HANDOUT

SPATIAL STATISTICS COURSE

UNIVERSITY OF WARWICK

SUMMER 2007

DR. FUENTES

Summer 2007
SPATIAL STATISTICS

Instructor:

Dr. Montserrat Fuentes
email: fuentes@stat.ncsu.edu
webpage: www.stat.ncsu.edu/people/fuentes/stwarwick/ (*include this site in your Bookmarks*)
Office Hours: by appointment

Description:

The course will cover the methodology and modern developments for spatial-temporal modeling estimation and prediction, Bayesian spatial statistics and hierarchical frameworks, spatial point processes, and spectral analysis of spatial processes. This course goes beyond standard practices and exposes the students to all the new developments and state of the art modern techniques for spatial data. All the methods presented will be introduced in the context of a specific dataset, then the motivation behind a particular method will be evident as it is developed. Some of the lectures will be held in the computer lab using the software R and WinBUGS, previous knowledge of R will not required.

Course prerequisites:

Linear Models and Variance Components.

Textbook:

The course material will be based on a set of notes being prepared by the instructor. This book is a good reference:

- *Hierarchical Modeling and Analysis for Spatial Data.* Banerjee, Carlin and Gelfand. Chapman and Hall.

Other recommended books:

- *Statistics for Spatial Data.* Noel Cressie. Wiley & Sons. 1993. (more complete and more advance level. It is a very good reference book but at an advance level.)
- *Interpolation of Spatial Data.* M. Stein. Springer, 1999. (Very advance level, this is a good reference book for spatial statistics in the spectral domain)

Schedule:

The lectures will be 1 hour and 20 minutes everyday at 10am from May 21 to June 1 (both days inclusive). No class on May 28th.

On May 23 and May 30 the class will be held on the computer lab.

Labs:

Sometimes the class will be held in the computer lab, The software used for this course is R and WinBUGS.

Lecture Notes:

Lecture notes and handouts will be available on the web,

webpage: `www.stat.ncsu.edu/people/fuentes/stwarwick/`

Objectives:

This course will cover a number of areas of spatial statistics and data assimilation applied to real, scientific and interesting problems. A tentative list of more specific topics is as follows:

- Introduction to spatial statistics:
  - Point level models
  - Areal (lattice) models
  - Spatial point processes.
- Estimation and modeling of spatial correlations:
  - estimating variogram
  - fitting parametric models: Matern class
  - maximum likelihood estimation
  - restricted maximum likelihood
- Prediction and Interpolation (kriging):
  - Spatial regression
  - Kriging
  - frequentist corrections for unknown covariance structure
  - model misspecification in kriging
- Bayesian spatial statistics:
  - Bayesian estimation
  - Bayesian kriging
  - Bayesian priors for covariance parameters
  - Hierarchical Bayesian methods.
- Spatial-temporal processes.
  - point-level modeling with continuous time
  - nonseparable models
  - dynamic space-time models
  - block-level modeling
  - misaligment problem.
- Nonstationary spatial processes:

- Bayesian deformation approaches
- eigenfunction expansion of the covariance (EOFs)
- kernel based methods
- mixing of process distributions

- Spectral domain:
  - Fourier Theory
  - Spectral Representation of a Spatial Process
  - Spectral Density and periodogram
  - Spectral methods to approximate the likelihood
  - Increasing domain asymptotics
  - Infill asymptotics

Some of the lectures will be conducted in the computer lab. Students will learn how to use existing software, the emphasis of the course is to learn the methodology needed to do research on spatial statistics and to analyze real data from the environmental, biomedical, geological and agricultural sciences. The methods will be introduced with examples.

Example of point processes

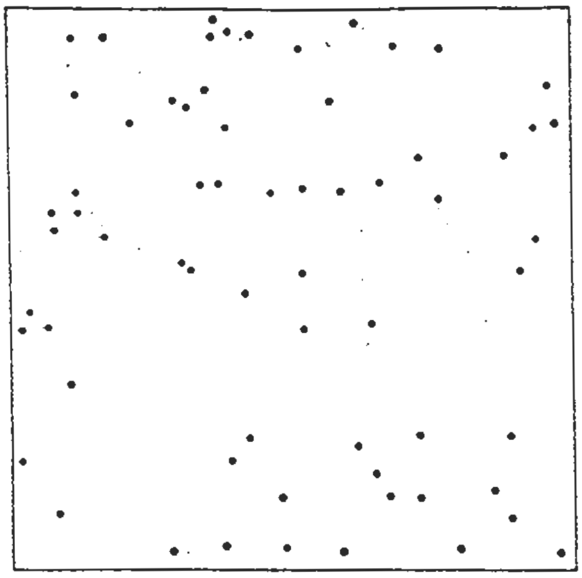


Fig. 1. Locations of 65 Japanese black pine saplings in a square of side 5.7 m (Numata, 1961).

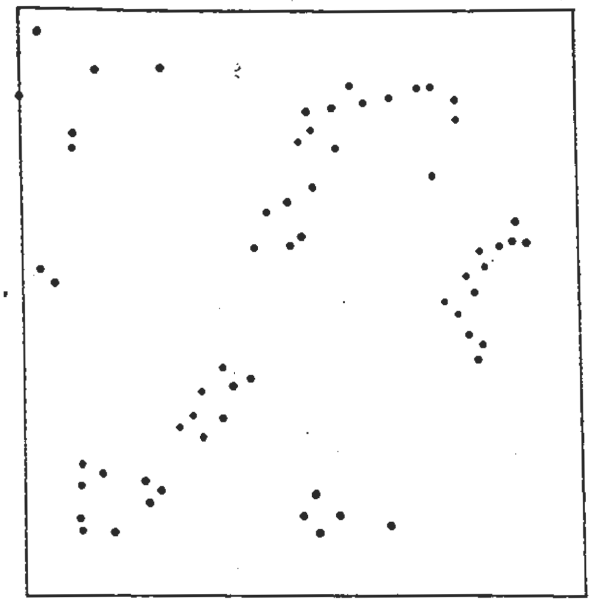


Fig. 2. Locations of 62 redwood seedlings in a square of side 23 m (Strauss, 1975; Ripley, 1977).

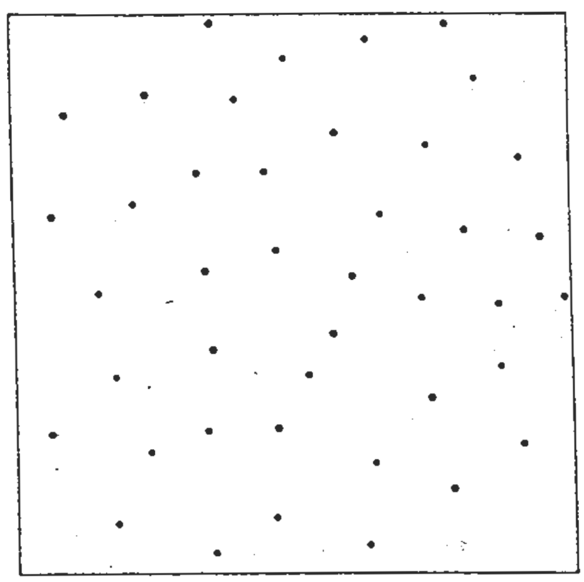
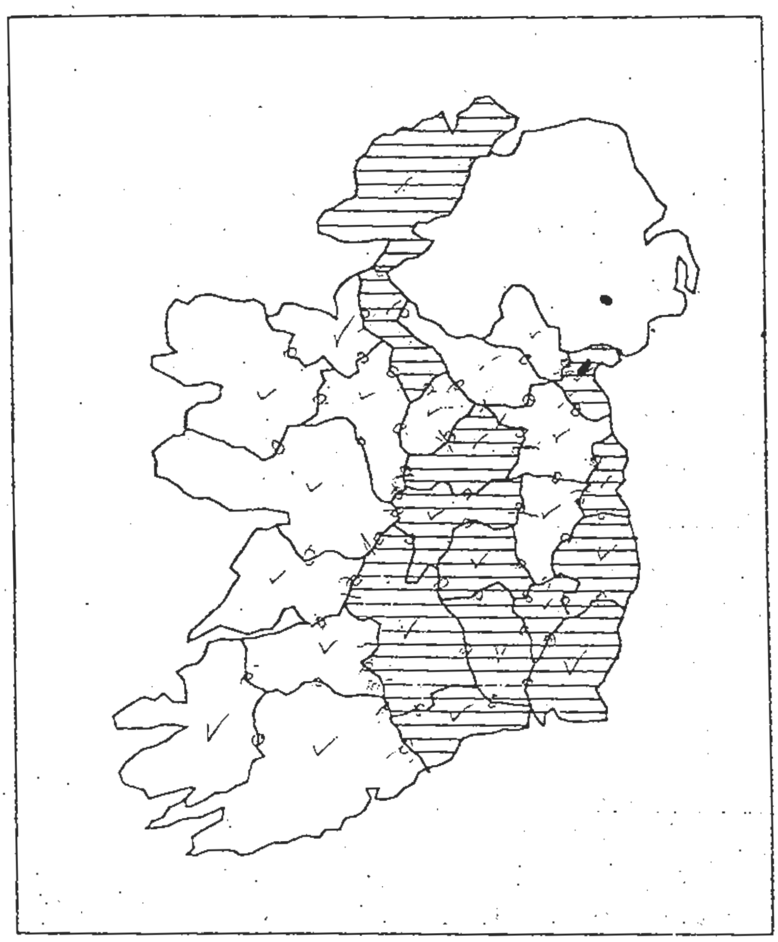


Fig. 3. Locations of 42 cell centres in a unit square (Ripley, 1977).

Fuentes

Example of areal data

2 6 counties of Eire



Key: ⊟ = Above median frequency (of A Allele gene)

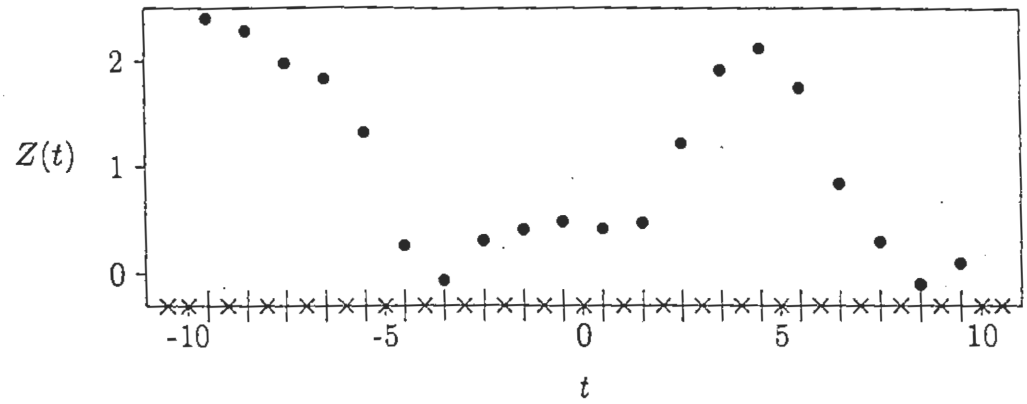
< 26 counties of ~~Eire~~>

— genetic relation

— irregulary spaced.

