Model Selection for Geostatistical Models

Richard A. Davis
Colorado State University

(http://www.stat.colostate.edu/~rdavis/lectures)

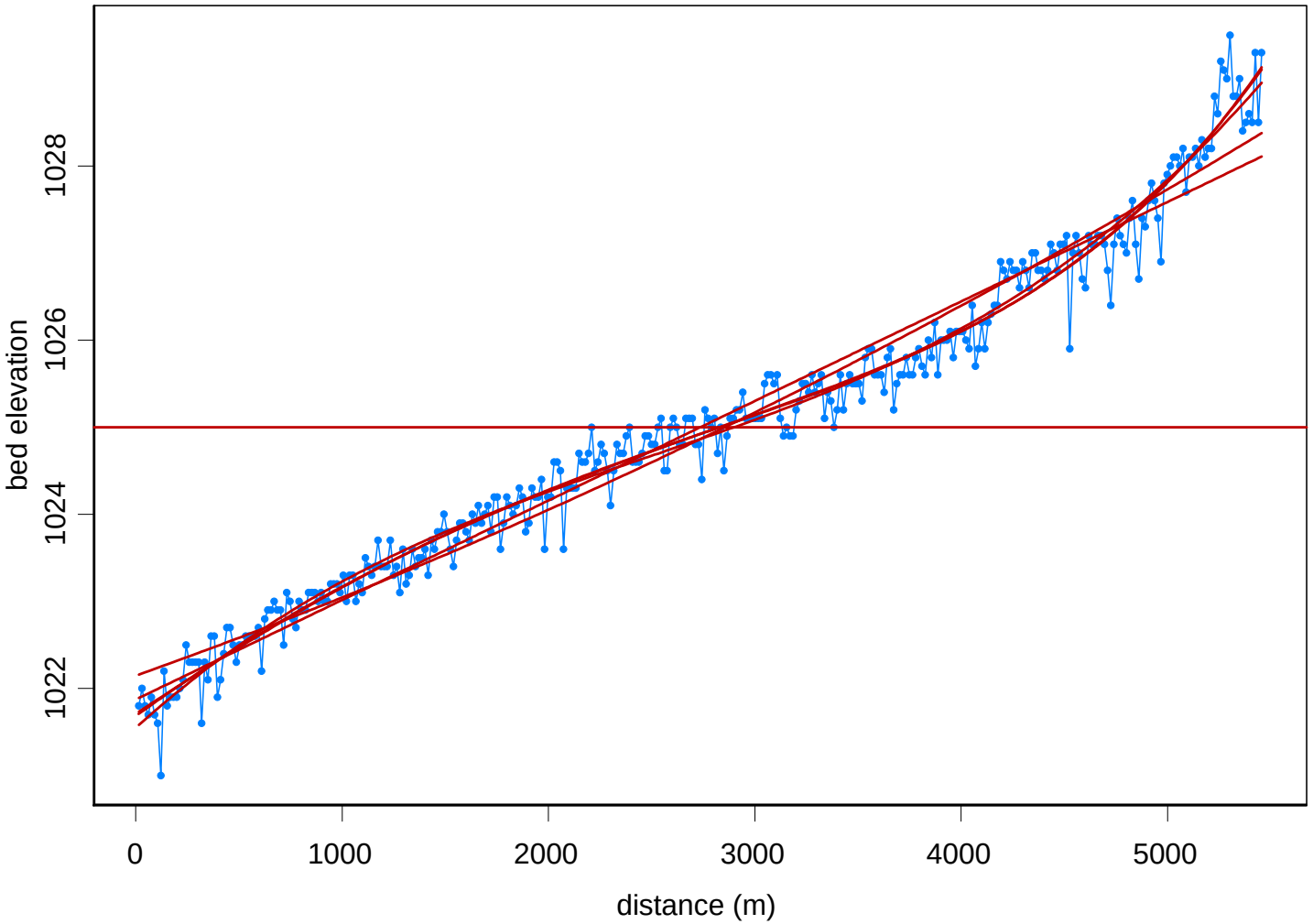
Joint work with:
Jennifer A. Hoeting, Colorado State University
Andrew Merton, Colorado State University
Sandy E. Thompson, Pacific Northwest National Lab

Partially supported by an *EPA funded project* entitled:
STARMAP: Space-Time Aquatic Resources Modeling and Analysis Program

The work reported here was developed un 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. EPA does not endorse any products or commercial services mentioned in this presentation.

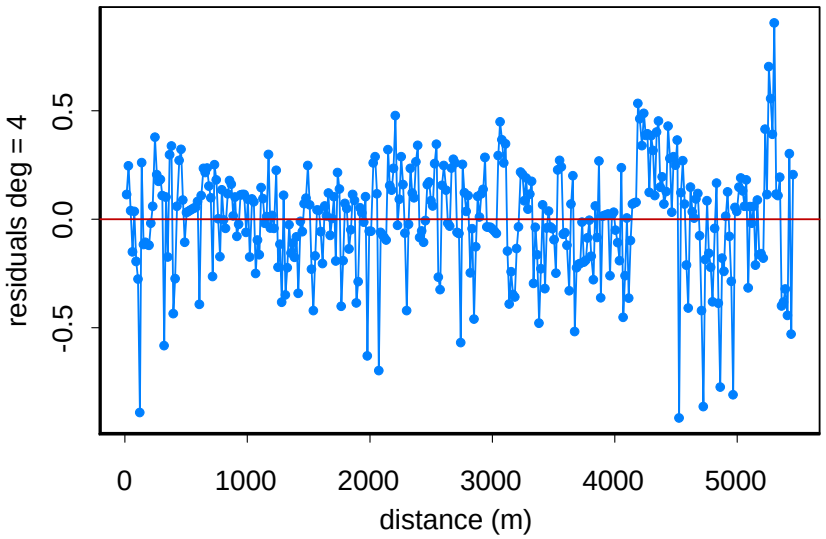
Muddy Creek- tributary to Sun River in Central Montana

Muddy Creek: surveyed every 15.24 meters, total of 5456m; 358 measurements



Degree	AIC_c
0	1455
1	294.3
2	251.3
3	47.1
4	34.0
5	35.5

Muddy Creek: residuals from poly(d=4) fit



Minimum AIC_c ARMA model: ARMA(1,1)

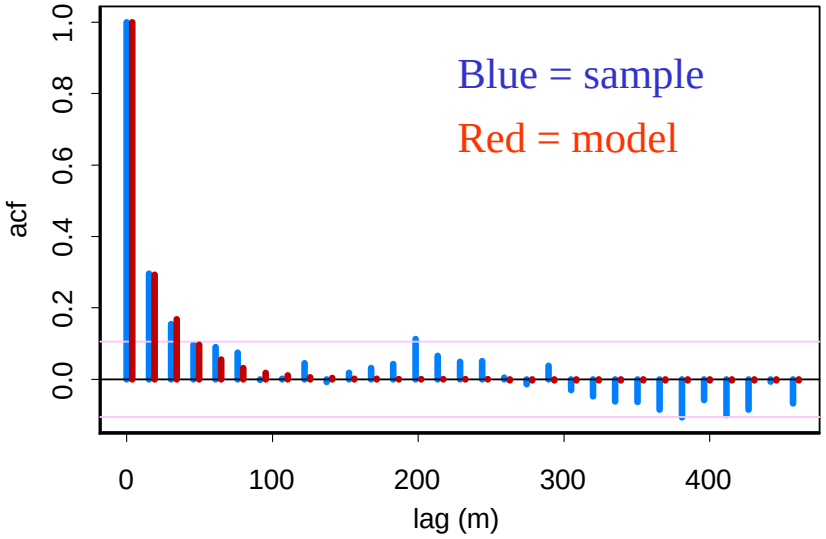
$Y_t = .574\ Y_{t-1} + \varepsilon_t - .311\ \varepsilon_{t-1}$, $\{\varepsilon_t\} \sim WN(0,.0564)$

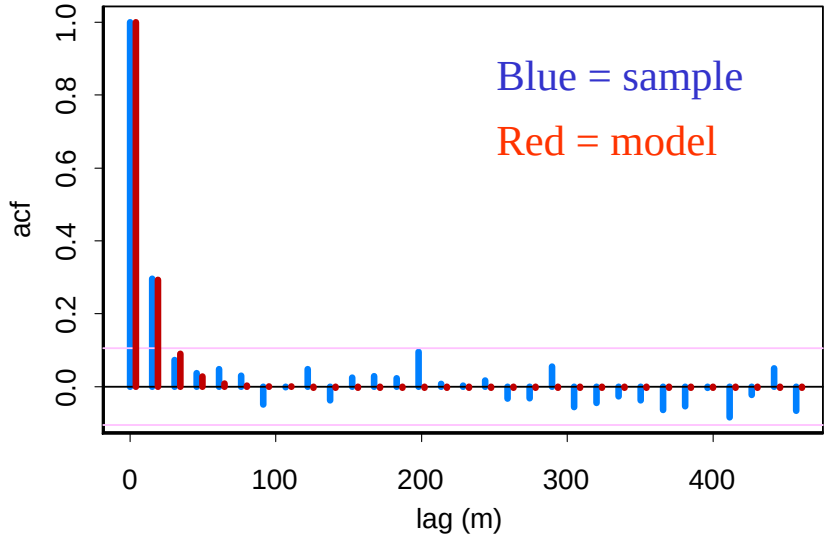
Some theory:

Normal ARMA(1,1) model:

$Y_t = 1.743\ Y_{t-1} + \varepsilon_t - .311\ \varepsilon_{t-1}$

- LS estimates of trend parameters are *asymptotically efficient*.
- LS estimates are *asymptotically indep* of cov parameter estimates.





