

Geostatistical Modeling: Model Selection and Parameter Estimation

Jennifer A. Hoeting
Department of Statistics
Colorado State University
www.stat.colostate.edu/~jah

This talk is based on work with two research groups which include

- Richard Davis and Andrew Merton, CSU
- Alix Gitelman and Kathi Georgitis, OSU

This work was partially funded by STAR Research Assistance Agreement CR-829095 awarded to Colorado State University by the U.S. Environmental Protection Agency (EPA). The views expressed here are solely those of authors. EPA does not endorse any products or commercial services mentioned here.

The Geostatistical Model

Model:

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

where

- $\mathbf{X}(s) = (1, X_1(s), \dots, X_p(s))'$ is a vector of explanatory variables observed at location s
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ is a $p + 1$ 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 zero and covariance function

$$\text{Cov}(\delta(s), \delta(t)) = \begin{cases} \sigma^2 + \tau^2 & \text{if } d = 0 \\ \tau^2 \rho(\boldsymbol{\theta}, d) & \text{if } d > 0 \end{cases}$$

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

Parameter estimation and model selection

Parameter estimation:

A variety of approaches are available including

- Bayesian parameter estimation
- Maximum (profile) likelihood estimation
- REML (Restricted maximum likelihood)

In this talk we focus on MLE and REML

Model selection:

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

GOALS OF THIS TALK

We are investigating the theoretical and practical implications of changing the

- covariance parameters
- sampling design
- sample size

on model selection and parameter estimation

Some of the related work in this area

Estimation (and prediction) for spatial models

- Mardia and Marshall (1984)
- Zimmerman and Zimmerman (1991)
- Zimmerman and Cressie (1992)
- Cressie (1993)
- Abt (1999)
- Lark (2000)
- Zimmerman (2005)

Asymptotics for spatial models

- Ying (1991 and 1993)
- Chen, Simpson, Ying (2000)
- Stein (1999)
- Zhang (2004)
- Zhang and Zimmerman (2005)

GOALS OF RESEARCH: Part 1

What are the implications of changing the

- covariance parameters:

- range

- ratio of nugget (sampling variance) to total variance (nugget + partial sill)

- sampling design: clustered, random, grid

- sample size: increasing sample size for a fixed domain (infill)

on parameter estimation?

Estimation: Simulation set-up

We consider the basic model:

$$Z(s) = \beta_0 + \delta(s),$$

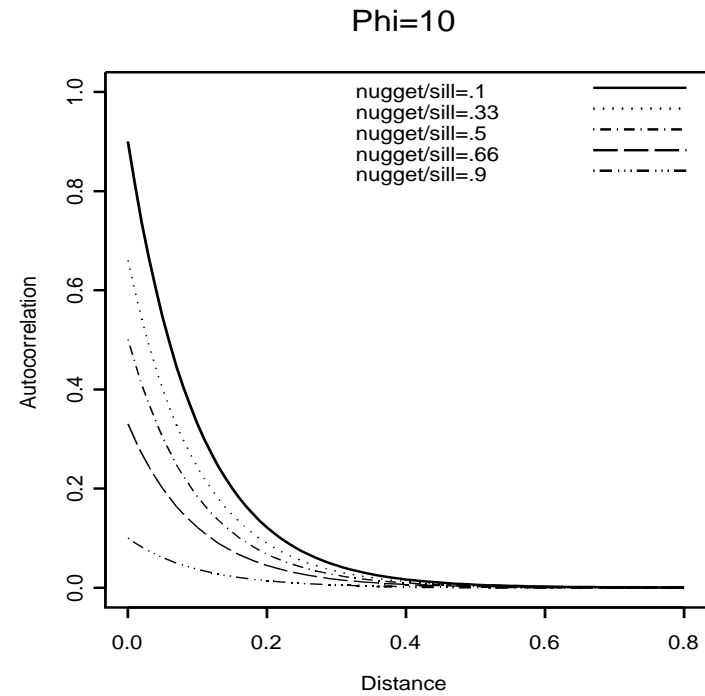
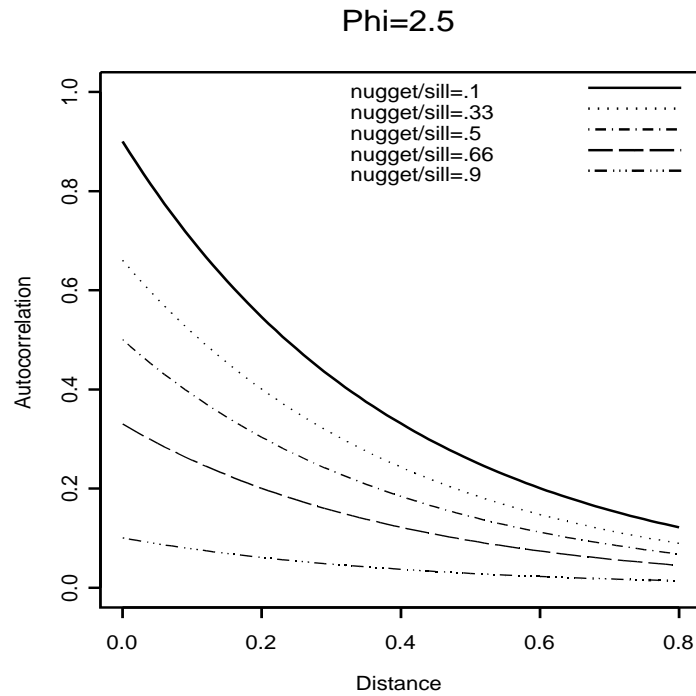
where $\delta(s)$ is a stationary, isotropic Gaussian process with mean zero and covariance function

