

# Local variable selection and parameter estimation for spatially varying coefficient models

Wesley Brooks

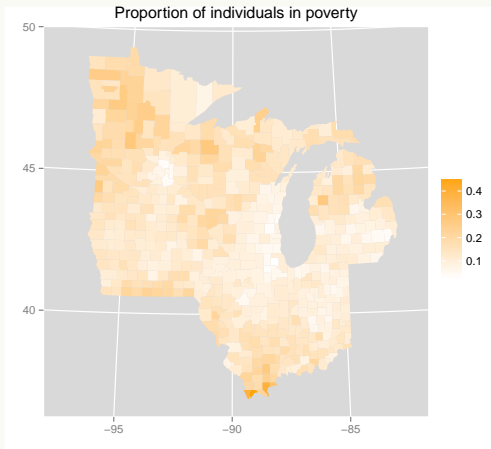
Department of Statistics  
University of Wisconsin–Madison

November 7, 2013

# Motivation

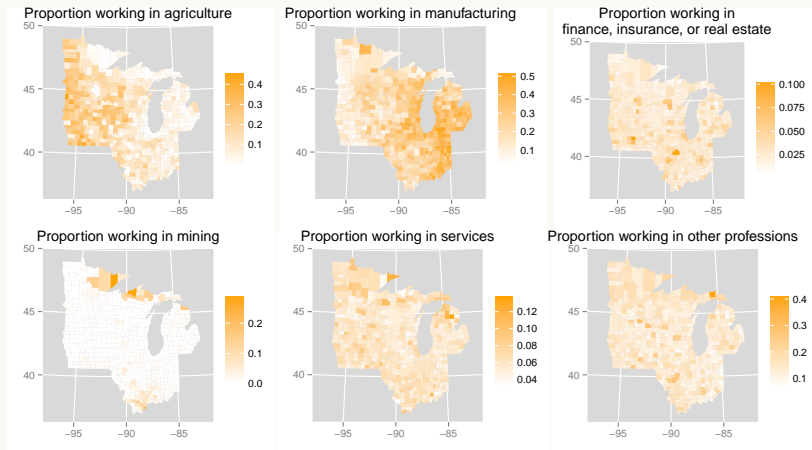
# Motivation

Take a look at some data



# Motivation

Take a look at some data



# Motivation

Sensible questions about the data

- ▶ Which of the economic-structure variables is associated with poverty rate?
- ▶ What are the sign and magnitude of that association?
- ▶ Is poverty rate associated with the same economic-structure variables across the entire region?
- ▶ Are the sign and magnitude of the associations constant across the region?

# Introduction

# Introduction

A review of existing methods

- ▶ Spatial regression
- ▶ Varying coefficient regression
  - Splines
  - Kernels
  - Wavelets
- ▶ Model selection via regularization

# Introduction

## Some definitions

- ▶ Univariate spatial response process  $\{Y(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$
- ▶ Multivariate spatial covariate process  $\{\mathbf{X}(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$
- ▶  $n$  = number of observations
- ▶  $p$  = number of covariates
- ▶ Location (2-dimensional)  $\mathbf{s}$
- ▶ Spatial domain  $\mathcal{D}$



# Introduction

## Further definitions

### ► Geostatistical data:

- Observations are made at sampling locations  $s_i$  for  $i = 1, \dots, n$
- E.g. elevation, temperature

### ► Areal data:

- Domain is partitioned into  $n$  regions  $\{D_1, \dots, D_n\}$
- The regions do not overlap, and they divide the domain completely:  $\mathcal{D} = \bigcup_{i=1}^n D_i$
- Sampling locations  $s_i$  for  $i = 1, \dots, n$  are the centroids of the regions
- E.g. poverty rate, population, spatial mean temperature

# Introduction

Spatial regression (Cressie, 1993)

- ▶ The typical spatial regression

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

$\text{cov}(W(\mathbf{s}), W(\mathbf{t}))$ : Matèrn class

- ▶  $W(\mathbf{s})$  is a spatial random effect that accounts for autocorrelation in the response variable
- ▶ The coefficients  $\boldsymbol{\beta}$  are constant
- ▶ Relies on *a priori* global variable selection