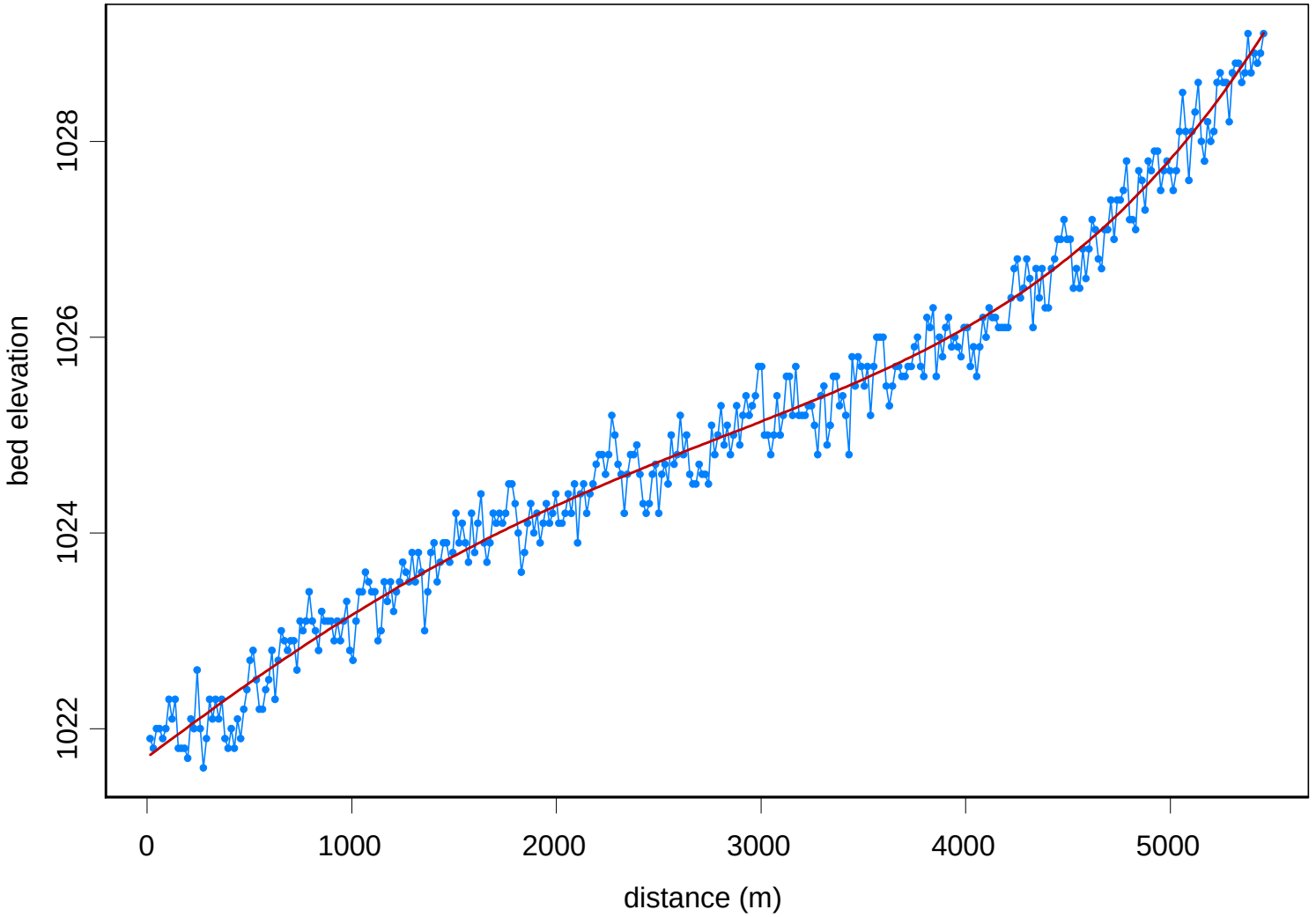
Muddy Creek (cont)

Summary of models fitted to Muddy Creek bed elevation:

Degree	AIC_c	ARMA	AIC_c
0	1455	(1,2)	59.67
1	294.3	(2,1)	26.98
2	251.3	(2,1)	26.30
3	47.1	(1,1)	7.12
4	**34.0**	**(1,1)**	**2.78**
5	35.5	(1,1)	4.68

Muddy Creek (cont)

Simulated series: polynomial degree 4 + ARMA(1,1):



Recap of Traditional Model Fitting

Suppose $\{Y_t\}$ follows the linear model with time series (or spatial) errors given by

$$Y_t = \beta_0 + X_{t1}\,\beta_1 + \ldots + X_{tp}\,\beta_p + W_t\,,$$

where $\{W_t\}$ is a stationary (ARMA) time series.

- Estimate $(\beta_0, \ldots, \beta_p)$ by ordinary least squares (OLS).
- Identify and estimate ARMA parameters using the estimated residuals,

$$Y_t - (\hat{\beta}_0 + X_{t1}\hat{\beta}_1 + \cdots + X_{tp}\hat{\beta}_p)$$

- Re-estimate $(\beta_0, \ldots, \beta_p)$ and ARMA parameters using full MLE.

Limitations of this approach for model selection:

- Ignore potential confounding between explanatory variables and correlation in $\{W_t\}$.
- Ignoring autocorrelation function can mask importance of explanatory variables.

PROGRAM

- Motivation
- Introduction
- Model Selection for Geostatistical Models
  - Basic setup
- AIC for Geostatistical Models
  - Basics
- Minimum Description Length (MDL) for Geostatistical Models
- Simulation Results
- Prediction
- Application to Orange-Throated Whiptail Lizard

Introduction

Model selection setup: Suppose $\{Z(s), s \in D\}$ is a random field where D is some subset of R^2.

Spatial data: observations $Z(s_1), \ldots, Z(s_n)$ at locations $s_1, \ldots, s_n$.

Explanatory variables (covariates): p explanatory variables $X_1(s), \ldots, X_p(s)$ are available at each location s.

Linear model for Z:

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

where $\delta(s)$ is a mean 0 stationary (isotropic) Gaussian random field.

Model selection issue.

- Which explanatory variables should be included in the model?
- What family of covariance functions should be used to model $\delta(s)$?

Model Selection

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

Some Objectives of Model Selection.

1. Choose the correct order model (*consistency*).
   - There exists a *true* model and the model selection procedure will choose the *correct* set of covariates and the right family of covariance functions as sample size increases.
2. Choose the model that performs *best* for prediction (*efficiency*).
   - Find the model that predicts (or interpolates) well at unobserved locations.
3. Choose the model that *maximizes* data compression.
   - Find a model that summarizes the data in the most compact fashion, yet retains the salient features present in the data.

The Geostatistical Model

Model :

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

- $X(s) = (1, X_1(s), \ldots, X_p(s))^T$ is a vector of explanatory variables observed at location s.
- $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is a $(p + 1)$ dimensional parameter vector.
- $\delta(s)$ is the unobserved regression error at location s.

Assumptions on $\delta(s)$.

- $\delta(s)$ is a stationary, isotropic Gaussian process with mean 0 and covariance function

$$\mathrm{Cov}(\delta(s), \delta(t)) = \sigma^2 \rho(\|s-t\|, \theta).$$

- σ^2 is the variance of the process
- $\|s-t\|$ is the euclidean distance between locations s and t.
- $\rho(.\,, \theta)$ is an isotropic correlation function parameterized by a k-dim'l vector θ.

