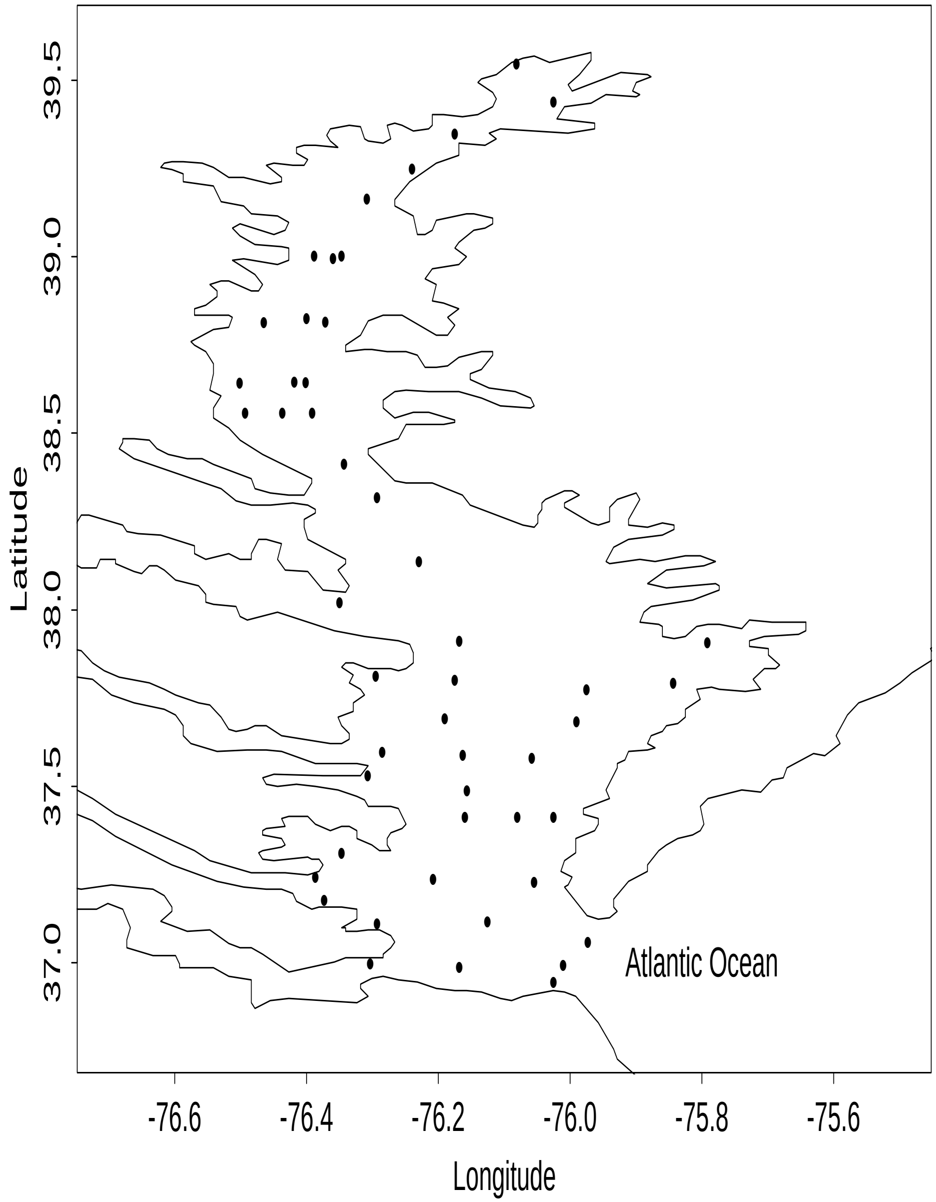
GEOSTATISTICAL ANALYSIS OF ENVIRONMENTAL DATA

Victor De Oliveira

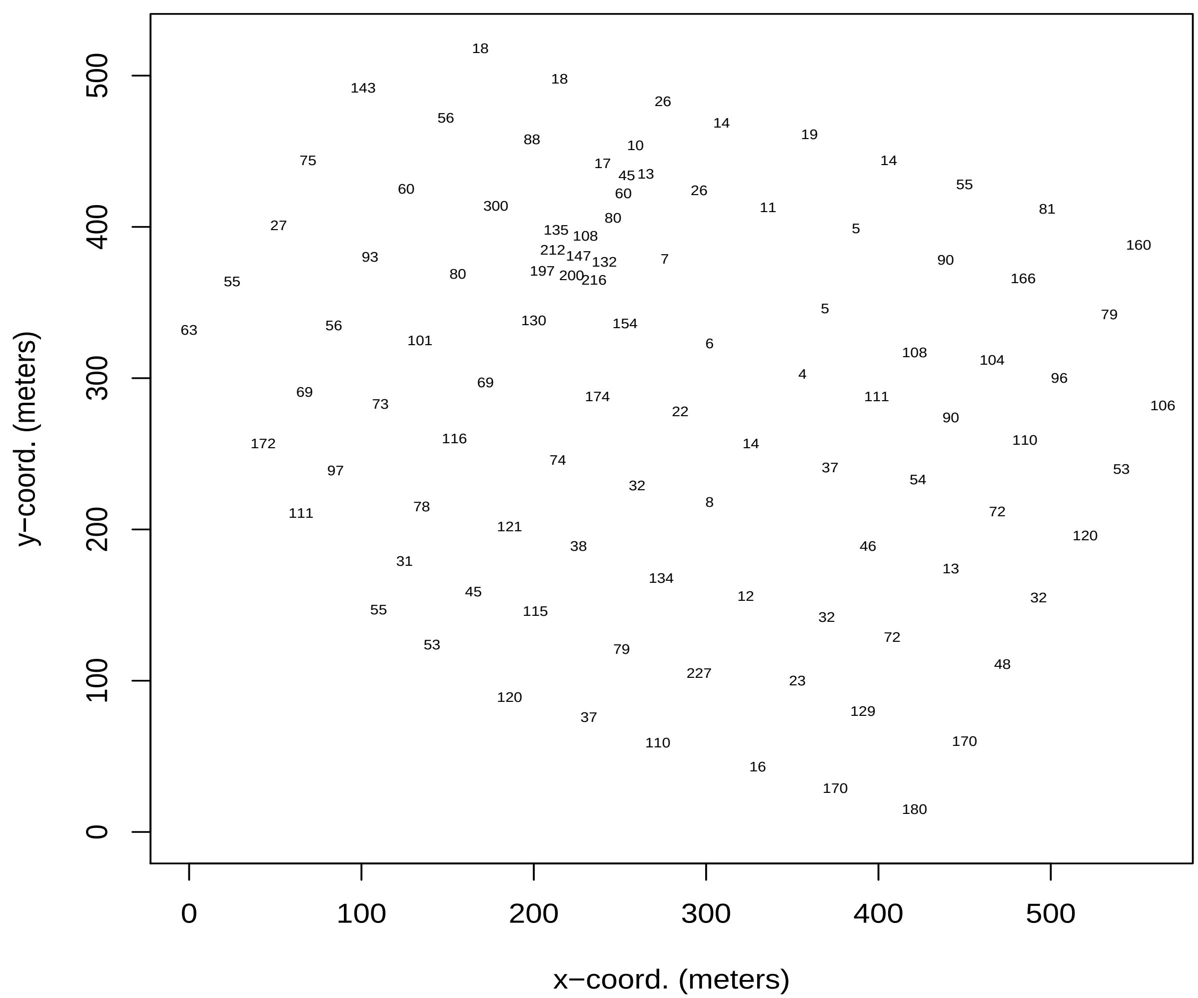
Department of Management Science and Statistics
The University of Texas at San Antonio
San Antonio, TX
USA
victor.deoliveira@utsa.edu
http://faculty.business.utsa.edu/vdeolive

2012 San Antonio Simulation and Vizualization Symposium
UTSA, November 13 2012

Example 1: Nitrogen data in Chesapeake Bay, Maryland



Example 2: Weed Data in Bjertorp farm, Sweden



Geostatistical Data

- Measurements or observations from a spatially varying phenomena in D (the domain of interest), $D \subset \mathbb{R}^2$

- Measurements or observations from the graph of unknown function

- Each datum is associated with a subset of D (a point or a larger set)

Some Geostatistical Problems

- Spatial estimation of pollution in air, water or soil
  (e.g. ozone, radon, SO_2)

- Determination of 'hot spots'

- Estimation of spatial averages

- Ecological monitoring
  (e.g. detecting decline of a species)

- Detection of temporal or spatial trends
  (e.g. global warming)

