.967**	0.986	0.969	0.976
11	0.880	0.789	0.840	0.791	**0.722**
12	0.707	0.496	0.605	0.496	**0.171**
13	1.003	**0.965**	0.981	0.966	0.968
14	0.880	0.796	0.842	0.797	**0.716**
15	0.711	0.504	0.609	0.504	**0.167**
16	1.009	0.976	0.990	**0.975**	0.979
17	0.881	0.798	0.840	0.796	**0.723**
18	0.692	0.482	0.588	0.481	**0.169**

In the first two columns, the color white is used to indicate areas where the nominal coverage frequency of 95% is achieved, while blue codes areas that exceeded 95% coverage and orange codes areas that fell short of 95% coverage. In the third column, the color white indicates areas where the relative efficiency is unity, while orange indicates areas where the relative efficiency was less than unity and blue indicates areas where the relative efficiency exceeded unity.

	original	preconditioned
1	**0.623**	0.622
2	**0.447**	0.430
3	**0.278**	0.238
4	0.560	**0.568**
5	**0.423**	0.411
6	**0.256**	0.207
7	0.502	**0.516**
8	**0.395**	0.387
9	**0.230**	0.204
10	0.541	**0.550**
11	**0.298**	0.293
12	**0.078**	0.060
13	0.535	**0.554**
14	**0.285**	0.279
15	**0.076**	0.060
16	0.458	**0.481**
17	**0.254**	0.250
18	**0.079**	0.065