Autocorrelation functions

Some standard autocorrelation functions *(these are a bit limiting).*

1. Exponential

$$\rho(d;\theta) = \exp(-d/\theta_1), \quad d > 0$$

2. Gaussian

$$\rho(d;\theta) = \exp(-d^2/\theta_1^2), \quad d > 0$$

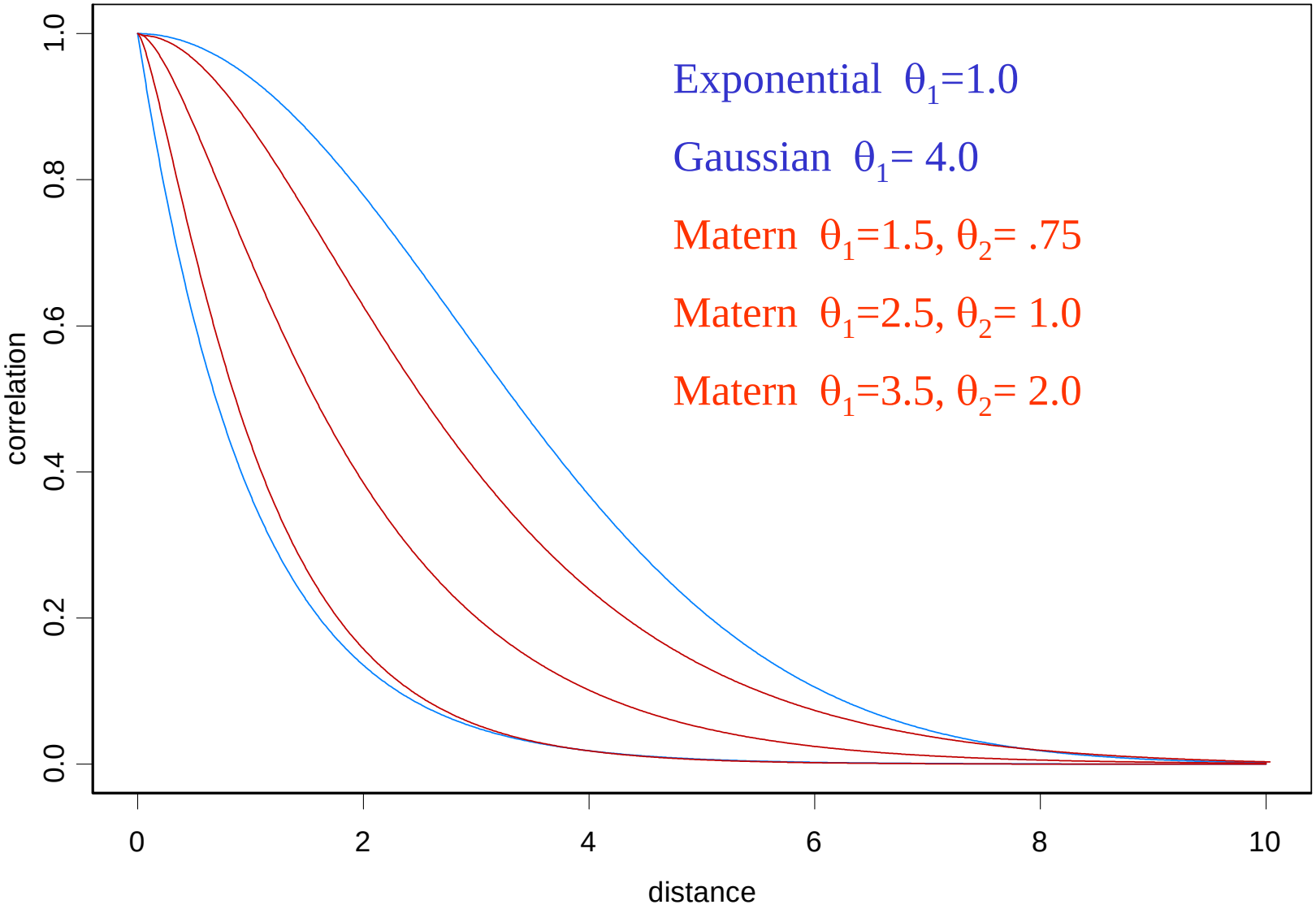
3. Matern

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

where $K_\theta(.)$ is the modified Bessel function.

- θ_1 is the *range parameter* controlling the rate of decay of correlations.
- θ_2 is the *smoothness parameter* controlling the smoothness of the random field.

Autocorrelation functions (cont)



Estimation

Log-likelihood: For the data $Z = (Z(s_1), \ldots, Z(s_n))^T$

$$\log L_Z(\beta,\theta,\sigma^2) = -.5 \log|\sigma^2\Omega| - .5\, \sigma^{-2}(Z - X\beta)^T \Omega^{-1} (Z - X\beta)^T,$$

where $\Omega = \rho(||s_i - s_j||;\theta)$ is the matrix of correlations for the data vector Z.

MLE estimate: $(\hat{\beta}, \hat{\theta}, \hat{\sigma}^2)$ maximizes the log-likelihood.

Note.

- *MLE estimates* can be difficult to compute for large sample sizes.
- *Restricted maximum likelihood* (REML) estimates often have more desirable sampling properties, but performance for model selection is not clear.

AIC for Geostatistical Models

Background on AIC (Burnham and Anderson (1998) and McQuarrie and Tsai (1998).):

Suppose

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

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

$$I(\psi) = \int -2\log\left(\frac{f(z \mid \psi)}{f_T(z)}\right) f_T(z)dz$$

has the interpretation as a measure of the

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

AIC for Geostatistical Models

By Jensen's inequality,

$$I(\psi) \geq -2\log\int\left(\frac{f(z\,|\,\psi)}{f_T(z)}\right)f_T(z)dz = 0$$

with equality holding if and only if $f(z\,;\psi) = f_T(z)$, a.e.

Basic idea: minimize the Kullback-Leibler index

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

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

Problem: Cannot compute $\Delta(\psi)$ or $\Delta(\hat{\psi})$, where $\hat{\psi}$ is the MLE of ψ.

Strategy: Search for an unbiased estimator of

$$E_\psi(\Delta(\hat{\psi}))$$

and find the candidate model which minimizes this statistic as a function of the model.

AIC for Geostatistical Models

Note: The expectation above can be written as the double expectation,

$$E_\psi(\Delta(\hat{\psi})) = E_\psi(E_T(-2\log L_Y(\hat{\psi}) \mid Z)) = E_\psi(-2\log L_Y(\hat{\psi})),$$

where Y is independent of Z with the same distribution. Thus $-2\log L_Y(\hat{\psi})$ is an unbiased estimate of $\Delta(\hat{\psi})$. But Y is unobserved and that is where the AIC correction factor comes into play.

Applied to the Geostatistical model: Parameter vector $\psi = (\beta,\theta,\sigma^2)^T$ and

$$-2\log L_Y(\hat{\beta},\hat{\theta},\hat{\sigma}^2) = -2\log L_Z(\hat{\beta},\hat{\theta},\hat{\sigma}^2) + \hat{\sigma}^{-2}S_Y(\hat{\beta},\hat{\theta}) - n,$$

where

$$S_Y(\beta,\theta) = (Y - X\beta)^T\Omega^{-1}(\theta)(Y - X\beta)$$

and

$$\hat{\sigma}^2 = (Y - X\hat{\beta})^T\Omega^{-1}(\hat{\theta})(Y - X\hat{\beta})/n.$$

So just need to approximate

$$E_\psi(\hat{\sigma}^{-2}S_Y(\hat{\beta},\hat{\theta})).$$

AIC for Geostatistical Models

In order to compute $E_\psi(\hat{\sigma}^{-2} S_Y(\hat{\beta},\hat{\theta}))$, we assume *standard asymptotics* hold:

- $(\hat{\beta},\hat{\theta})$ is AN$((\beta,\theta)^T, I_n^{-1})$, where I_n is the Fisher information.
- For large n, I_n can be approximated by

$$V(\beta,\theta) = 1/2\sigma^{-2} E_\psi\left(\frac{\partial^2 S_Y}{\partial(\beta,\theta)^T \partial(\beta,\theta)^T}(\beta,\theta)\right).$$

- For n large, $n\hat{\sigma}^2 = S_Z(\hat{\beta},\hat{\theta})$ is approximately distributed as $\sigma^2\chi^2(n-p-1-k)$.
- $\hat{\sigma}^2$ is approximately independent of $(\hat{\beta},\hat{\theta})$.

Then

$$\begin{aligned} E_\psi(\hat{\sigma}^{-2} S_Y(\hat{\beta},\hat{\theta})) - n &\sim \sigma^2(n+1+p+k)E_\psi\hat{\sigma}^{-2} - n \\ &\sim \sigma^2(n+1+p+k)\left(\sigma^2\frac{n-p-1-k-2}{n}\right)^{-1} - n \\ &= \frac{2(p+2+k)n}{n-p-k-3}. \end{aligned}$$

AIC for Geostatistical Models

The quantity, referred to as the corrected AIC, is then

$$AIC_c = -2\log L_Z(\hat{\beta}, \hat{\theta}, \hat{\sigma}^2) + \frac{2(p+2+k)n}{n-p-k-3}.$$

The standard AIC statistic is given by

$$AIC = -2\log L_Z(\hat{\beta}, \hat{\theta}, \hat{\sigma}^2) + 2(p+2+k).$$

Model Selection Using Minimum Description Length

Basics of MDL:

Choose the model which maximizes the compression of the data or, equivalently, select the model that minimizes the code length of the data (i.e., amount of memory required to encode the data).