1

Example by M. Stein.



* Empirical semivariogra

FIGURE 9. Simulated realization of Gaussian process with mean 0 and autocovariance function $K(t) = e^{-0.4|t|}(1 + 0.4|t|)$. The xs on the horizontal axis are the locations of predictands and the |s are the locations of the observations.

TABLE 4. Simulated values of process pictured in Figure 9. The last three rows are the additional observations used towards the end of this section.

t	$Z(t)$	t	$Z(t)$
−9.5	2.3956811	0.5	0.4109609
−8.5	2.2767195	1.5	0.4647669
−7.5	1.9736058	2.5	1.2113779
−6.5	1.8261141	3.5	1.9055446
−5.5	1.3136954	4.5	2.1154852
−4.5	0.2550507	5.5	1.7372076
−3.5	−0.0741740	6.5	0.8333657
−2.5	0.2983559	7.5	0.2932142
−1.5	0.4023333	8.5	−0.1024508
−0.5	0.4814850	9.5	0.0926624
−0.25	0.4267716		
0.0	0.4271087		
0.25	0.4461579		

Additional observations

PARAMETERS:

$\sigma^2 = 1$ (sill)

$C_0 = 0$ (nugget)

$\theta_2 = 1.5$ (smoothing)

$\theta_1 = 6.12$ (Matern spatial scale parameter "range")

$^{-.4|t|}(1 + .4|t|)$

tein
variance

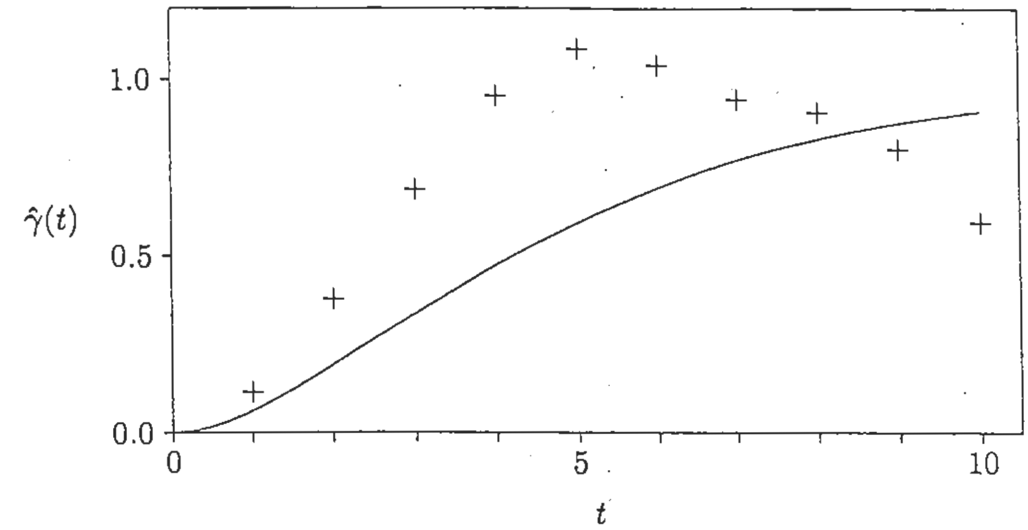


FIGURE 10. Empirical and actual semivariograms for data shown in Figure 9. Smooth curve is the actual semivariogram and +s are the empirical semivariogram.

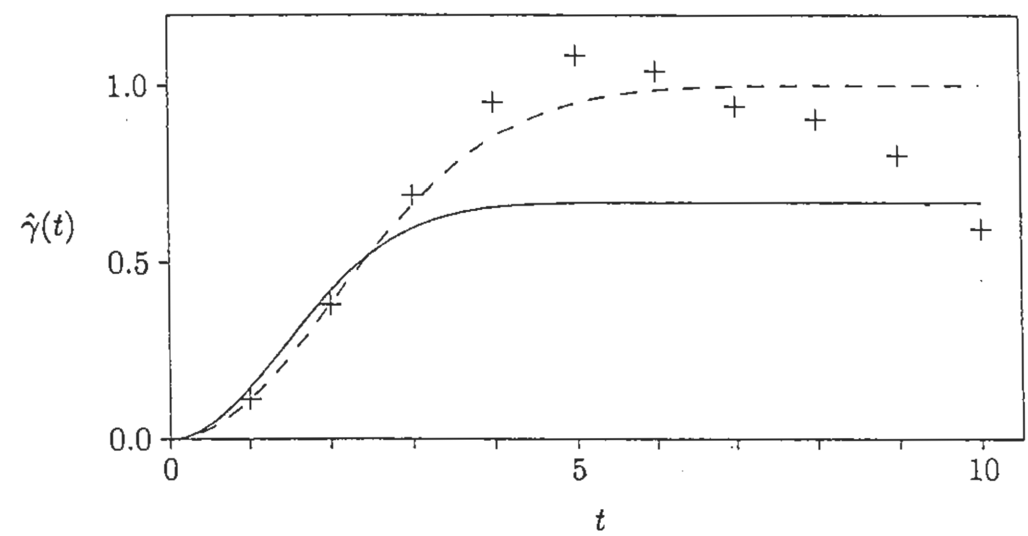


FIGURE 11. Empirical and estimated semivariograms for data shown in Figure 9. Using the Gaussian model for the semivariogram, solid line indicates REML estimate and dashed line indicates eyeball estimate.

eye-ball estimate matches the empirical semivariogram distinctly better than the REML estimate at the shorter distances. Furthermore, comparing Figures 10 and 11 shows that the eyeball estimate is visually closer to the true semivariogram than the REML estimate. Is this evidence that the REML estimate is inferior to the eyeball fit in this example?

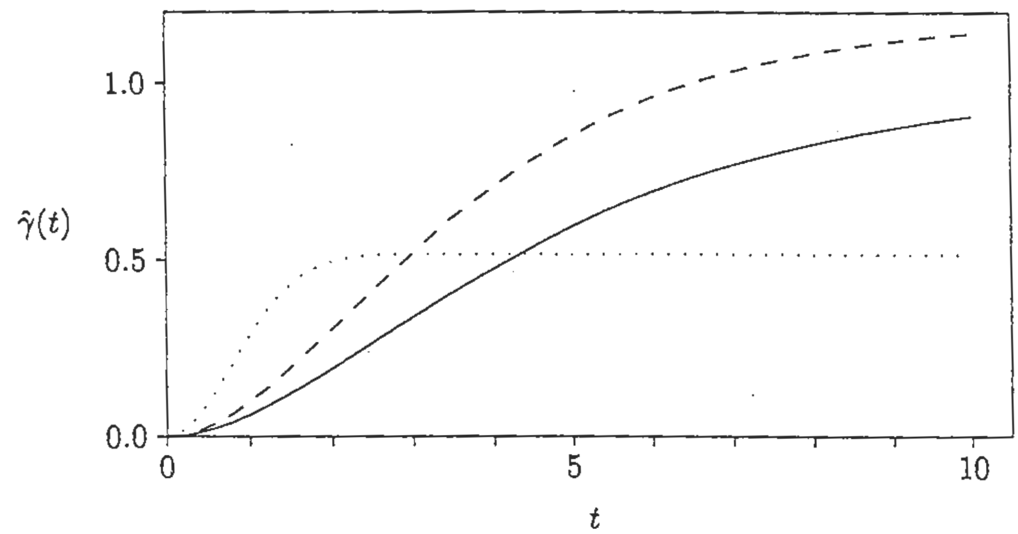


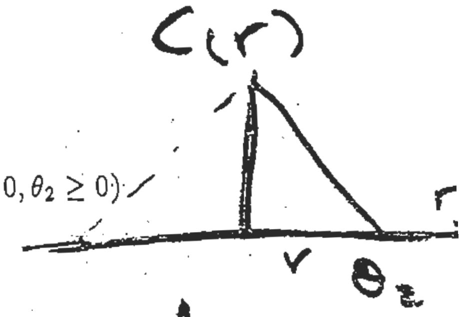
FIGURE 15. True and estimated semivariograms using 3 additional observations. The solid line indicates the truth, the dashed line the REML estimate under the Matérn model and the dotted line the REML estimate under the Gaussian model.

Dr. Fuentes

Let us now give some examples of isotropic, parametric covariance function models (letting $r = \|\mathbf{h}\|$). Note that each model has a corresponding parameter space for which the covariance function is valid.

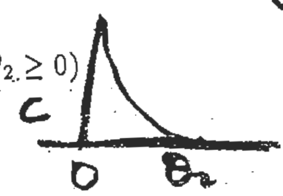
- *Tent (triangular, piecewise linear) model* (valid in R^1 only)

$$C(r;\theta) = \begin{cases} \theta_1(1 - r/\theta_2) & \text{for } 0 \le r \le \theta_2 \\ 0 & \text{for } r > \theta_2 \end{cases} \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$



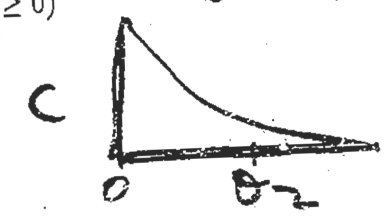
- *Spherical model*

$$C(r;\theta) = \begin{cases} \theta_1\left(1 - \frac{3r}{2\theta_2} + \frac{r^3}{2\theta_2^3}\right) & \text{for } 0 \le r \le \theta_2 \\ 0 & \text{for } r > \theta_2 \end{cases} \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$



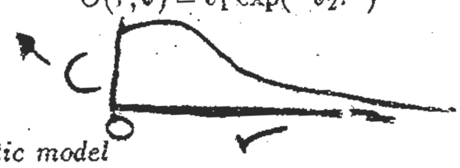
- *Exponential model*

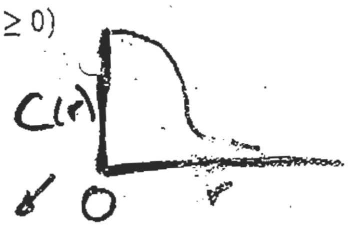
$$C(r;\theta) = \theta_1 \exp(-\theta_2 r) \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$