$$\text{Cov}(\delta(s), \delta(t)) = \begin{cases} \sigma^2 + \tau^2 & \text{if } d = 0 \\ \tau^2 \exp\{-\phi d\} & \text{if } d > 0 \end{cases}$$

Investigations (to date) include 100 realizations for different spatial correlations and ranges:

- sill is fixed: $\tau^2 + \sigma^2 = 3$
- range = 0.4 and 0.1, so $\phi = 2.5$ and 10
- nugget/sill ratio = $\frac{\sigma^2}{\sigma^2 + \tau^2} = \frac{1}{10}, \frac{1}{3}, \frac{1}{2}, \frac{2}{3}, \frac{9}{10}$

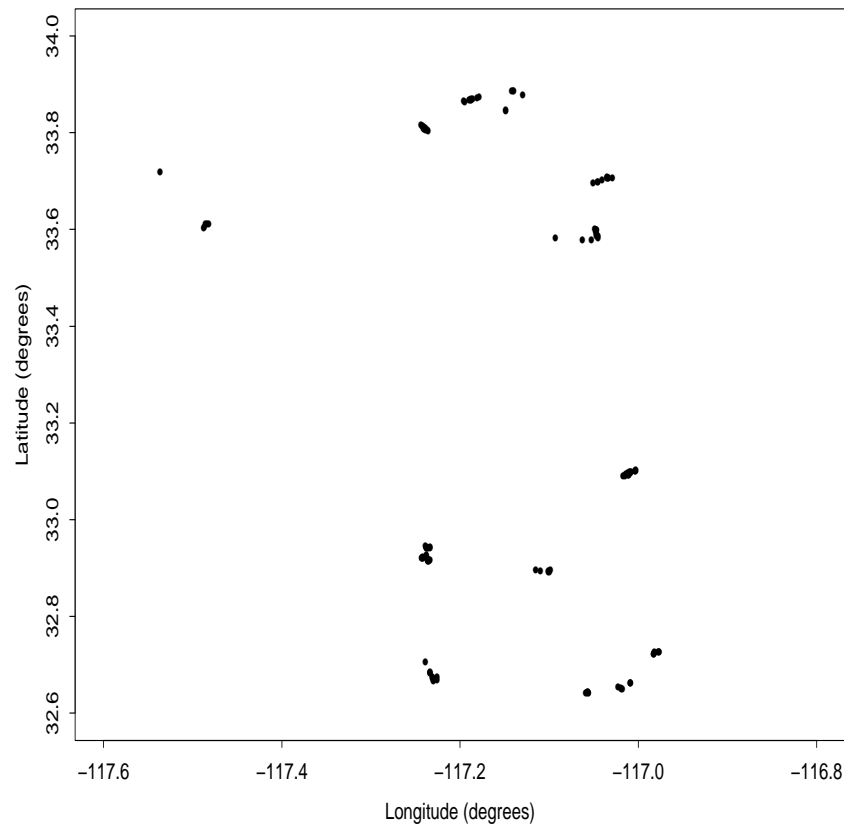
Autocorrelation functions used in simulations