$\boldsymbol{M}$ = class of operating models for $y = (y_1, \ldots, y_n)$

$L_{\boldsymbol{F}}(y)$ = code length of y relative to $\boldsymbol{F} \in \boldsymbol{M}$

Typically, this term can be decomposed into two pieces (two-part code),

$$L_{\boldsymbol{F}}(y) = L(\hat{\boldsymbol{F}}|y) + L(\hat{e}\,|\,\hat{\boldsymbol{F}}),$$

where

$L(\hat{\boldsymbol{F}}|y)$ = code length of the fitted model for $\boldsymbol{F}$

$L(\hat{e}|\hat{\boldsymbol{F}})$ = code length of the residuals based on the fitted model

Illustration Using a Simple Regression Model (see T. Lee `01)

Encoding the data: $(x_1,y_1), \ldots, (x_n,y_n)$

1. "Naïve" case

$$L("naive") = L(x_1,\ldots,x_n) + L(y_1,\ldots,y_n)$$
$$= L(x_1)+\cdots+L(x_n) + L(y_1)+\cdots+L(y_n)$$

2. Linear model; suppose $y_i = a_0 + a_1 x_i$, $i = 1, \ldots, n$. Then

$$L("p=1") = L(x_1,\ldots,x_n) + L(a_0,a_1)$$
$$= L(x_1)+\cdots+L(x_n) + L(a_0)+L(a_1)$$

3. Linear model with noise; suppose $y_i = a_0 + a_1 x_i + \varepsilon_i$, $i = 1, \ldots, n$, where $\{\varepsilon_i\}$~IID $N(0,\sigma^2)$. Then

$$L("p=1") = L(x_1)+\cdots+L(x_n) + \underbrace{L(\hat{a}_0)+L(\hat{a}_1)+L(\hat{\sigma}^2)+L(\hat{\varepsilon}_1,\ldots,\hat{\varepsilon}_n \mid \hat{a}_0,\hat{a}_1,\hat{\sigma}^2)}_{A}$$

If $A < L(y_1) + \ldots + L(y_n)$, then "p=1" encoding scheme dominates the "naïve" scheme.

Model Selection Using Minimum Description Length (cont)

Applied to the Geostatistical model:

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

First term $L(\hat{\boldsymbol{F}}|y)$:

$$L(\hat{\boldsymbol{F}}|y) = L(\hat{\beta}_0,\ldots,\hat{\beta}_p) + L(\hat{\theta}_1,\ldots,\hat{\theta}_k)$$

Encoding:

integer I : $\log_2 I$ bits (if I unbounded)

$\log_2 I_U$ bits (if I bounded by I_U)

MLE $\hat{\theta}$: ½ $\log_2 N$ bits (where N = number of observations used to compute $\hat{\theta}$; Rissanen (1989))

Model Selection Using Minimum Description Length (cont)

So,

$$L(\hat{F}|y) = \sum_{j=0}^{p} \frac{1}{2}\log_2 n + \sum_{j=1}^{k} \frac{1}{2}\log_2 n = \frac{1}{2}(p+1+k)\log_2 n$$

Second term $L(\hat{e}\,|\,\hat{F})$: Using Shannon's classical results on information theory, Rissanen demonstrates that the code length of $\hat{e}$ can be approximated by the negative of the log-likelihood of the fitted model, i.e., by

$$L(\hat{e}\,|\,\hat{F}) \approx -\log L_Z(\hat{\beta}, \hat{\theta}, \hat{\sigma}^2)$$

MDL for the model F is then

$$MDL = 1/2(p+1+k)\log_2 n - \log_2 L_Z(\hat{\beta}, \hat{\theta}, \hat{\sigma}^2)$$

The strategy is to find the model that minimizes *MDL*.

Note:

$$MDL = 1/2\left(-2\log_2 L_Z(\hat{\beta}, \hat{\theta}, \hat{\sigma}^2) + (p+1+k)\log_2 n\right)$$

Penalty coefficient $(\log_2 n)$ is larger than that (2) for AIC.

Model Selection and Spatial Correlation

Traditional approach to model selection:

1. *Select explanatory variable* to model the large scale variation. Here one might use AIC_c assuming independence in the noise term.

2. *Estimate correlation function* parameters using residuals from model in previous step.

3. *Re-estimate regression* parameters using GLS.

4. *Iterate steps 2 and 3* until convergence.

Limitations of this approach:

- Ignore potential confounding between explanatory variables and correlation in spatial process.

- Ignoring autocorrelation function can mask importance of explanatory variables.

Model Selection: Simulation Set-up

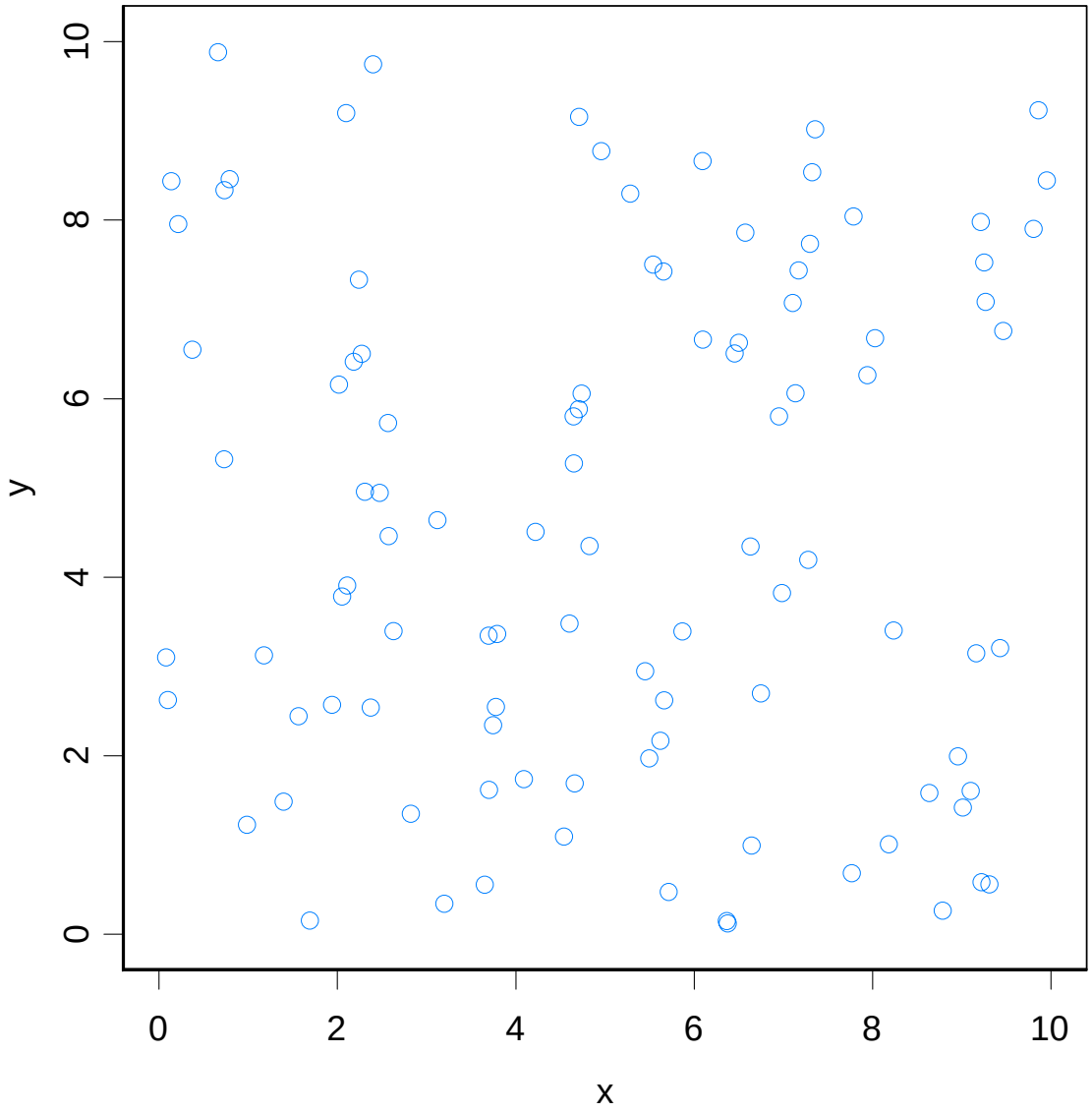
Goal of simulation study: compare model selection performance of AIC versus traditional method.

1. Sampling design: 100 sites randomly located on [0,10]×[0,10].
2. Explanatory variables: Five potential explanatory variables,
   $X_1(s_j), X_2(s_j), \ldots, X_5(s_j)$, j =1, . . . , 100, IID sqrt(12/10)*t_{12}
3. Response variables:

   $$Z = 2 + 0.75\, X_1 + 0.50\, X_2 + 0.25\, X_3 + \delta,$$

   where δ is Gaussian with mean 0, σ^2 =50, and Matern autocorrelation function with parameters θ_1=4 and θ_2=1.
4. Replicates: 500 replicates were simulated with a new Gaussian random field generated for each replicate.
5. AIC: Computed for 2^5=32 possible models per replicate.

