

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

Wesley Brooks

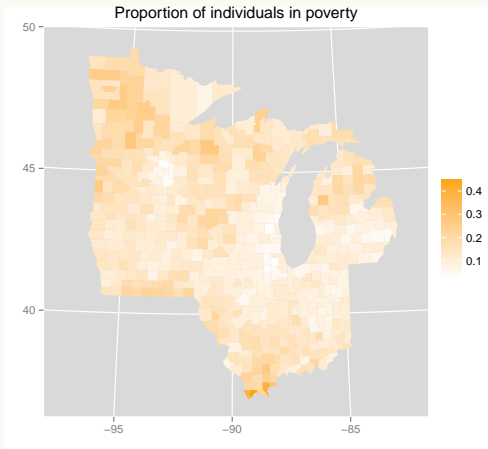
Department of Statistics  
University of Wisconsin–Madison

October 31, 2013

# Motivation

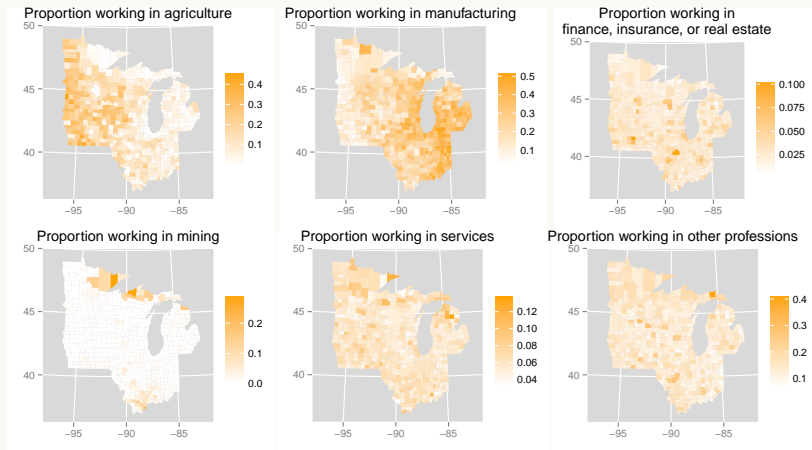
# Motivation

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

$$\begin{aligned}Y(\mathbf{s}) &= \mathbf{X}(\mathbf{s})'\boldsymbol{\beta} + W(\mathbf{s}) + \varepsilon(\mathbf{s}) \\ \text{cov}(W(\mathbf{s}), W(\mathbf{t})) &= \Gamma(\delta(\mathbf{s}, \mathbf{t})) \\ \delta(\mathbf{s}, \mathbf{t}) &= \sqrt{\|\mathbf{t} - \mathbf{s}\|_2} \\ \text{E.g. } \Gamma(\delta(\mathbf{s}, \mathbf{t})) &= \exp\{-\phi^{-1}\delta(\mathbf{s}, \mathbf{t})\}\end{aligned}\tag{1}$$

- $W(\mathbf{s})$  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

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

- ▶ The spatial random effect has been incorporated into the spatially varying intercept
- ▶  $\{\beta_1(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}, \dots, \{\beta_p(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$  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
- ▶ Splines can be incorporated into a generalized additive model (GAM):
  - $E\{Y(t)\} = f\{X_1(t)\} + \cdots + f\{X_p(t)\}$
- ▶ It is possible to parameterize a VCR model with splines for the coefficient functions:
  - $E\{Y(t)\} = \beta_1(t)X_1(t) + \cdots + \beta_p(t)X_p(t)$

# Introduction

Global selection in spline-based VCR models (Wang, Li, and Huang, 2008; Antoniadis, Gijbels, and Verhasselt, 2012)

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
  - Non-negative garrote penalty

# Introduction

Wavelet methods for VCR models (Shang, 2011; J. Zhang and Clayton, 2011)

- ▶ Wavelet methods: decompose coefficient function into local frequency components
- ▶ Selection of nonzero local frequency components with nonzero coefficients:
  - Bayesian variable selection
  - Lasso
- ▶ 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  $s_1, \dots, s_n$
- ▶  $y(s_i) = y_i$  the univariate response at location  $s_i$
- ▶  $x(s_i) = x_i$  the  $(p + 1)$ -variate vector of covariates at location  $s_i$
- ▶ Assume  $y_i = x_i' \beta_i + \varepsilon_i$  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:

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

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 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 elastic net (GWEN)

where  $\sum_{i'=1}^n w_{ii'} (y_{i'} - \mathbf{x}_{i'}' \boldsymbol{\beta}_i)^2$  is the weighted sum of squares minimized by traditional GWR, and  $\mathcal{J}_2(\boldsymbol{\beta}_i) = \alpha_i \lambda_i^* \sum_{j=1}^p |\beta_{ij}| / \gamma_{ij} + (1 - \alpha_i) \lambda_i^* \sum_{j=1}^p (\beta_{ij} / \gamma_{ij})^2$  is the AEN penalty.

# Local variable selection and parameter estimation

## Geographically weighted elastic net (GWEN)

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

$$\begin{aligned}\text{BIC}_i &= -2 \sum_{i'=1}^n \ell_{ii'} + \log \left( \sum_{i'=1}^n w_{ii'} \right) \text{df}_i \\ &= -2 \sum_{i'=1}^n \log \left[ \left( 2\pi \hat{\sigma}_i^2 \right)^{-1/2} \exp \left\{ -\frac{1}{2} \hat{\sigma}_i^{-2} \left( y_{i'} - \mathbf{x}_{i'}' \hat{\boldsymbol{\beta}}_i \right)^2 \right\} \right]^{w_{ii'}} \\ &\quad + \log \left( \sum_{i'=1}^n w_{ii'} \right) \text{df}_i\end{aligned}\tag{3}$$

# Local variable selection and parameter estimation

Geographically weighted elastic net (GWEN)

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

## Simulation study

# Simulation study

## Simulating covariates

Five covariates  $\tilde{X}_1, \dots, \tilde{X}_5$  were simulated by Gaussian random fields on the domain  $[0, 1] \times [0, 1]$  on a  $30 \times 30$  grid of sampling locations:

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

Where the covariates were simulated with colinearity, the colinearity was induced by multiplying the design matrix by the square root of the colinearity matrix:

$$\begin{aligned}\text{diag}(\Omega_{5 \times 5}) &= 1 \\ \text{off-diag}(\Omega_{5 \times 5}) &= \rho \\ X &= \tilde{X}R\end{aligned}\tag{4}$$

Where  $\Omega_{5 \times 5} = R'R$  is the Cholesky decomposition.

# Simulation study

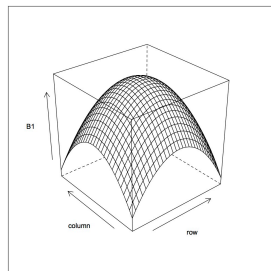
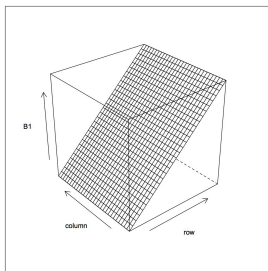
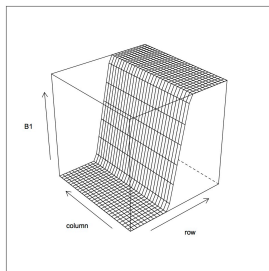
## Simulating the response

- ▶  $Y(\mathbf{s}) = X(\mathbf{s})'\boldsymbol{\beta}(\mathbf{s}) = \sum_{j=1}^5 \beta_j(\mathbf{s})X_j(\mathbf{s}) + \varepsilon(\mathbf{s})$
- ▶  $\varepsilon \sim iid \ N(0, \sigma^2)$
- ▶  $\beta_1(\mathbf{s})$ , 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