# Introduction

Spatially varying coefficient process (Gelfand et al., 2003)

- ▶ Making model more flexible: coefficients in a spatial regression model can be allowed to vary

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

- ▶  $\{\beta_1(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}, \dots, \{\beta_p(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$  are stationary spatial processes with Matérn covariance functions
- ▶ Still relies on *a priori* global variable selection

# Introduction

Varying coefficients regression (Hastie and Tibshirani, 1993)

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

- ▶ Assume an effect modifying variable  $S$
- ▶ Coefficients are functions of  $S$

# Introduction

Spline-based VCR models (Wood, 2006)

- ▶ Splines are a way to parameterize smooth functions
- ▶ Estimate the varying coefficients via splines:
  - $E\{Y(t)\} = \beta_1(t)X_1(t) + \cdots + \beta_p(t)X_p(t)$

# Introduction

Global selection in spline-based VCR models

Regularization methods for global variable selection in VCR models:

- ▶ The integral of a function squared (e.g.  $\int \{f(t)\}^2 dt$ ) is zero if and only if the function is zero everywhere.
- ▶ Use regularization to encourage coefficient functions to be zero
  - SCAD penalty (Wang, Li, and Huang, 2008)
  - Non-negative garrote penalty (Antoniadas, Gijbels, and Verhasselt, 2012)

# Introduction

## Wavelet methods for VCR models

- ▶ Wavelet methods: decompose coefficient function into local frequency components
- ▶ Selection of nonzero local frequency components with nonzero coefficients:
  - Bayesian variable selection (Shang, 2011)
  - Lasso (J. Zhang and Clayton, 2011)
- ▶ Sparsity in the local frequency components; not in the local covariates

## Geographically weighted regression



# Geographically weighted regression

(Brundson, S. Fotheringham, and Martin Charlton, 1998;  
A. Fotheringham, Brunsdon, and M. Charlton, 2002)

- ▶ Consider observations at sampling locations  $\mathbf{s}_1, \dots, \mathbf{s}_n$
- ▶  $y(\mathbf{s}_i) = y_i$  the univariate response at location  $\mathbf{s}_i$
- ▶  $\mathbf{x}(\mathbf{s}_i) = \mathbf{x}_i$  the  $(p + 1)$ -variate vector of covariates at location  $\mathbf{s}_i$
- ▶ Assume  $y_i = \mathbf{x}_i' \boldsymbol{\beta}_i + \varepsilon_i$  where  $\varepsilon \stackrel{iid}{\sim} \mathcal{N}(0, \sigma^2)$

# Geographically weighted regression

(Brundson, S. Fotheringham, and Martin Charlton, 1998;

A. Fotheringham, Brunsdon, and M. Charlton, 2002)

- ▶ The total log likelihood is

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

- ▶ With  $n$  observations and  $n(p+1)$  free parameters, the model is not identifiable.
- ▶ Estimate parameters by borrowing strength from nearby observations

# Geographically weighted regression

Local regression (Loader, 1999)

Local regression uses a kernel function at each sampling location to weight observations based on their distance from the sampling location.

$$L_i = \prod_{i'=1}^n (L_{i'})^{w_{ii'}}$$
$$\ell_i = \sum_{i'=1}^n w_{ii'} \left\{ \log \sigma_i^2 + \sigma_i^{-2} (y_{i'} - \mathbf{x}_{i'}' \boldsymbol{\beta}_i)^2 \right\}$$

Given the weights, a local model is fit at each sampling location using the local likelihood

# Geographically weighted regression

Local likelihood(Loader, 1999)

Weights are calculated via a kernel, e.g. the bisquare kernel:

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

Where

- ▶  $\phi$  is a bandwidth parameter
- ▶  $\delta_{ii'} = \delta(\mathbf{s}_i, \mathbf{s}_{i'}) = \|\mathbf{s}_i - \mathbf{s}_{i'}\|_2$  is the Euclidean distance