Model Selection: Random Pattern Sampling Design

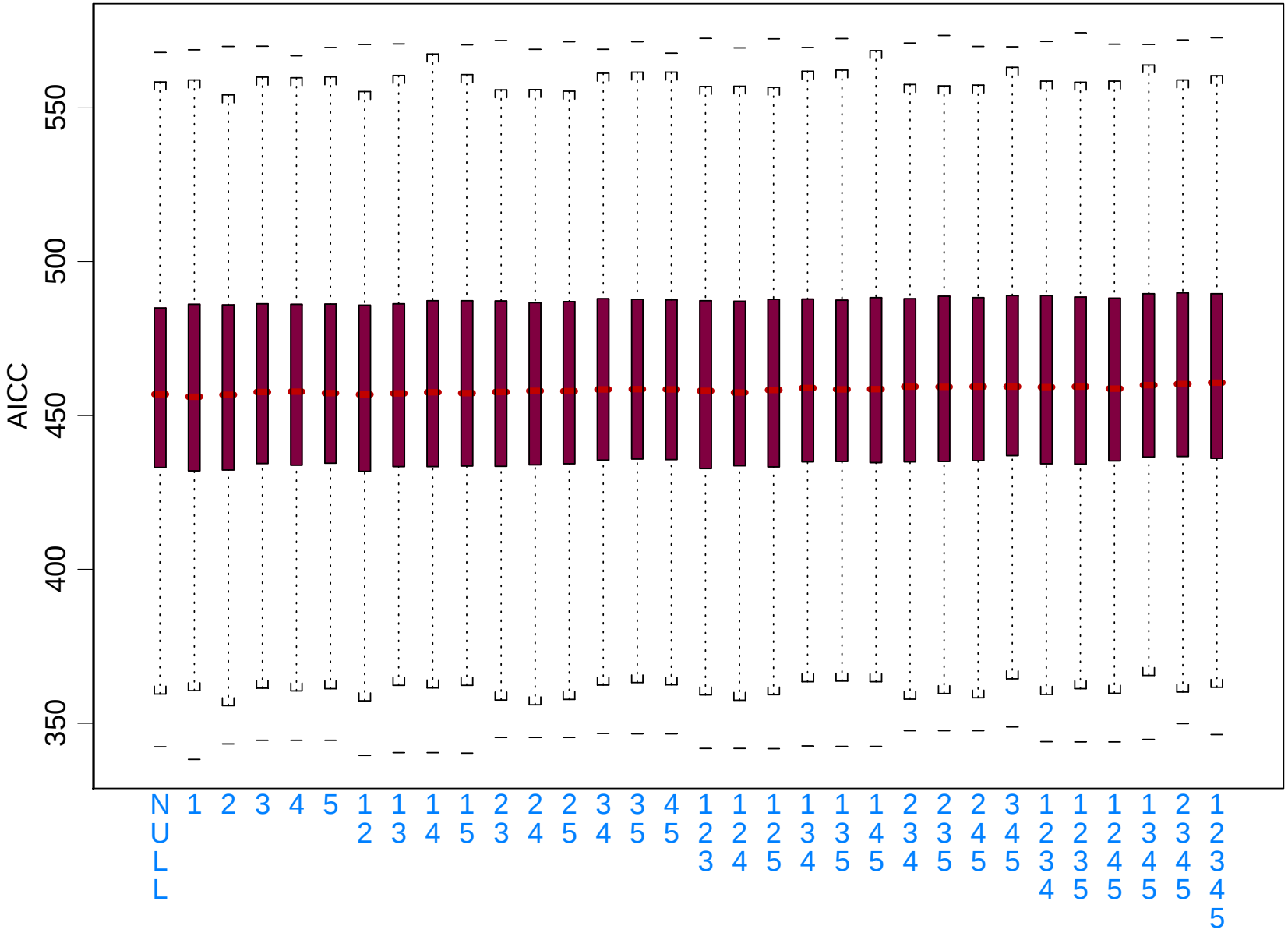


Model Selection: Simulation Results for the Random Pattern

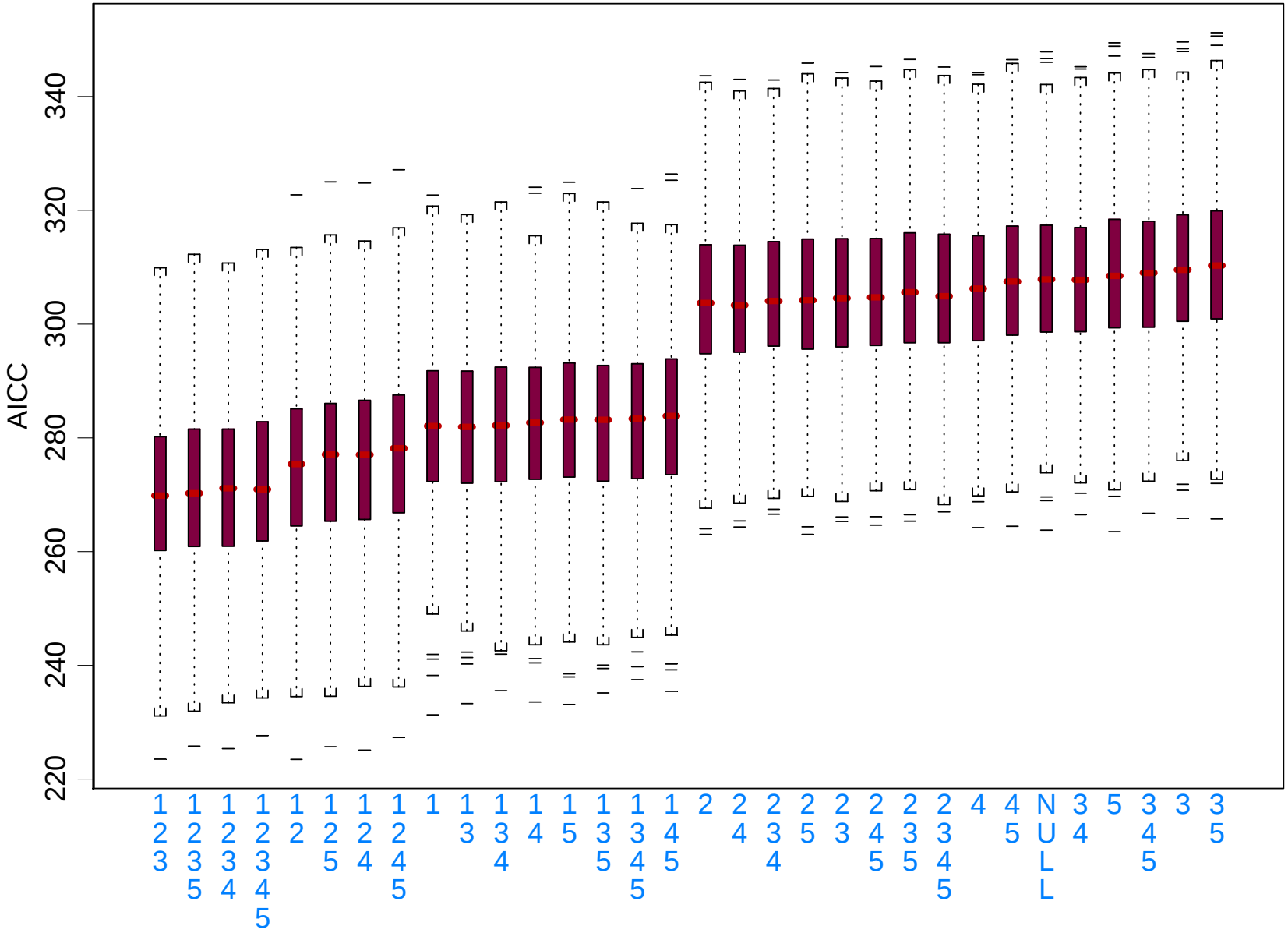
- Independent AIC and Spatial AIC report the percentage of simulations that each model was selected.
- Of the 32 possible models, the results given here include only those with 10% or more support for one of the models.

Variables in Model	Spatial AIC	Independent AIC	MDL
X_1, X_2, X_3	56.0	2.4	40.4
X_1, X_2, X_3, X_5	14.4	0.2	4.2
X_1, X_2, X_3, X_4	10.8	0.2	0.8
X_1, X_2	10.2	8.4	46.4
Intercept only	0.0	26.8	0.0
X_1	0.4	14.2	1.2
X_2	0.0	13.8	0.2