Main Features of Spatial Data

- Can be discrete or continuous. Most often non-negative
- Each datum has associated a 'unit' of space (support)
  - ▷ Rain measured by a tipping bucket (point support)
  - ▷ Rain 'measured' by a radar (areal support)
- Stochastically dependent: Observations measured at nearby locations tend to be more alike than observations measured at far away locations
- Often the variable of interest is not directly measured, but only a surrogate (an inverse problem)

Spatial Prediction/Interpolation Problem

Variable of interest varies spatially over a certain region of the plane according to an unknown function $z(\mathbf{s}) : D \subset \mathbb{R}^2 \to \mathbb{R}$

Variable measured at finite set of locations, $\mathbf{s}_1, \ldots, \mathbf{s}_n \in D$
Data vector is $\mathbf{z} = (z(\mathbf{s}_1), \ldots, z(\mathbf{s}_n))$

Other related spatial variables may also be available
(i.e. covariates)

<u>Goal</u>: make statistical inference about $\mathbf{z}_o = (z(\mathbf{s}_{01}), \ldots, z(\mathbf{s}_{0k}))$
where $\mathbf{s}_{01}, \ldots, \mathbf{s}_{0k} \in D$ ate locations with no measurements
For every $\mathbf{s}_{0j}$ we would like to compute $(\widehat{z}(s_{0j}), \widehat{\sigma}(s_{0j}))$

Random Fields/Spatial Processes

A random field on the region $D \subset \mathbb{R}^2$, $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$, is a collection of random variables indexed by the elements of D (often an infinite set)

The stochastic approach to the solution of spatial prediction/interpolation problems starts with the assumption that the graph of the unknown function, $\{(\mathbf{s}, z(\mathbf{s})) : \mathbf{s} \in D\}$ is a realization of a random field

Basic Components of a Random Field

- Mean function: $\mu(\mathbf{s}) = E\{Z(\mathbf{s})\}$ (spatial trend)

- Covariance function: $C(\mathbf{s}, \mathbf{u}) = \text{cov}\{Z(\mathbf{s}), Z(\mathbf{u})\}$
(spatial similarity)

- From these can compute variance function $\sigma^2(\mathbf{s}) = \text{var}\{Z(\mathbf{s})\}$ $(= C(\mathbf{s}, \mathbf{s}))$ and correlation function

$$K(\mathbf{s}, \mathbf{u}) = \frac{C(\mathbf{s}, \mathbf{u})}{\sigma(\mathbf{s})\sigma(\mathbf{u})}$$

- Closely related to $C(\mathbf{s}, \mathbf{u})$ is the semivariogram function
(spatial dissimilarity)

$$\begin{aligned} \gamma(\mathbf{s}, \mathbf{u}) &= \frac{1}{2}\text{var}\{Z(\mathbf{s}) - Z(\mathbf{u})\} \\ &= \frac{1}{2}[\sigma^2(\mathbf{s}) + \sigma^2(\mathbf{u}) - 2C(\mathbf{s}, \mathbf{u})] \end{aligned}$$

Mean Function

Any function $\mu : D \to \mathbb{R}$ can be the mean function of a random field. Typical examples:

- β_1 (constant)
- $\beta_1 \mathbf{1}_{D_1}(\mathrm{s}) + \beta_2 \mathbf{1}_{D_2}(\mathrm{s});\quad D = D_1 \cup D_2,\ D_1 \cap D_2 = \emptyset$
- $\beta_1 + \beta_2 x + \beta_3 y \quad (\mathrm{s} = (x, y))$
- $\beta_1 + \beta_2 X(\mathrm{s});\quad X(.)$ a related process

All of the above are examples of linear mean functions

$$\mu(\mathrm{s}) = \sum_{j=1}^{p} f_j(\mathrm{s})\beta_j$$

Covariance Function

On the other hand, not any function $C : D \times D \to \mathbb{R}$ can be a covariance function of a random field.

Fact. $C(\mathbf{s}, \mathbf{u})$ is the covariance function of some random field if and only if it is symmetric ($C(\mathbf{s}, \mathbf{u}) = C(\mathbf{u}, \mathbf{s})$) and positive semi-definite, meaning that

$\forall\ m \in N$, $\mathbf{s}_1, \ldots, \mathbf{s}_m \in D$ and $a_1, \ldots, a_m \in \mathbb{R}$:

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j C(\mathbf{s}_i, \mathbf{s}_j) \geq 0$$

Stationarity (Invariance)

$Z(.)$ is <u>strictly stationary</u> if $\forall\ m \in \mathbb{N}$, $\mathbf{s}_1, \ldots, \mathbf{s}_m \in D$, and $\mathbf{h} \in \mathbb{R}^2$

$$(Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_m)) \stackrel{\text{d}}{=} (Z(\mathbf{s}_1 + \mathbf{h}), \ldots, Z(\mathbf{s}_m + \mathbf{h}))$$

$Z(.)$ is <u>weakly (2nd order) stationary</u> if

$$\mu(\mathbf{s}) = \mu \quad \text{and} \quad C(\mathbf{s}, \mathbf{u}) = \tilde{C}(\mathbf{s} - \mathbf{u})$$

in which case $C(\mathbf{s}, \mathbf{u}) = \sigma^2 \tilde{K}(\mathbf{s} - \mathbf{u})$

$Z(.)$ is <u>intrinsically stationary</u> if

$$\mu(\mathbf{s}) = \mu \quad \text{and} \quad \gamma(\mathbf{s}, \mathbf{u}) = \tilde{\gamma}(\mathbf{s} - \mathbf{u})$$

A covariance function is <u>isotropic</u> if $C(\mathbf{s}, \mathbf{u}) = \bar{C}(\|\mathbf{s} - \mathbf{u}\|)$

Basic Covariance Models

For isotropic covariance functions $C(\mathbf{s}, \mathbf{s}) = \sigma^2$ (constant) and $C(\mathbf{s}, \mathbf{u}) = \sigma^2 K(d)$ where $d = \|\mathbf{s} - \mathbf{u}\|$

A few examples:

Spherical Model

$$K_{\vartheta}^{S}(d) = \begin{cases} 1 - \frac{3}{2}\left(\frac{d}{\theta_1}\right) + \frac{1}{2}\left(\frac{d}{\theta_1}\right)^3, & \text{if } 0 \le d \le \theta_1 \\ 0, & \text{if } d > \theta_1 \end{cases} \quad ; \qquad \theta_1 > 0$$