# Geographically weighted regression

Bandwidth estimation via the  $AIC_c$  (Hurvich, Simonoff, and Tsai, 1998)

- ▶ Smaller bandwidth: less bias, more flexible coefficient surface
- ▶ Large bandwidth: less variance, less flexible coefficient surface
- ▶ Estimate the degrees of freedom used in estimating the coefficient surface:
  - $\hat{y} = Hy$
  - $\nu = \text{tr}(H)$
- ▶ Then the corrected AIC for bandwidth selection is:
$$AIC_c = 2n \log \sigma + n \left\{ \frac{n+\nu}{n-2-\nu} \right\}$$

## Local variable selection and parameter estimation

# Geographically weighted regression

Geographically weighted Lasso (Wheeler, 2009)

Within a GWR model, using the Lasso for local variable selection is called the geographically weighted Lasso (GWL).

- ▶ The GWL requires estimating a Lasso tuning parameter for each local model
- ▶ Wheeler, 2009 estimates the local Lasso tuning parameter at location  $s_i$  by minimizing a jackknife criterion:  $|y_i - \hat{y}_i|$
- ▶ The jackknife criterion can only be calculated where data are observed, making it impossible to use the GWL to impute missing data or to estimate the value of the coefficient surface at new locations
- ▶ Also, the Lasso is known to be biased in variable selection and suboptimal for coefficient estimation

# Local variable selection and parameter estimation

## Geographically weighted adaptive Lasso (GWAL)

- ▶ Local variable selection in a GWR model using the adaptive Lasso (AL) (Zou, 2006)
- ▶ Under suitable conditions, the AL has an oracle property for selection

$$\begin{aligned}\mathcal{S}(\beta_i) &= -2\ell_i(\beta_i) + \mathcal{J}_2(\beta_i) \\ &= \sum_{i'=1}^n w_{ii'} \left\{ \log \sigma_i^2 + (\sigma_i^2)^{-1} (y_{i'} - \mathbf{x}_{i'}' \beta_i)^2 \right\} \\ &\quad + \lambda_i \sum_{j=1}^p |\beta_{ij}| / \gamma_{ij}\end{aligned}$$



# Local variable selection and parameter estimation

Geographically weighted adaptive elastic net (GWAL)

Note:

- ▶  $\sum_{i'=1}^n w_{ii'} (y_{i'} - \mathbf{x}_{i'}' \boldsymbol{\beta}_i)^2$  is the weighted sum of squares minimized by traditional GWR
- ▶  $\mathcal{J}_1(\boldsymbol{\beta}_i) = \lambda_i^* \sum_{j=1}^p |\beta_{ij}| / \gamma_{ij}$  is the AL penalty.

# Local variable selection and parameter estimation

## Geographically weighted adaptive elastic net (GWEN)

- ▶ Local variable selection in a GWR model using the adaptive elastic net (AEN) (Zou and H. Zhang, 2009)
- ▶ Under suitable conditions, the AEN has an oracle property for selection

$$\begin{aligned}\mathcal{S}(\beta_i) &= -2\ell_i(\beta_i) + \mathcal{J}_2(\beta_i) \\ &= \sum_{i'=1}^n w_{ii'} \left\{ \log \sigma_i^2 + (\sigma_i^2)^{-1} (y_{i'} - \mathbf{x}_{i'}' \beta_i)^2 \right\} \\ &\quad + \alpha_i \lambda_i \sum_{j=1}^p |\beta_{ij}| / \gamma_{ij} \\ &\quad + (1 - \alpha_i) \lambda_i^* \sum_{j=1}^p (\beta_{ij} / \gamma_{ij})^2\end{aligned}$$

# Local variable selection and parameter estimation

Geographically weighted adaptive elastic net (GWEN)