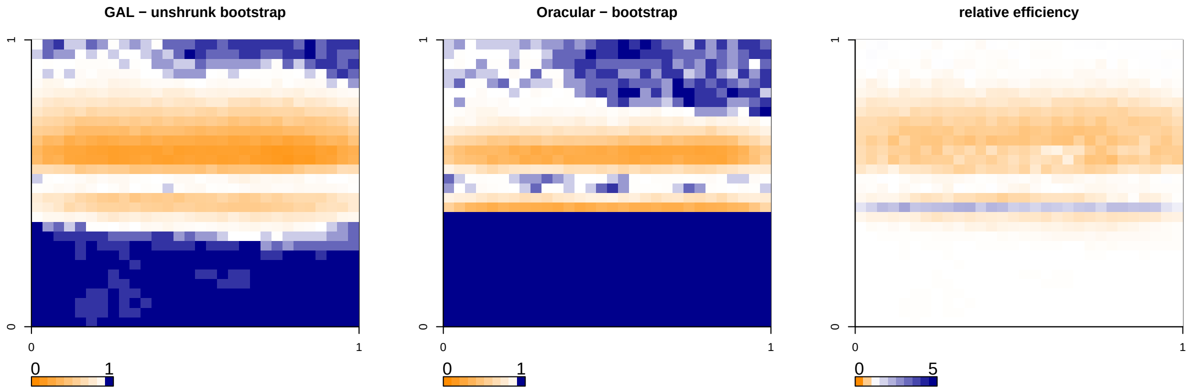

Figure 1: Coverage frequency of 95% CIs: setting 1

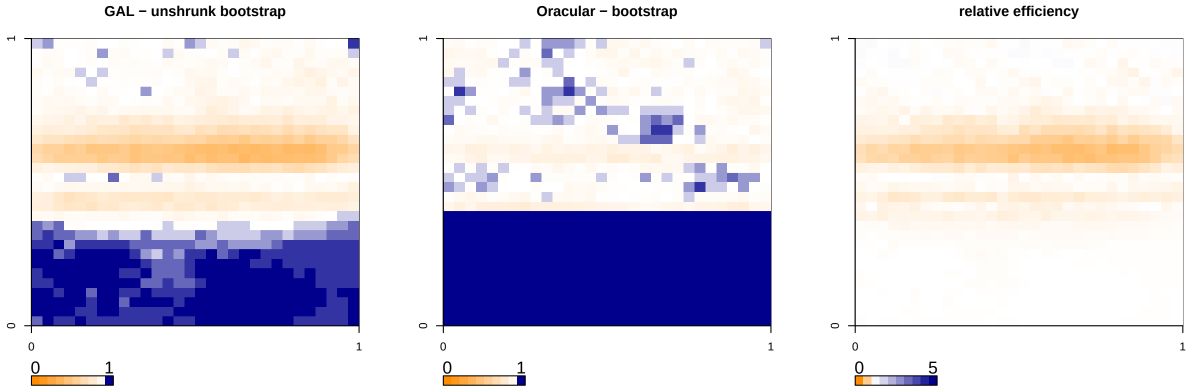

Figure 2: Coverage frequency of 95% CIs: setting 2

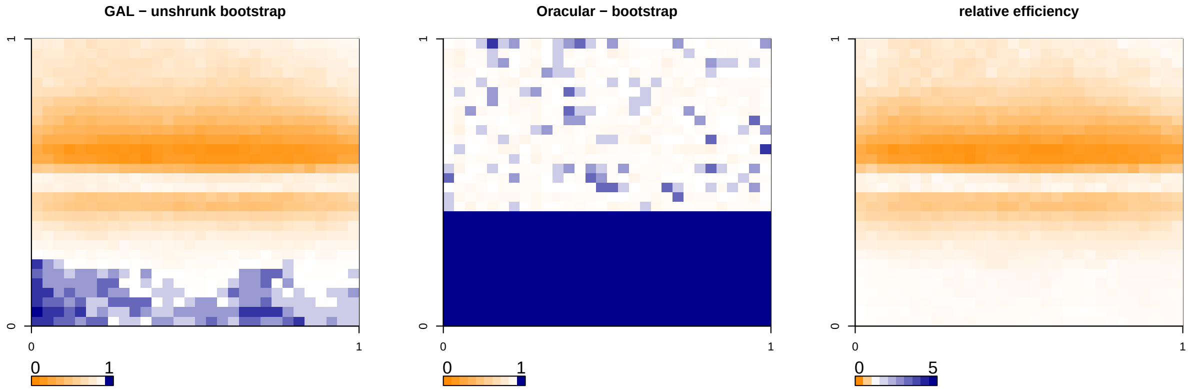

Figure 3: Coverage frequency of 95% CIs: setting 3

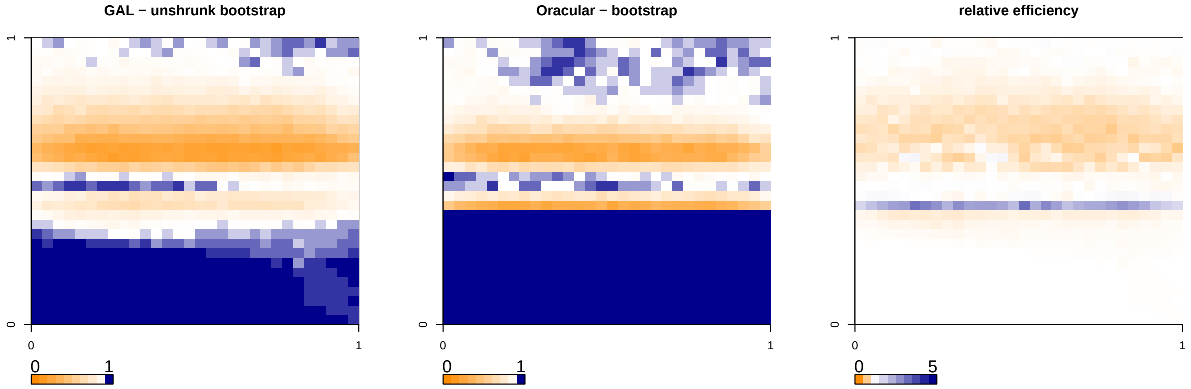

Figure 4: Coverage frequency of 95% CIs: setting 4

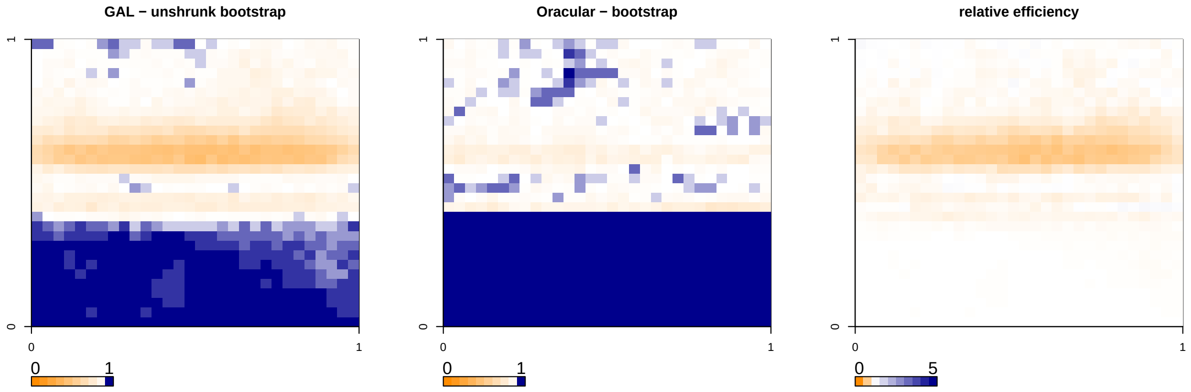

Figure 5: Coverage frequency of 95% CIs: setting 5

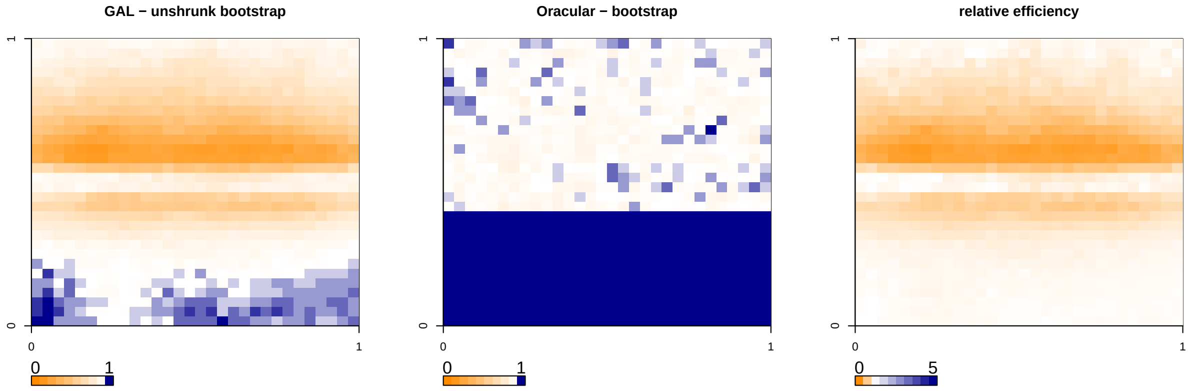

Figure 6: Coverage frequency of 95% CIs: setting 6

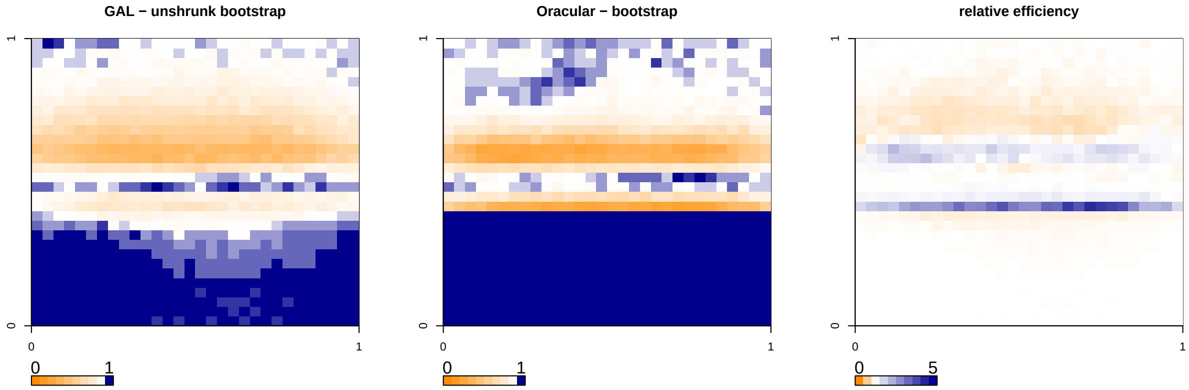

Figure 7: Coverage frequency of 95% CIs: setting 7

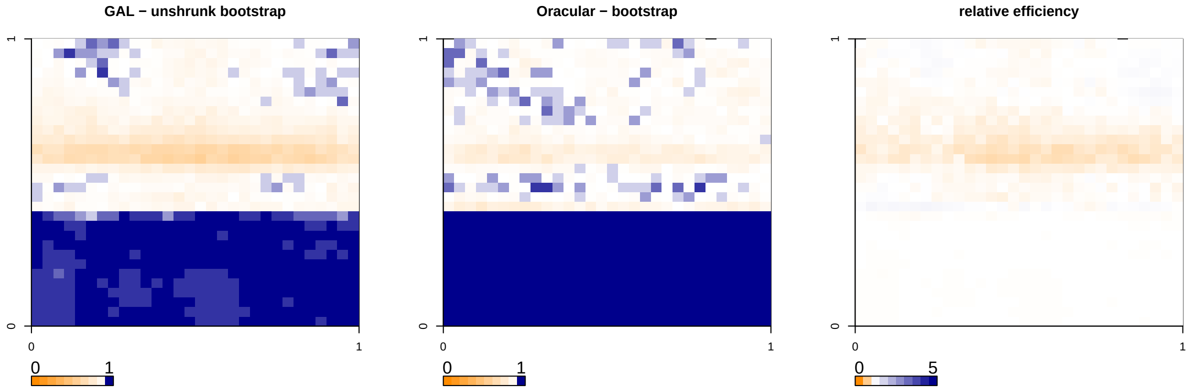

Figure 8: Coverage frequency of 95% CIs: setting 8