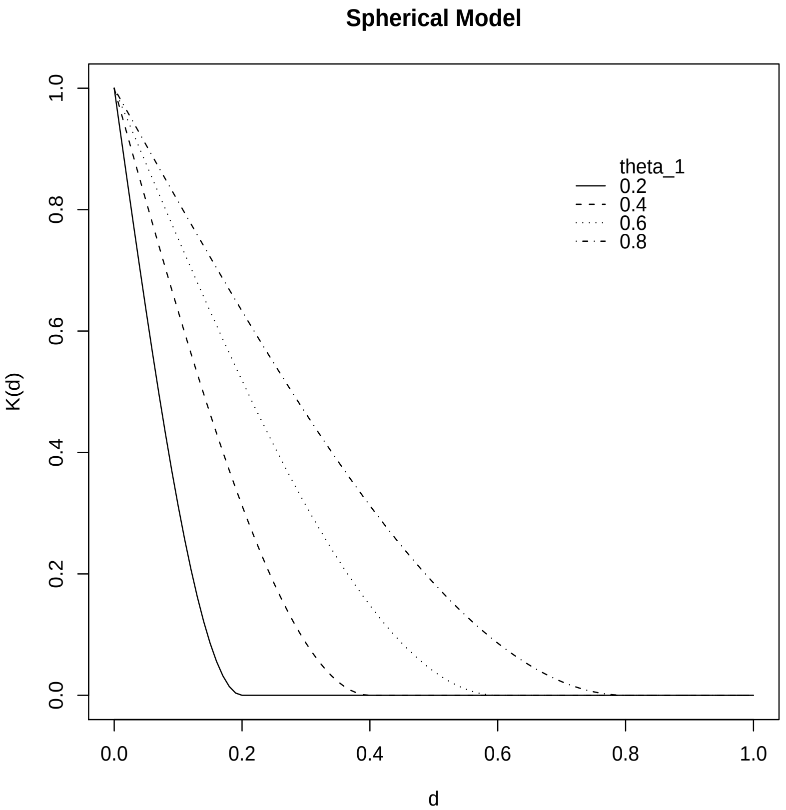
Power Exponential

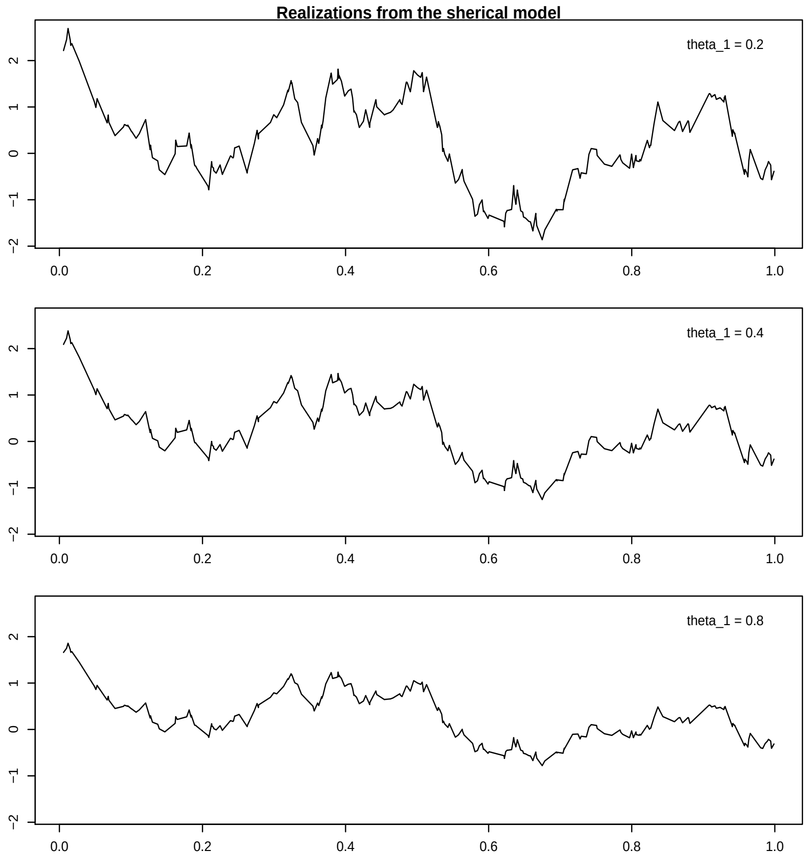
$$K_{\vartheta}^{PE}(d) = \exp\left(- \left(\frac{d}{\theta_1}\right)^{\theta_2} \right); \qquad \theta_1 > 0,\ \theta_2 \in (0, 2]$$

Wave Effect

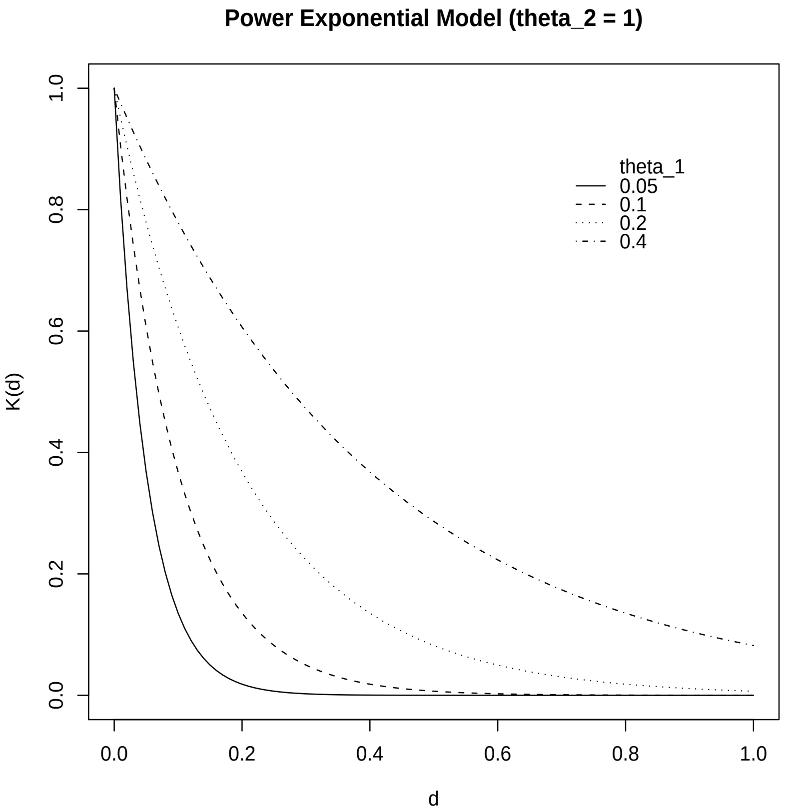
$$K_{\vartheta}^{WE}(d) = \frac{\theta_1}{d} \sin\left(\frac{d}{\theta_1}\right); \qquad \theta_1 > 0.$$