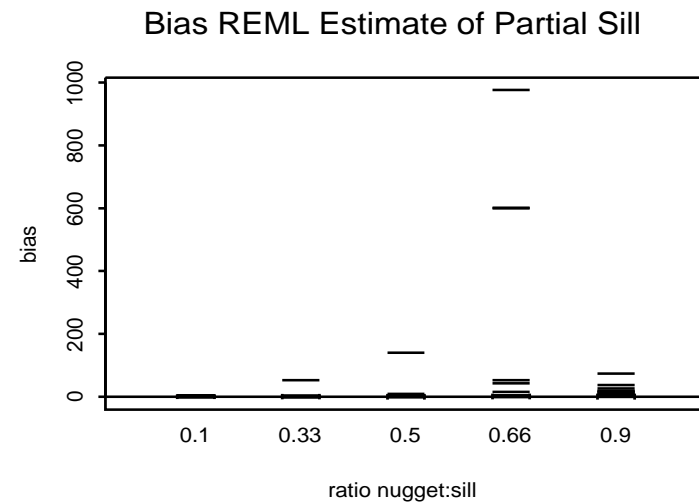
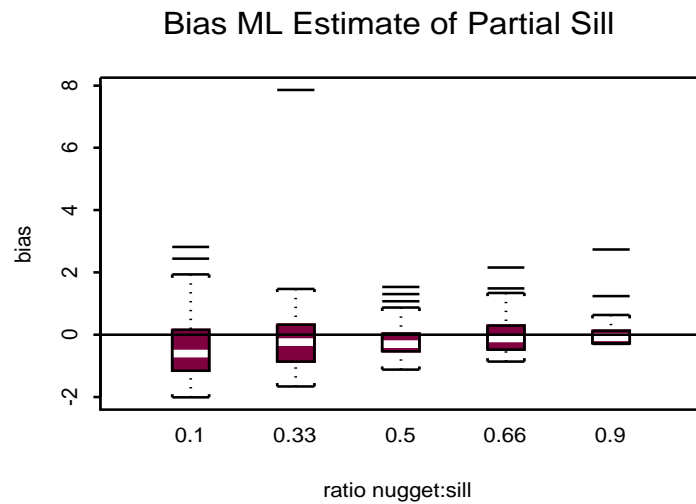
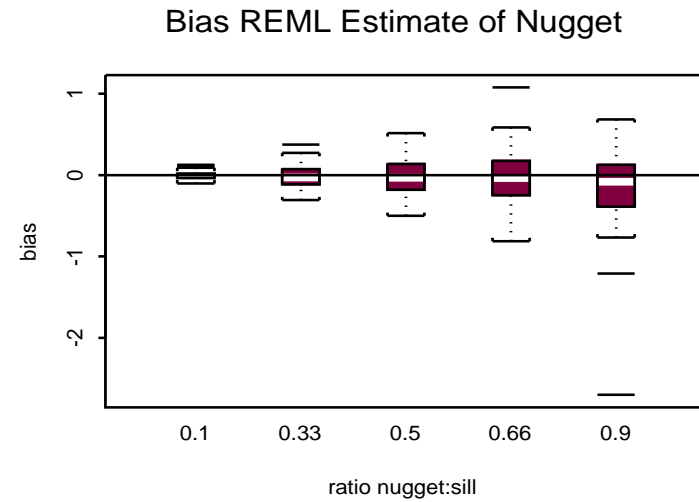
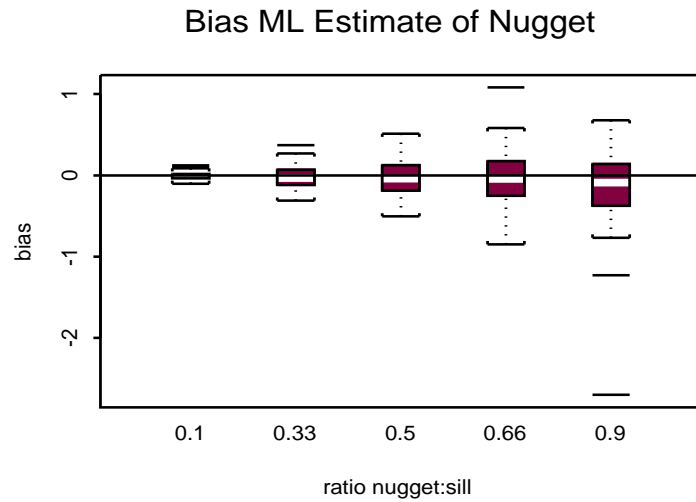
Sampling design for simulations

Based on a study for the orange-throated whiptail lizard abundance at 147 sites in southern California

Ver Hoef, Cressie, Fisher, Case (2001)

