

Model Selection for Geostatistical Models

Andrew A. Merton

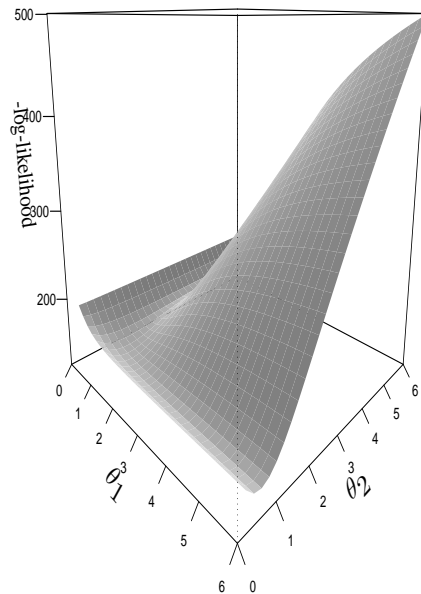
Jennifer A. Hoeting, Richard A. Davis

DEPARTMENT OF STATISTICS, COLORADO STATE UNIVERSITY

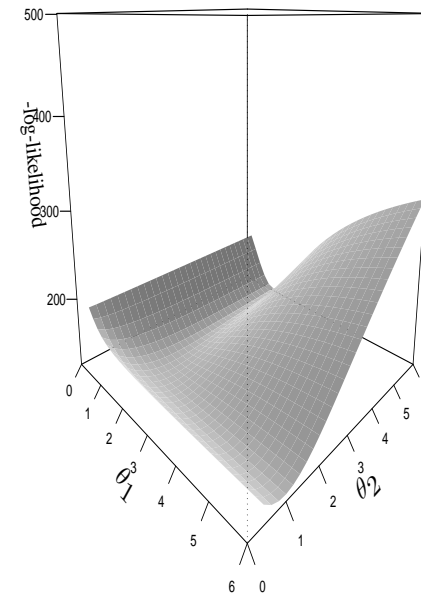
paper: www.stat.colostate.edu/~jah/papers/spavarsel.pdf

software: www.stat.colostate.edu/~jah/software/matern.R

NULL Model



True Model



Model Selection for Geostatistical Models

MOTIVATION

- Many investigators have looked at model selection but not in context of geostatistical models.
- We will show that ignoring the spatial dependence in the error structure can have a profound effect on the model selected.

Consider a problem where we observe some response Z at n locations such that

$$\mathbf{Z} = (Z(s_1), \dots, Z(s_n))'.$$

- At each location we also observe $p - 1$ explanatory variables, *i.e.*, at location s we observe $X_1(s), \dots, X_{p-1}(s)$.

A linear model for Z is given by

$$Z(s) = \beta_0 + X_1(s)\beta_1 + \dots + X_{p-1}(s)\beta_{p-1} + \delta(s).$$

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

The Geostatistical Model

Let $\mathbf{Z} = (Z(s_1), \dots, Z(s_n))'$ 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

- $\mathbf{X}(s) = (1, X_1(s), \dots, X_{p-1}(s))'$ is a p -vector of explanatory variables observed at location s ,
- $\boldsymbol{\beta}$ is a p -vector of unknown coefficients, and
- $\delta(s)$ is the unobserved regression error at location s .

We assume that the error process $\delta(s)$ is a stationary, isotropic Gaussian process with mean zero and covariance function

$$\text{Cov}(\delta(s), \delta(t)) = \sigma^2 \rho(d, \boldsymbol{\theta}),$$

where

- σ^2 is the variance of the process,
- $d = \|s - t\|$ is the Euclidean distance between locations s and t , and
- $\rho(\cdot, \boldsymbol{\theta})$ is an isotropic correlation function depending on a k dimensional parameter vector $\boldsymbol{\theta}$.