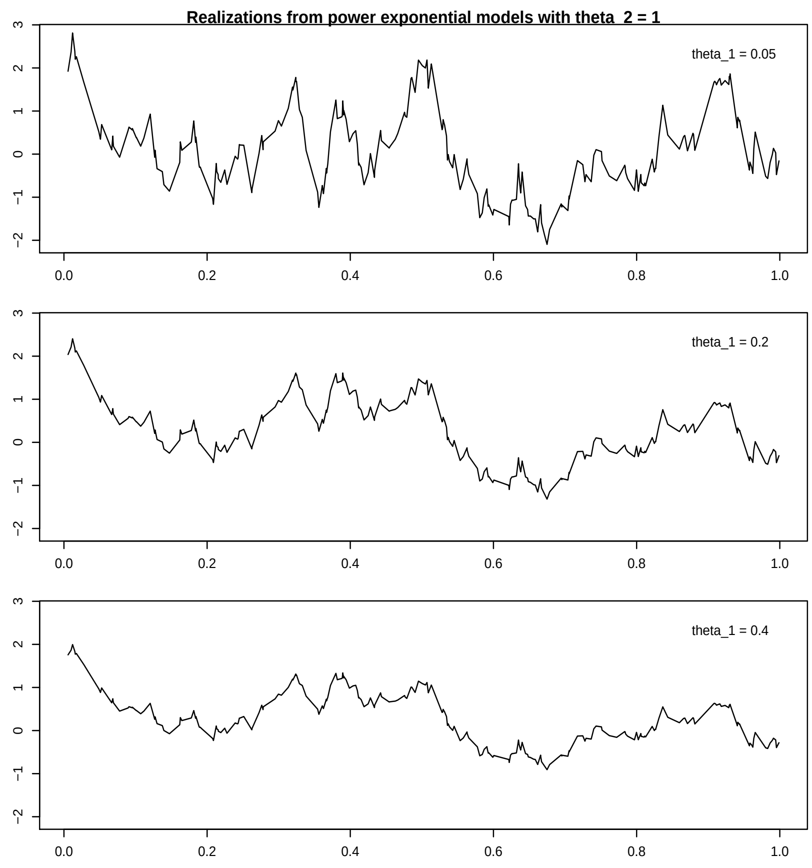
Spherical


Spherical Model
theta_1
0.2
0.4
0.6
0.8
K(d)
d
0.0
0.2
0.4
0.6
0.8
1.0

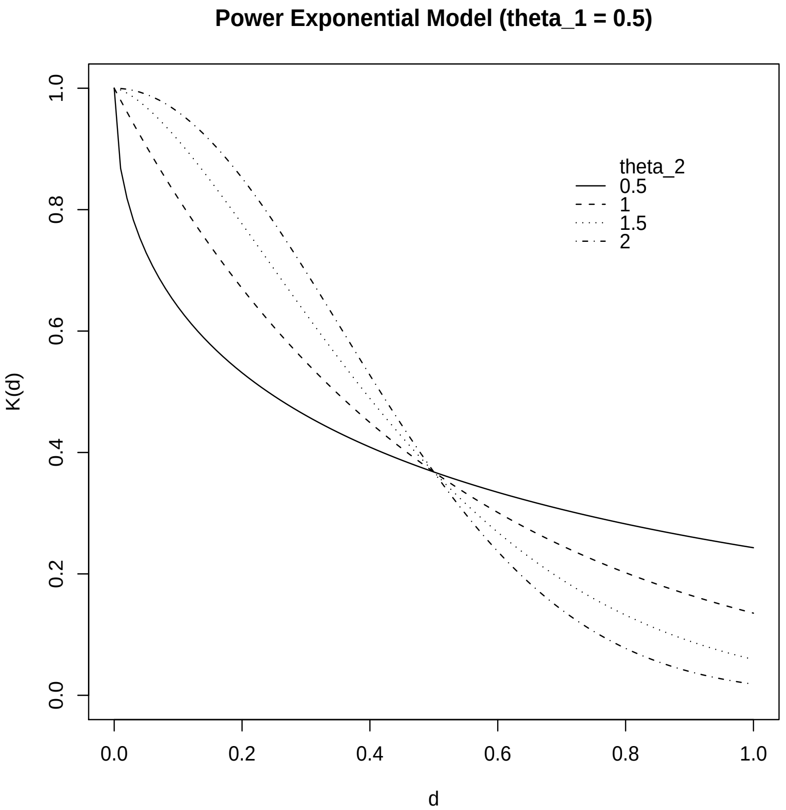

Realizations from the sherical model
theta_1 = 0.2
theta_1 = 0.4
theta_1 = 0.8

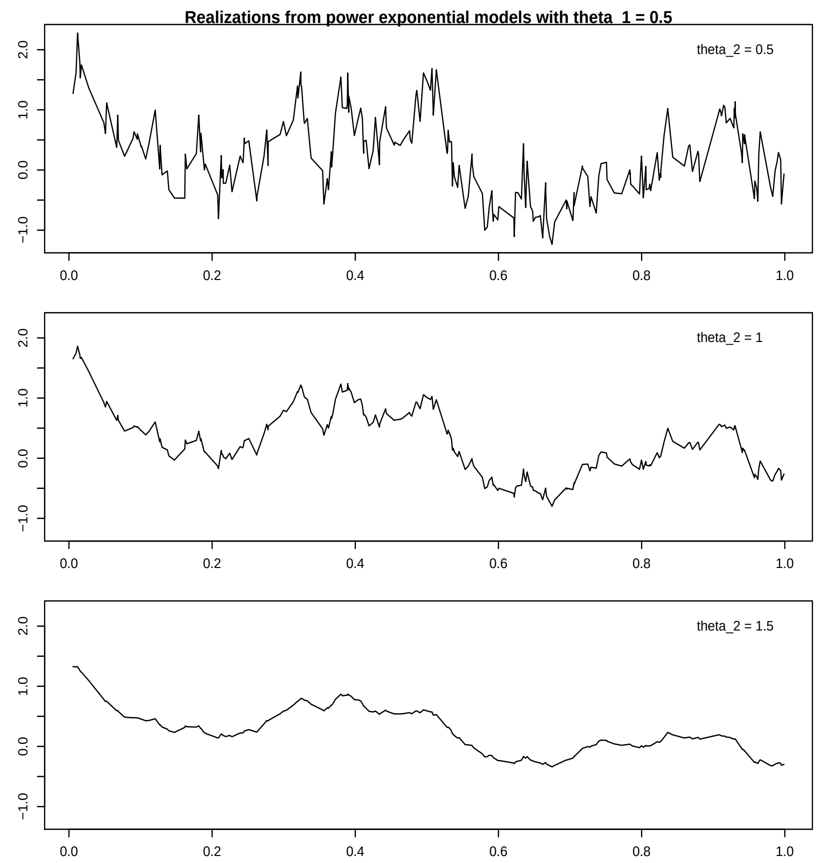
Power Exponential



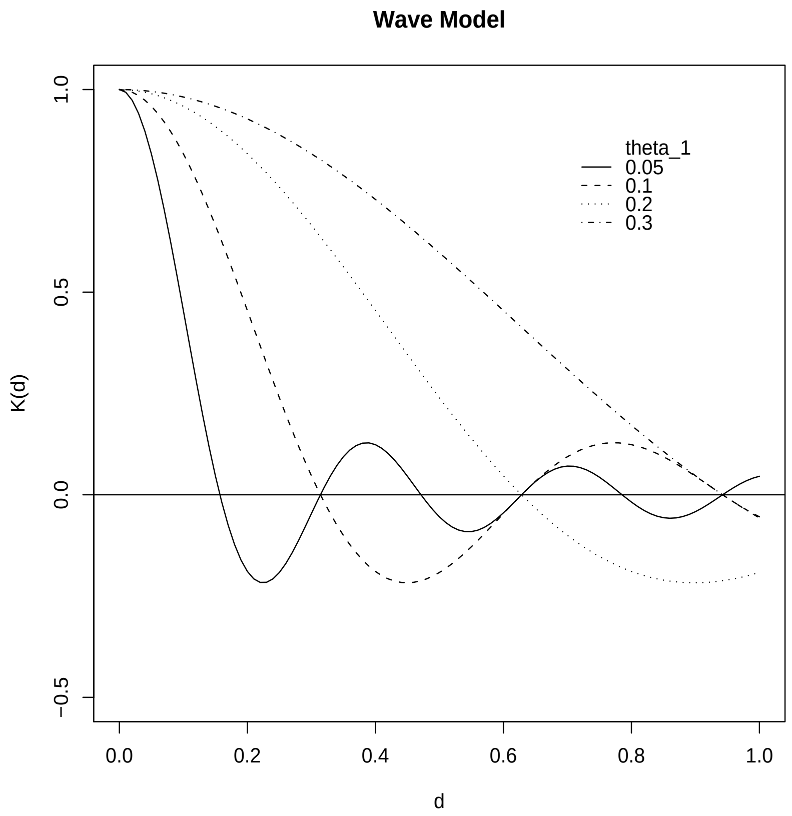


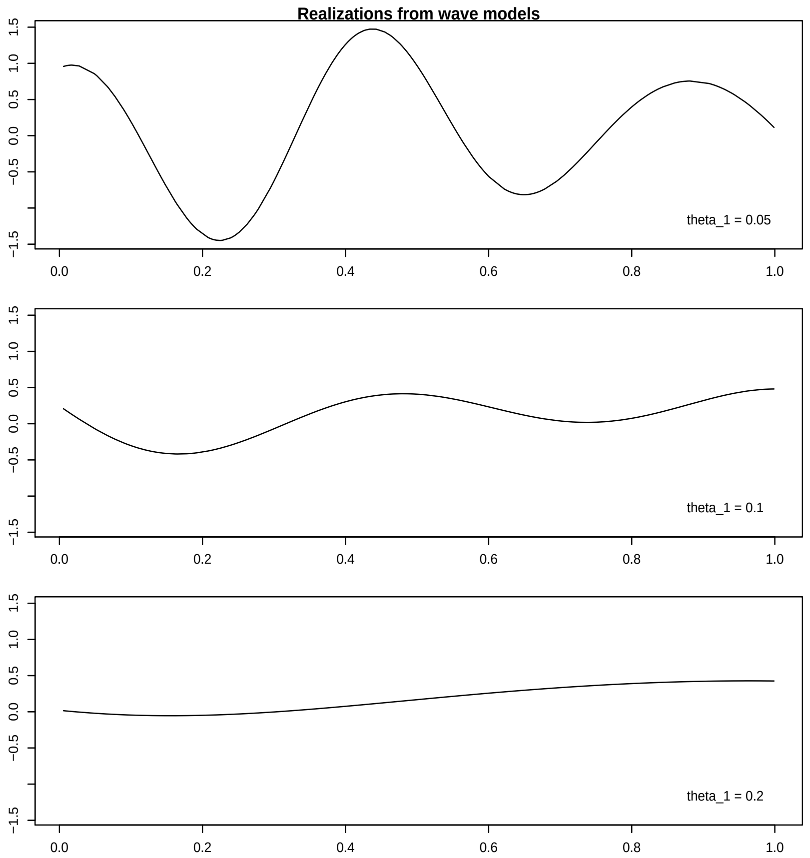
Power Exponential (cont.)





Wave Effect


Wave Model
theta_1
0.05
0.1
0.2
0.3
K(d)
d
1.0
0.5
0.0
−0.5
0.0
0.2
0.4
0.6
0.8
1.0


Realizations from wave models
theta_1 = 0.05
theta_1 = 0.1
theta_1 = 0.2
1.5
1.0
0.5
0.0
−0.5
−1.5
0.0
0.2
0.4
0.6
0.8
1.0

Specifying Geostatistical Models

- Complete Specification: Family of finite-dimensional distributions

$$F_{\mathbf{s}_1,\ldots,\mathbf{s}_m}(x_1,\ldots,x_n) = P\{Z(\mathbf{s}_1) \le x_1,\ldots,Z(\mathbf{s}_m) \le x_m\}$$

$\forall\ m \in \mathbb{N}$ and $\mathbf{s}_1,\ldots,\mathbf{s}_m \in D$.