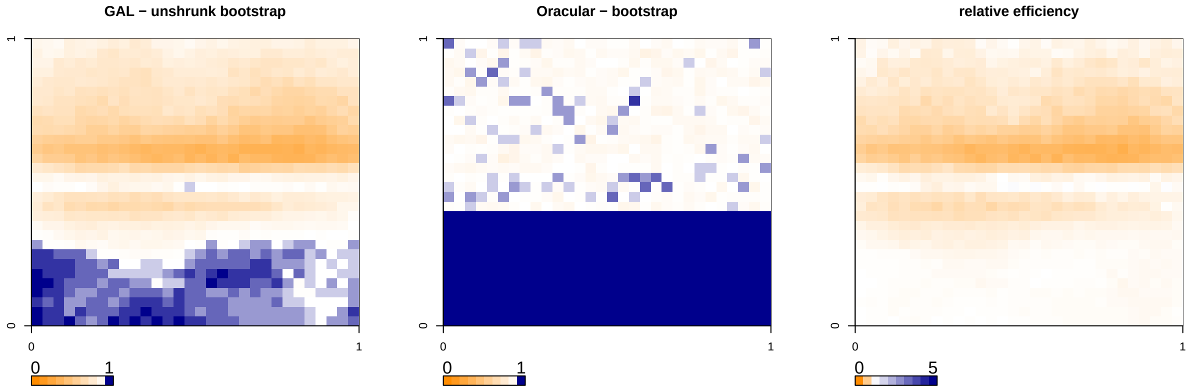

Figure 9: Coverage frequency of 95% CIs: setting 9

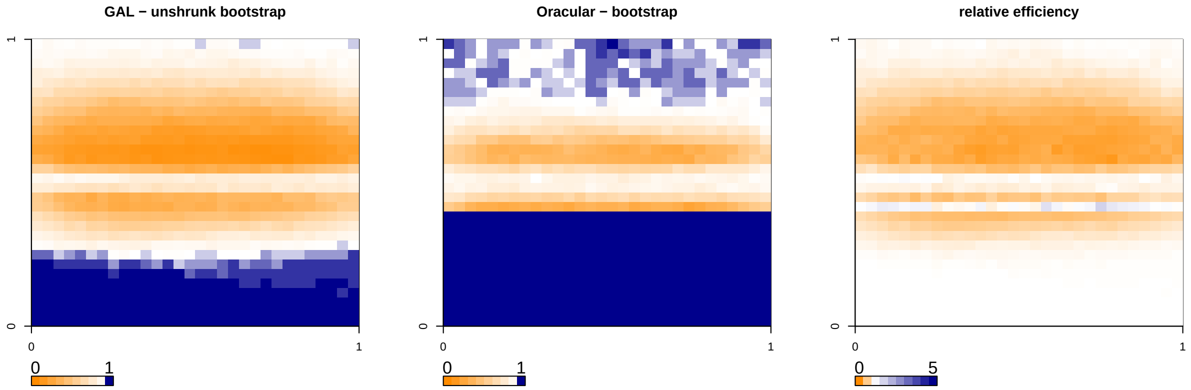

Figure 10: Coverage frequency of 95% CIs: setting 10

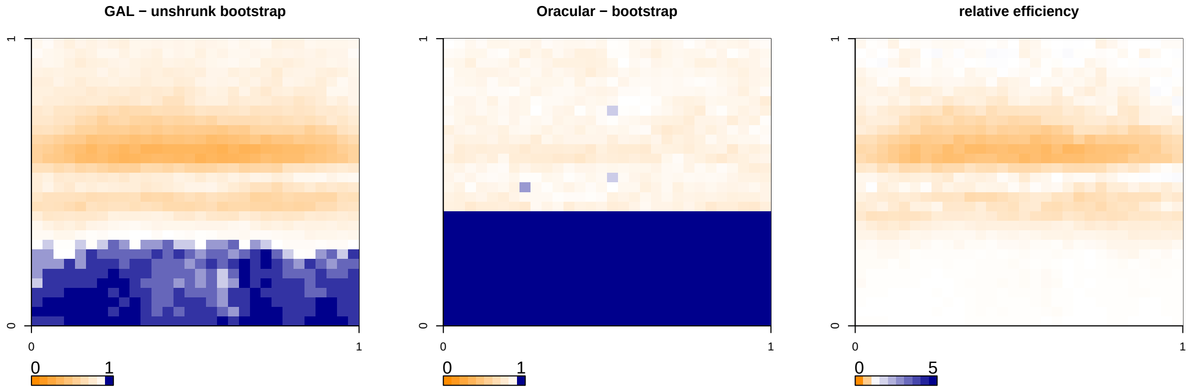

Figure 11: Coverage frequency of 95% CIs: setting 11

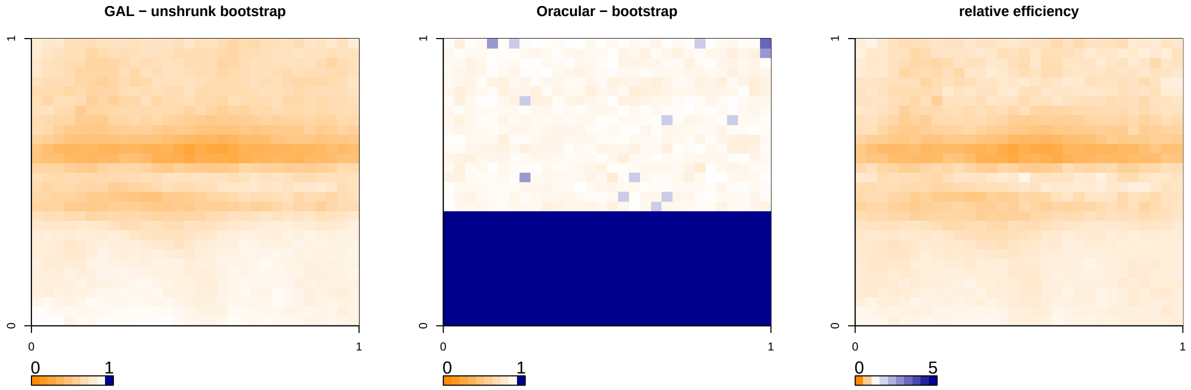

Figure 12: Coverage frequency of 95% CIs: setting 12

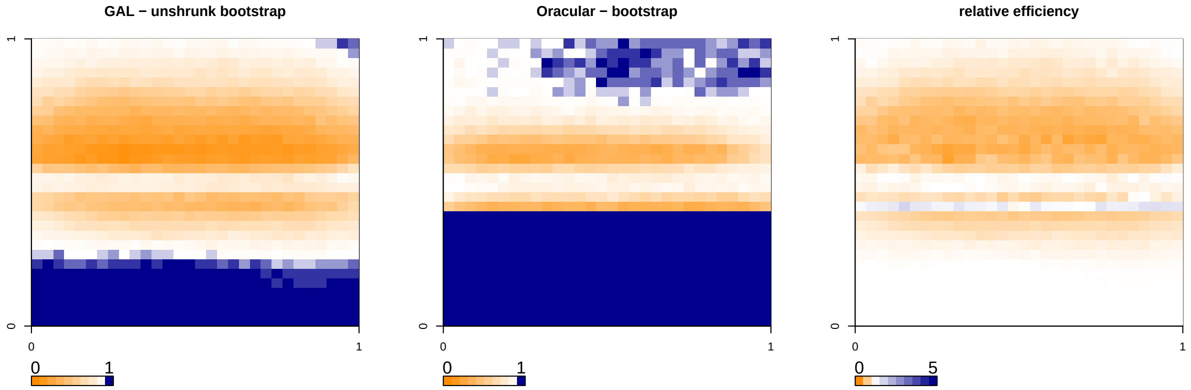

Figure 13: Coverage frequency of 95% CIs: setting 13

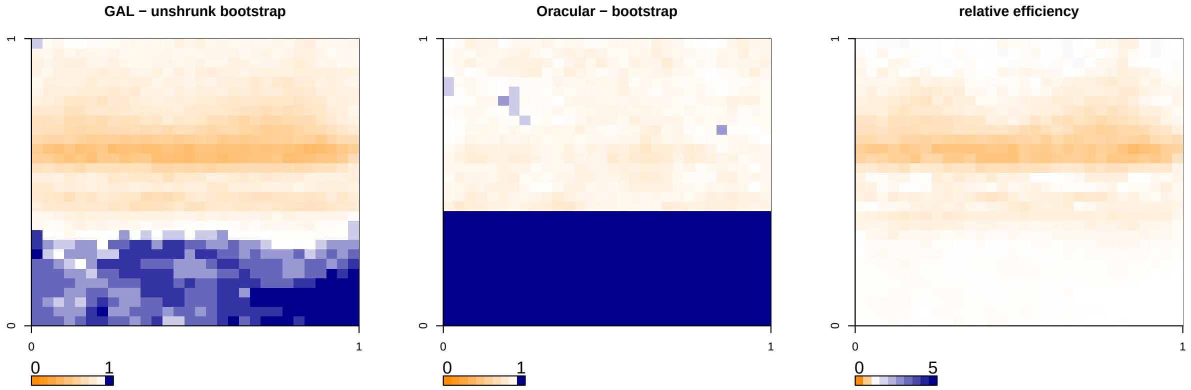

Figure 14: Coverage frequency of 95% CIs: setting 14

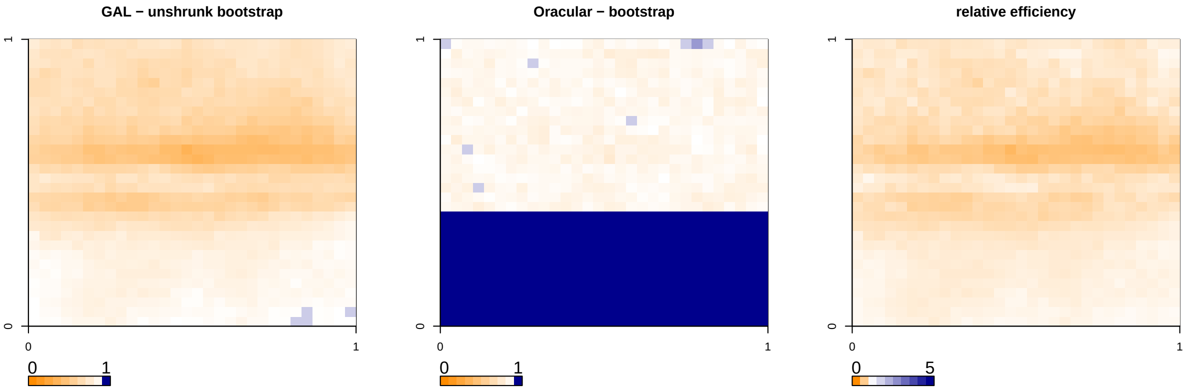

Figure 15: Coverage frequency of 95% CIs: setting 15

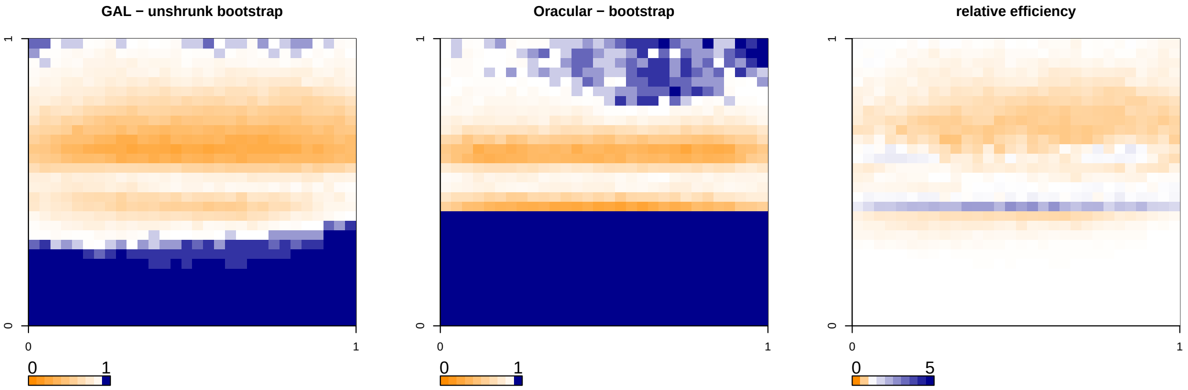

Figure 16: Coverage frequency of 95% CIs: setting 16