Independent Model AIC values

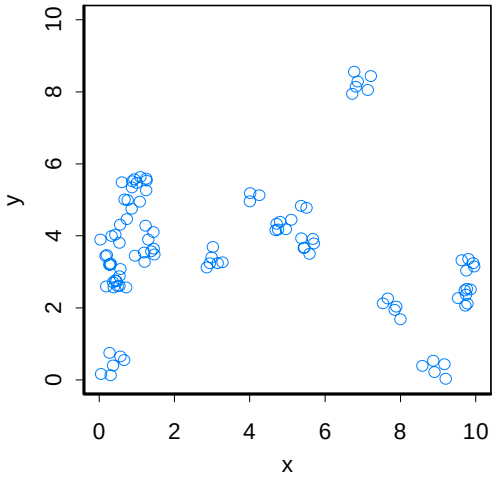


Spatial Model AIC values

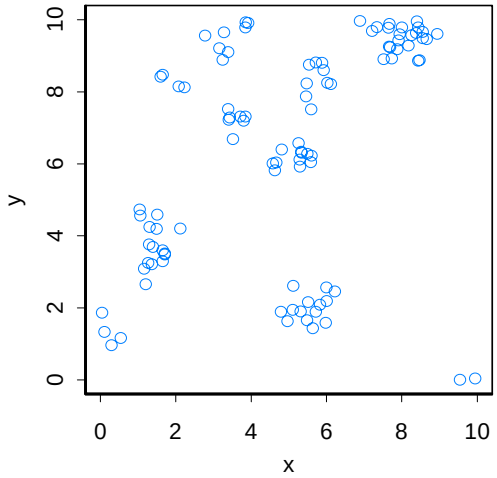


Sampling Patterns

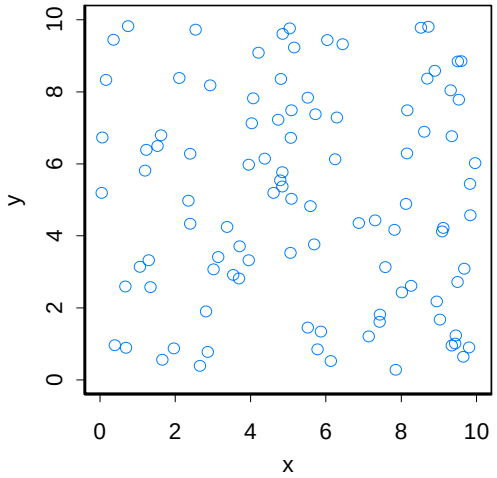
Highly Clustered



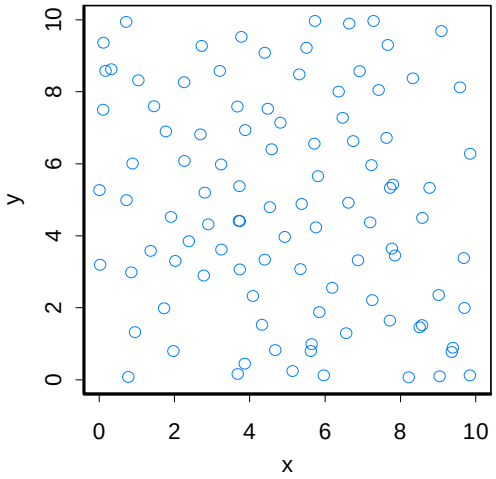
Lightly Clustered



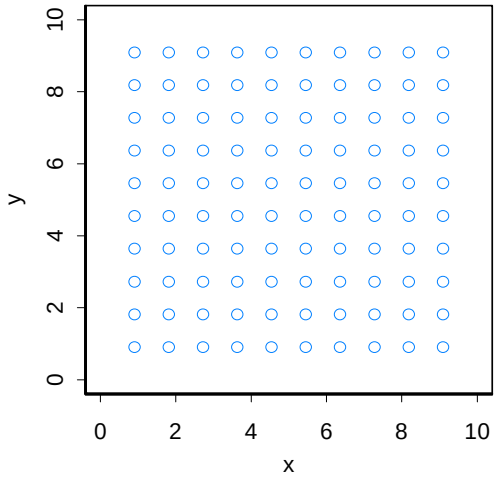
Random Pattern



Regular Pattern



Grid Design



Effect of Sampling Design

Variables in Model	Highly Clustered	Lightly Clustered	Random	Regular Pattern	Grid Design
X_1, X_2, X_3	73	65	46	43	16
X_1, X_2	0	2	18	21	35
X_1, X_2, X_3, X_4	12	13	8	8	3
X_1, X_2, X_3, X_4, X_5	10	13	11	7	7

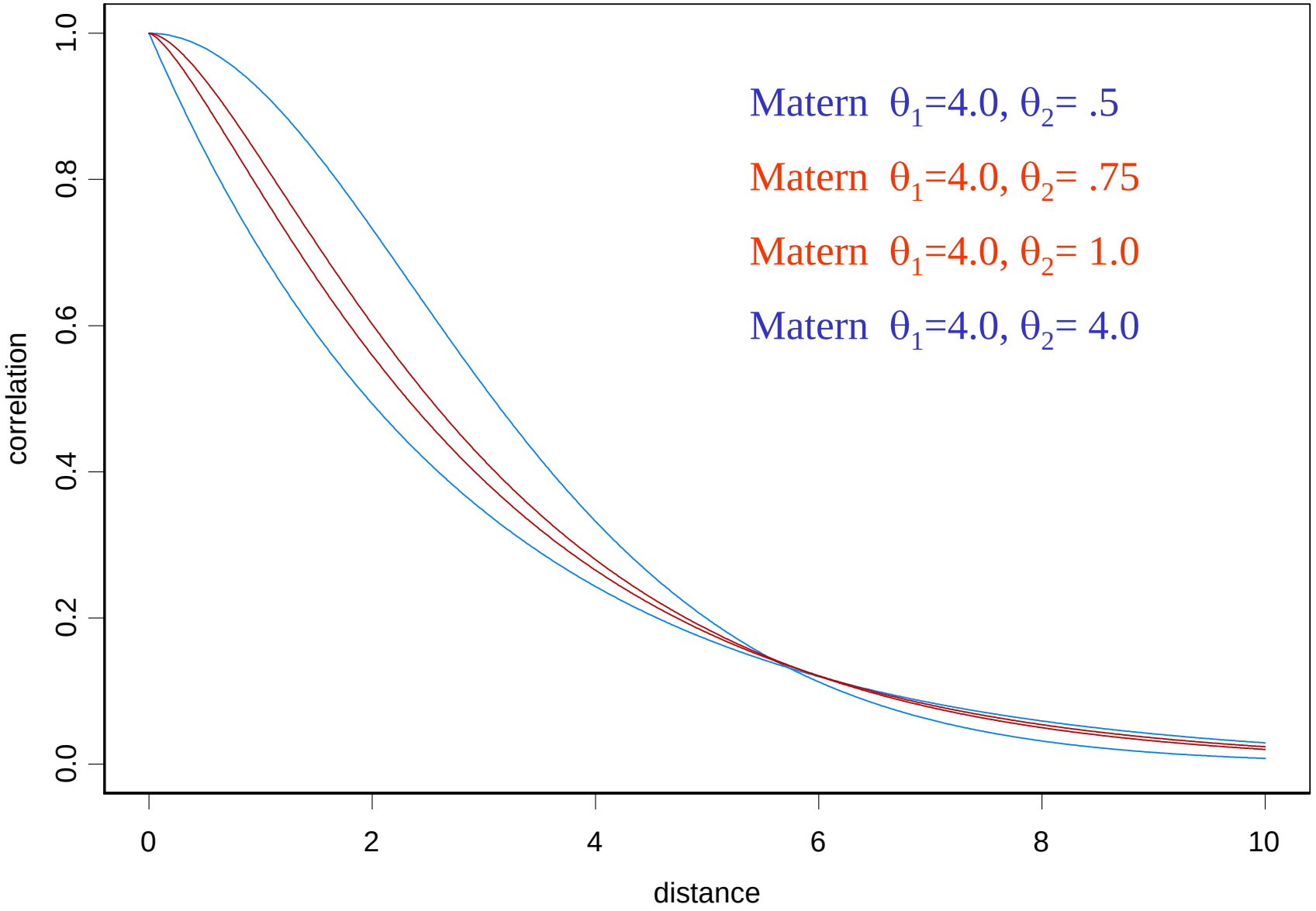
- Each column reports the percentage of simulations that each model was selected.
- Of the 32 possible models, the results given here include only those with 10% or more support for one of the models.

Effect of Smoothness Parameter

Variables in Model	$\theta_2 = 0.50$		$\theta_2 = 0.75$		$\theta_2 = 1.00$		$\theta_2 = 4.00$	
	spat	ind	spat	ind	spat	ind	spat	ind
X_1, X_2, X_3	15	3	35	1	56	2	62	3
X_1, X_2, X_3, X_5	1	0	3	0	14	0	14	0
X_1, X_2, X_3, X_4	3	0	7	0	11	0	18	1
X_1, X_2	22	7	20	4	10	8	0	8
Intercept only	4	30	0	32	0	27	0	22
X_1	17	17	12	20	0	14	0	20
X_2	4	12	0	11	0	14	0	8

- 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 one of the models.

