Model Selection for Geostatistical Models

Jennifer A. Hoeting, Richard A. Davis, Andrew Merton
Colorado State University

Sandra E. Thompson
Pacific Northwest National Lab

The work reported here was developed under STAR Research Assistance Agreements CR-829095 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University. This presentation has not been formally reviewed by EPA.

Model Selection for Geostatistical Models

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

- Which explanatory variables should be included?
- What is the form of $\delta(s)$?

Model Selection for Geostatistical Models

$$Z(s) = \underbrace{\beta_0 + X_1(s)\beta_1 + \cdots + X_p(s)\beta_p}_{\text{deterministic}} + \underbrace{\delta(s)}_{\text{stochastic}}$$

- Which explanatory variables should be included?
- What is the form of $\delta(s)$?

Model Selection for Geostatistical Models

Problem: How does one choose the "best" set of covariates and family of covariance functions?

Potential Objectives of Model Selection

1. Choose the correct model (consistency)
   - There exists a "true" finite-dimensional model.
   - If not a finite-dimensional model, at least include the key explanatory variables.
2. Choose the model that is best for prediction (efficiency)
   - Find a model that predicts well at un-observed locations.
3. Choose the model that maximizes data compression.
   - Find a model that summarizes the data in the most compact fashion.

The Geostatistical Model

Let $\mathbf{Z} = (Z(s_1), \ldots, Z(s_n))'$ be a partial realization of a random field $\mathbf{Z}(s)$, where $s \in D$, a fixed finite area under study.

A model for the random field at any location s is given by

$$Z(s) = \mathbf{X}'(s)\boldsymbol{\beta} + \delta(s),$$

where

- $\mathbf{X}(s) = (1, X_1(s), \ldots, X_p(s))'$ is a vector of explanatory variables observed at location s,
- $\boldsymbol{\beta}$ is a $p+1$ vector of unknown coefficients
- We assume that the error process $\delta(s)$ is a stationary, isotropic Gaussian process with mean zero and covariance function $\mathrm{Cov}(\delta(s), \delta(t)) = \sigma^2 \rho(||s-t||, \boldsymbol{\theta})$, where σ^2 is the variance of the process, $\rho(\cdot, \boldsymbol{\theta})$ is an isotropic correlation function, and $||\cdot||$ denotes Euclidean distance.

Autocorrelation Functions

Some of the standard autocorrelation functions:

1. **Exponential**

$$\rho(d) = \exp\left(\frac{-d}{\theta_1}\right)$$

2. **Gaussian**

$$\rho(d) = \exp\left(\frac{-d^2}{\theta_1^2}\right)$$

3. **Matern**

$$\rho(d) = \frac{1}{2^{\theta_2 - 1}\Gamma\left(\theta_2\right)}\left(\frac{2d\sqrt{\theta_2}}{\theta_1}\right)^{\theta_2} \mathcal{K}_{\theta_2}\left(\frac{2d\sqrt{\theta_2}}{\theta_1}\right), \qquad \theta_1 > 0,\ \theta_2 > 0,$$

where $\mathcal{K}_{\theta_2}(\cdot)$ is the modified Bessel function.

- Range parameter, θ_1, controls the rate of decay of the correlation between observations as distance increases.
- Smoothness parameter, θ_2, controls the smoothness of the random field.

AIC for Spatial Models

Background on AIC

Burnham and Anderson (1998), and McQuarrie and Tsai (1998)

Suppose

- $\boldsymbol{Z} \sim f_T$
- $\{f(\cdot;\psi),\ \psi \in \Psi\}$ is a family of candidate probability density functions

The Kullback-Leibler information between $f(\cdot;\psi)$ and f_T

$$I(\psi) = \int -2\log\left(\frac{f(\boldsymbol{z}\,|\psi)}{f_T(\boldsymbol{z})}\right) f_T(\boldsymbol{z})d\boldsymbol{z}\,.$$

- distance between $f(\cdot;\psi)$ and f_T
- similar to the notion of relative entropy
- loss of information when $f(\cdot;\psi)$ is used instead of f_T.

AIC for Spatial Models

By Jensen's inequality,

$$I(\psi) \;\geq\; 0 \quad \text{if and only if} \quad f(\boldsymbol{z};\psi) = f_T(\boldsymbol{z}) \quad \text{a.e. } [f_T]$$

Basic idea: minimize the Kullback-Leibler index

$$\begin{aligned}\Delta(\psi) \;&=\; \int -2\log\left(f\left(\boldsymbol{z}\,|\psi\right)\right) f_T(\boldsymbol{z}) d\boldsymbol{z} \\ &=\; E_T(-2\log L_Z(\psi))\,,\end{aligned}$$

where $L_Z(\psi)$ is the likelihood based on the data $\boldsymbol{Z}$.