- A random field is said to be Gaussian if all members of the above family if distributions are multivariate normal

- Gaussian random fields are completely specified by their mean and covariance functions

- For Gaussian random fields, strong and weak stationarity are the same

2nd Order Random Field Specification

Let $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ be the random field of interest, with

$$
\begin{aligned}
E\{Z(\mathbf{s})\} &= \sum_{j=1}^{p} \beta_j f_j(\mathbf{s}) \\
\text{cov}\{Z(\mathbf{s}), Z(\mathbf{u})\} &= \sigma^2 K_{\vartheta}(\mathbf{s}, \mathbf{u}) \quad (= C(\mathbf{s}, \mathbf{u}))
\end{aligned}
$$

- $\underline{f}(\mathbf{s}) = (f_1(\mathbf{s}), \ldots, f_p(\mathbf{s}))$ location-dependent covariates
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ unknown regression parameters
- $\sigma^2 = \text{var}\{Z(\mathbf{s})\}$
- $K_{\vartheta}(\mathbf{s}, \mathbf{u})$ correlation function on $\mathbb{R}^2$
- $\boldsymbol{\vartheta}$ correlation parameters controlling geometric and other features of random field (e.g. differentiability).

Spatial Prediction/Interpolation

- Suppose want to predict $Z(\mathbf{s}_0)$, $\mathbf{s}_0 \in D$ unsampled location
- The kriging predictor is the one that minimizes

$$\text{MSPE}(\hat{Z}(\mathbf{s}_0)) = E\{(Z(\mathbf{s}_0) - \hat{Z}(\mathbf{s}_0))^2\}$$

over the class of linear unbiased predictors

$$\hat{Z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i(\mathbf{s}_0) Z(\mathbf{s}_i)$$

that are unbiased

$$E\{\hat{Z}(\mathbf{s}_0)\} = E\{Z(\mathbf{s}_0)\}$$

This is also know as the BLUP predictor

Spatial Prediction/Interpolation (cont.)

- The optimal coefficients (weights) $\boldsymbol{\lambda}(\mathrm{s}_0) = (\lambda_1(\mathrm{s}_0), \ldots, \lambda_n(\mathrm{s}_0))$ are obtained as the solution of the linear system of equations

$$\left\{ \begin{array}{rcll} \sum_{j=1}^{n} \lambda_j C(\mathrm{s}_i, \mathrm{s}_j) - \sum_{j=1}^{p} m_j f_j(\mathrm{s}_i) & = & C(\mathrm{s}_0, \mathrm{s}_i) \; ; & i = 1, \ldots, n \\ \sum_{i=1}^{n} \lambda_i f_j(\mathrm{s}_i) & = & f_j(\mathrm{s}_0) \; ; & j = 1, \ldots, p \end{array} \right.$$

- An uncertainty measure is

$$\begin{aligned} \widehat{\sigma}^2(\mathrm{s}_0) &= \mathrm{MSPE}(\widehat{Z}^K(\mathrm{s}_0)) \\ &= C(\mathrm{s}_0, \mathrm{s}_0) - \sum_{j=1}^{n} \lambda_j C(\mathrm{s}_0, \mathrm{s}_j) + \sum_{j=1}^{p} m_j f(\mathrm{s}_j) \end{aligned}$$

- Repeat for many $\mathrm{s}_0 \in D$ to get estimate of graph of $z(\mathrm{s})$

Spatial Prediction/Interpolation (cont.)

When the random field is (approximately) Gaussian:

- $\hat{Z}^K(\mathrm{s}_0)$ agrees with best unbiased predictor

- A nominal 95% prediction interval for $Z(\mathrm{s}_0)$ is

$$\hat{Z}^K(\mathrm{s}_0) \pm 1.96 \cdot \hat{\sigma}(\mathrm{s}_0)$$

- These classical methods are implemented in the `R` package `geoR`

Comments

- The above kriging predictor is an 'interpolator'

$$\hat{Z}^K(\mathbf{s}_i) = Z(\mathbf{s}_i) \qquad\qquad (\text{and} \quad \hat{\sigma}^2(\mathbf{s}_i) = 0)$$

- $\hat{\sigma}^2(\mathbf{s}_0)$ does not depend (directly) on the data $\mathbf{z}$