Effect of smoothness (cont)

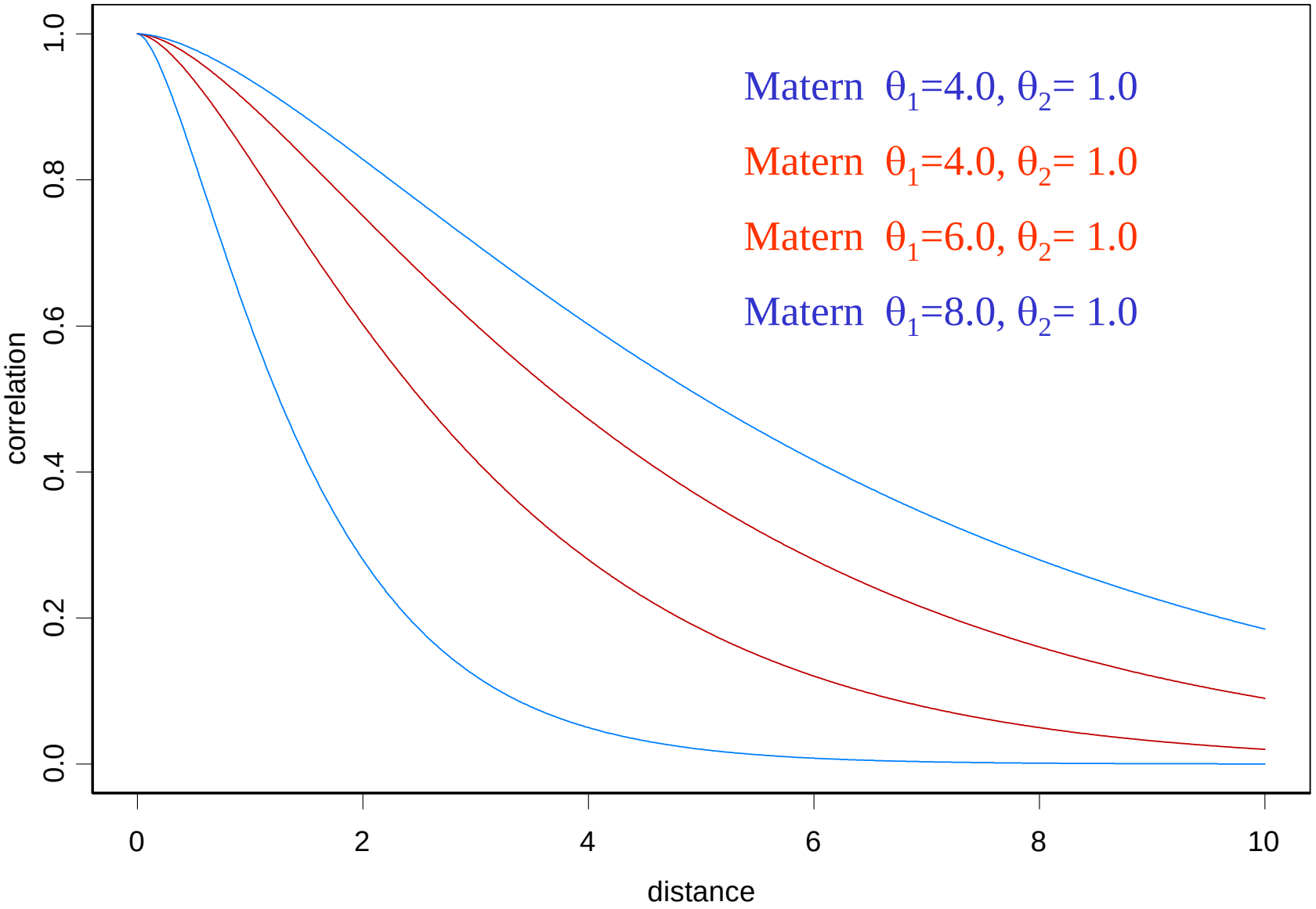


Effect of Range

Variables in Model	$\theta_1 = 2$		$\theta_1 = 4$		$\theta_1 = 6$		$\theta_1 = 8$	
	spat	ind	spat	ind	spat	ind	spat	ind
X_1, X_2, X_3	15	2	56	2	66	5	71	4
X_1, X_2, X_3, X_5	6	0	14	0	11	0	10	0
X_1, X_2, X_3, X_4	7	0	11	0	19	2	14	1
X_1, X_2	22	8	10	8	0	14	0	21
Intercept only	0	22	0	27	0	25	0	12
X_1	14	20	0	14	0	15	0	23
X_2	1	8	0	14	0	11	0	9

- 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 one of the models.

Effect of range (cont)



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 Error

Simulations:

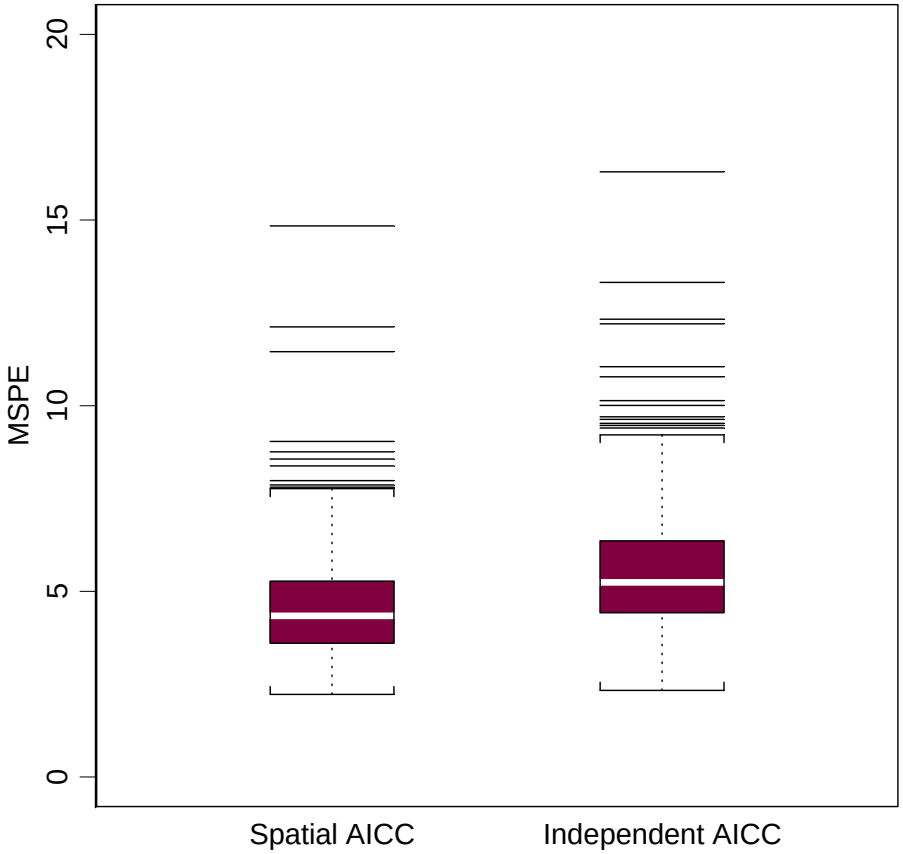
- Model selection and estimation using 100 observations.
- Used fitted model to predict at 100 additional locations.
- Evaluated performance using mean square prediction error

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

where $\hat{Z}_j$ is the universal kriging predictor for the j^{th} prediction location.

MSPE

- Spatial AICC is 16.9% improvement-methods agreed only 11 times.
- Spatial AICC is 39.6% better over independent AICC with indep noise.
- Predictive coverage was about the same (.92 for spatial AIC, .95 for indep error AIC)