- *Gaussian model*

$$C(r;\theta) = \theta_1 \exp(-\theta_2 r^2) \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$





- *Rational quadratic model*

$$C(r;\theta) = \theta_1\left(\theta_2 - \frac{r^2}{1 + r^2/\theta_2}\right) \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$

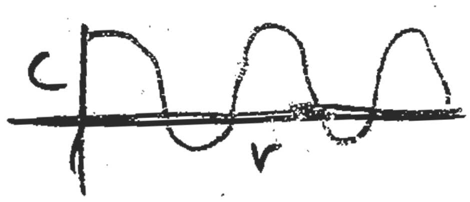
- *Matern class of models*

$$C(r;\theta) = \theta_1 \frac{1}{2^{\theta_3 - 1}\Gamma(\theta_3)}\left(\frac{2r\sqrt{\theta_3}}{\theta_2}\right)^{\theta_3} K_{\theta_3}\left(\frac{2r\sqrt{\theta_3}}{\theta_2}\right) \qquad (\theta_1 \ge 0, \theta_2 \ge 0, \theta_3 > 0)$$

- *Cosine model*

$$C(r;\theta) = \theta_1 \cos(r/\theta_2) \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$

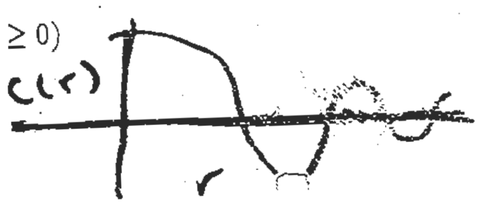
not realistic



- *Wave or hole-effect model*

$$C(r;\theta) = \theta_1 \theta_2 \frac{\sin(r/\theta_2)}{r} \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$

eg. crop yield data
(have periodic behavior)



Examples of valid isotropic semivariograms (with corresponding parameter spaces):

- *Tent* (valid in R^1 only).

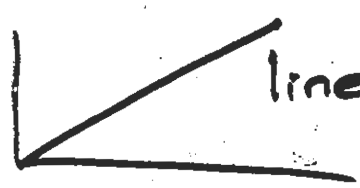
$$\gamma(r;\theta) = \begin{cases} \theta_1 r/\theta_2 & \text{for } 0 \le r \le \theta_2 \\ \theta_1 & \text{for } r > \theta_2 \end{cases} \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$

- *Exponential.* $\gamma(r;\theta) = \theta_1\{1 - \exp(-\theta_2 r)\}$ $\quad (\theta_1 \ge 0, \theta_2 \ge 0)$
- *Spherical.*

$$\gamma(r;\theta) = \begin{cases} \theta_1\left(\frac{3r}{2\theta_2} - \frac{r^3}{2\theta_2^3}\right) & \text{for } 0 < r \le \theta_2 \\ \theta_1 & \text{for } r > \theta_2 \end{cases} \qquad (\theta_1 \ge 0, \theta_2 \ge 0)$$

- *Gaussian.* $\gamma(r;\theta) = \theta_1\{1 - \exp(-\theta_2 r^2)\}$ $\quad (\theta_1 \ge 0, \theta_2 \ge 0)$
- *Rational quadratic.* $\gamma(r;\theta) = \theta_1 \frac{r^2}{1 + r^2/\theta_2}$ $\quad (\theta_1 \ge 0, \theta_2 \ge 0)$
- *Matern.* $\gamma(r;\theta) = \theta_1\left(1 - \frac{1}{2^{\theta_3 - 1}\Gamma(\theta_3)}\left(\frac{2r\sqrt{\theta_3}}{\theta_2}\right)^{\theta_3} K_{\theta_3}\left(\frac{2r\sqrt{\theta_3}}{\theta_2}\right)\right)$ $\quad (\theta_1 \ge 0, \theta_2 \ge 0, \theta_3 > 0)$
- *Cosine.* $\gamma(r;\theta) = \theta_1\{1 - \cos(r/\theta_2)\}$ $\quad (\theta_1 \ge 0, \theta_2 \ge 0)$
- *Wave or hole-effect model.* $\gamma(r;\theta) = \theta_1(1 - \theta_2 \sin(r/\theta_2))$ $\quad (\theta_1 \ge 0, \theta_2 \ge 0)$.
- *Linear.* $\gamma(r;\theta_1) = \theta_1 r$ $\quad (\theta_1 \ge 0)$
- *Power.* $\gamma(r;\theta) = \theta_1 r^{\theta_2}$ $\quad (\theta_1 \ge 0, 0 \le \theta_2 < 2)$

> only exists for intrinsic stationary processes



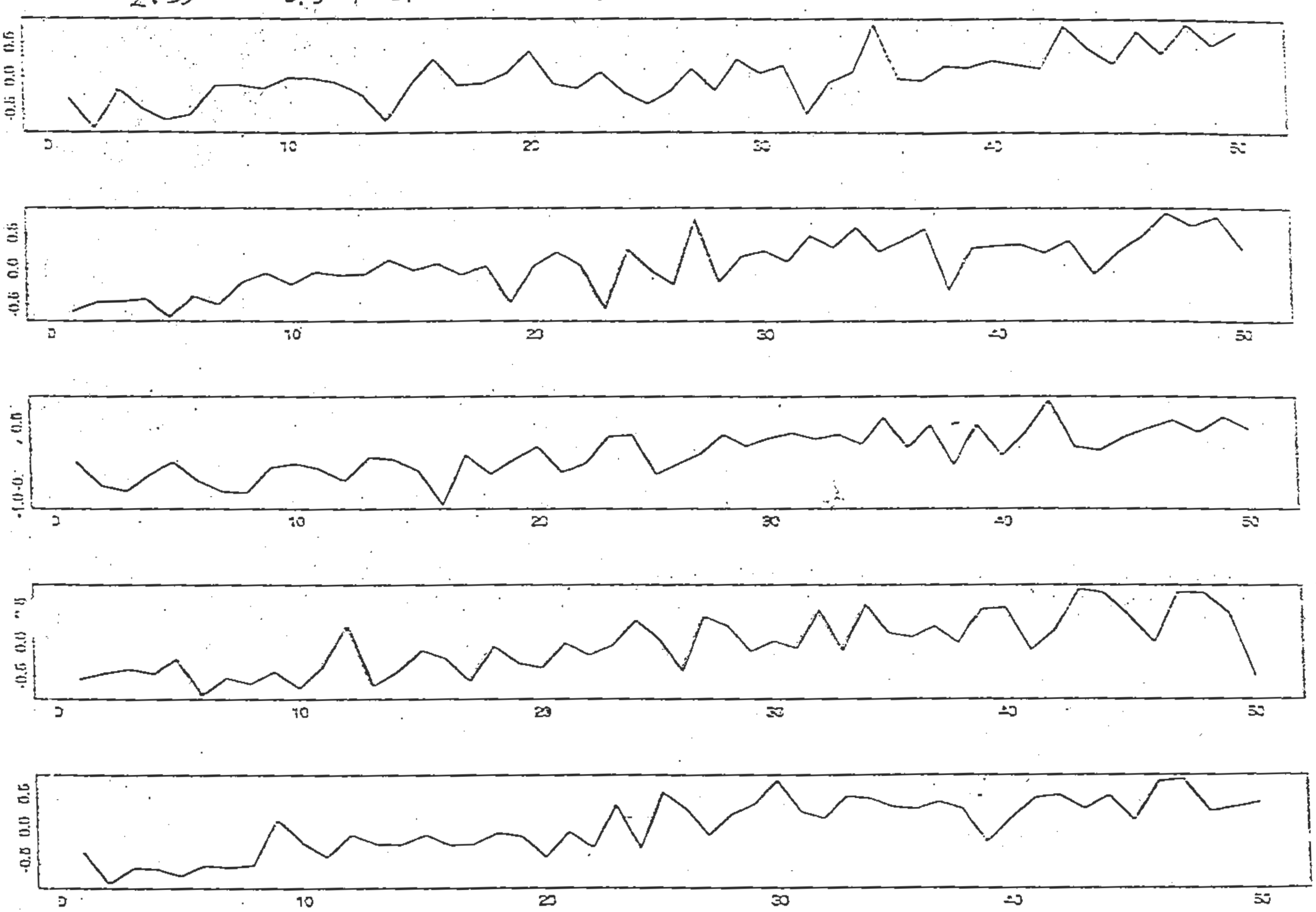
linear

(a) Spatial trend, no correlation

iid $N(0,1)$

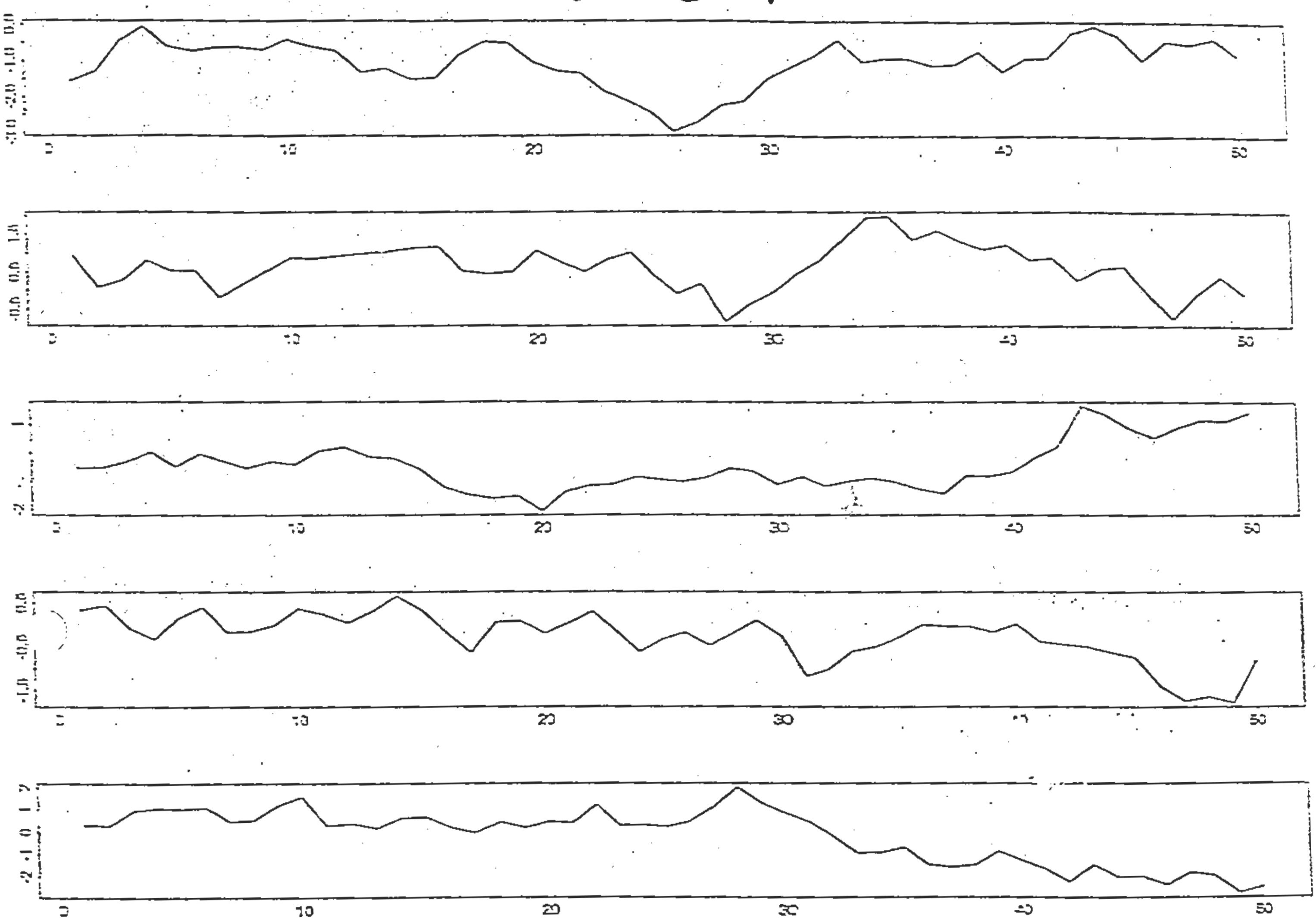
— 1 dim — 50 obs.