Note:



$$\begin{aligned}\mathcal{J}_2(\beta_i) = & \\ & \alpha_i \lambda_i^* \sum_{j=1}^p |\beta_{ij}| / \gamma_{ij} + \\ & (1 - \alpha_i) \lambda_i^* \sum_{j=1}^p (\beta_{ij} / \gamma_{ij})^2\end{aligned}$$

- ▶ This is the AEN penalty

# Local variable selection and parameter estimation

## Tuning parameter estimation

It is necessary to estimate an AL (or AEN) tuning parameter for each local model. Using the local BIC allows fitting a local model at any location within the domain

$$\begin{aligned}\text{BIC}_i &= -2 \sum_{i'=1}^n \ell_{ii'} + \log \left( \sum_{i'=1}^n w_{ii'} \right) \text{df}_i \\ &= \sum_{i'=1}^n w_{ii'} \left\{ \log(2\pi) + \log \hat{\sigma}_i^2 + \hat{\sigma}_i^{-2} \left( y_{i'} - \mathbf{x}_{i'}' \hat{\boldsymbol{\beta}}_{i'} \right)^2 \right\} \\ &\quad + \log \left( \sum_{i'=1}^n w_{ii'} \right) \text{df}_i\end{aligned}$$

# Local variable selection and parameter estimation

## Bandwidth parameter estimation

- ▶ Traditional GWR:

- $\hat{y} = Hy$
- So traditional GWR is a linear smoother
- $\nu = \text{tr}(H)$  is the degrees of freedom for the model

- ▶ GWAL:

- $\hat{y} = H^\dagger y + T^\dagger \gamma$

- ▶ GWEN:

- $\hat{y} = H^* y + T^* \gamma$

- ▶ Neither GWEN nor GWAL is a linear smoother

- df not equal to trace of projection matrix for GWAL, GWEN

- ▶ Solution: use GWEN or GWAL for selection then fit local model for the selected variables via traditional GWR

- Now  $\text{df} = \nu = \text{tr}(H)$

# Local variable selection and parameter estimation

## Locally linear coefficient estimation

- ▶ GWR, GWEN, GWAL: coefficients locally constant
  - as in Nadaraya-Watson kernel smoother
  - Leads to bias where there is a gradient at the boundary
- ▶ Solution: local polynomial modeling
  - First-order polynomial: locally linear coefficients
- ▶ Augment with covariate-by-location interactions
  - Two-dimensional
  - Augment with selected covariates only

## Simulation study

# Simulation study

## Simulating covariates

- ▶ Five covariates  $\tilde{X}_1, \dots, \tilde{X}_5$
- ▶ Gaussian random fields
- ▶  $30 \times 30$  grid on  $[0, 1] \times [0, 1]$
- ▶

$$\begin{aligned}\tilde{X}_j &\sim N(0, \Sigma) \text{ for } j = 1, \dots, 5 \\ \{\Sigma\}_{i,i'} &= \exp\{-\tau^{-1}\delta_{ii'}\} \text{ for } i, i' = 1, \dots, n\end{aligned}$$

- ▶ Colinearity:  $\rho$



# Simulation study

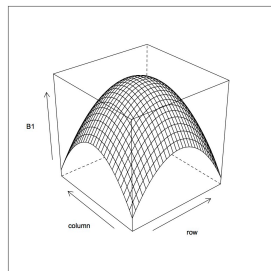
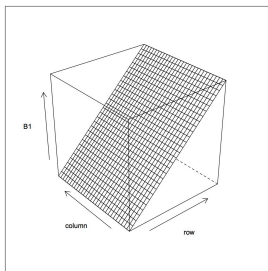
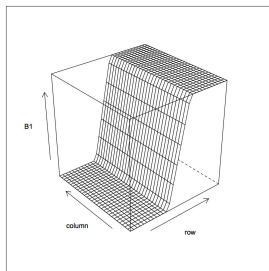
## Simulating the response