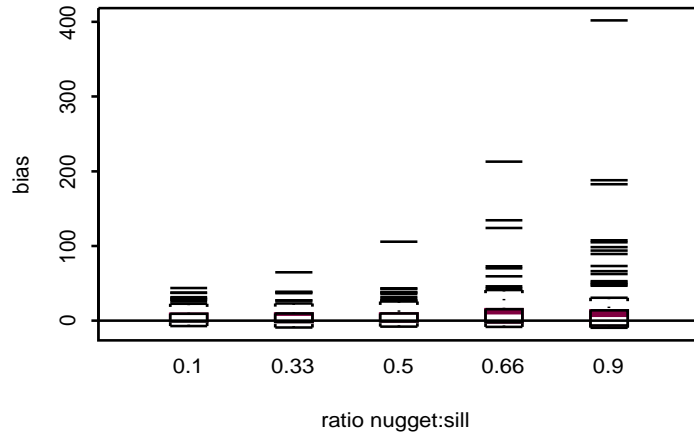


Estimation plots: $\phi = 10$

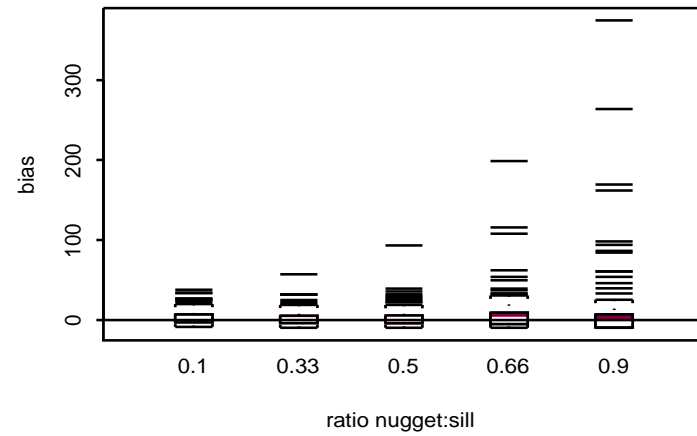


Estimation plots: $\phi = 10$

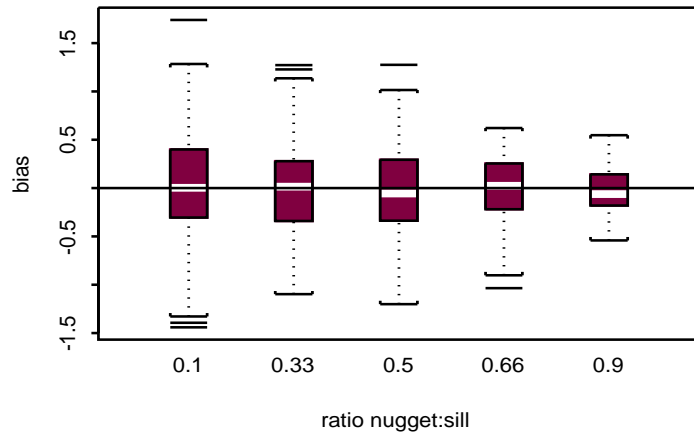
Bias ML Estimate of Phi



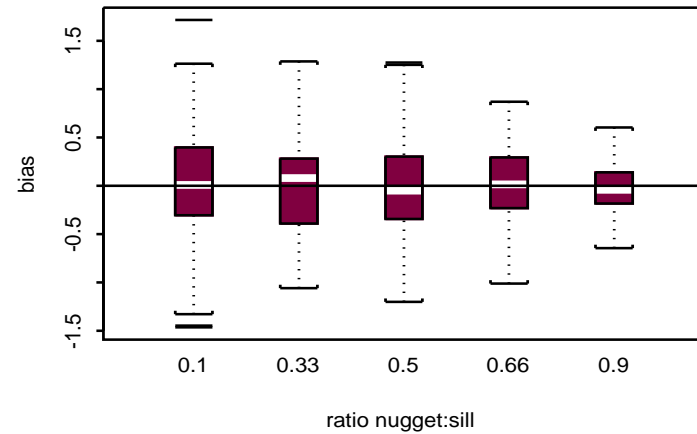