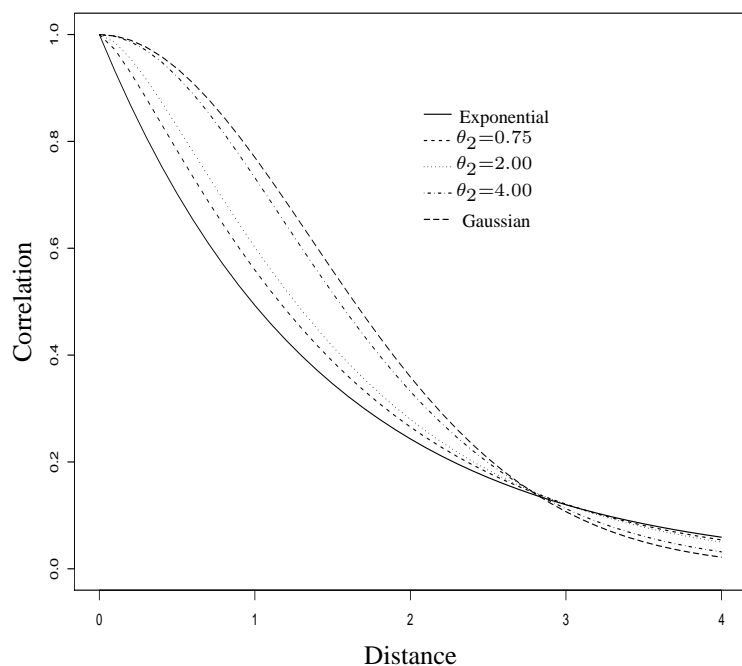
Matérn Autocorrelation Function

The Matérn autocorrelation function is defined as

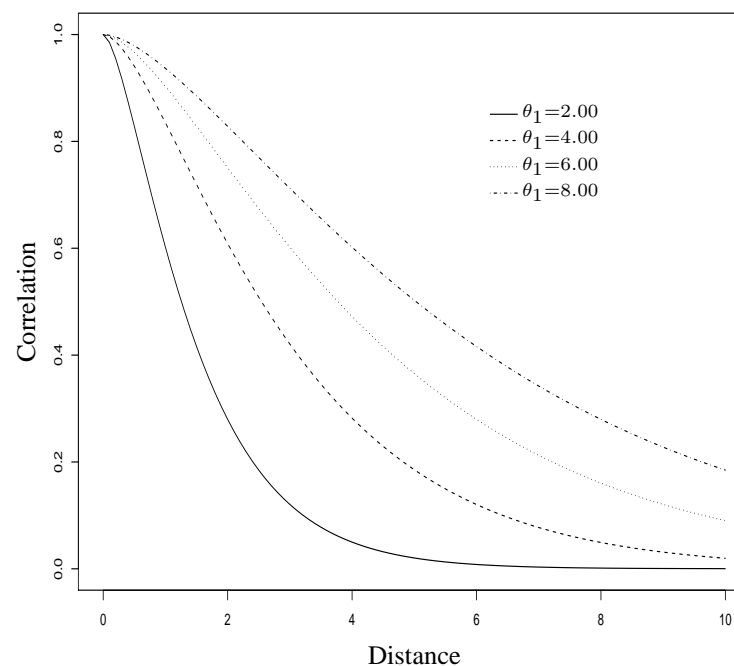
$$\rho(d, \boldsymbol{\theta}) = \frac{1}{2^{\theta_2-1} \Gamma(\theta_2)} \left(\frac{2d\sqrt{\theta_2}}{\theta_1} \right)^{\theta_2} \mathcal{K}_{\theta_2} \left(\frac{2d\sqrt{\theta_2}}{\theta_1} \right), \quad \theta_1, \theta_2 > 0,$$

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

Fixed range parameter, $\theta_1 = 4.00$



Fixed smoothness parameter, $\theta_2 = 1.00$



Estimation

Parameter estimation can proceed using one of several likelihood based approaches or a Bayesian approach. Here we consider the former.

The log-likelihood of the parameters $(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2)$ given the data, \mathbf{Z} , is

$$\log L_{\mathbf{Z}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) \propto -\frac{1}{2} \log |\sigma^2 \boldsymbol{\Omega}| - \frac{1}{2\sigma^2} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})' \boldsymbol{\Omega}^{-1} (\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}),$$

where $\boldsymbol{\Omega} = [\rho(d, \boldsymbol{\theta})]$ represents the matrix of correlations between all pairs of observations.