- ▶  $Y(\mathbf{s}) = X(\mathbf{s})'\boldsymbol{\beta}(\mathbf{s}) = \sum_{j=1}^5 \beta_j(\mathbf{s})X_j(\mathbf{s}) + \varepsilon(\mathbf{s})$
- ▶  $\varepsilon \sim iid \ N(0, \sigma^2)$
- ▶  $\beta_1(\mathbf{s})$ , the coefficient function for  $X_1$ , is nonzero in part of the domain.
- ▶ Coefficients for  $X_2, \dots, X_5$  are zero everywhere

# Simulation study

Coefficient functions

Call these functions step, gradient, and parabola:



# Simulation study

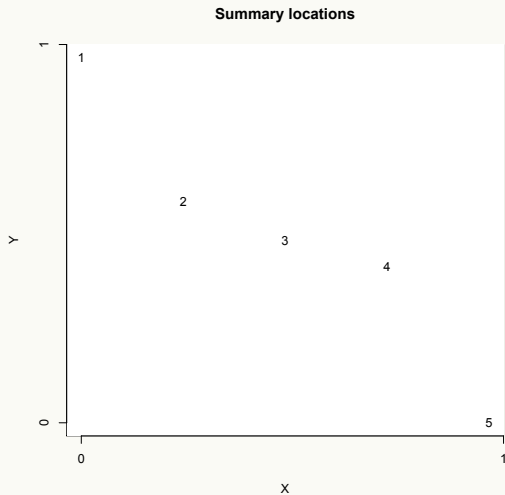
Simulation settings

Each setting simulated 100 times:

Setting	function	$\rho$	$\sigma^2$
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

# Simulation results

Selection



# Simulation results

## Selection performance

- ▶ Non-ambiguous locations (80):
  - 52 saw no false negatives
  - 72 had no false positives
  - 26 neither false positives nor false negatives
- ▶ Noise variance had more effect on selection than colinearity
- ▶ No difference between GWEN, GWAL

# Simulation results

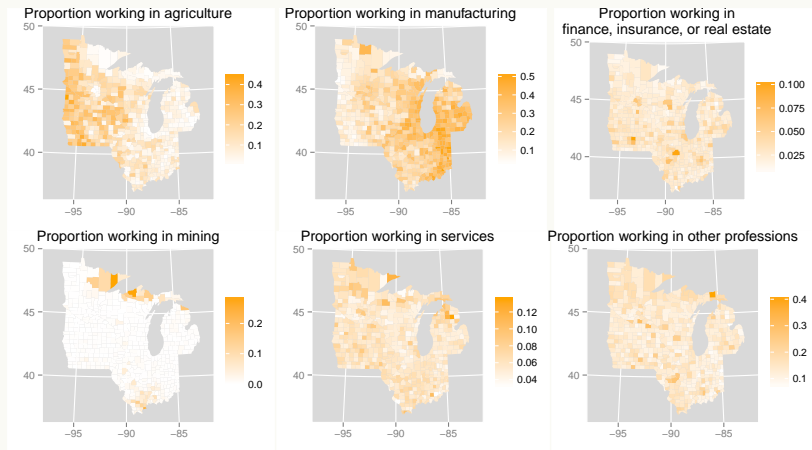
## Estimation performance

- ▶ Oracular selection
  - best  $\text{MSE}(\hat{\beta}_1)$  in 41 of the 60 cases
- ▶ Oracular selection
  - best  $\text{MSE}(\hat{\beta}_1)$  in 41 of the 60 cases
- ▶ Generally small difference between GWR, oracular, GWEN-LLE, and GWAL-LLE
- ▶ Noise variance had more effect on selection than collinearity
- ▶ No difference between GWEN, GWAL
- ▶ Fitting  $\hat{y}$ : best MSE split evenly between