Bias REML Estimate of Phi



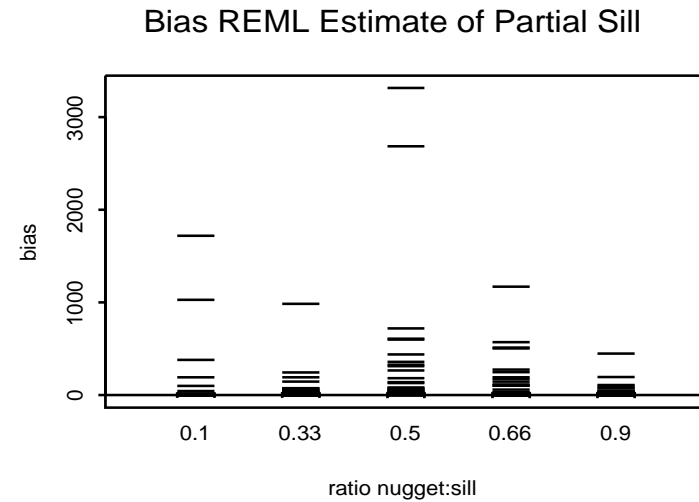
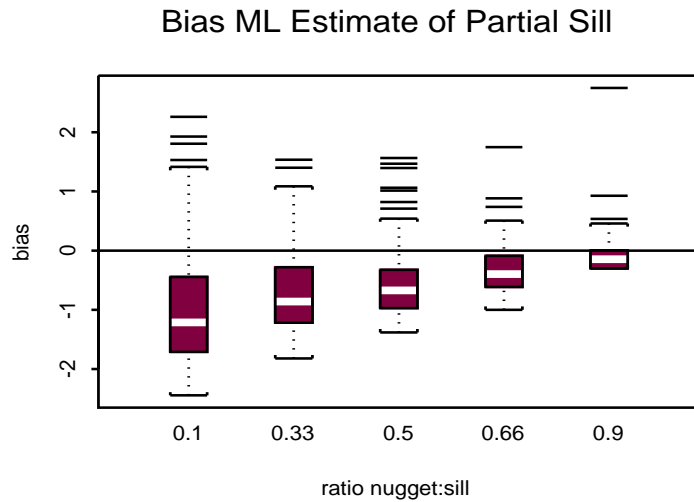
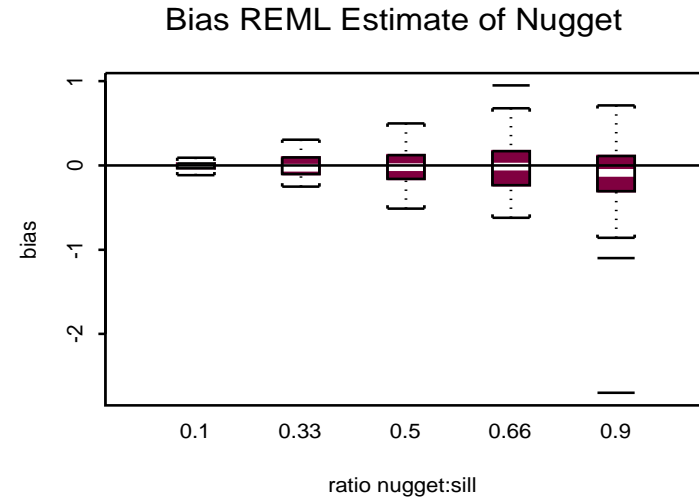
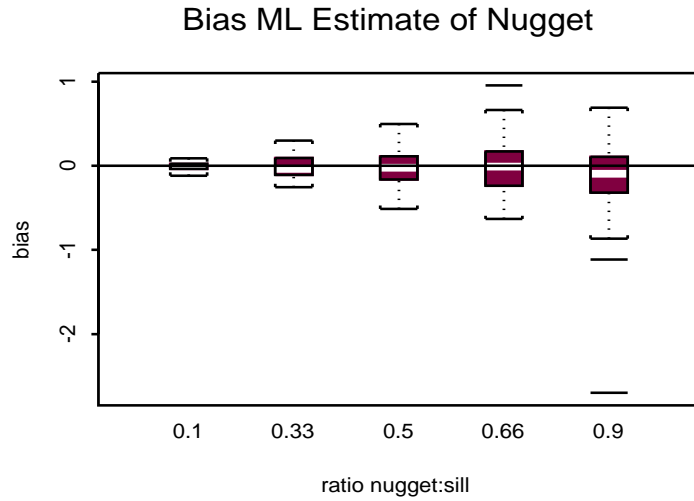
Bias ML Estimate of Mean