$Z(s) = -0.5 + 0.02s + \epsilon(s)$



(b) Spatial correlation: no trend

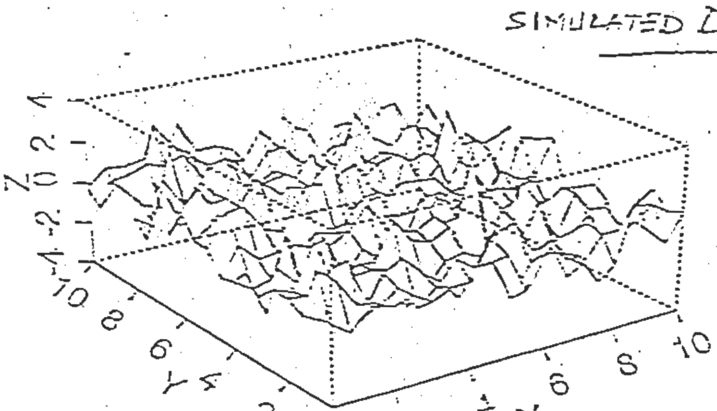
$Z(s) = 0 + \epsilon(s)$ → $C(r) = e^{-r/10}$



ST 733

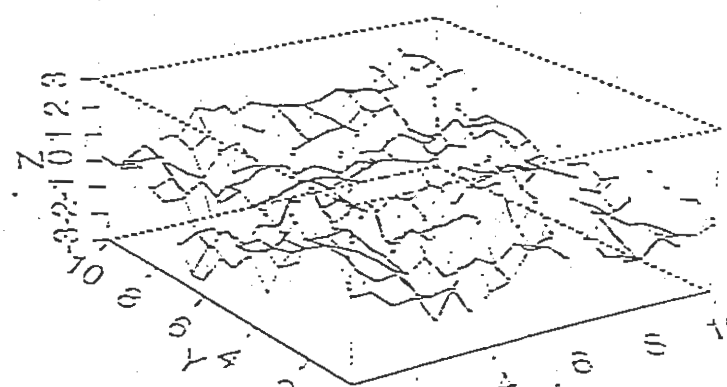
Dr Fuentes

SIMULATED DATA



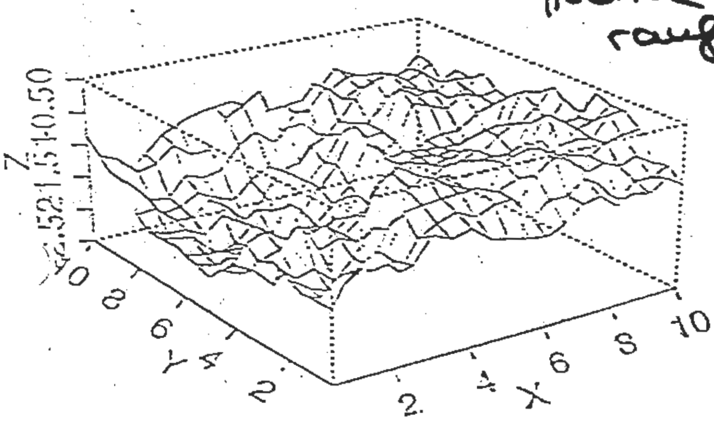
$\theta_0 = 0$; $\theta_2 = \underline{0.01}$

effective range



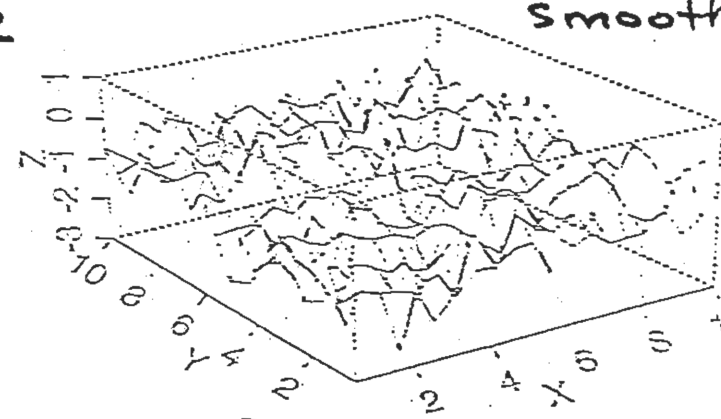
$\theta_0 = 0$. $\theta_2 = \underline{2.0}$

smoother



$\theta_0 = 0$, $\theta_2 = \underline{20.0}$

a lot more smoother



$\overline{\theta_0 = 0.25}$, $\theta_2 = 20.0$

nugget (noise)

θ_0 adds additional variation

$\left(\frac{.25}{1.+.25} = 20\% \text{ variation} \right)$

SIMULATED DATA HAVING AN EXPONENTIAL COVARIANCE.

DATA LOCATIONS LIE ON THE GRID

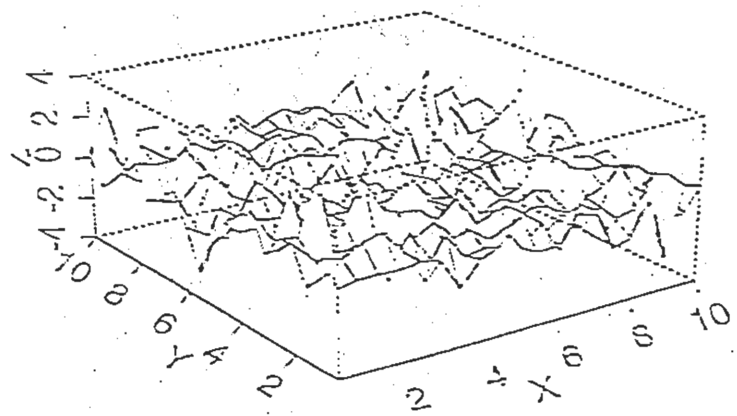
$\{ x, y : x = .5, 1, 1.5, \ldots, 10; \; y = .5, 1, \ldots, 10 \}$

θ_0 = nugget

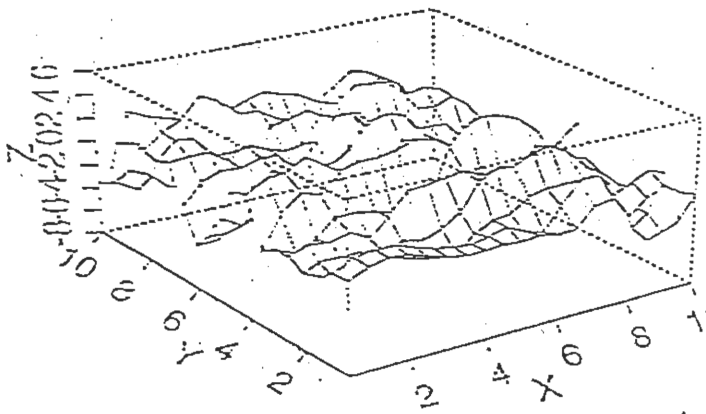
θ_1 = sill = 1 = Var(Z)

θ_2 = 'effective range'

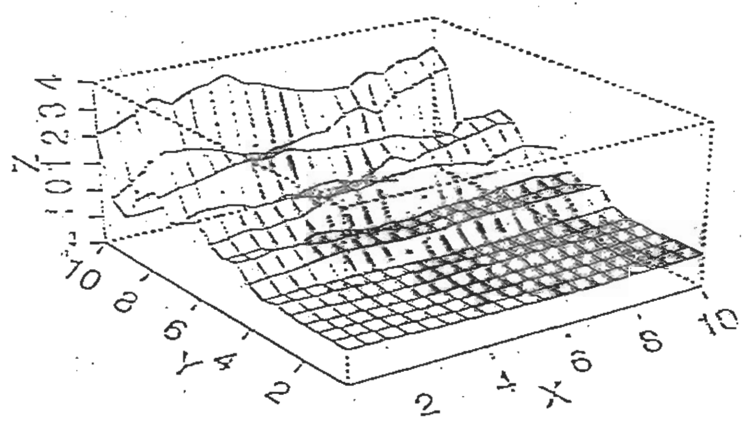
SIMULATED DATA



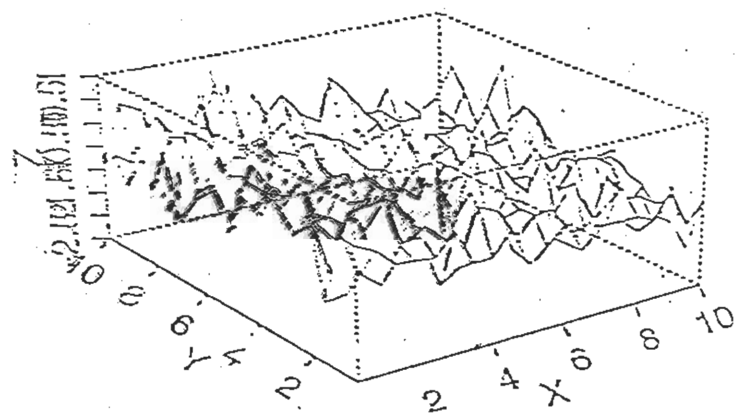
$\theta_0 = 0.\ \theta_2 = 0.01$



$\theta_0 = 0.\ \underline{\theta_2 = 2.0}$ ← effective range



$\theta_0 = 0;\ \theta_2 = \underline{20.0}$



$\theta_0 = 0.25,\ \theta_2 = 20.0$

large range: more highly correlated

SIMULATED DATA HAVING A GAUSSIAN COVARIANCE.