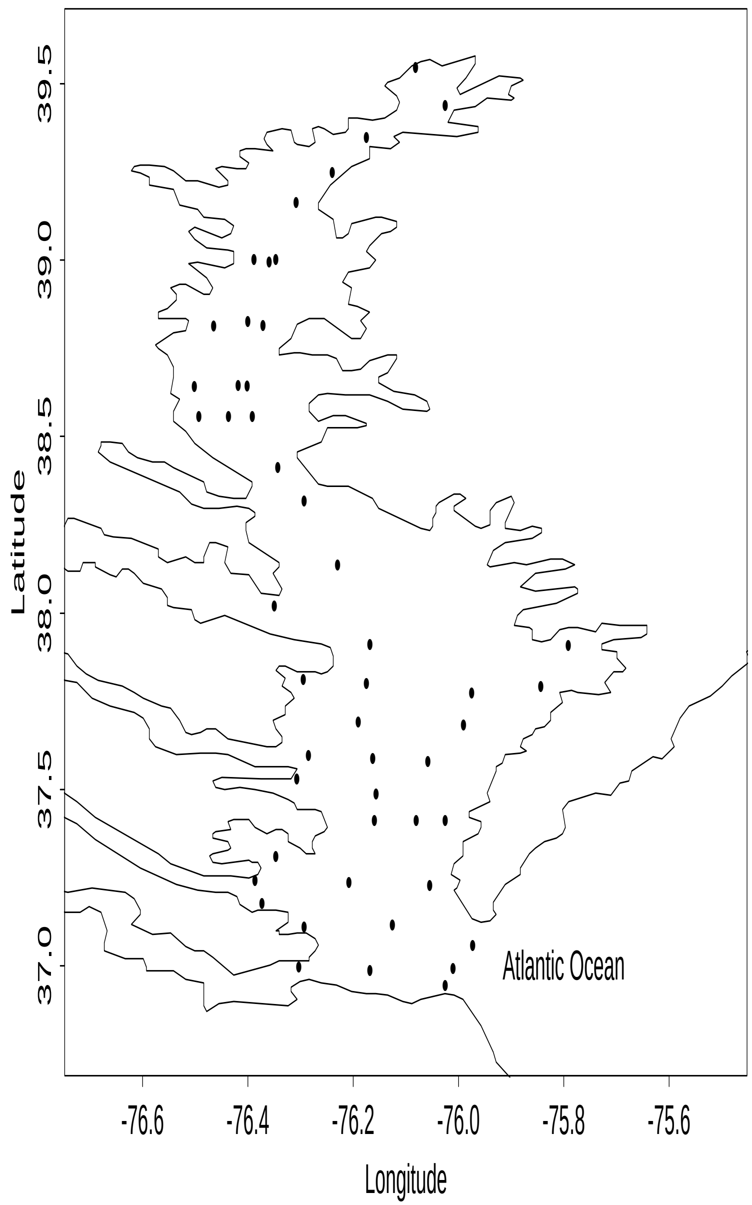
- Data often contain measurement error

$$Z_{i,\text{obs}} = Z(\mathbf{s}_i) + \epsilon_i, \qquad i = 1, \ldots, n$$

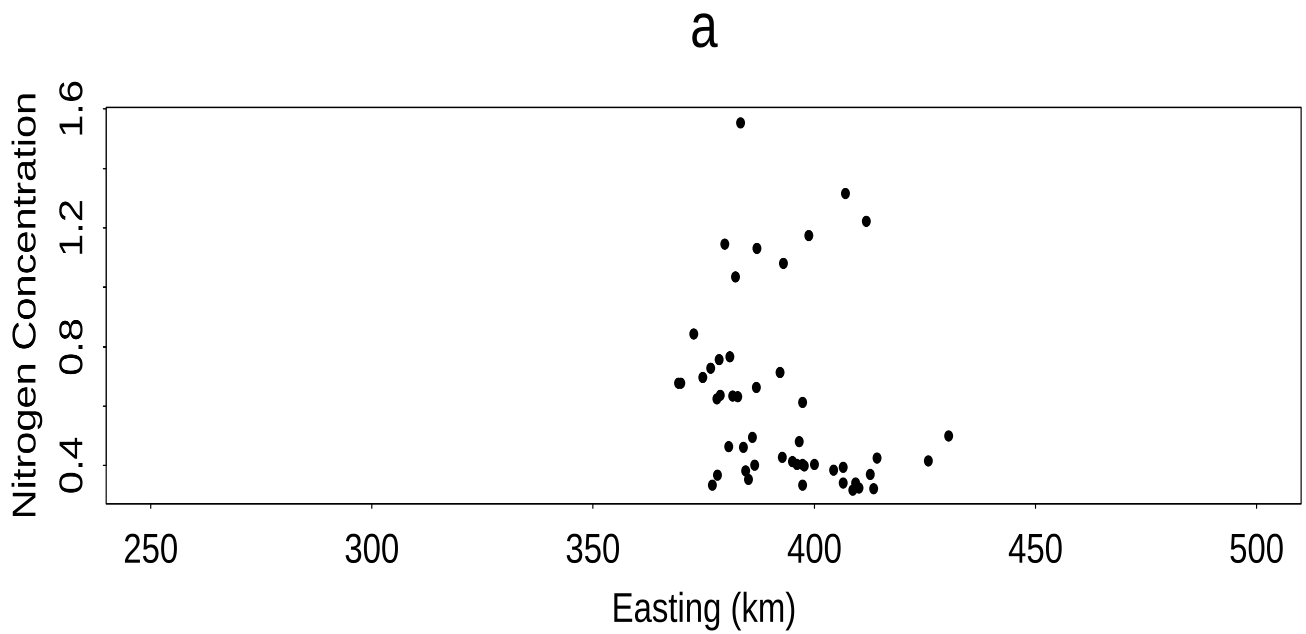
$\epsilon_1, \ldots, \epsilon_n$ i.i.d with mean 0 and variance σ_ϵ^2. In this case

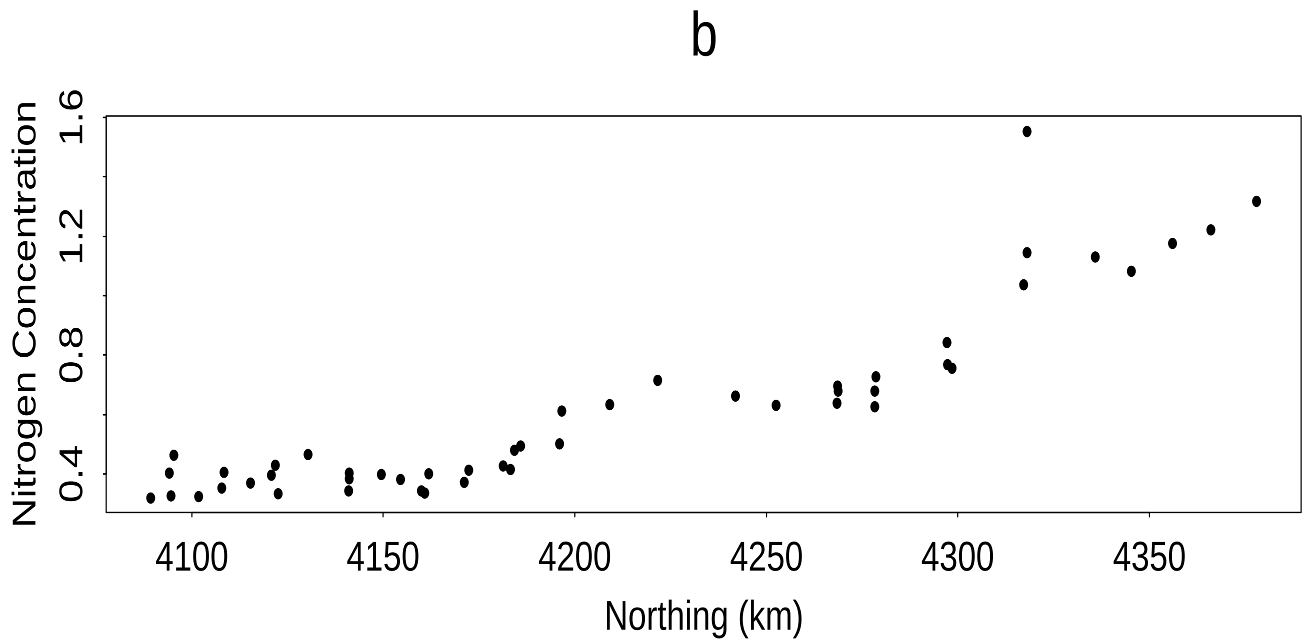
- ▷ $\hat{Z}^K(\mathbf{s}_0)$ is a 'smoother' rather than an interpolator
- ▷ $\hat{Z}^K(\mathbf{s}_0)$ remains the same for $\mathbf{s}_0 \neq \mathbf{s}_i$
- ▷ $\hat{\sigma}^2(\mathbf{s}_0)$ increases

Example 1: Nitrogen in Chesapeake Bay (cont.)