Bias REML Estimate of Mean

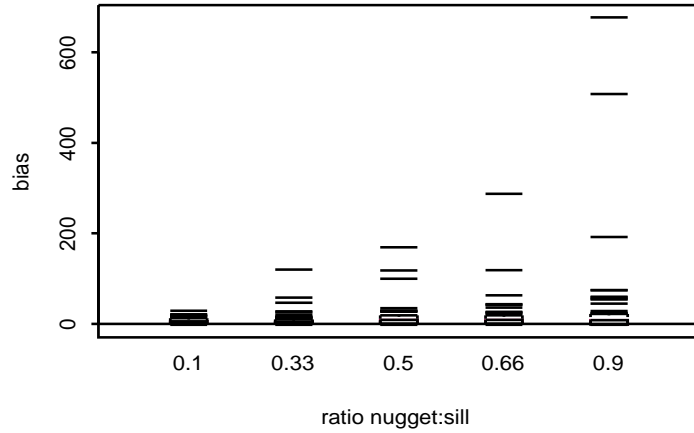


Estimation plots: $\phi = 2.5$

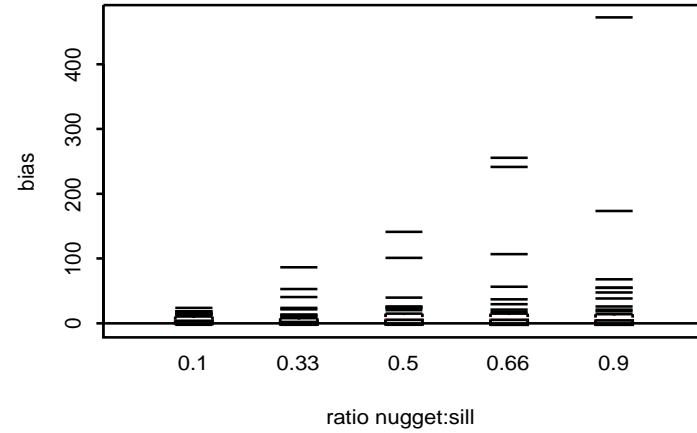


Estimation plots: $\phi = 2.5$

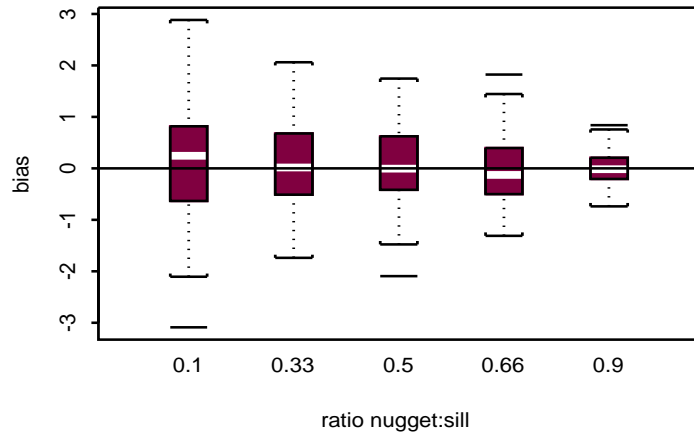
Bias ML Estimate of Phi



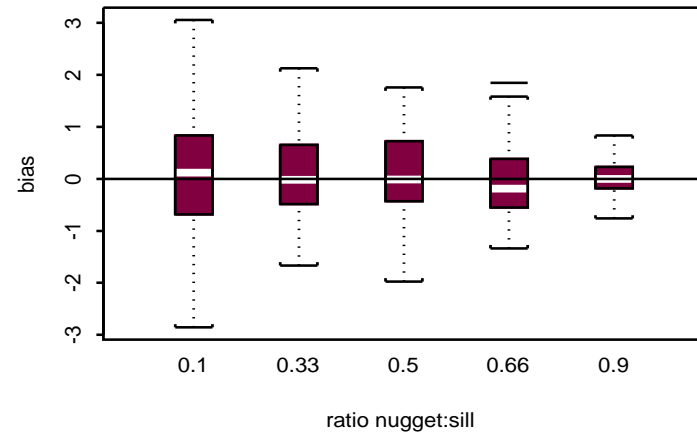
Bias REML Estimate of Phi



Bias ML Estimate of Mean



Bias REML Estimate of Mean



Simulation findings to date

$\sigma^2 = \text{Nugget}$

- Similar results for $\phi = 2.5$ and 10
- ML and REML estimates show more variability as the spatial signal decreases

$\tau^2 = \text{Partial sill}$

- ML: bias becomes more negative as range decreases (stronger spatial signal)
- REML: does poorly and influenced by outliers

$\phi = 1/\text{Range}$

- ML and REML estimates are highly skewed, especially as ϕ increases
- The variability of ML and REML estimates increases as ϕ decreases

GOALS OF RESEARCH: Part 2

What are the theoretical and practical implications of changing the

- covariance parameters
- sampling design: 5 designs from clustered to grid
- sample size: infill and increasing domain asymptotics

on model selection and parameter estimation?

Model Selection: AIC for Geostatistical Models

Let

k = number of parameters in covariance function

p = number of explanatory variables

The quantity, referred to as the corrected AIC,

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

is an approximately unbiased estimate of the expected Kullback-Leibler information evaluated at the maximum likelihood estimates of the parameters.

The standard AIC statistic is given by

$$AIC = -2 \log L_Z(\hat{\beta}, \hat{\theta}, \hat{\sigma}^2, \hat{\tau}^2) + 2(p + k + 1).$$

Initial conclusions related to model selection

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. Sampling patterns that offer observation pairs at small and larger distances may be advantageous for model selection.

Available at my webpage: www.stat.colostate.edu/~jah

- J.A. Hoeting, R. A. Davis, A. A. Merton, and S. E. Thompsen (2005) “Model Selection for Geostatistical Models”, to appear in *Ecological Applications*.
- R software for geostatistical model selection

Additional considerations for model selection

Motivating question: The AIC statistic is based on a number of asymptotic approximations. Are these valid for the geostatistical model?

Asymptotic frameworks for spatial data

Infill asymptotics: observations are taken ever more densely in a fixed and bounded domain

Increasing domain: Minimum distance between sampling points is bounded away from 0 and the spatial domain of observation is unbounded

For the remainder of this talk:

- Increasing n corresponds to infill.
- Increasing m corresponds to increasing domain.

First order auto-regressive process

Consider the continuous AR(1) process in one-dimensional space. The solution to the stochastic differential equation

$$dY_t = -\phi Y_t dt + \sigma \sqrt{2\phi} e^{\phi t} dB_t,$$

where dB_t is a standard Brownian motion, is known as the *Ornstein-Uhlenbeck process*, is

$$Y_t = e^{-\phi(t-s)} Y_s + \sigma \sqrt{2\phi} \int_s^t e^{-\phi(t-u)} dB_u$$

where $s < t$ and $\phi > 0$.

It can be shown that the discrete AR(1) process exactly coincides with the continuous case at the observed locations. So, let $\{Y_t\}$ be AR(1) process that satisfies the recursions

$$Y_{t_i} = e^{-\phi/n} Y_{t_{i-1}} + \varepsilon_{t_i}, \quad t_i = \{i/n; i = 1, \dots, mn\}$$

where $\{\varepsilon_{t_i}\} \stackrel{\text{iid}}{\sim} \mathbf{N}(0, \sigma^2(1 - e^{-2\phi/n}))$.

Estimation under infill and increasing domain asymptotics

Many authors have investigated the consequences of infill or increasing domain alone, we are interested in investigating the distribution of an estimator of ϕ as both as both $n \rightarrow \infty$ and $m \rightarrow \infty$.

To simplify notation, define

$$\alpha_n = \exp \{ -\phi/n \}$$

Mann and Wald (1943) showed that for with $t_0=0$ the least squares estimator for α_n is given by

$$\hat{\alpha}_n = \frac{\sum_{i=2}^{mn} Y_{t_i} Y_{t_{i-1}}}{\sum_{i=2}^{mn} Y_{t_{i-1}}^2}$$

Theorem

Let Y be an mn -vector of observations generated by the continuous AR(1) process as defined by equation () such that $Y_i = Y_{t_i}$ where $t_i = \{i/n; i = 1, \dots, mn\}$. Define $\hat{\phi} = -n \log \hat{\alpha}_n$ where $\hat{\alpha}_n$ is the maximum likelihood estimate of $\alpha_n = e^{-\phi/n}$ given above.