DATA LOCATIONS LIE ON THE GRID

$\{x, y : x = .5, 1, 1.5, \ldots, 10 ; y = .5, 1, \ldots, 10\}$.

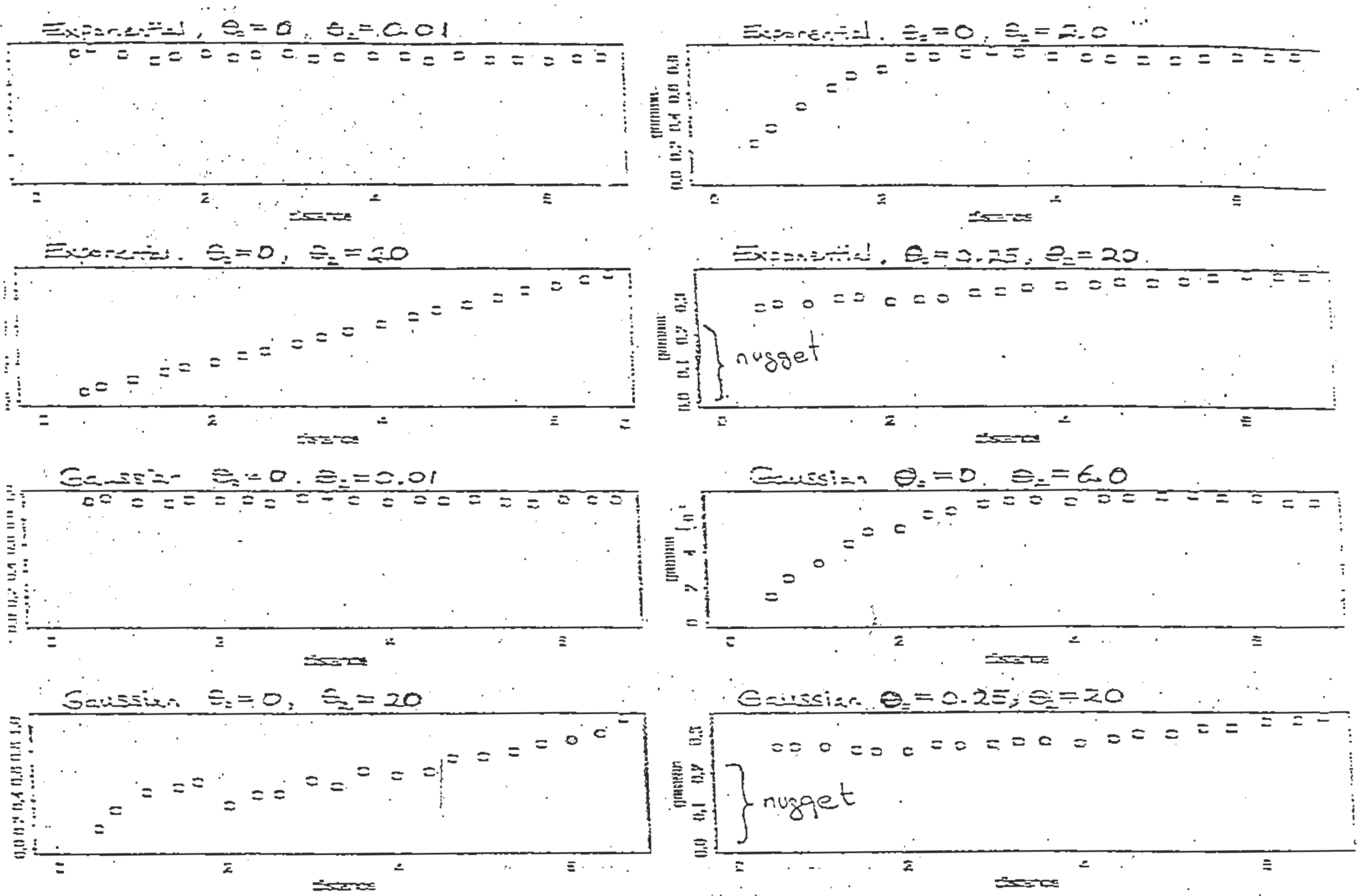
θ_0 = nugget

θ_1 = sill = 1

θ_2 = effective range

Gaussian smoother than exponential

SAMPLE VARIOGRAMS

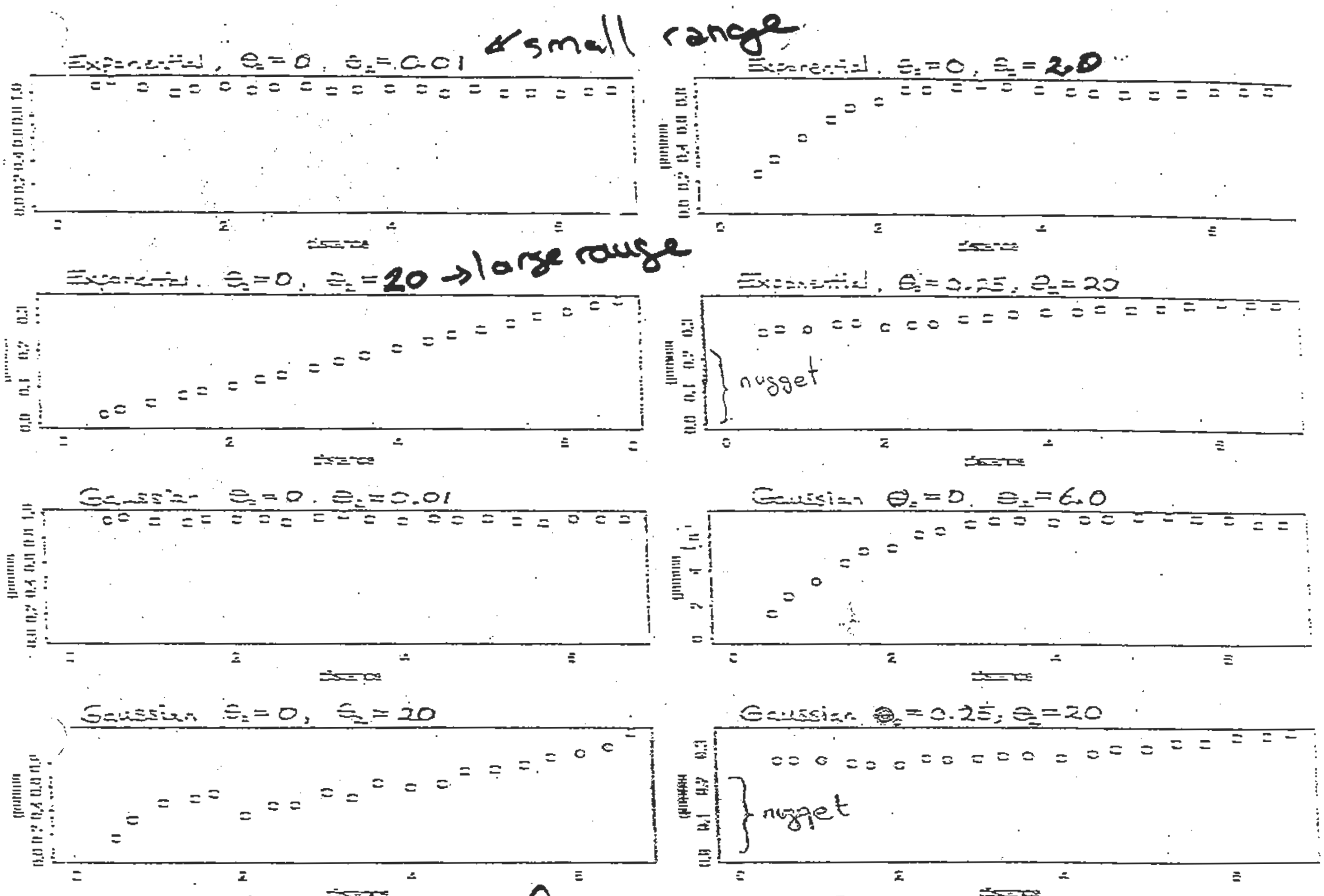


θ_0 = nugget

θ_1 = sill = 1

θ_2 = 'effective range' or range.

SAMPLE VARIOGRAMS



smooth at origin: Gaussian semivgram.

θ_0 = nugget

θ_1 = sill = 1

θ_2 = 'effective range' or range.

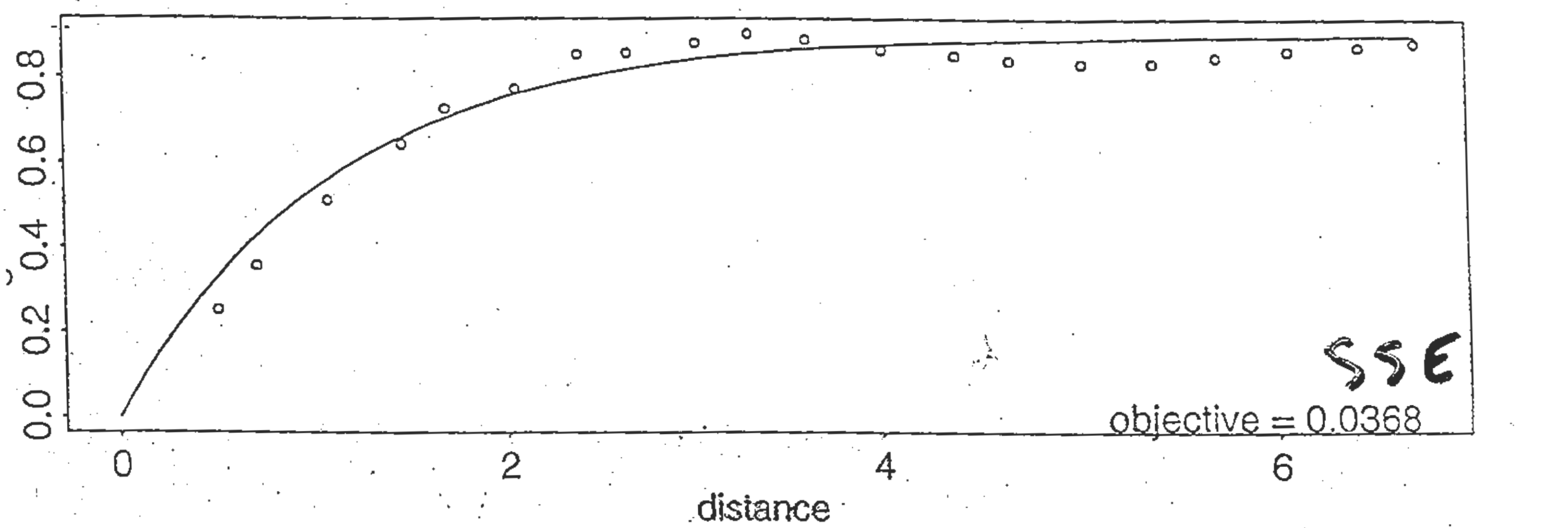
Fitted ("by eye") semivariogram models

(Data were simulated with exponential covariance function, $\theta_1 = 1$, $\theta_0 = 0$, $\theta_2 = 2.0$).

