

# *Bayesian Model Selection for Geostatistical Regression Data*

Devin S. Johnson

*Department of Mathematics and Statistics*

*University of Alaska Fairbanks*

JSM August 7 – 11, 2005  
Minneapolis, MN

## Sponsor

*The work reported here was developed under the STAR Research Assistance Agreement CR-829095 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University. This presentation has not been formally reviewed by EPA. The views expressed here are solely those of presenter and the STARMAP, the Program he represents. EPA does not endorse any products or commercial services mentioned in this presentation.*



## Introduction

---

- Environmental data is often collected within a spatial domain (obtained through GIS layers)
- Failing to take spatial correlation into account can affect regression model selection results.
- Previous methods to account for spatial correlation in model selection
  - Ver Hoef et al. (2001), Spatial stepwise selection
  - Thompson (2001), Bayes factor approx.
  - Hoeting et al. (2005), Spatial AICC

## Benefits of a Reversible Jump MCMC approach

---

- Allows inclusion of expert knowledge in selection of regression covariates
  - Spatial stepwise and AICC methods treat all covariates equally
- Selection over large model spaces
  - In both the Bayes factor and AICC methods parameters in each model must be separately estimated
- A sample from the joint model and parameter posterior is obtained
- Straightforward extension to spatial generalized linear mixed models (model based geostatistics)

# Model based geostatistical regression

---

## Data Model

$$Y(\mathbf{s})|Z(\mathbf{s}) \sim \text{i.i.d. } P(\ell^{-1}\{Z(\mathbf{s})\}),$$

where

$$E[Y(\mathbf{s})|Z(\mathbf{s})] = \ell^{-1}\{Z(\mathbf{s})\}$$

## Parameter model

$$\mathbf{Z} = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))' \sim N_n(\mathbf{X}\boldsymbol{\beta}; \boldsymbol{\Sigma})$$

where  $\boldsymbol{\Sigma}$  is defined by a geostatistical covariance

## Covariance function

---

$$\text{Cov}\{Z(\mathbf{s}), Z(\mathbf{s}')\} = \sigma^2 \rho(\mathbf{h}; \boldsymbol{\phi})$$

$$\text{Var}\{Z(\mathbf{s})\} = \sigma^2$$

where,

- $\mathbf{h} = \mathbf{s} - \mathbf{s}'$
- $\sigma^2$  is the sill ( $0 < \sigma^2 < \infty$ )
- $\boldsymbol{\phi}$  are the spatial correlation parameters
- $\rho(\mathbf{h}; \boldsymbol{\phi})$  is a nonnegative correlation function  
e.g.  $\rho(\mathbf{h}; \boldsymbol{\phi}) = \exp\{-(\mathbf{h}'\boldsymbol{\phi}\mathbf{h})^{1/2}\}$

## Bayesian model selection

- Model incorporated as another parameter,  $M$  with sample space  $\mathcal{M} = \{m_0, \dots, m_K\}$
- For each  $m_k$  we have  $\boldsymbol{\vartheta}_k = (\boldsymbol{\beta}_k, \sigma^2, \phi, \mathbf{Z})$
- Inference for the model can be made through the posterior model probability (PMP)

$$\begin{aligned} P(m_k | \mathbf{Y}) &\propto \int P(\mathbf{Y} | \boldsymbol{\vartheta}_k, m_k) P(\boldsymbol{\vartheta}_k | m_k) P(m_k) d\boldsymbol{\vartheta}_k \\ &= P(\mathbf{Y} | m_k) P(m_k) \end{aligned}$$

## Model prior distribution

---

A classic model prior is derived by treating inclusion of the  $p$  coefficients as a series of independent Bernoulli trials with probability  $\pi_j$ . The result is the following prior

$$P(m_k) = \prod_{j=1}^{p_k} \pi_j^{I_{k_j}} (1 - \pi_j)^{1 - I_{k_j}},$$

where  $I_{k_j}$  is the indicator that  $\beta_{k_j} \neq 0$

## Reversible Jump MCMC

---

Objective: Draw a sample from  $P(\boldsymbol{\vartheta}_k, m_k | \mathbf{Y})$

For current state  $\mathbf{q} = (\boldsymbol{\vartheta}_k, m_k)$ :