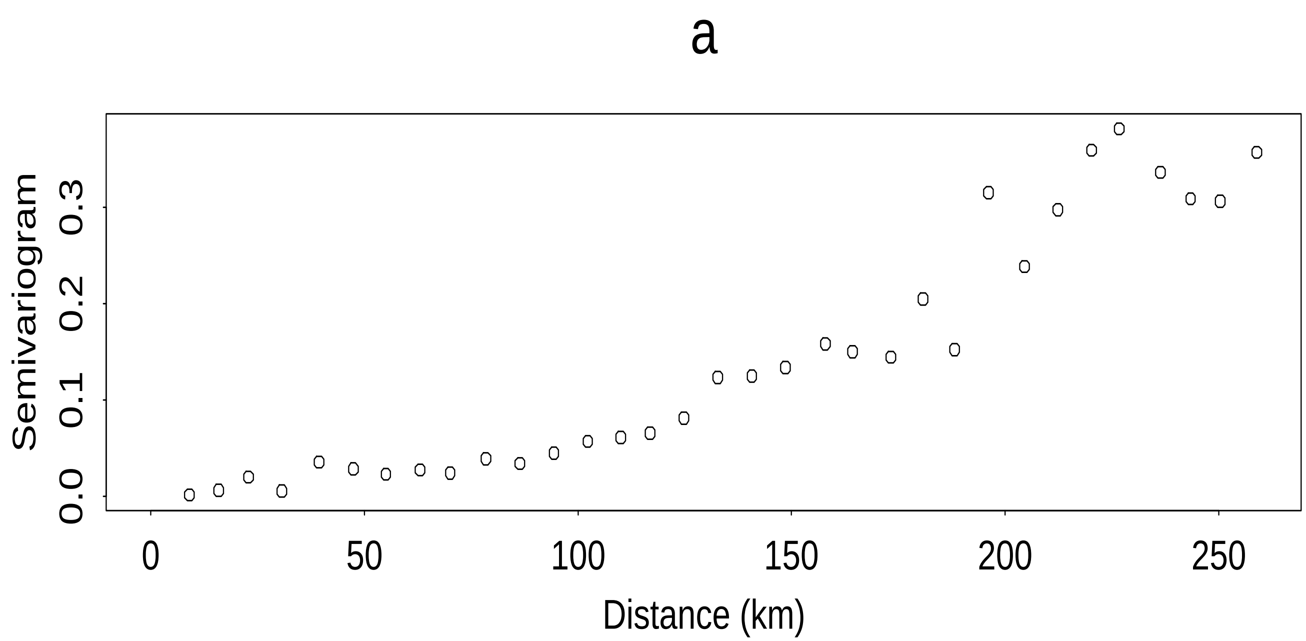


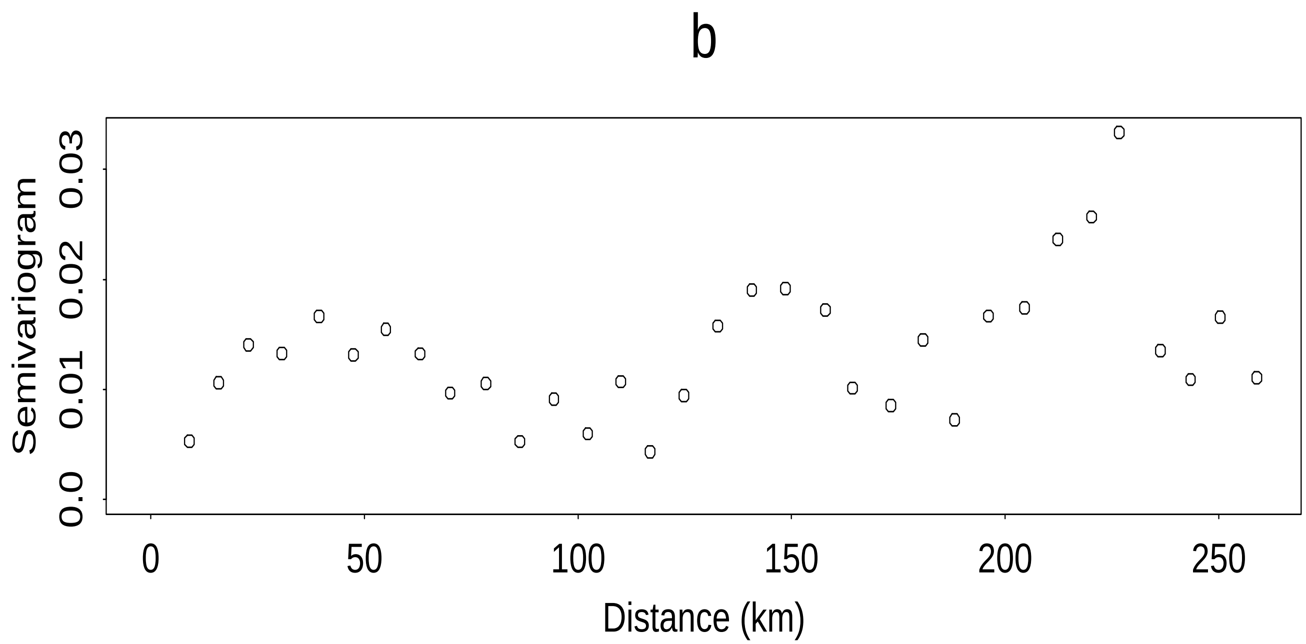
Exploratory Analysis





Exploratory Analysis (cont.)





Proposed Model

- Data: $\mathbf{Z}_{obs} = (Z_{1,obs}, \ldots, Z_{49,obs})$, where

$$Z_{i,obs} = Z(\mathbf{s}_i) + \epsilon_i; \quad i = 1, \ldots, 49$$

$$\begin{aligned} E\{Z(\mathbf{s})\} &= \beta_1 + \beta_2 y, \quad \mathbf{s} = (x, y) \\ \text{cov}\{Z(\mathbf{s}), Z(\mathbf{u})\} &= \sigma^2 \frac{\theta}{d} \sin\left(\frac{d}{\theta}\right), \quad d = \|\mathbf{s} - \mathbf{u}\| \end{aligned}$$

$\epsilon_1, \ldots, \epsilon_n$ represent "measurement errors" (i.i.d.) with mean 0 and variance σ_ϵ^2

- Unknown parameters: $\boldsymbol{\eta} = (\beta_1, \beta_2, \sigma^2, \sigma_\epsilon^2, \theta)$

Hot Spot Estimation

Based on scientific and/or regulatory considerations define "hot spots" as

$$H = \{\mathbf{s} \in D : Z(\mathbf{s}) > c_\eta(\mathbf{s})\}$$

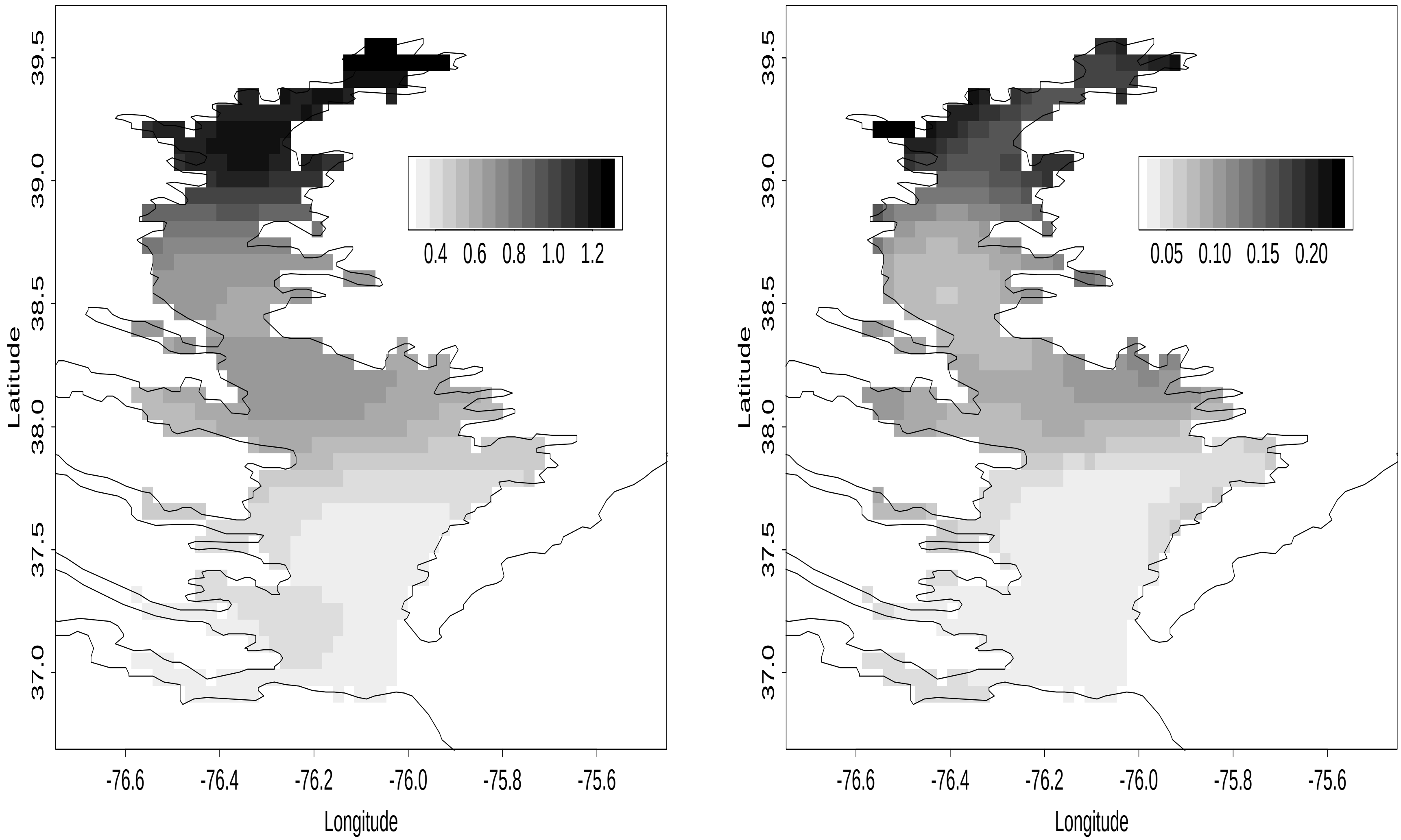
for some threshold function $c_\eta(\mathbf{s})$

Estimate H by

$$\widehat{H} = \{\mathbf{s} \in D : P\Big(Z(\mathbf{s}) > c_\eta(\mathbf{s}) \mid \mathbf{z}_{\text{obs}}\Big) > p\}$$

