



Bayesian Selection of Geostatistical Regression Models

Devin S. Johnson, Ph.D.

Department of Mathematics and Statistics

and

Institute of Arctic Biology

University of Alaska Fairbanks

WNAR/IMS June 21 – 24, 2005
Fairbanks, AK



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 RJMCMC 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
- Straightforward extension to spatial GLMMs

Geostatistical regression models

$$Y(s) = \beta_0 + X_1(s)\beta_1 + \cdots + X_p(s)\beta_p + \delta(s)$$

where

- $s \in \mathcal{D} \subset \mathbb{R}^2$ is a spatial location
- $Z(s)$ is the response variable of interest
- $X_i(s)$ is an explanatory variable, $i = 1, \dots, p$
- $\{\delta(s); s \in \mathcal{D}\}$ is a Gaussian random field with covariance function that decreases with distance

Covariance function

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

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

where,

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

Spatial GLMMs



Data Model

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

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

Parameter model

$$Z = (Z(s_1), \dots, Z(s_n))' \sim N_n(X_k\beta_k; \Sigma)$$

where Σ is defined by a geostatistical covariance



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 $\vartheta_k = (\beta_k, \sigma^2, \phi, Z)$
- Inference for the model can be made through the posterior model probability (PMP)

$$\begin{aligned} P(m_k|Y) &\propto \int P(Y|\vartheta_k, m_k)P(\vartheta_k|m_k)P(m_k)d\vartheta_k \\ &= P(Y|m_k)P(m_k) \end{aligned}$$

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 π_j . The result is the following prior

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

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

Markov Chain Monte Carlo

- Appropriate for large model spaces
- Objective: Draw a sample $(\vartheta_M^{(1)}, M^{(1)}), \dots, (\vartheta_M^{(N)}, M^{(N)})$ from $P(\vartheta_k, m_k | Y)$
 - Construct a Markov chain with stationary distribution $P(\vartheta_k, m_k | Y)$
 - Approximate $P(m_k | Y)$ by the proportion of $M^{(i)}$ that equal m_k
 - Posterior Coefficient Probability (PCP)
$$P(\beta_j \neq 0 | Y) = \sum_{k: \beta_{kj} \neq 0} P(m_k | Y)$$

Reverse Jump MCMC



For current state $x = (\vartheta_k, m_k)$:

1. Propose move of type i to $m_{k'}$ from distribution $J_i(x)$
2. Draw $\vartheta_{k'}$ from $G_i(x, m_{k'})$
3. Accept new state x' with probability

$$\min \left\{ 1, \frac{P(x'|Y)J_i(x')G_i(x')}{P(x|Y)J_i(x)G_i(x)} \right\}$$



Difficulty with RJMCMC

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

Acceptance probability for $P(\vartheta_{k'} | m_{k'}, Y)$

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

Partial analytic RJMCMC

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

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

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

Acceptance ratio for GLMM

• Suppose $P(\beta_k | m_k) = N_p(\mu_k, V_k)$

• Since $Z = X\beta_k + \delta$,

$$Z | m_k, \sigma^2, \phi \sim N_n(X_k \mu_k, X_k V_k X_k' + \Sigma)$$

$$P(m_k | Y, Z, \sigma^2, \phi) = P(m_k | Z, \sigma^2, \phi)$$

$$\propto \exp \left\{ -\frac{1}{2} (Z - X_k \mu_k)' (X_k V_k X_k' + \Sigma)^{-1} (Z - X_k \mu_k) \right\} \\ \times P(m_k)$$

Sampler for spatial regression

1. Update σ^2 from $P(\sigma^2 | \dots)$
2. Update ϕ from $P(\phi | \dots)$
3. Update Z from $P(Z | \dots)$ (If GLMM is used)
4. Update m_k using partial analytic RJMCMC
5. Update β_k from $P(\beta_k | \dots)$
6. goto step 1

Fish abundance in MAHA region

- In 1994 – 1995, $n = 119$ stream sites were sampled by EPA in Mid-Atlantic highlands region of U.S. (MAHA)

- *Abundance of pollution intolerant fish* important indicators of stream health

- Environmental Covariates:

Strahler order

Elevation

Watershed area

Dissolved O₂ conc.