1. Propose move of type  $i$  to  $m_{k^*}$  from distribution  $J_i(\mathbf{q})$
2. Draw  $\boldsymbol{\vartheta}_{k^*}$  from  $G_i(\mathbf{q}, m_{k^*})$
3. Accept new state  $\mathbf{q}^*$  with probability

$$\min \left\{ 1, \frac{P(\mathbf{q}^* | \mathbf{Y}) J_i(\mathbf{q}^*) G_i(\mathbf{q}^*)}{P(\mathbf{q} | \mathbf{Y}) J_i(\mathbf{q}) G_i(\mathbf{q})} \right\}$$

## Difficulty with RJMCMC

- Low acceptance rate: Even if the appropriate model is chosen, bad parameter proposals will hinder mixing
- Conjecture: Proposals distributions  $G(\mathbf{q})$  close to  $P(\boldsymbol{\vartheta}_{k^*} | m_{k^*}, \mathbf{Y})$  will produce the best results

Acceptance probability for  $P(\boldsymbol{\vartheta}_{k^*} | m_{k^*}, \mathbf{Y})$

$$\min \left\{ 1, \frac{P(m_{k^*} | \mathbf{Y}) J(m_{k^*})}{P(m_k | \mathbf{Y}) J(m_k)} \right\}$$

## Partial analytic RJMCMC (PARJ)

- Godsill (2001) for AR order selection
- Use parameter proposal
  - Propose  $\beta_{k^*} \sim P(\beta_{k^*} | m_{k^*}, \sigma^2, \phi, \mathbf{Z}, \mathbf{Y})$
  - Set  $(\sigma_{k^*}^2, \phi_{k^*}, \mathbf{Z}_{k^*}) = (\sigma^2, \phi, \mathbf{Z})$
- Acceptance probability

$$\min \left\{ 1, \frac{P(m_{k^*} | \sigma^2, \phi, \mathbf{Z}, \mathbf{Y}) J(m_{k^*})}{P(m_k | \sigma^2, \phi, \mathbf{Z}, \mathbf{Y}) J(m_k)} \right\}$$

No need to actually simulate  $\beta_{k^*}$  values

## Acceptance ratio for PARJ

- Suppose  $P(\boldsymbol{\beta}_k | m_k) = N_{p_k}(\boldsymbol{\mu}_k, \mathbf{V}_k)$

- Since  $\mathbf{Z} = \mathbf{X}\boldsymbol{\beta}_k + \boldsymbol{\delta}$ ,

$$\mathbf{Z} | m_k, \sigma^2, \boldsymbol{\phi} \sim N_n(\mathbf{X}_k \boldsymbol{\mu}_k, \mathbf{X}_k \mathbf{V}_k \mathbf{X}_k' + \boldsymbol{\Sigma})$$

$$P(m_k | \mathbf{Y}, \mathbf{Z}, \sigma^2, \boldsymbol{\phi}) = P(m_k | \mathbf{Z}, \sigma^2, \boldsymbol{\phi})$$

$$\propto \exp \left\{ -\frac{1}{2} (\mathbf{Z} - \mathbf{X}_k \boldsymbol{\mu}_k)' (\mathbf{X}_k \mathbf{V}_k \mathbf{X}_k' + \boldsymbol{\Sigma})^{-1} (\mathbf{Z} - \mathbf{X}_k \boldsymbol{\mu}_k) \right\} \\ \times P(m_k)$$

## Fish abundance in the Appalachian region

---

- In 1994 – 1995,  $n = 119$  stream sites were sampled by EPA in Mid-Atlantic highlands region of U.S. (MAHA)
- $Z(s)$  = Abundance of pollution intolerant fish – important indicators of stream health
- Environmental Covariates:

Strahler order	Road density
Elevation	% watershed disturbed
Watershed area	Habitat quality index
	% fish cover
	Dissolved O <sub>2</sub> conc.
	% fine sediments

## Parameters and priors

---

- Model:  $\pi_j = 0.75$  for Strahler Order, Elevation, and Watershed Area and  $\pi_j = 0.5$  for others (a uniform prior was also used)
- $\beta_k \sim N_{p_k}(\mathbf{0}, 100\sigma^2(\mathbf{X}'_k\mathbf{X}_k)^{-1})$  ( $N_{p_k}(\cdot, \cdot)$  update)
- $\theta_1 = \log \sigma^2 \sim N(0, 10)$
- $\phi = \eta^{-1}\mathbf{I}$ ,  $\theta_2 = \log \eta \sim N(0, 1)$