Estimation can proceed via an iterative maximum (profile) likelihood approach or via a restricted maximum likelihood (REML) approach.

- Although REML estimates often have more desirable sampling properties, their performance for model selection is not clear.

The resulting log *profile* likelihood is

$$\ell_{profile}(\boldsymbol{\theta}; \hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \mathbf{Z}) = -\frac{1}{2} \log |\boldsymbol{\Omega}| - \frac{n}{2} \log (\hat{\sigma}^2) - \frac{n}{2},$$

where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{Z} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}(\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}})' \boldsymbol{\Omega}^{-1}(\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}).$$

AIC for Geospatial Models

Suppose

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

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

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

- Measures the distance between $f(\cdot; \psi)$ and f_T .
- Quantifies the loss of information when $f(\cdot; \psi)$ is used as the model for the data instead of f_T .

The quantity

$$AICC = -2 \log L_Z(\hat{\beta}, \hat{\theta}, \hat{\sigma}^2) + 2n \frac{p + k + 1}{n - p - k - 2},$$

referred to as the corrected AIC, is an approximately unbiased estimate of the expected Kullback-Leibler information.

Model Selection and Spatial Correlation

Traditional approach to model selection:

1. Select explanatory variables to model the large scale variation.
 - Select the best set of explanatory variables *without* considering the dependence between locations.
2. Estimate correlation function parameters using residuals from model in step 1.
3. Re-estimate regression parameters using GLS.
4. Iterate steps 2 and 3.

Limitations:

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

Proposed Solution – Fit all candidate models assuming dependence between locations and select the model(s) with the smallest spatial AICC.

Model Selection: Simulation Set-up

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

1. **Sampling Design:** 100 locations simulated in a random pattern.
2. **Explanatory Variables:** Five possible explanatory variables such that $X_1, X_2, X_3, X_4, X_5 \stackrel{iid}{\sim} \sqrt{\frac{10}{12}} t_{12}$.

3. **Response:**

$$\mathbf{Z} = 2 + 0.75\mathbf{X}_1 + 0.50\mathbf{X}_2 + 0.25\mathbf{X}_3 + \boldsymbol{\delta},$$

where $\boldsymbol{\delta}$ is a Gaussian random field with mean zero, $\sigma^2 = 50$, and autocorrelation Matérn with parameters $\theta_1 = 4$ and $\theta_2 = 1$.