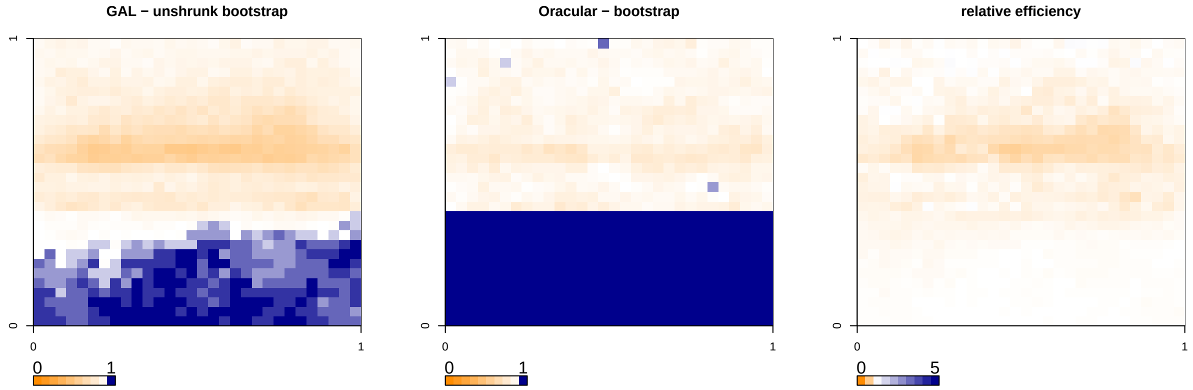

Figure 17: Coverage frequency of 95% CIs: setting 17

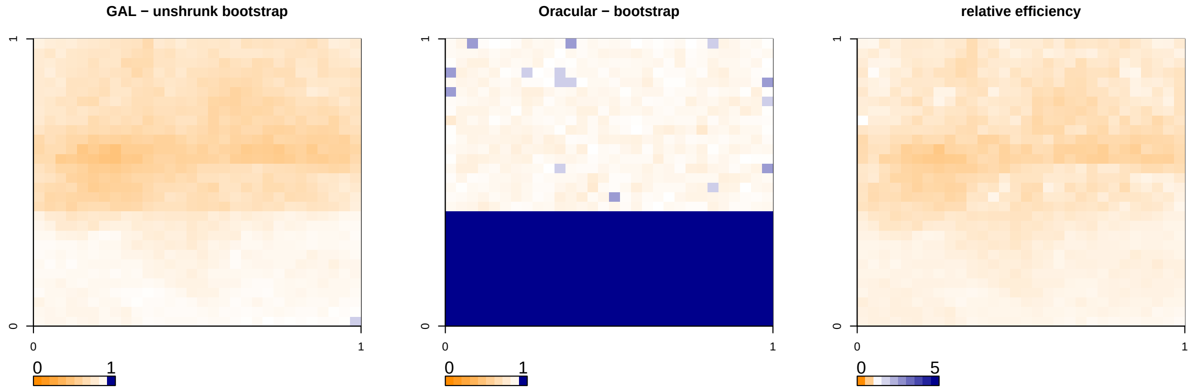

Figure 18: Coverage frequency of 95% CIs: setting 18

5. Data analysis

5.1. Census poverty data

We present the following analysis to demonstrate one possible application of the geographically-weighted lasso in a linear regression context. We use county-level data from the US Census Bureau to select the social and demographic variables that are important predictors of the county-level poverty rate in the upper midwest, and to estimate the coefficients associated with these predictors. Data are from six censuses - the decennial censuses from 1960 to 2000, and from the American

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 2: Description of the variables used in the census-data example

Community Survey in 2006. Selection and estimation are done for each census individually (no attempt is made here to borrow strength across years). The outcome of interest (poverty rate) is a proportion and so takes values on $[0, 1]$, but to demonstrate the GAL in a linear regression context, we model the logit-transformed poverty rate. Our data set covers all counties in the states of Minnesota, Iowa, Wisconsin, Illinois, Indiana, and Michigan. The potential predictors are described in Table 2.

5.2. Figures

The coefficient estimates are plotted on maps of the upper midwest in Figures 19 - 24. It is immediately apparent that the estimated coefficient surfaces are non-constant for most variables.