Model Selection and Spatial Correlation

Traditional approach to model selection:

1. Select explanatory variables to model the large scale variation.
2. Estimate parameters using residuals from model in step 1.
3. Iterate.

Limitations:

- Ignores potential confounding between explanatory variables and correlation in spatial process
- Ignoring autocorrelation function can mask importance of explanatory variables

Simulations: Compare model selection performance of AIC for independent error regression model and geostatistical model

Model Selection: Simulation Set-up

1. **Sampling Design:** 100 locations simulated in a random pattern.

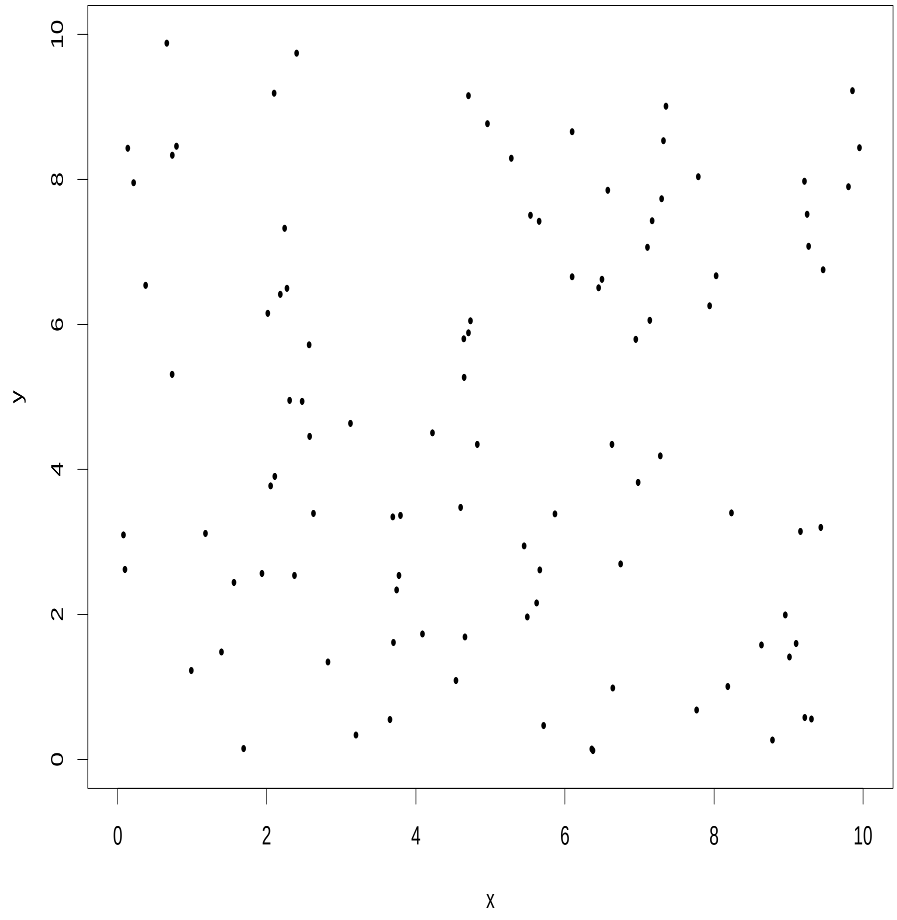
2. **Explanatory Variables:** Five possible explanatory variables:
   $X_1, X_2, X_3, X_4, X_5 \sim \sqrt{\frac{12}{10}}\, t_{12}$

3. **Response:**
   $$\boldsymbol{Z} = 2 + 0.75\boldsymbol{X}_1 + 0.50\boldsymbol{X}_2 + 0.25\boldsymbol{X}_3 + \boldsymbol{\delta},$$
   where $\boldsymbol{\delta}$ is a Gaussian random field with mean zero, $\sigma^2 = 50$, and autocorrelation Matern with parameters $\theta_1 = 4$ and $\theta_2 = 1$.

4. **Replicates:** 500 replicates were simulated with a new Gaussian random field generated for each replication.