## Sampler for spatial regression

---

Step	Update Type	Proposal Distribution <sup>a</sup>
1. Update $\log \sigma^2$	Metropolis	Gaussian
2. Update $\log \eta$	Metropolis	Gaussian
3. Update $m_k$	PARJ	Discrete random walk
4. Update $\beta_k$	Gibbs	Gaussian <sup>b</sup>
5. Update $\mathbf{Z}$	Langevin-Hastings	Gaussian

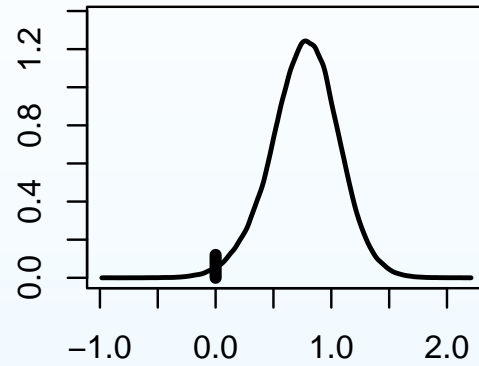
<sup>a</sup> Metropolis proposal distributions are centered on the current parameter value

<sup>b</sup> Full conditional distribution

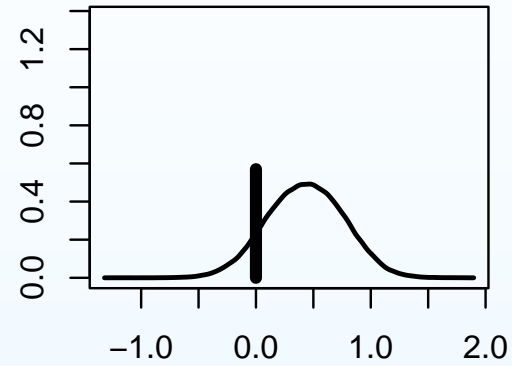
## Model chain summary

Covariate	PIP	PMP				
		0.12	0.06	0.05	0.05	0.04
Strahler order	0.88	●	●	●	●	●
Elevation	0.29				●	
Area	0.43		●			
Road density	0.38			●		●
% Disturbance	0.79	●	●		●	●
Habitat quality	0.74	●	●	●	●	●
Dissolved O <sub>2</sub>	0.15					
% Fish cover	0.10					
% Fine sed.	0.13					

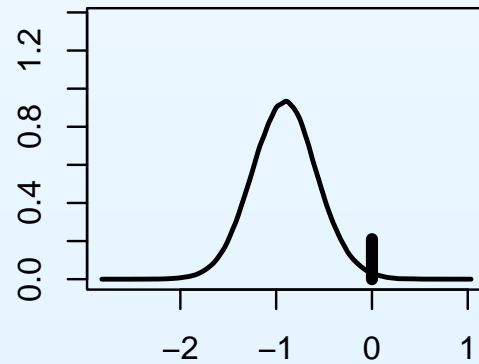
# Marginal coefficient posterior distributions



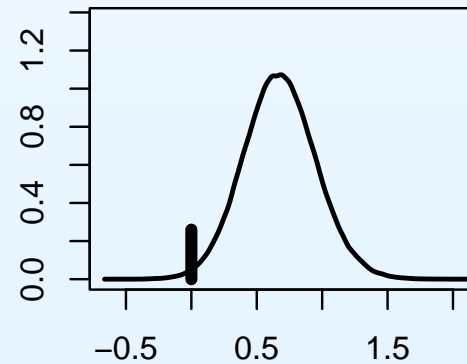
Strahler Order



Watershed Area



% Watershed Disturbed



Habitat Quality Index

## Comments / Future work

---

- Partial analytic RJMCMC provides a straightforward method of model update in an MCMC sampler
  - Simple addition to a standard Gibbs sampler
  - Hierarchical centering allows partial analytic approach
- Future investigations
  - Transformed Gaussian models possible
  - Covariate based model proposals
  - Robust hierarchical centering