1. For fixed n and $m \rightarrow \infty$,

$$\sqrt{m}(\hat{\phi} - \phi) \xrightarrow{d} N\left(0, n(e^{2\phi/n} - 1)\right).$$

2. As $n \rightarrow \infty$ and $m \rightarrow \infty$,

$$\sqrt{m}(\hat{\phi} - \phi) \xrightarrow{d} N(0, 2\phi).$$

Do these asymptotic results hold for realistic sample sizes?

Simulation set-up

Step 1:

- Generate the t_i 's: Generate complete grid for all simulations (all $m \times n$ points)

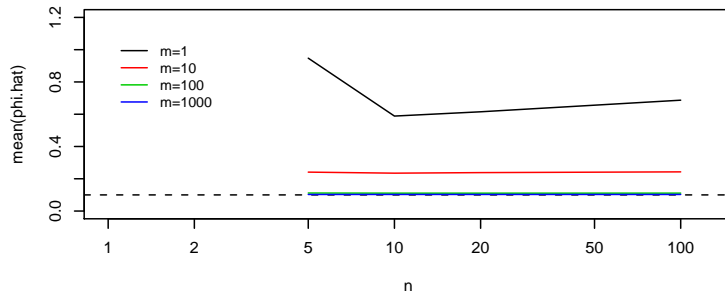
Step 2:

- Generate the Z 's: Generate all realizations at all $m \times n$ points
- Subsample for each appropriate n and m combination