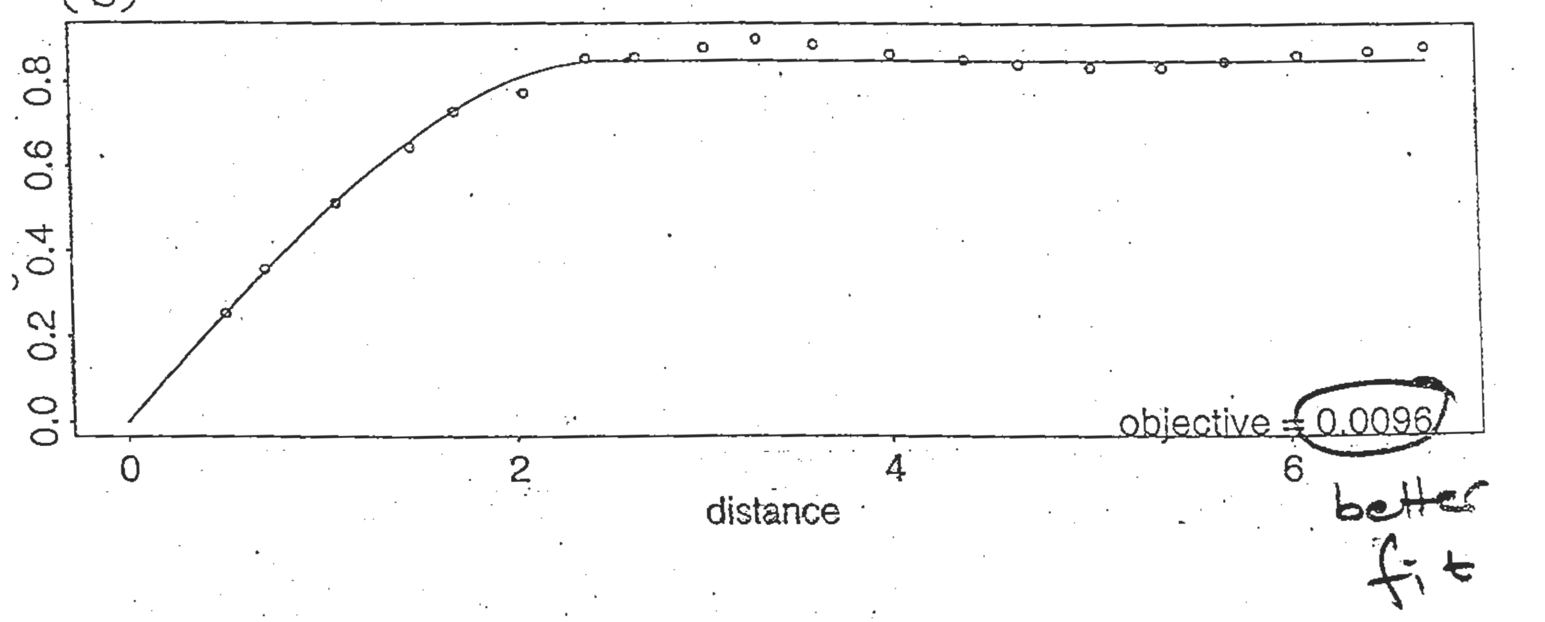
(a) Exponential model with $\theta_1 = 0.9$, $\theta_0 = 0$, $\theta_2 = 2.8$ → effective range

(b) Spherical model with $\theta_1 = 0.85$, $\theta_0 = 0$, $\theta_2 = 2.5$ → real range

(a)

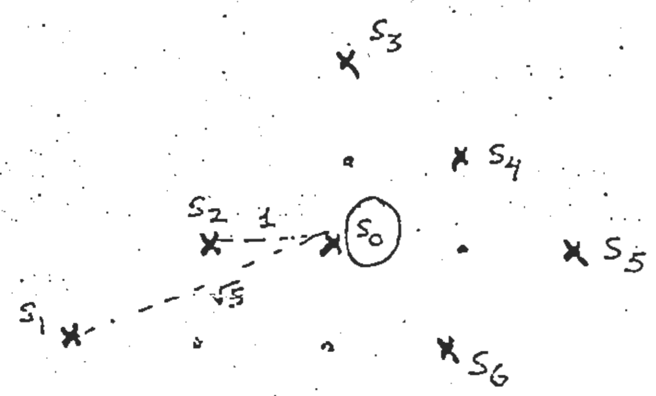


(b)



ORDINARY KRIGING.

Example:



Observations: $Z(s_1), Z(s_2), \ldots, Z(s_6)$. We want to predict $Z(s_0)$

Take $\gamma(\|h\|) = 1 - \exp(-\|h\|/2)$.

$$\gamma_O = \begin{bmatrix} 1-\exp(-\sqrt{5}/2) \\ 1-\exp(-1/2) \\ 1-\exp(-1) \\ 1-\exp(-\sqrt{2}/2) \\ 1-\exp(-1) \\ 1-\exp(-\sqrt{2}/2) \\ 1 \end{bmatrix} \qquad \begin{array}{l} \gamma(s_1-s_0) \\ \gamma(s_2-s_0) \\ \gamma(s_3-s_0) \\ \vdots \end{array}$$

$$\Gamma_O = \begin{array}{c} \begin{array}{ccccccc} & \gamma(s_1-s_2) & \gamma(s_1-s_3) & & & & \end{array} \\ \begin{bmatrix} 0 & 1-\exp(-\sqrt{2}/2) & 1-\exp(-\sqrt{13}/2) & \cdots & & 1 \\ & 0 & 1-\exp(-\sqrt{5}/2) & \cdots & & 1 \\ & & 0 & & & 1 \\ & & & \ddots & & \vdots \\ & & & & 0 & 1 \\ & & & & & 0 \end{bmatrix} \end{array}$$

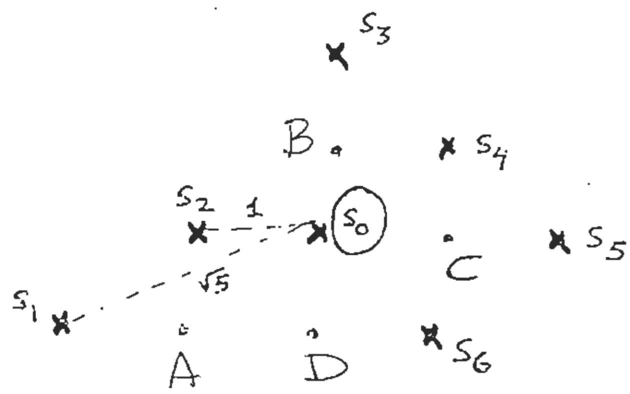
$$\lambda_O = \Gamma_O^{-1}\gamma_O = \begin{bmatrix} .017 \\ .422 \\ .065 \\ .218 \\ .031 \\ .246 \\ .004 \end{bmatrix}, \qquad \sigma^2_{OK}(s_0) = \lambda_O'\gamma_O = .478$$

$m \leftarrow .004$

only depends on locations

* Sampling Design Example

Example:



Observations: $Z(s_1), Z(s_2), \ldots, Z(s_6)$. We want to predict $Z(s_0)$

Take $\gamma(\|\mathbf{h}\|) = 1 - \exp(-\|\mathbf{h}\|/2)$.

Suppose we wish to minimize the kriging variance at $\mathbf{s}_0$ (the circled site), and we have sufficient resources to take an observation at any one of the four remaining unsampled sites (excluding $\mathbf{s}_0$).

The kriging variances at $\mathbf{s}_0$ corresponding to the addition of each of the sites A, B, C, and D are as follows:

Additional site	$\sigma^2_{OK}(\mathbf{s}_0)$
A	.4687
B	.4366
C	.4368
D	.4347

Thus, the best additional site is D.