5. **AIC:** Computed for $2^5 = 32$ possible models per replicate

Model Selection: Random Pattern Sampling Design

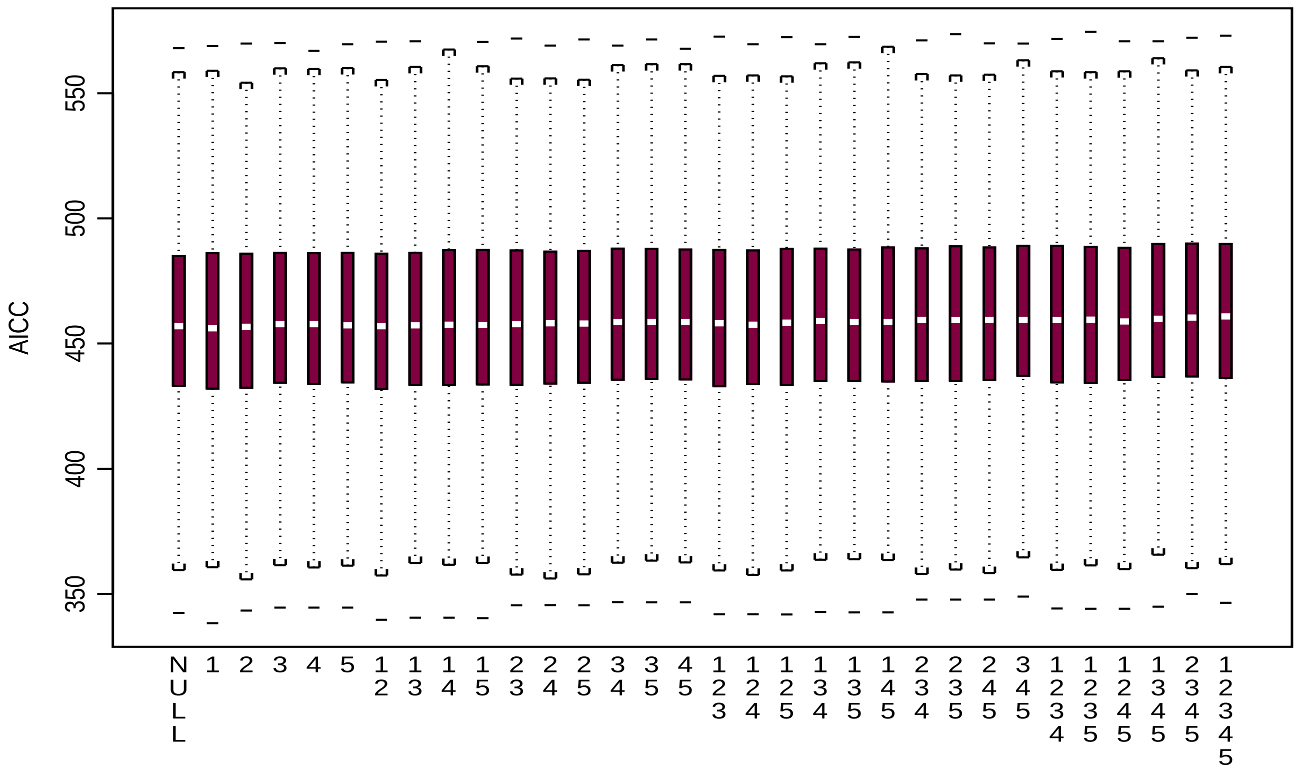


Model Selection: Simulation Results for the Random Pattern

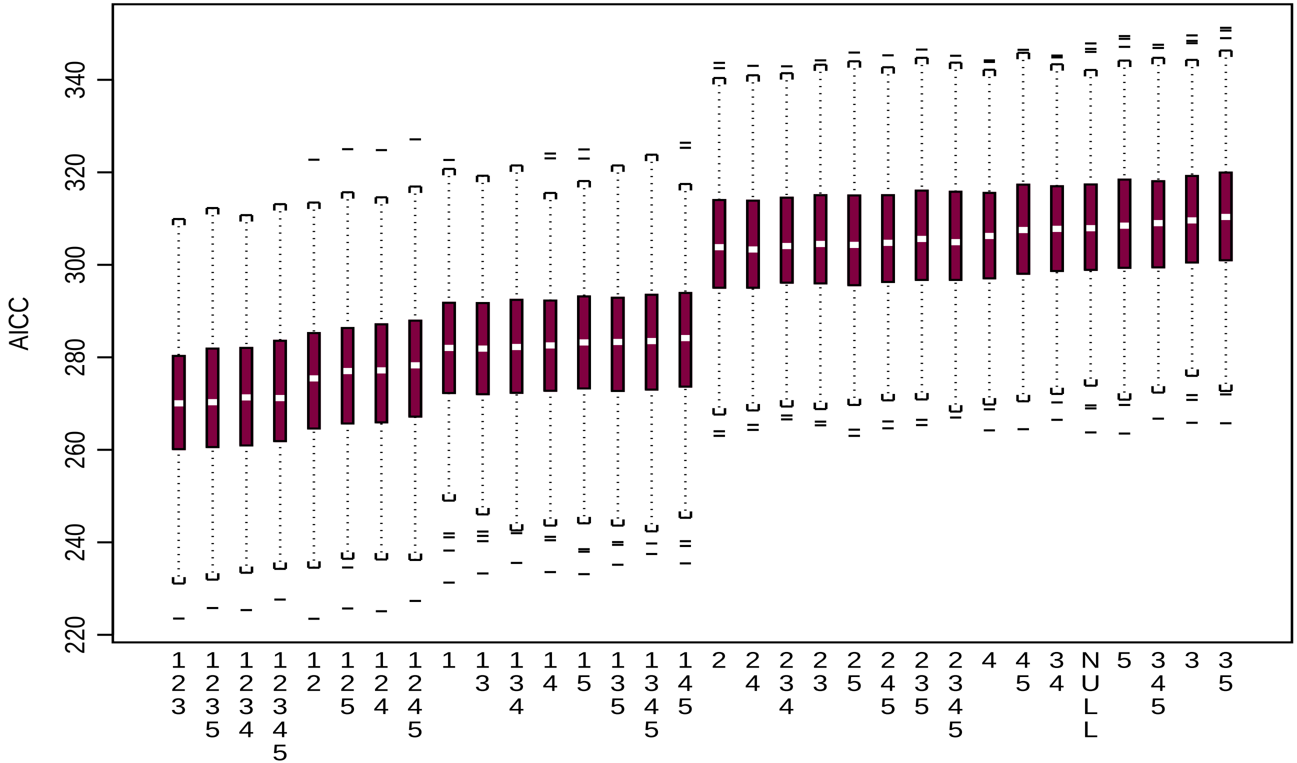
- Independent AIC and Spatial AIC report the percentage of simulations that each model was selected.
- Of the 32 possible models, the results given here include only those with 10% or more support for one of the models.

Variables in Model	Spatial AIC	Independent AIC
$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$	46	0
$\boldsymbol{X}_1, \boldsymbol{X}_2$	18	6
$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_5$	11	0
Intercept only	0	37
$\boldsymbol{X}_1$	1	18
$\boldsymbol{X}_2$	0	12