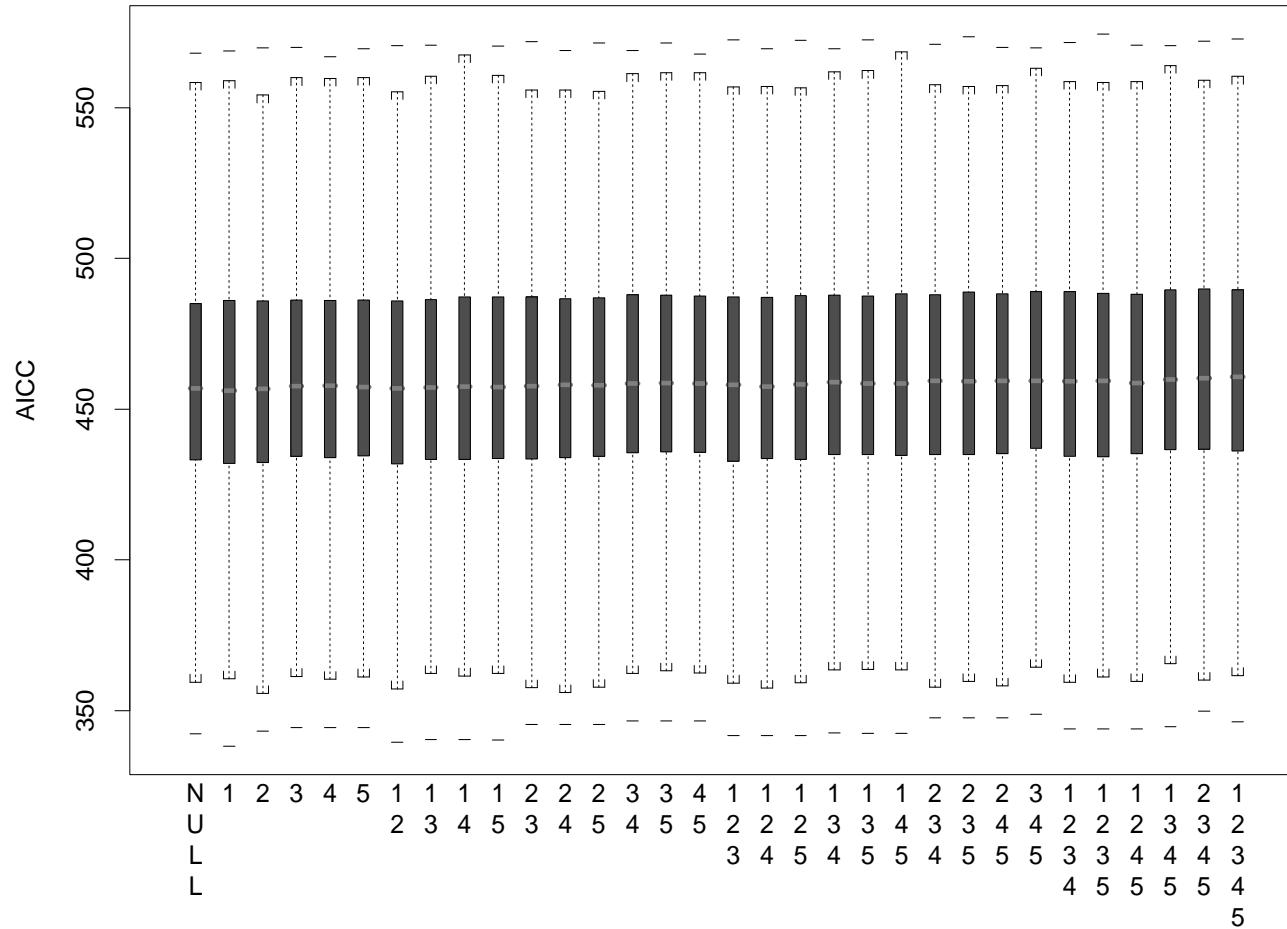
4. **Replicates:** 500 replicates were simulated with a new Gaussian random field generated for each replication.
5. **AICC:** Computed for $2^5 = 32$ possible models per replicate.

Model Selection: Simulation Results for the Random Pattern

- Independent AICC and Spatial AICC 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 AICC | Independent AICC |
|----------------------|--------------|------------------|
| X_1, X_2, X_3 | 56.0 | 2.4 |
| X_1, X_2, X_3, X_5 | 14.4 | 0.2 |
| X_1, X_2, X_3, X_4 | 10.8 | 0.2 |
| X_1, X_2 | 10.2 | 8.4 |
| Intercept only | 0.0 | 26.8 |
| X_1 | 0.4 | 14.2 |
| X_2 | 0.0 | 13.8 |