6. References

References

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

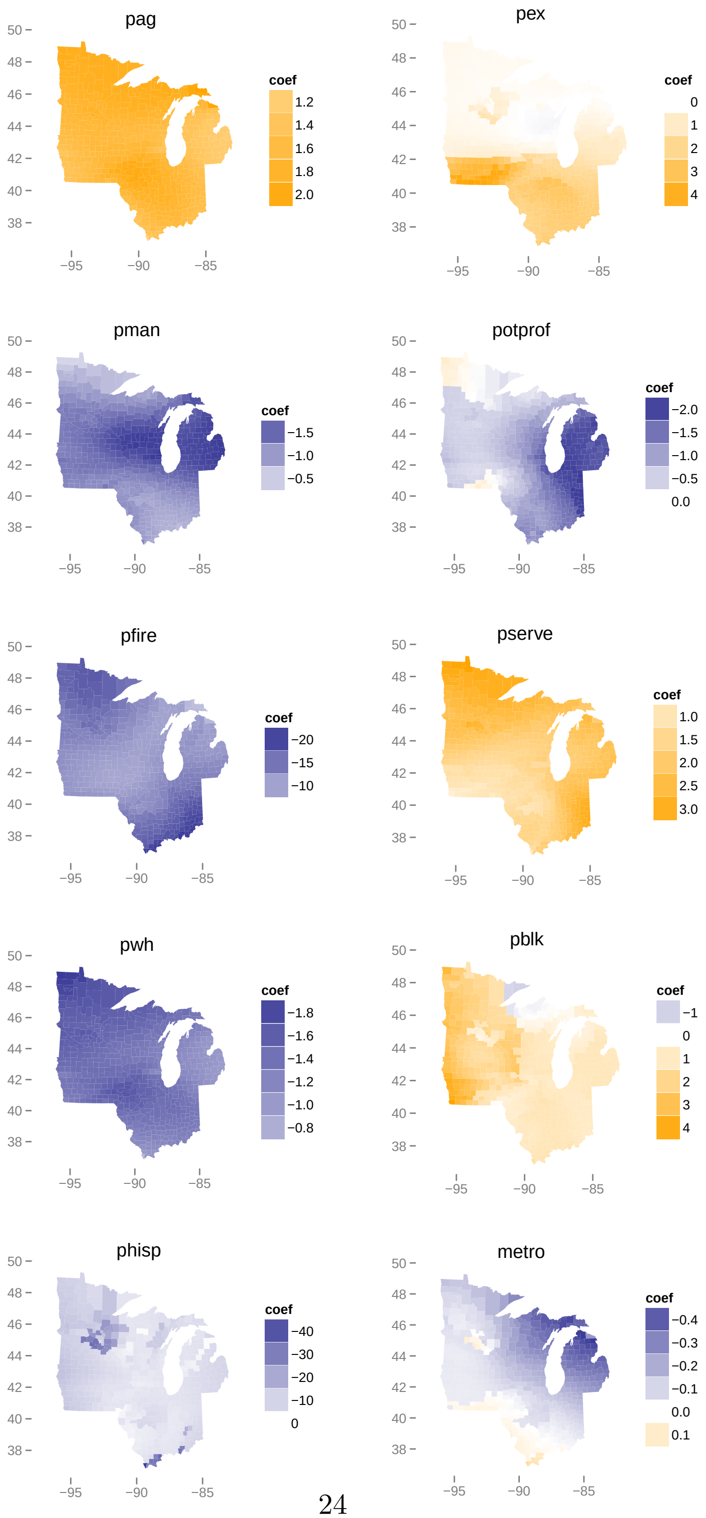

Figure 19: Estimated coefficient surfaces for the 1960 census.

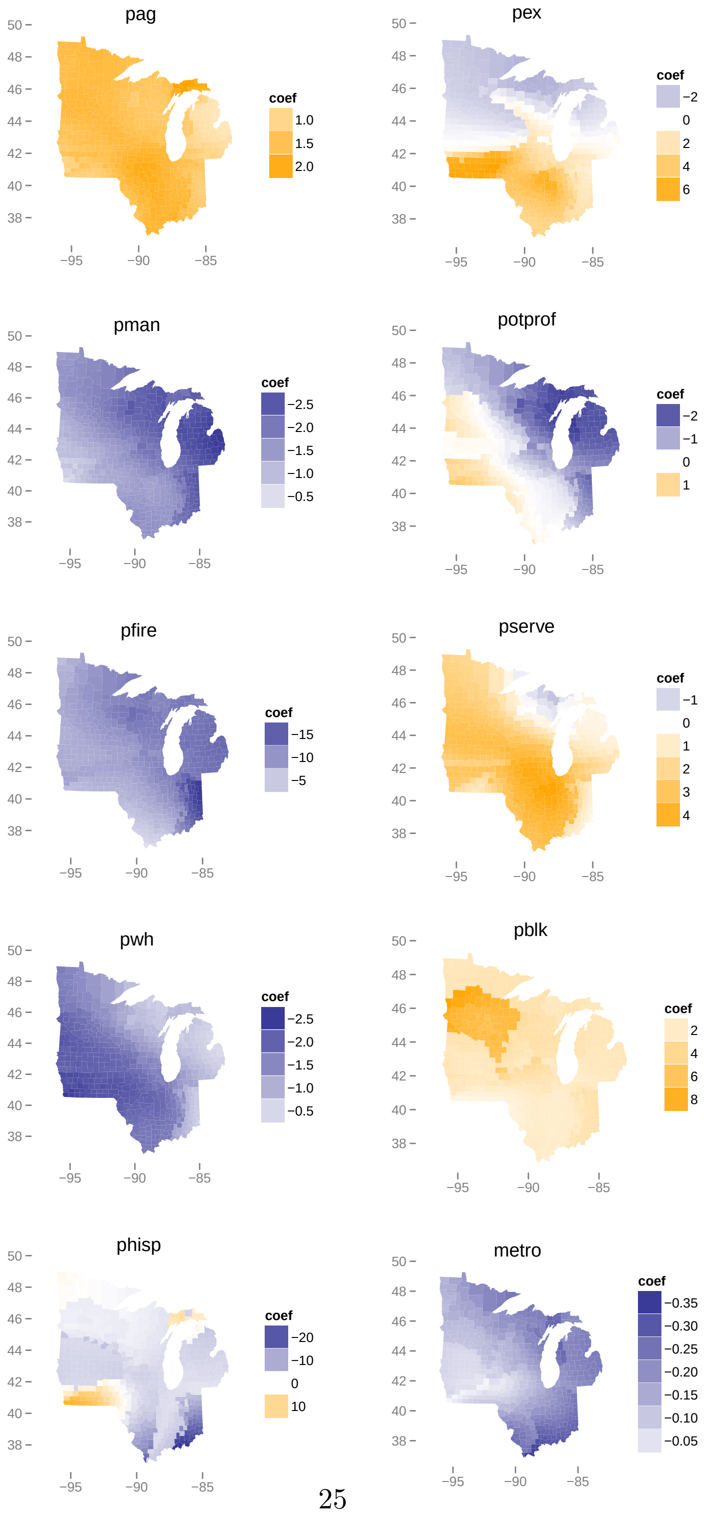

Figure 20: Estimated coefficient surfaces for the 1970 census.