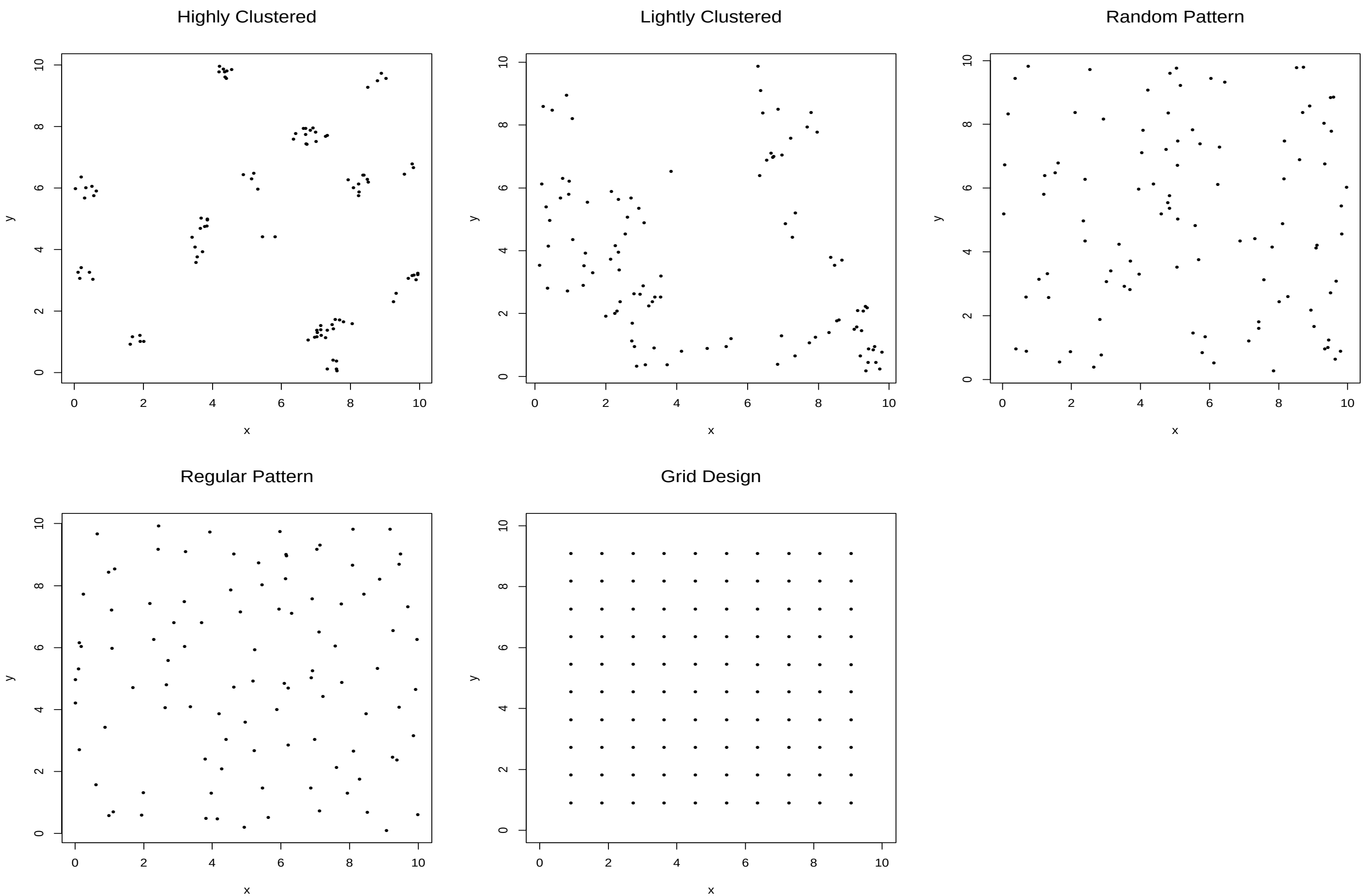
Model Selection: Independent model AIC Values



Model Selection: Spatial model AIC Values



Sampling Patterns



Model Selection: Effect of Sampling Design

Summary of model selection results for 5 different sampling patterns

Variables in Model	Highly Clustered	Lightly Clustered	Random	Regular Pattern	Grid Design
$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$	73	65	46	43	16
$\boldsymbol{X}_1, \boldsymbol{X}_2$	0	2	18	21	35
$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_4$	12	13	8	8	3
$\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3, \boldsymbol{X}_5$	10	13	11	7	7

- Each column reports the percentage of simulations that each model was selected.
- Of the 32 possible models, the results given here include only those with 10% or more support for at least one of the sampling patterns.

Prediction

Efficient prediction

- Time series (Shibata (1980), Brockwell and Davis (1991)). AIC is an efficient order selection procedure for autoregressive models.
- Regression (see McQuarrie and Tsai (1998)).
- Other notions of efficiency, e.g., Kullback-Leibler efficiency and L_2 efficiency (see McQuarrie and Tsai (1998)).

Prediction: Prediction Error

Simulations:

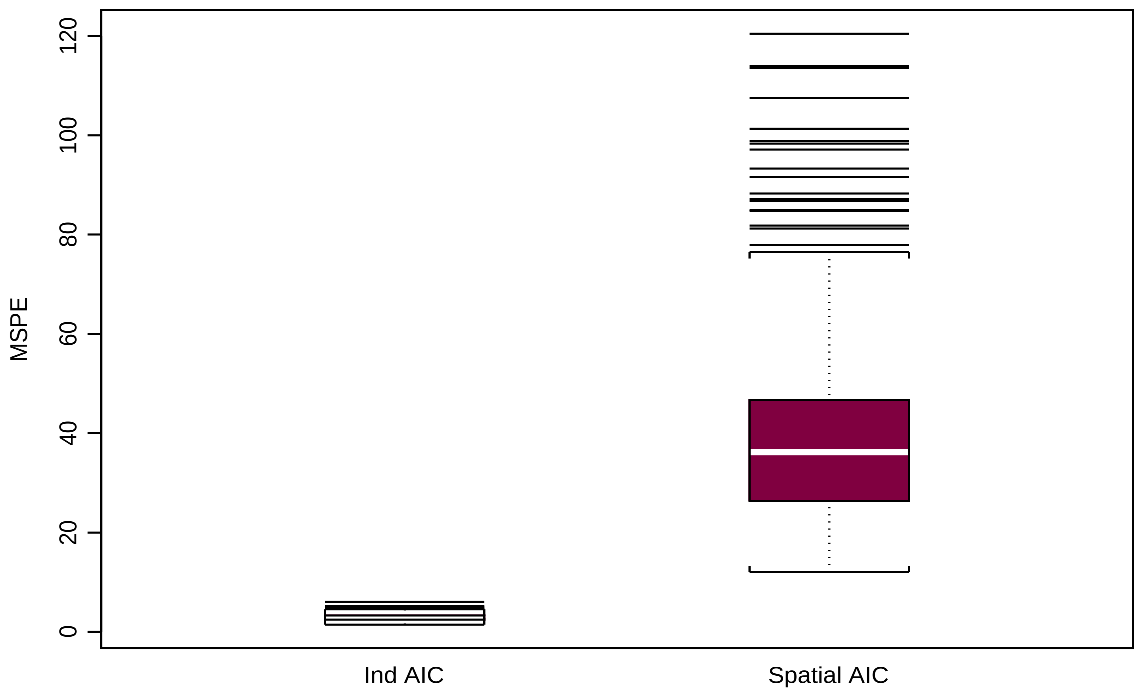
- Performed model selection and estimation using 100 observations and evaluated prediction performance using 100 additional observations simulated as above.
- Evaluated predictive performance

Mean Square Prediction Error:

$$\mathrm{MSPE} = \frac{1}{100} \sum_{j=1}^{100} \left(Z_j - \hat{Z}_j \right)^2$$

where $\tilde{Z}_j$ is the universal kriging predictor for the j^{th} prediction location using the true parameter values.

Prediction: MSPE



Prediction: Predictive Coverage

Predictive Coverage: for a 95% prediction interval, do 95% of the observed data fall in their corresponding prediction intervals?

Simulations:
For each of the 500 simulations, we compute predictive coverage. Then, over all 500 simulations, we examine:

- Mean predictive coverage
- Standard deviation of predictive coverage

Model	Mean	Std Dev
Independent error AIC	0.95	0.18
Spatial error AIC	0.92	0.25

Example: Lizard abundance

Abundance for the orange-throated whiptail lizard in southern California
Ver Hoef et al. (2001)

Data:

- 147 locations
- $Z = \log(\text{ave \# of lizards caught per day})$
- Explanatory variables: ant abundance (three levels), log(% sandy soils), elevation, barerock indicator, % cover, log(% chapparal plants)

Example: Lizard abundance

- Explanatory variables:
  ant abundance (three levels), log(% sandy soils), % cover, elevation, barerock indicator, log(% chapparal plants)
- 160 possible models

Predictors	AIC	Spatial Rank	Ind Rank
Ant_1, % sand	54.8	1	66
Ant_1,Ants_2, % sand	54.8	2	56
Ant_1, % sand, % cover	55.7	3	59
Ant_1,Ant_2, % sand, % cover, elevation, barerock, % chaparral	92.2	41	1
Ant_1,Ant_2, % sand, % cover, elevation, barerock, % chaparral	95.5	33	2
Ant_1, % sand, % cover, elevation, barerock,	95.7	38	3

Some Other Approaches to Model Selection and Prediction

- Bayesian Model Averaging
  - Model uncertainty is typically ignored in inference
  - Protect from over-confident inferences by averaging over models
- Minimum Description Length (MDL)
  - Goal: Find model that achieves maximum data compression.

    The code length (CL) of the data (Lee 2001) is the amount of memory required to store the data. Decomposition of CL:

    $$CL(\text{“}data\text{”}) = CL(\text{“}fitted\ model\text{”}) + CL(\text{“}data\ given\ fitted\ model\text{”}).$$

    Here $CL(\text{“}fitted\ model\text{”})$ might be interpreted as the code length of the model parameters and $CL(\text{“}data\ given\ fitted\ model\text{”})$ as the code length of the residuals from the fitted model.

Conclusions

- Ignoring spatial correlation can influence model selection results for both covariate selection and prediction
- Sampling patterns that offer observation pairs at small and larger distances may be advantageous for model selection
- Preliminary results suggest that accounting for spatial correlation can have large effects on prediction errors, but perhaps smaller impacts on predictive coverage.