* Example by Paulo Ribeiro (2000), Tech. Report, Univ. Lancaster

* BAYESIAN KRIGING:

The first simulated data set was generated in a 1x1 square, with 256 data points randomly located within the area (irregular grid). The simulated values were generated for model (1) with no covariates, zero mean ($\beta = 0$) and covariance parameters $(\sigma^2, \phi) = (1, 0.3)$. Ten prediction locations were chosen with one point intentionally located outside of the 1x1 square area to illustrate the behaviour of the predictions in a extrapolation problem. The data points and locations to be predicted can be seen in Figure 1.

$$(1) \qquad Y = X\beta + \epsilon$$

$$\epsilon \sim N(0, \Sigma)$$

$\beta = 0$ $\qquad \alpha = \sigma^2 =$ partial sill $\qquad$ nugget $= 0$

$\phi =$ range

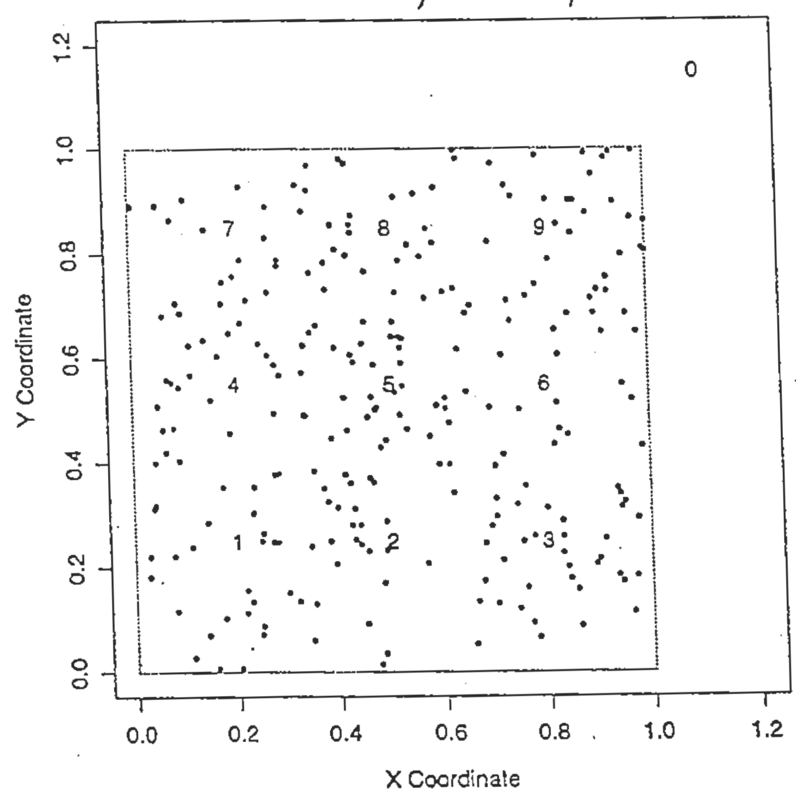


Figure 1: Data points locations (dots) and the 10 locations where prediction is aimed (numbers)

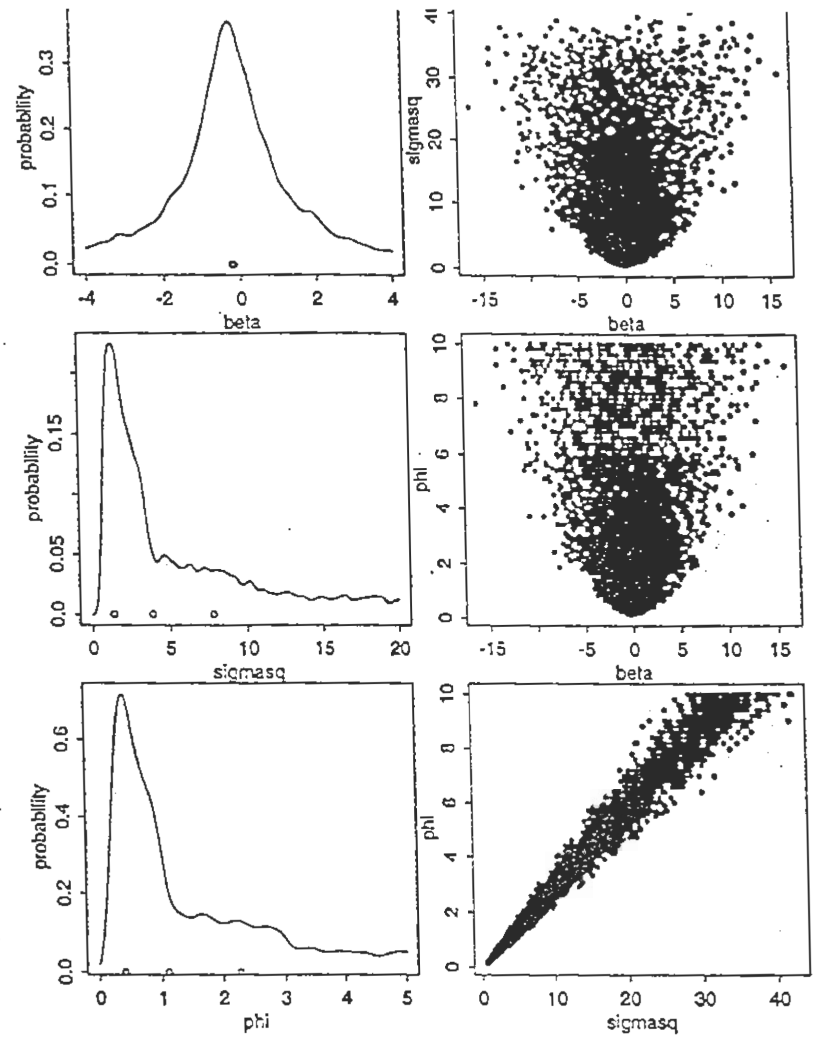


Figure 2. Posterior distributions for model parameters (β, σ^2, ϕ) on the left plots and correlations between parameters using the samples from the posterior distribution

priors:

$$p(\beta, \sigma^2, \phi) \propto \frac{1}{\sigma^2}$$

- uniform for β
- discrete uniform for range σ^2

(151 values between 0 and 10, higher concentration of points for small values of ϕ.

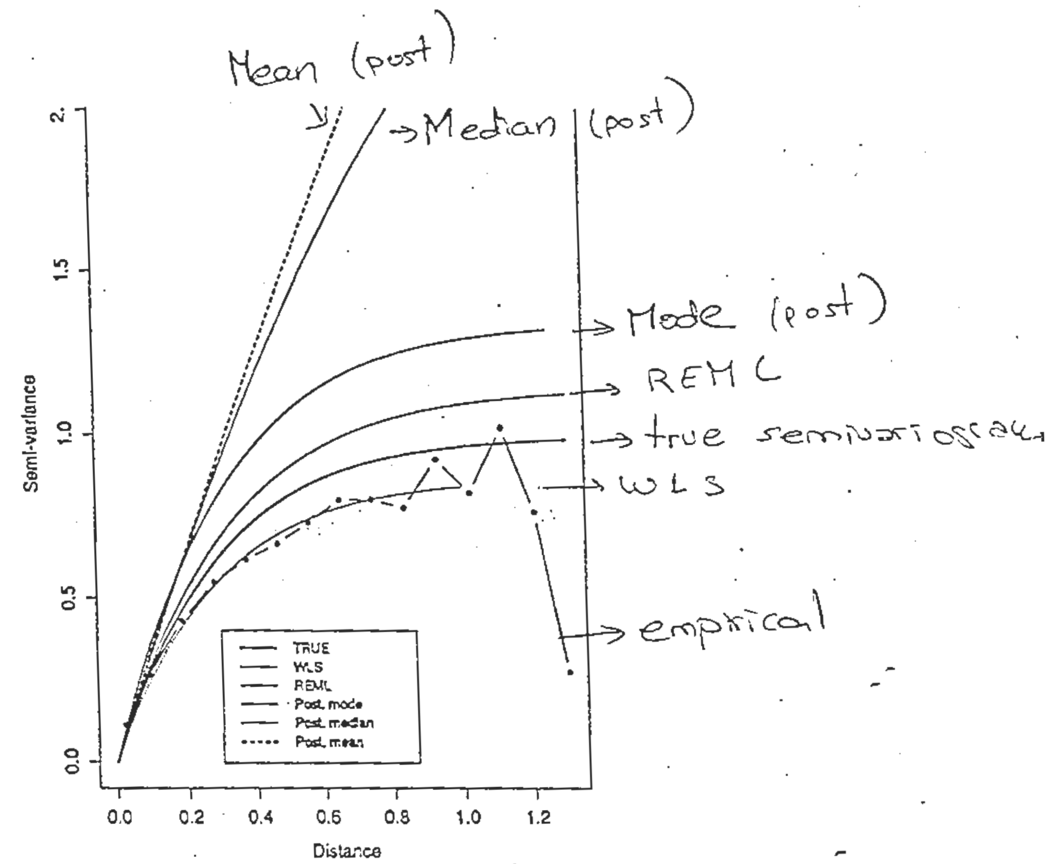


Figure 3: Empirical variogram and the variogram models fitted by different estimates

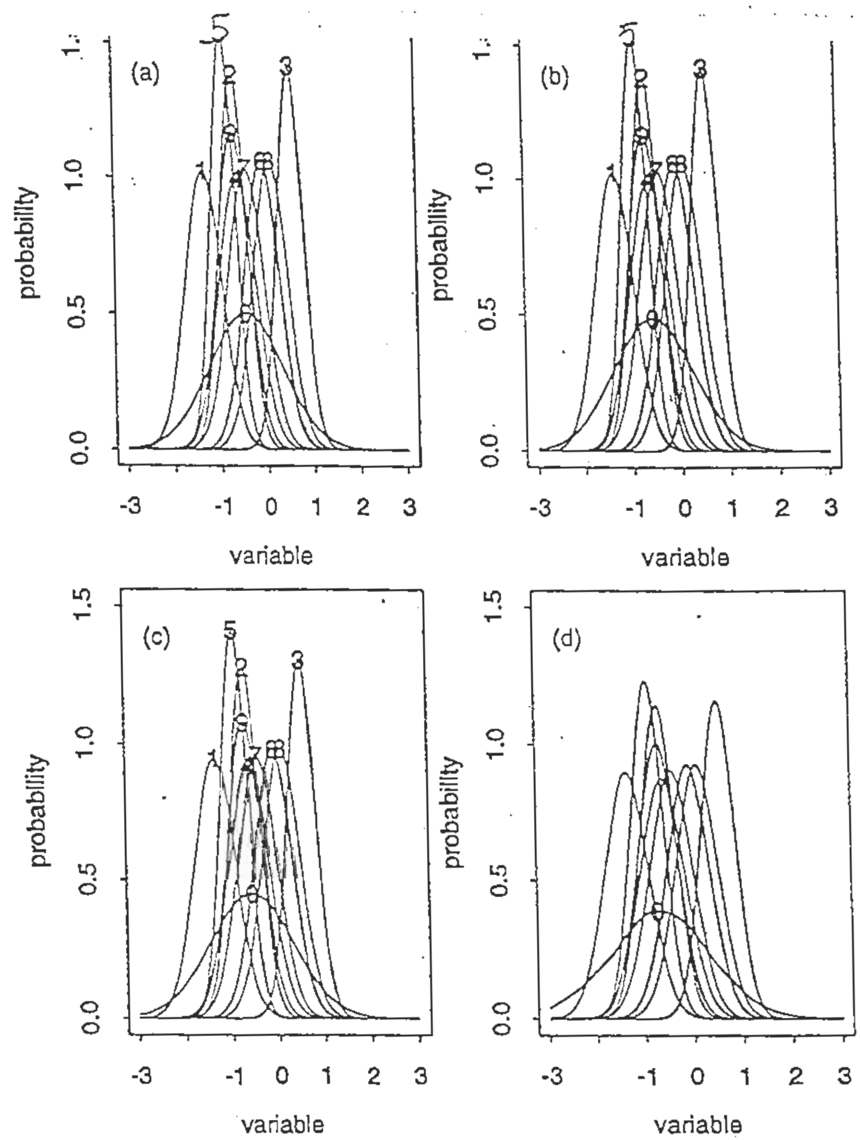


Figure 4: Predictive distributions at the 10 locations for (a) known parameters, (b) mean unknown, (c) mean and scale unknown and (d) all parameters unknown