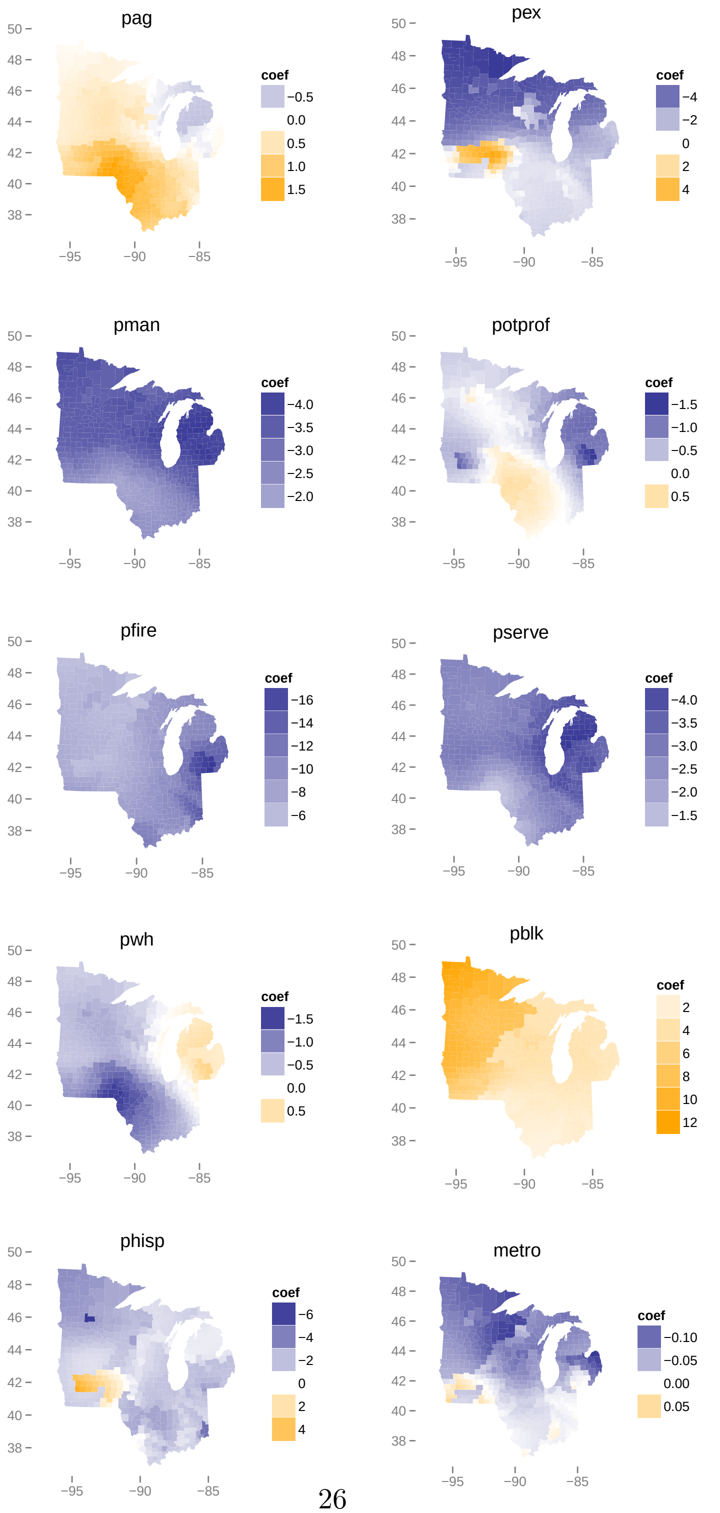

Figure 21: Estimated coefficient surfaces for the 1980 census.

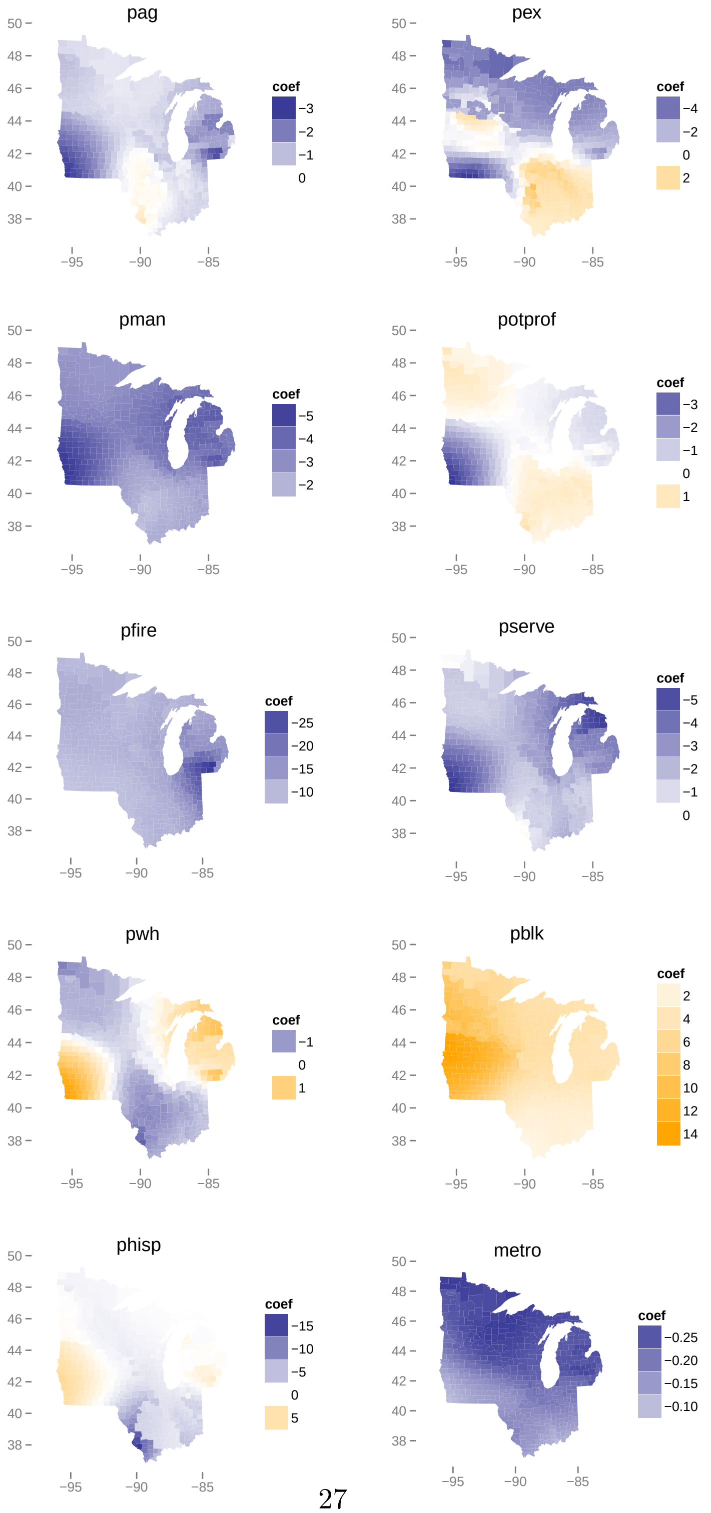

Figure 22: Estimated coefficient surfaces for the 1990 census.