Model Selection: Independent Model AICC Values

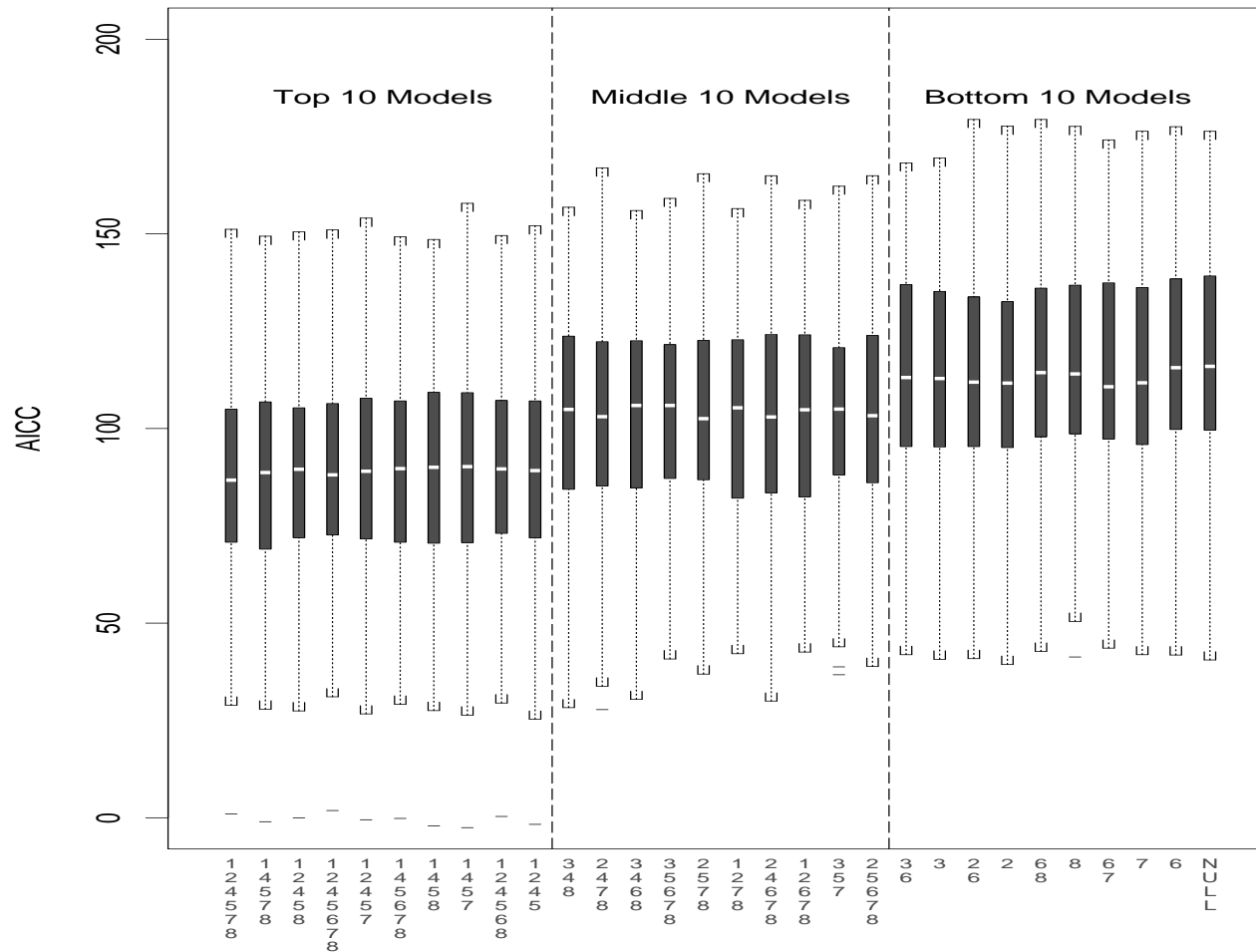


Application: Lizard Abundance

- Applied the model selection strategy to the orange-throated whiptail lizard data set (Ver Hoef *et al.* (2001)).
 - Response variable is $\log(\text{lizard abundance})$ at 148 highly clustered sites throughout southern California.
 - Explanatory variables include a single categorical variable and five continuous variables for a total of $5 \times 2^5 = 160$ unique models.
- The model selected coincides with that of Ver Hoef *et al.*
 - The selected model contains variables 1 and 4 only.
- Generated 100 simulated response data sets using the selected model and applied the spatial AICC model selection procedure.

Model Selection: Lizard Data Simulation

Independent Model AICC Values



Conclusions

- We advise against ignoring spatial correlation when selecting explanatory variables.
 - Model choice for prediction should involve *joint* selection of the explanatory variables and the form of the autocorrelation function.
- The model selection methodology presented here can be easily adapted for use with other information criterion.
 - Bayesian Information Criterion (BIC).
 - Minimum Description Length (MDL).
- The sampling pattern can severely impact model selection (complete results not presented here).
 - Simulation studies demonstrate that this selection methodology is even more successful for lightly and heavily clustered sampling patterns.

Paper, Software, & Acknowledgements

- We have made a copy of the paper and the simulation software available.
 - Paper: www.stat.colostate.edu/~jah/papers/spavarsel.pdf
email: merton@stat.colostate.edu
 - Software: www.stat.colostate.edu/~jah/software/matern.R
- A special thank you to my co-advisors, Jennifer A. Hoeting and Richard A. Davis (**Colorado State University**), and our co-author 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.