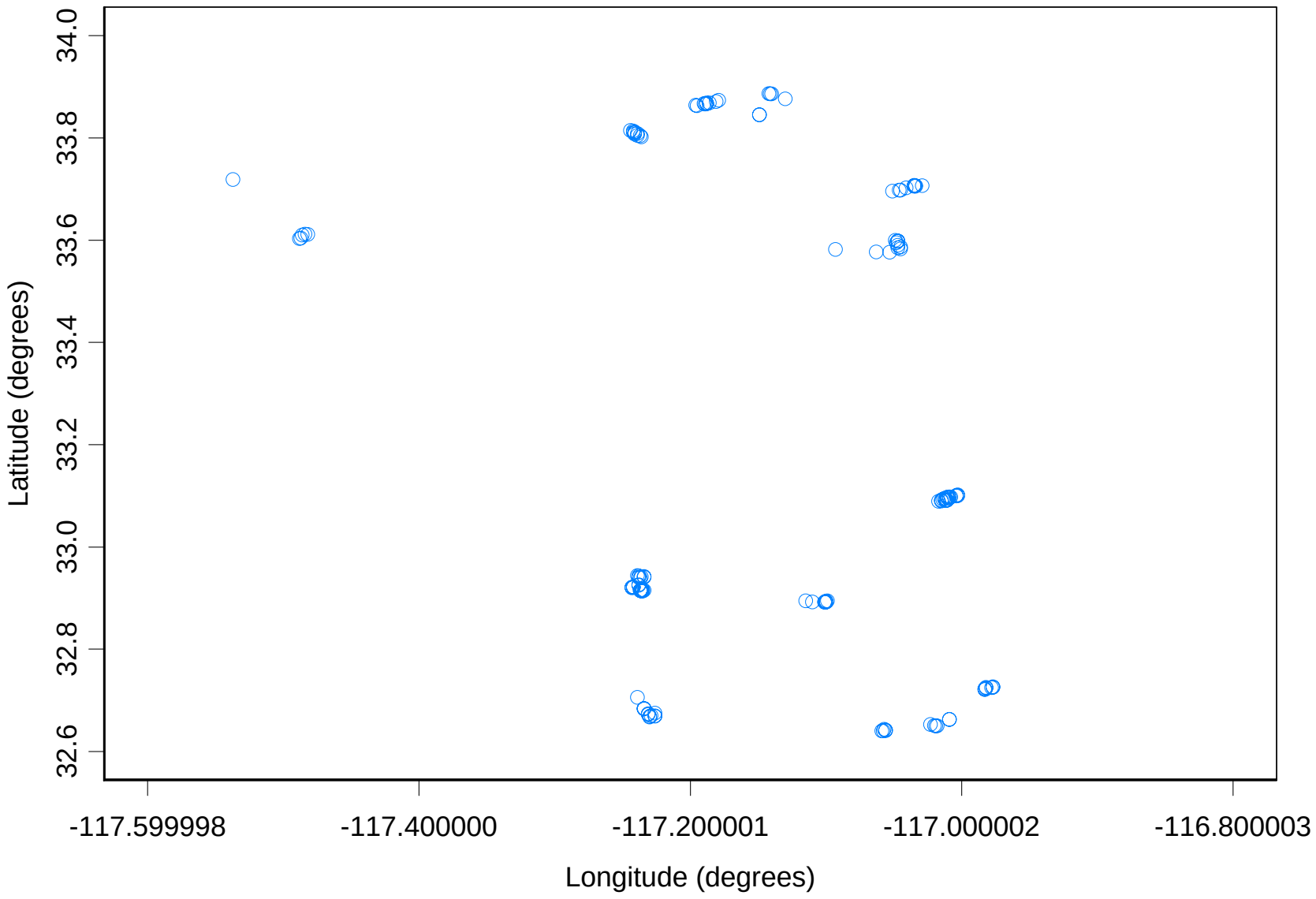
Example: Lizard abundance

Abundance for the orange-throated whiptail lizard in southern California
Ver Hoef, Cressie, Fisher, Case (2001).

Data:

- 147 locations
- $Z(s_i)$ = log (average number of lizards caught per day at location s_i)
- Explanatory variables. 37 variables ($2^{37} = 1.374*10^{11}$ models) reduced to the following 6.
  - ant abundance (three levels)
  - log (% sandy soils)
  - elevation
  - barerock indicator
  - % cover
  - log(% chaparral plants)
- 160 possible models

Locations (Southern California) of Whiptail Lizards



Lizard abundance (cont)

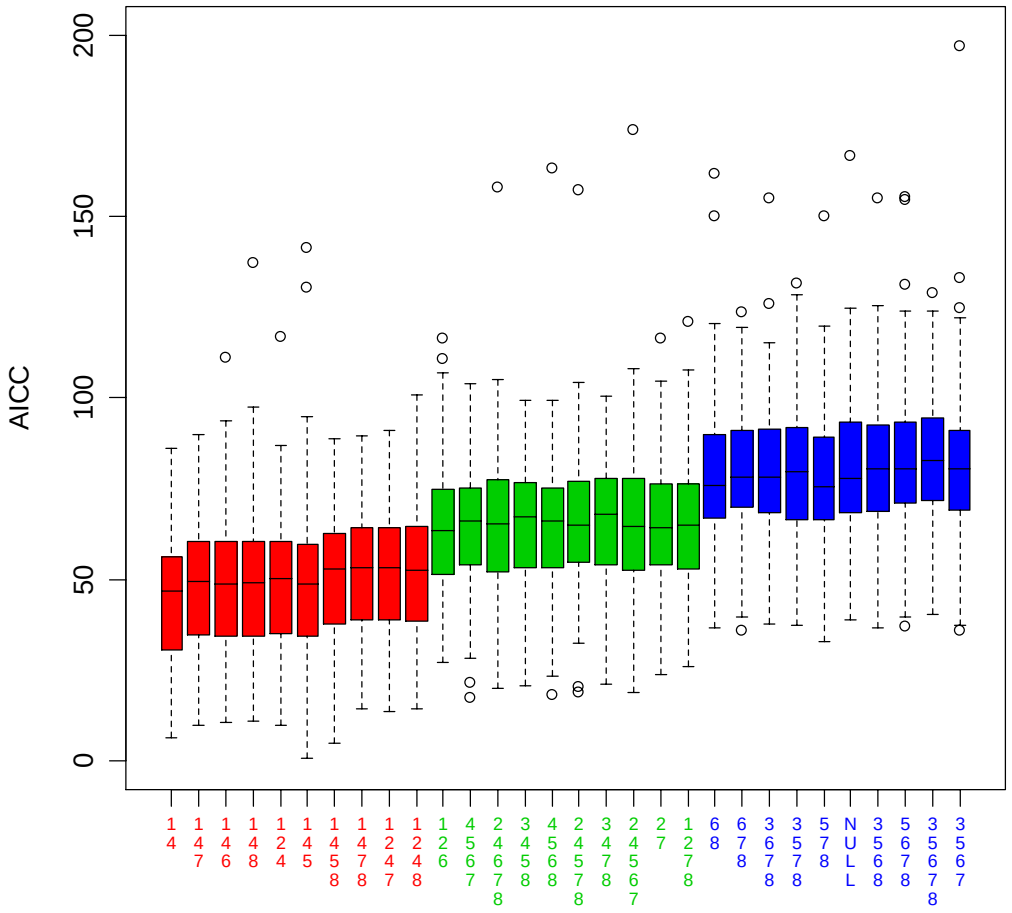
Predictors	AIC	Spatial Rank	Ind Rank
Ant_1, % sand	54.1	1	66
Ant_1, $Ants_2$, % sand	54.8	2	56
Ant_1, % sand, % cover	55.7	3	59
Ant_1, Ant_2 , % sand, % cover, elevation, barerock, % chaparral	92.2	41	1
Ant_1, Ant_2 , % sand, elevation, barerock, % chaparral	95.3	33	2
Ant_1, % sand, % cover, elevation, barerock	95.7	38	3

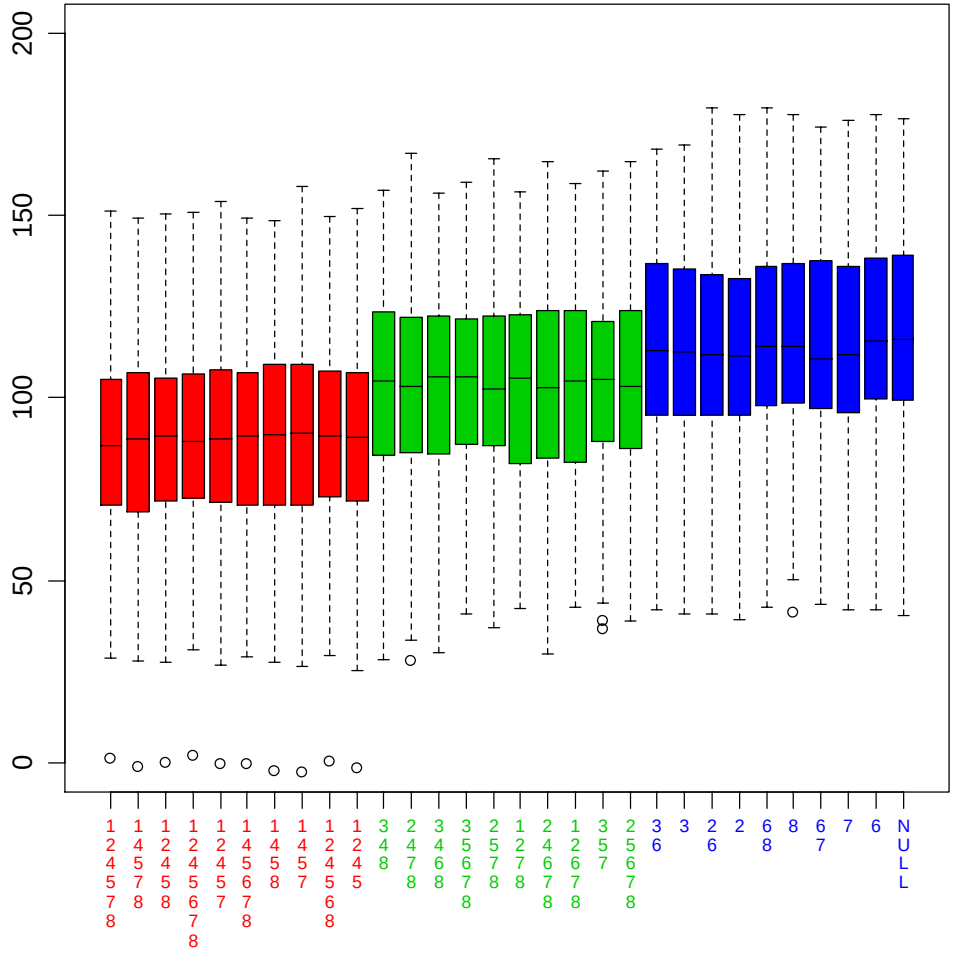
Note: MDL chooses the same model as AICC.

Lizard abundance—simulation

Simulation of model: use covariates Ant_1, % sand, with Matern covariance function

Key: **1 = ant$_1$**, 2 = ant$_2$, 3 = ant$_3$, **4 = % sand**, 5 = elevation,
6 = barerock, 7 = cover, 8 = chaparral plants





Summary Remarks

1. *Spatial correlation* should not be ignored when selecting explanatory variables.

2. *Model choice* for prediction should involve *joint selection* of the explanatory variables and the form of the autocorrelation function.

3. *Sampling patterns* can severely impact *model selection*.

4. *MDL* offers a nice philosophical alternative to AIC for modeling complex phenomena.