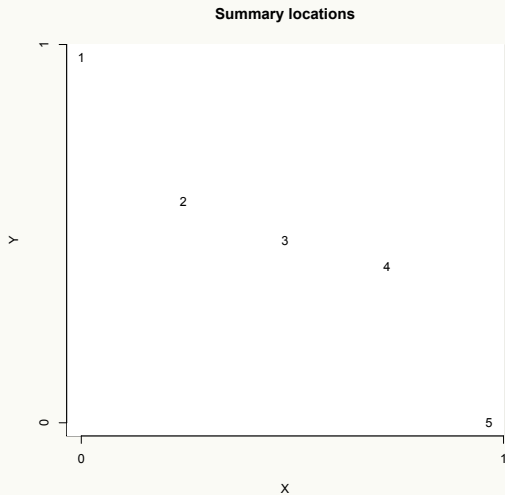
Setting	function	$\rho$	$\sigma^2$
1	step	0	0.25
2	step	0	1
3	step	0.5	0.25
4	step	0.5	1
5	gradient	0	0.25
6	gradient	0	1
7	gradient	0.5	0.25
8	gradient	0.5	1
9	parabola	0	0.25
10	parabola	0	1
11	parabola	0.5	0.25
12	parabola	0.5	1

Table : Simulation parameters for each setting.



# Simulation results

Selection



## Simulation results

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

Selection performance was more affected by an increase in the noise variance from  $\sigma_\varepsilon^2 = 0.25$  to  $\sigma_\varepsilon^2 = 1$  than by an increase in collinearity from  $\rho = 0$  to  $\rho = 0.5$ . Of the 44 cases where model selection recovered exactly the

Data example: poverty rate in the upper  
midwest

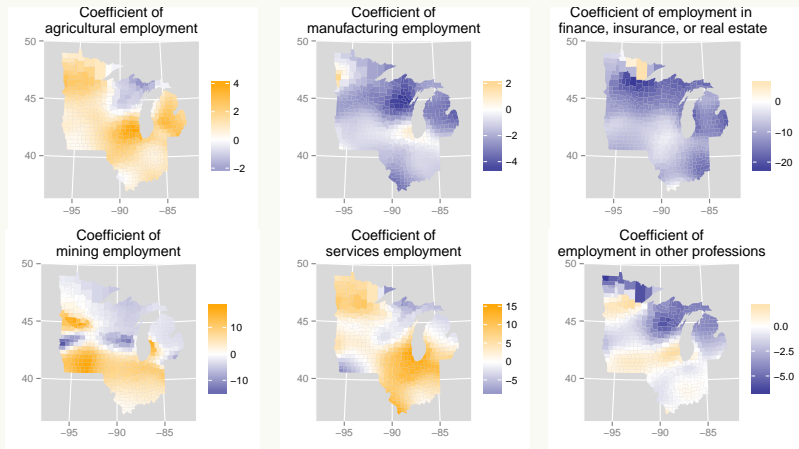
# Data example: poverty rate in the upper midwest

## 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

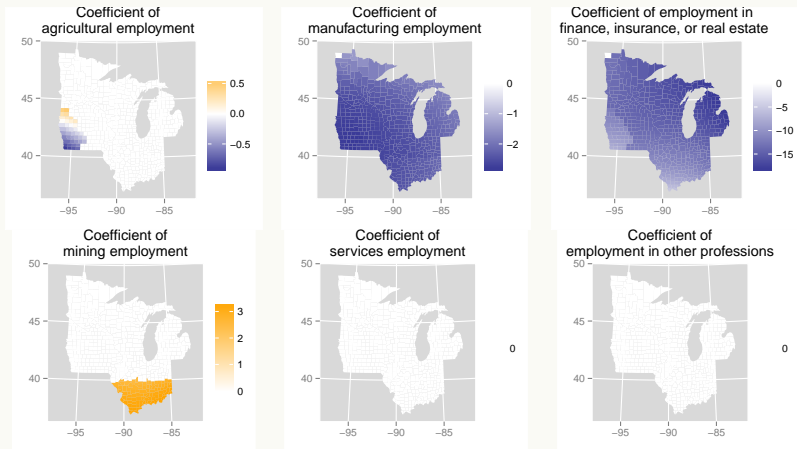
# Data example: poverty rate in the upper midwest

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

The coefficient of employment in finance, insurance, and real estate has a larger magnitude than the other covariates (minimum value near -20, as opposed to the next-largest-magnitude coefficient, that of manufacturing employment, with a minimum near -3), indicating that

## 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