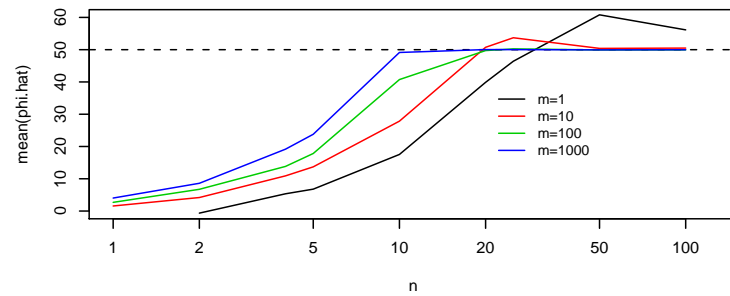
Repeat step 2 100 times

Effects of infill and increasing domain

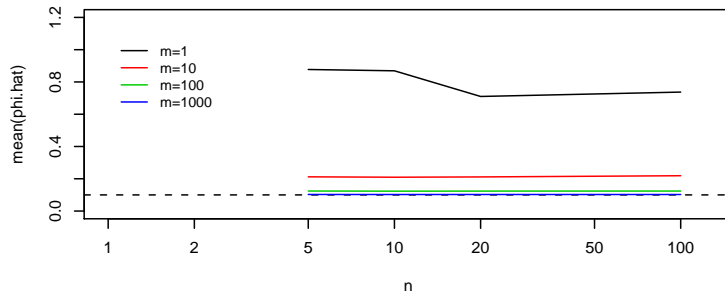
$\phi = 0.1, \sigma^2 = 0.25$



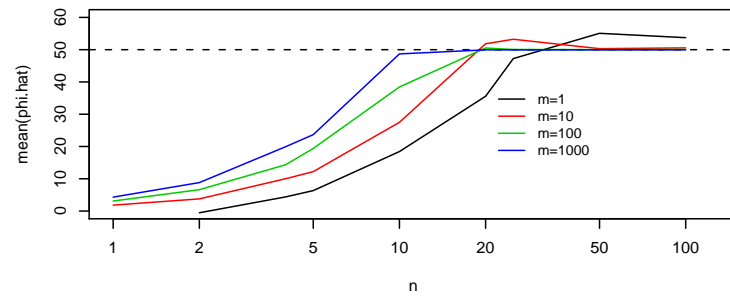
$\phi = 50.0, \sigma^2 = 0.25$



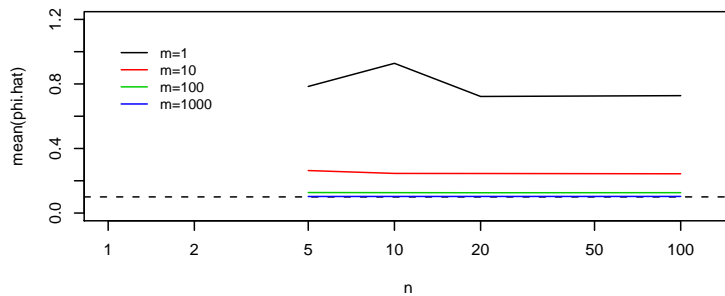
$\phi = 0.1, \sigma^2 = 1$



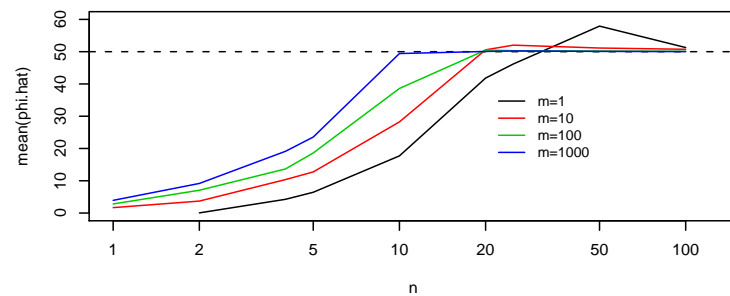
$\phi = 50.0, \sigma^2 = 1$



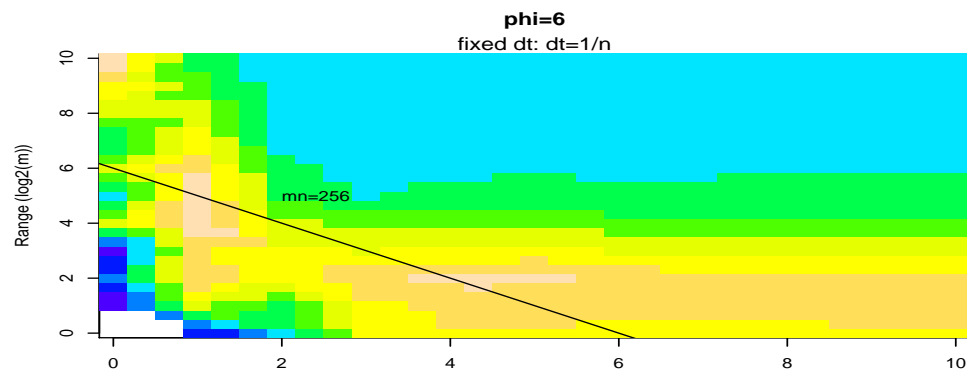
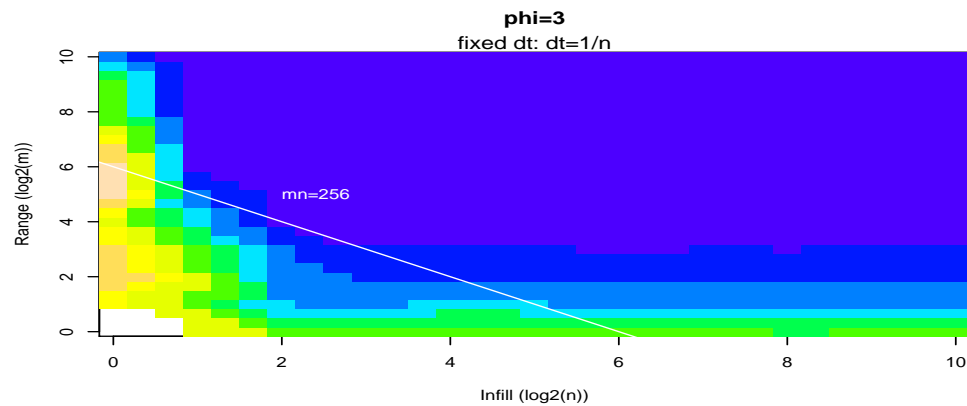
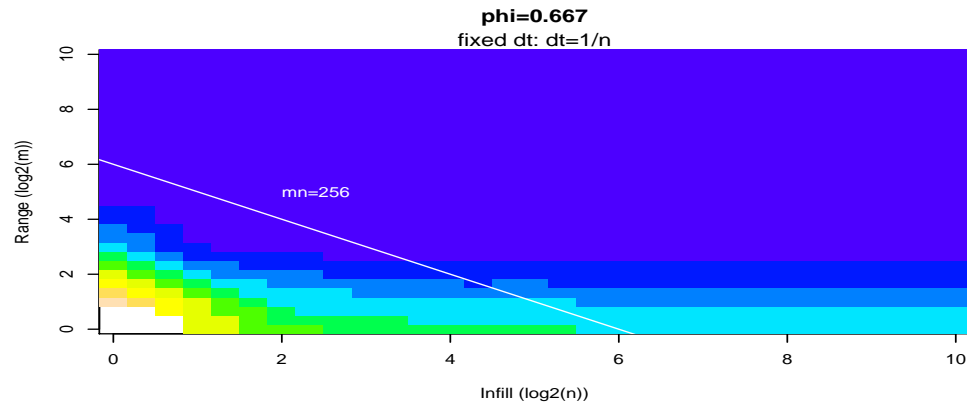
$\phi = 0.1, \sigma^2 = 4$



$\phi = 50.0, \sigma^2 = 4$



Effects of infill and increasing domain: Plot of $\hat{\phi} - \phi$



Simulation results to investigate effects of infill and increasing domain

Simulation results based on the AR(1) model described above have indicated the following

- For a fixed domain m , increasing infill (n) has little impact on improving the bias
- Increasing the domain m rapidly reduces the bias towards zero for fixed n
- For large ϕ , the number of sites per unit length (n) influences the rate at which the bias tends toward zero.

Future work

Much more work to be done and some of it is already in progress.

Stay tuned for upcoming papers and PhD dissertations by Andrew Merton and Kathi Georgitis!