Data example: poverty rate in the upper  
midwest

# Data example: poverty rate in the upper midwest

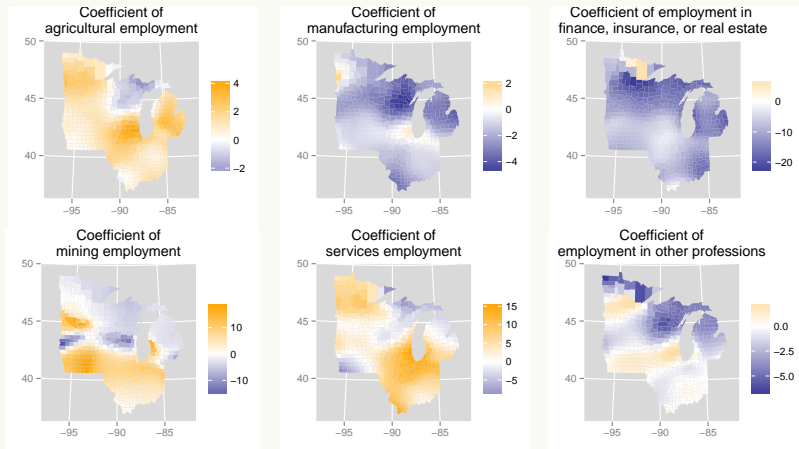
## Revisiting the introductory example





# Data example: poverty rate in the upper midwest

## Results from traditional GWR



# Data example: poverty rate in the upper midwest

## Data description

- ▶ Response: logit-transformed poverty rate in the Upper Midwest states of the U.S.
  - Minnesota, Iowa, Wisconsin, Illinois, Indiana, Michigan
- ▶ Covariates: employment structure (raw proportion employed in:)
  - agriculture
  - finance, insurance, and real estate
  - manufacturing
  - mining
  - services
  - other professions
- ▶ Data source: U.S. Census Bureau's decennial census of 1970

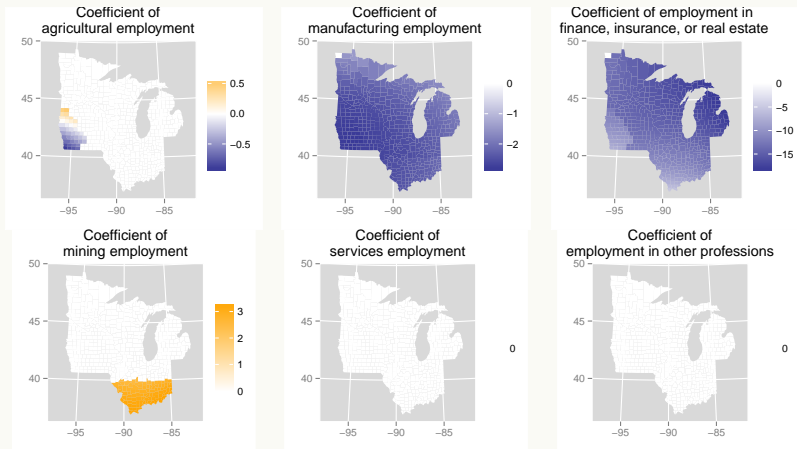
# Data example: poverty rate in the upper midwest

## Data description

- ▶ Data aggregated to the county level
  - counties are areal units
- ▶ county centroid treated as sampling location

# Data example: poverty rate in the upper midwest

## Results from GWEN



# Data example: poverty rate in the upper midwest

## Results from GWEN-LLE

- ▶ Relatively constant compared to GWR
- ▶ Services, "other professions" do not affect the poverty rate
- ▶ Manufacturing: negative coefficient everywhere
- ▶ Finance, insurance, and real estate negative coefficient everywhere
  - Largest magnitude (min: -20, next-largest: -3)
  - GWR comparable to GWEN-LLE
- ▶ Manufacturing: negative coefficient everywhere
  - GWR: coefficient greater than zero near Chicago and in NW Minnesota
- ▶ Agriculture: nonzero in western Iowa
  - North-south gradient to coefficient
  - ranges positive to negative
- ▶ Mining: nonzero in parts south
  - Associated with increased poverty rate
  - Comparable to GWR within far southern range

## Future work

## Future work

- ▶ Apply the GWEN to data with non-Gaussian response variable
- ▶ Incorporate spatial autocorrelation in the model (simulated errors were iid)

# Acknowledgements