* Large variance when predicting at 0
* Small variances predicting at 5, 2, and 3

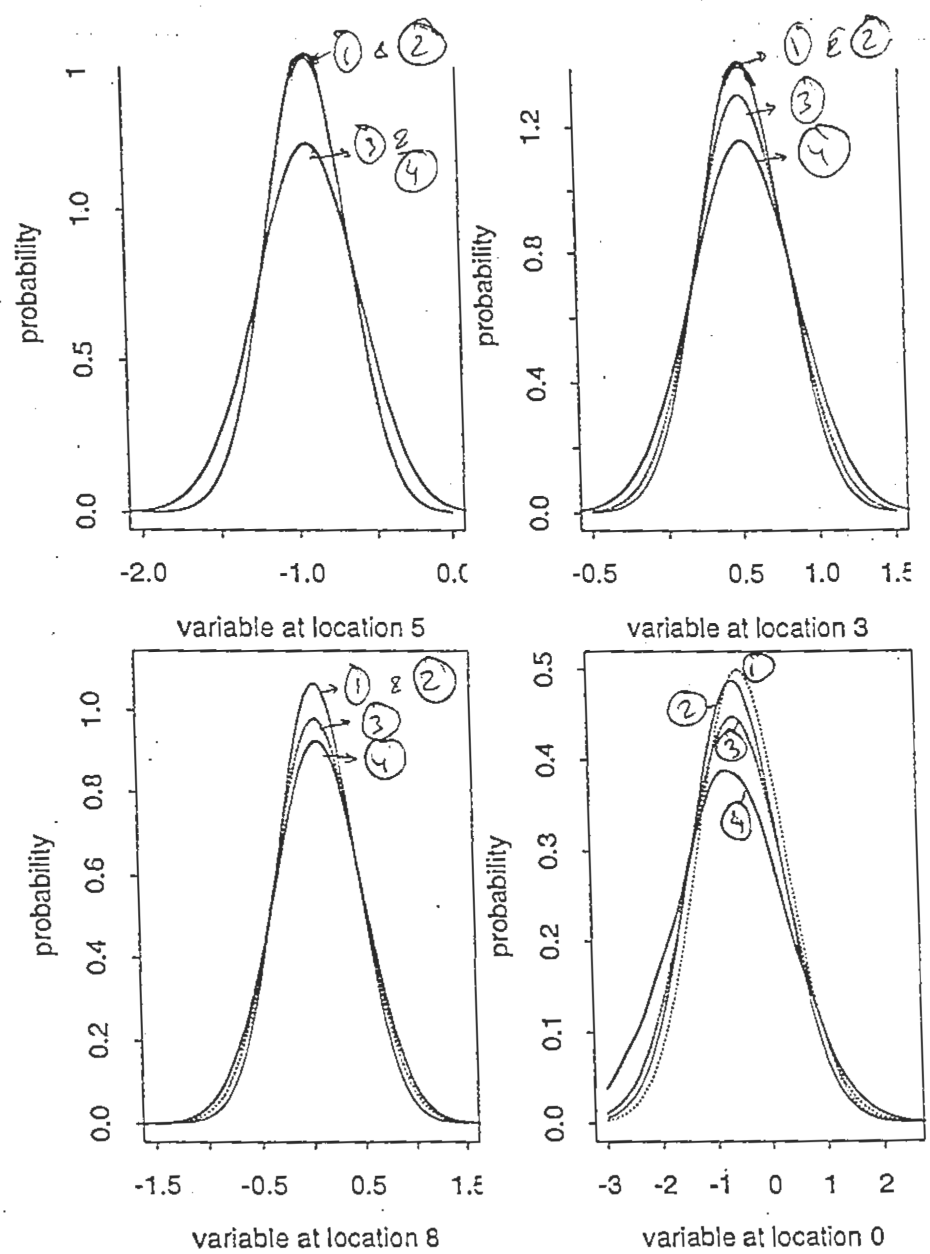


Figure 5: Predictive distributions for at some selected locations. SK (dashed line), OK (solid line), uncertainty in mean and variance (thick dashed line) and uncertainty in all parameters (thick solid line)

1 SK = simple Kriging ($\mu = 0$)
2 OK = ordinary Kriging (constant mean)
3 Bayesian : uncertainty in mean and σ^2
4 Bayesian : uncertainty in all parameters

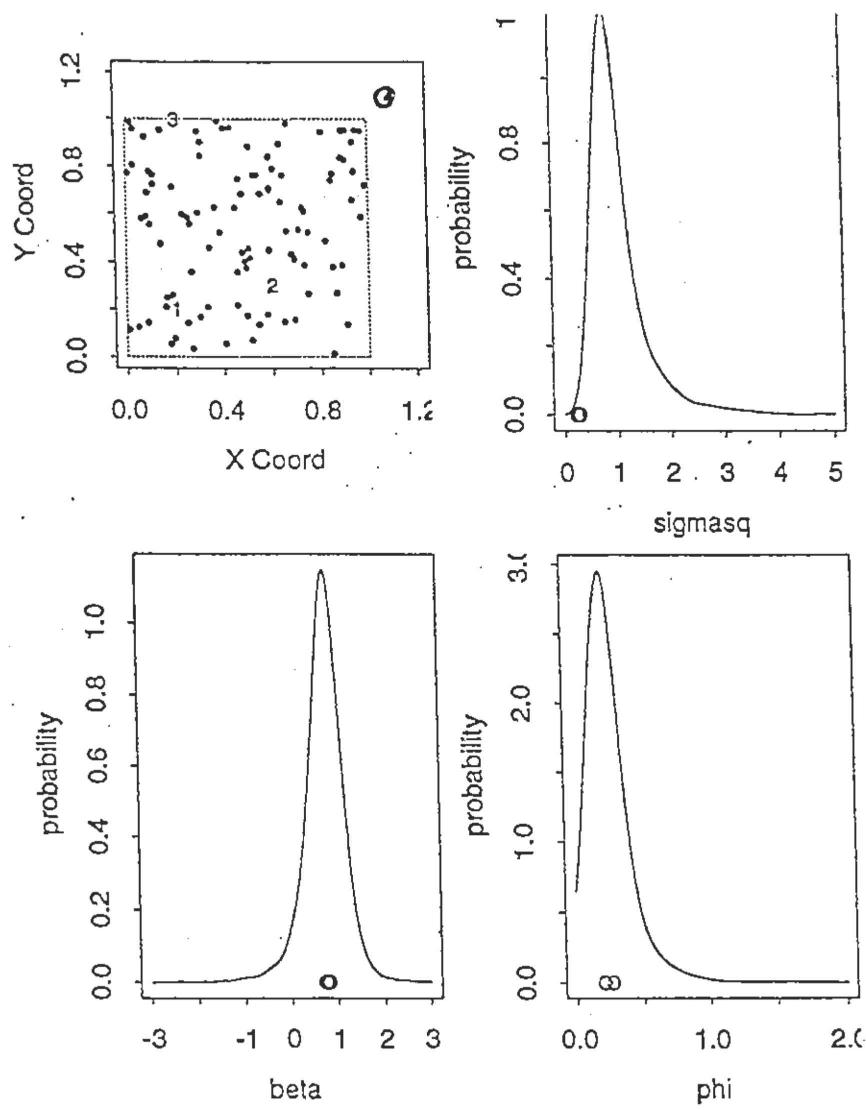


Figure 6 Data and prediction location for the second simulated data set with n=100 (top left) and the posterior distributions for model parameters (1000 samples from posterior)

Smaller dataset

from posterior

* The prior for the range here gives more weight for small values (not uniform)

$$\text{prior}(\beta, \sigma^2, \phi) \propto \frac{1}{\phi^2}\frac{1}{\sigma^2}$$

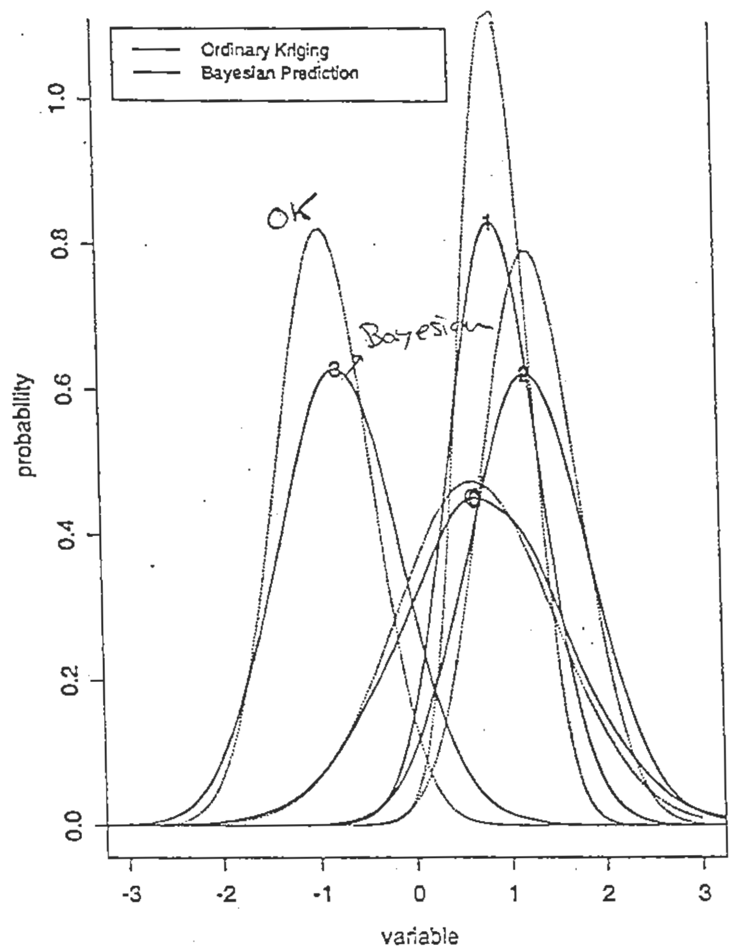


Figure 7: Predictive distributions at the 4 prediction locations for the second simulated data set

OK: Parameters estimated by WLS.
Bayesian: all parameters consider unknown.

- location 0: outside the area.
- location 1: central location (with neighbours)
- location 2: central location (with not many neighbours
- location 3: border of the area