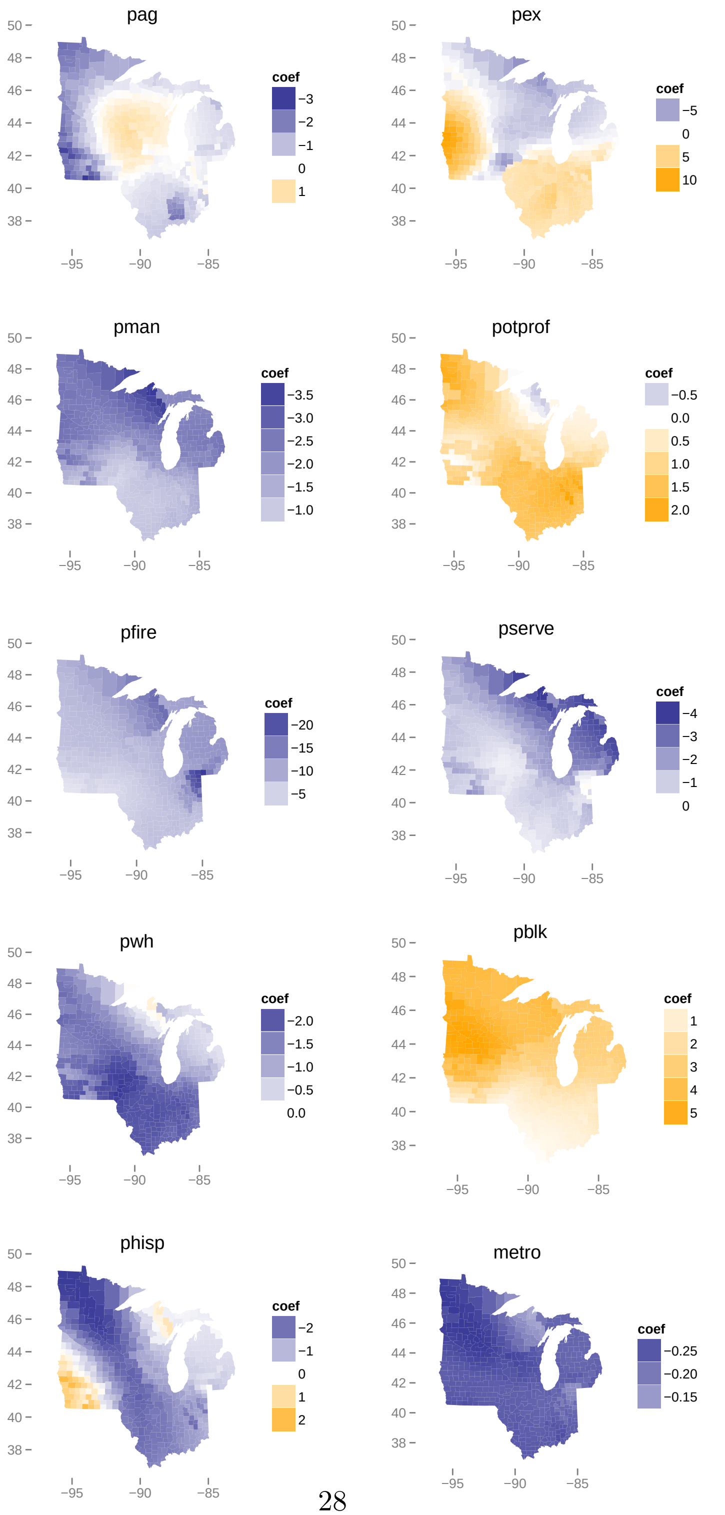

Figure 23: Estimated coefficient surfaces for the 2000 census.

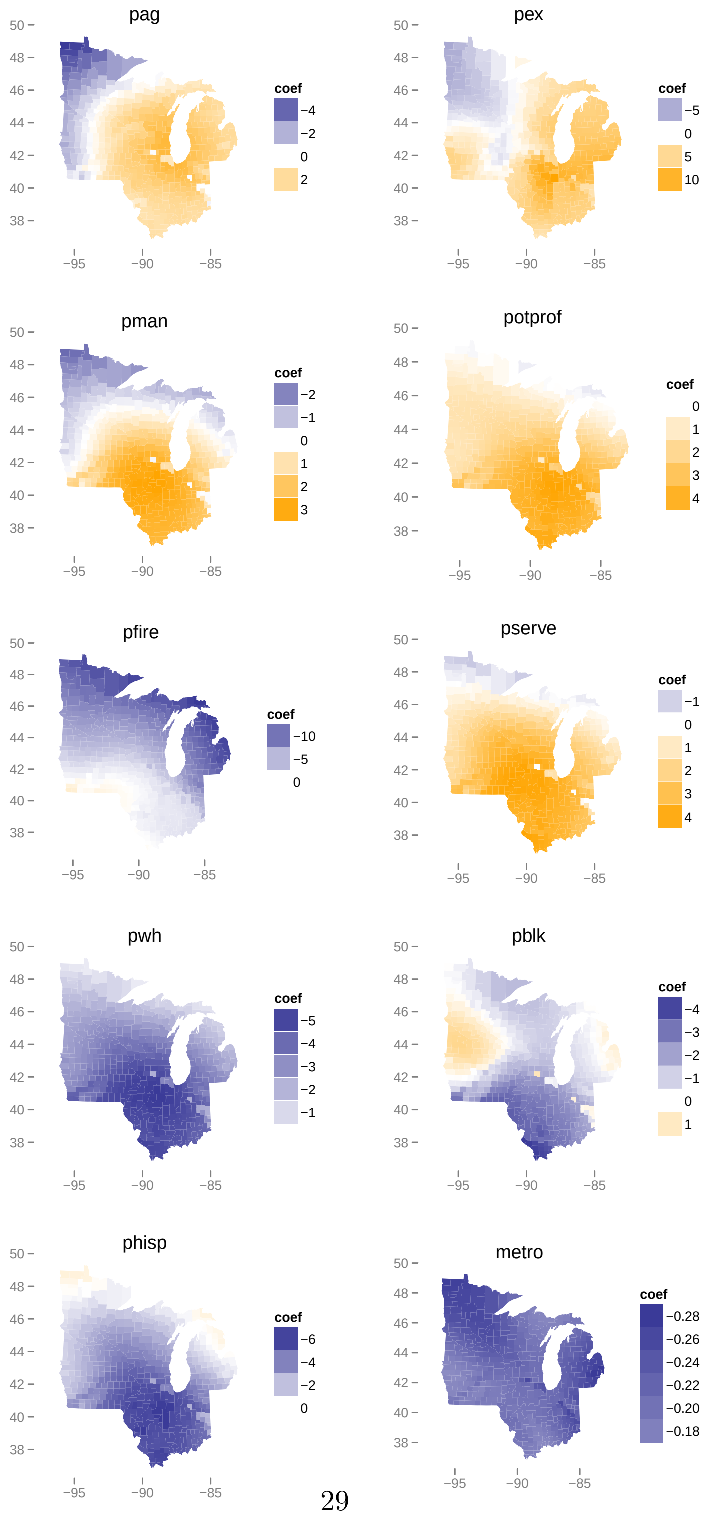

Figure 24: Estimated coefficient surfaces for the 2006 census.

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 wregression 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.

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.

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.