% fine sediments

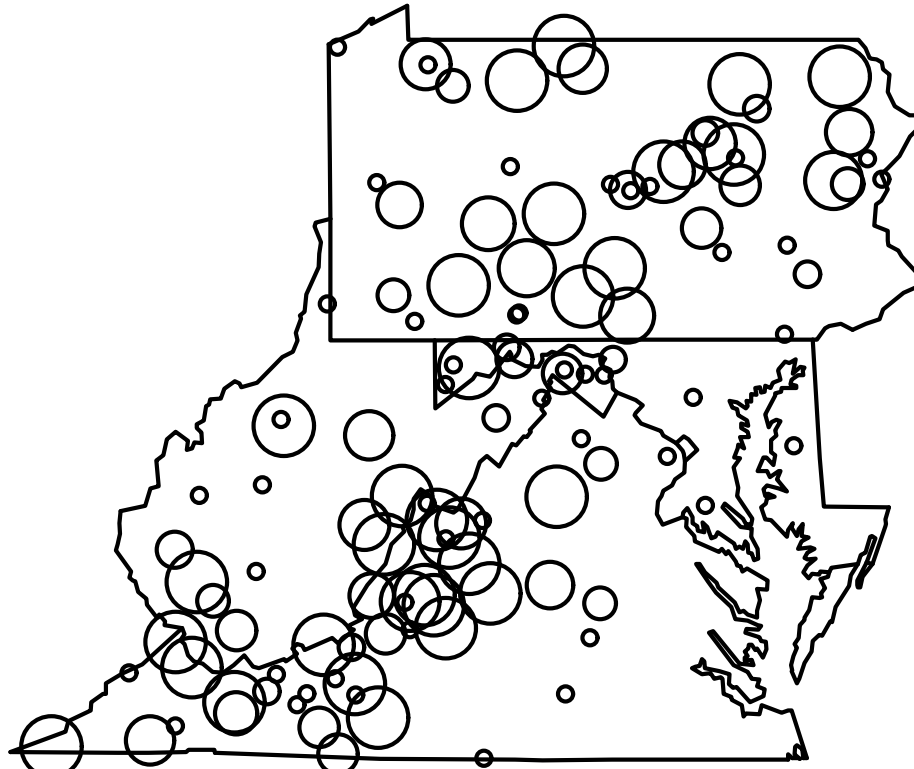
Road density

% watershed disturbed

Habitat quality index

% fish cover

Fish sampling locations



Circles are proportional to abundance of pollution intolerant fish

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)
- $\beta_k \sim N_p(0, 100\sigma^2(X'_k X_k)^{-1})$ ($N_p(\cdot, \cdot)$ update)
- $\theta_1 = \log \sigma^2 \sim N(0, 10)$
- $\phi = \eta^{-1}I$, $\theta_2 = \log \eta \sim N(0, 1)$

Proposal details

For current state Z , σ^2 , ϕ , β_k , m_k :

- **Normal** proposals used for spatial parameters
- **Langevin-Hastings** proposal for Z :
Draw $Z' \sim N_n \left(Z + \frac{h}{2} \frac{\partial}{\partial Z} \log P(Z | \dots), hI \right)$
- **Random walk** proposal for model jumps:
Propose to add or drop a randomly selected covariate ($J(m_{k'}) / J(m_k) = 1$)
- **Gibbs** update for β_k

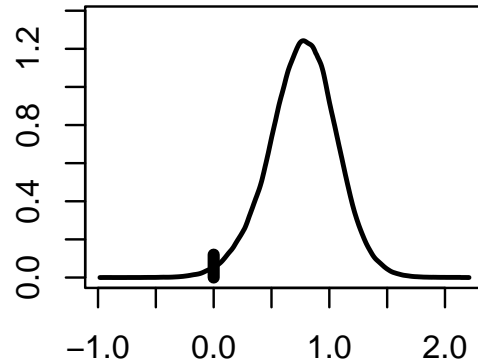
Model chain summary

Covariate	PCP	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 ₂	0.15					
% Fish cover	0.10					
% Fine sed.	0.13					

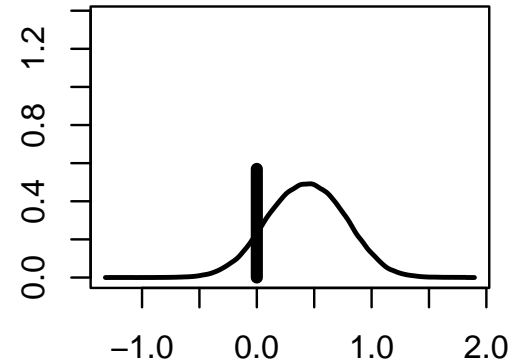
419 out of 512 models visited



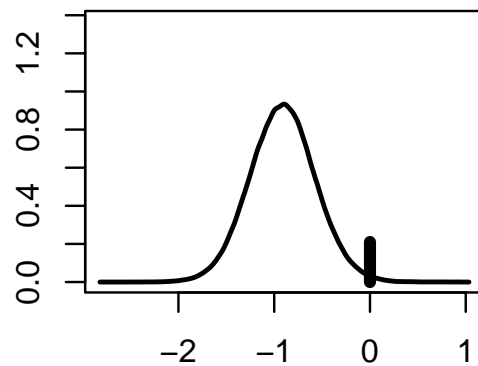
Model Averaged Coefficients



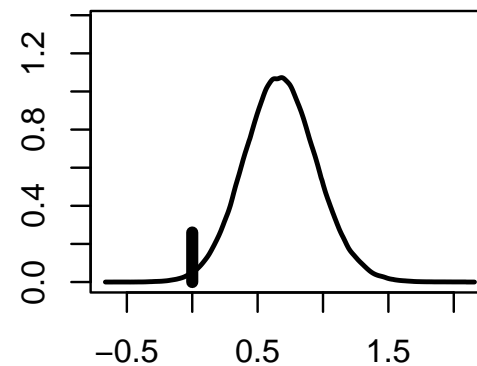
Strahler Order



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
 - Future: Transformed Gaussian models possible
 - Future: Covariate based model proposals
- Straightforward extension to generalized linear spatial models
 - Hierarchical centering allows partial analytic approach
 - Future: Robust hierarchical centering