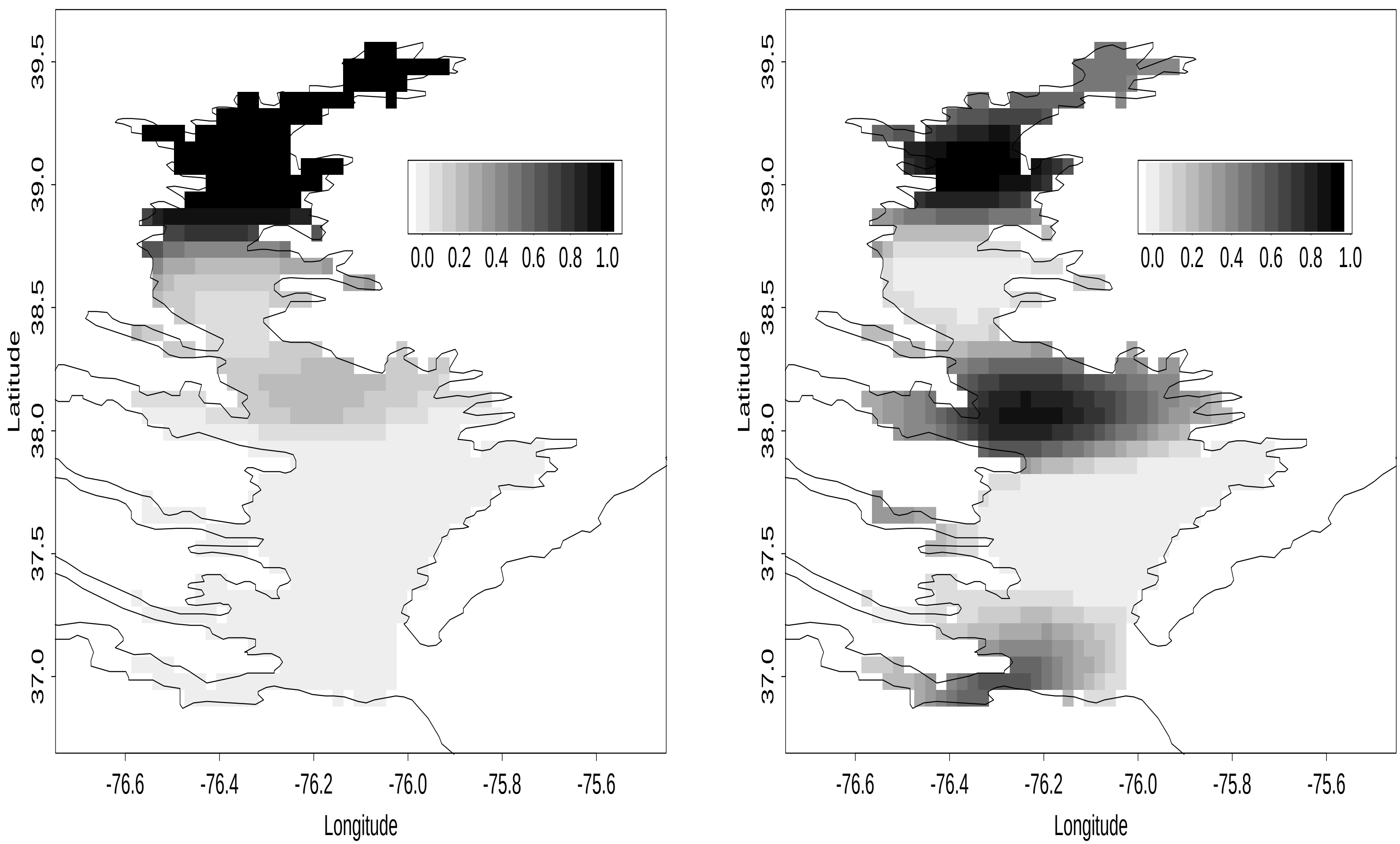
with p given

Estimated Maps



Maps of estimated nitrogen concentration (left) and uncertainty (right)

Detecting Hot Spots



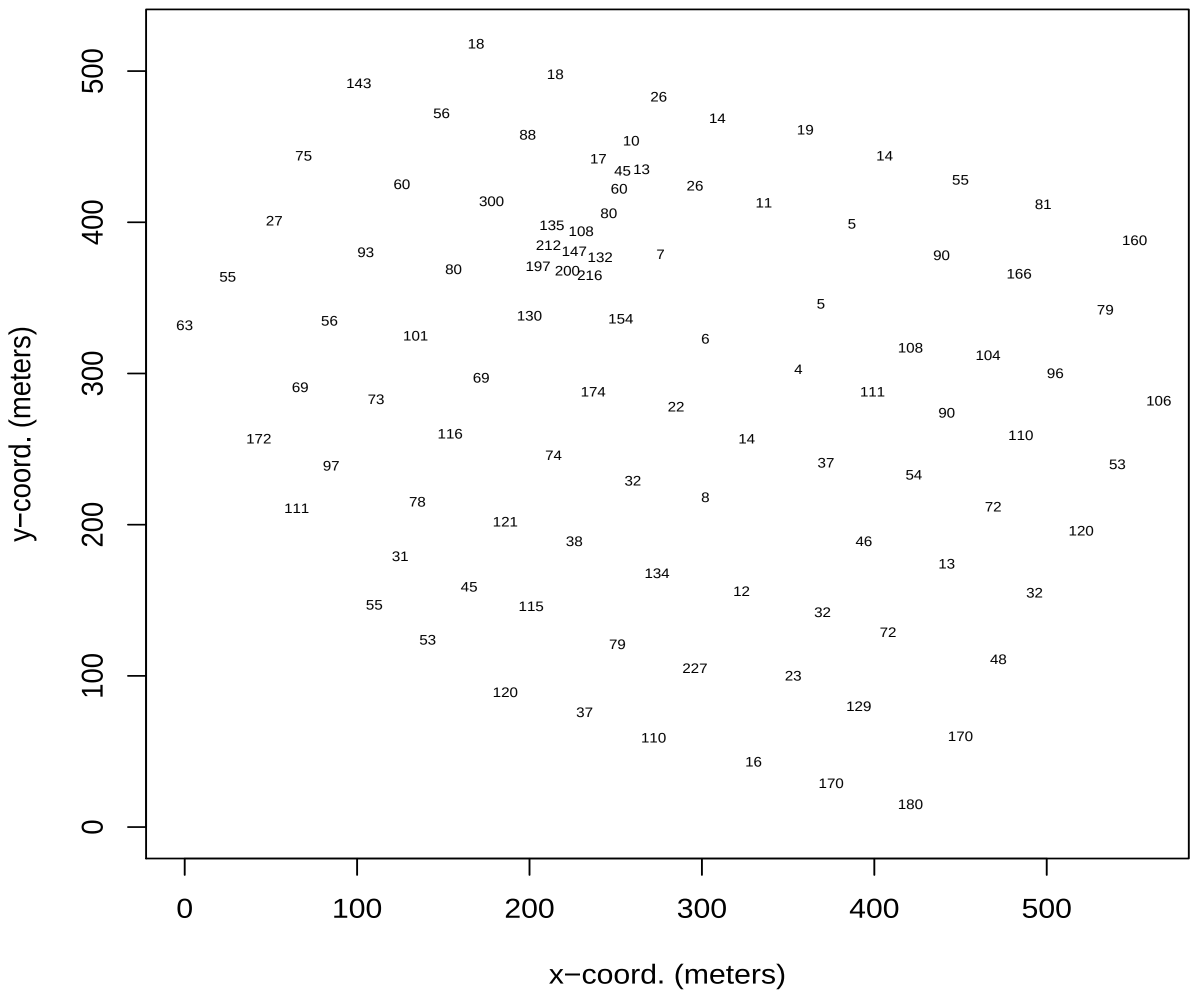
Maps of estimated $P\{Z(\mathbf{s}) > 0.75 \mid \mathbf{z}_{obs}\}$ (left) and $P\{Z(\mathbf{s}) > \mu(\mathbf{s}) + 0.05 \mid \mathbf{z}_{obs}\}$ (right)

Non-Gaussian Data

Many geostatistical datasets are markedly non-Gaussian:

- Data with skewed distributions and/or heavy tails
- Binary data (e.g. presence/absence data)
- Count data

Example 2: Weed Data in Bjertorp farm, Sweden



Description of Data and Process

- $\{\Lambda(\mathbf{s}) : \mathbf{s} \in D\}$ *positive* random field describing variation of quantity of interest; *not* observable

- To learn about $\Lambda(\cdot)$ spatial count variables $Z_1, \ldots, Z_n$ are collected having mean values related to $\Lambda(\cdot)$

- For weed data:
$\Lambda(\mathbf{s})$ = intensity of weed occurrence at $\mathbf{s}$
Z_i = number of weeds observed within a rectangle of area t_i centered at location $\mathbf{s}_i$

- The main goal is prediction of $\Lambda(\cdot)$ based on the data $\mathbf{z} = (Z_1, \ldots, Z_n)$ and the covariate information (if available).

Poisson Kriging Model

(1) Data: $Z_1, \ldots, Z_n$ are conditionally independent given $\mathbf{\Lambda} = (\Lambda(\mathbf{s}_1), \ldots, \Lambda(\mathbf{s}_n))$, and

$$\mathrm{E}\{Z_i \mid \mathbf{\Lambda}\} = \mathrm{var}\{Z_i \mid \mathbf{\Lambda}\} = t_i \Lambda(\mathbf{s}_i), \qquad i = 1, \ldots, n$$

with $t_i > 0$ known representing "sampling effort" at $\mathbf{s}_i$

(2) Latent process: $\Lambda(\mathbf{s}) = \mu(\mathbf{s})\epsilon(\mathbf{s})$, with $\mu(\mathbf{s}) > 0$ spatial trend and $\{\epsilon(\mathbf{s}) : \mathbf{s} \in D\}$ a positive random field with

$$\mathrm{E}\{\epsilon(\mathbf{s})\} = 1 \quad \text{and} \quad \mathrm{cov}\{\epsilon(\mathbf{s}), \epsilon(\mathbf{u})\} = C_\epsilon(\mathbf{s} - \mathbf{u})$$

To complete model specification, assume

$$\mu(\mathbf{s}) = \exp(\boldsymbol{\beta}'\mathbf{f}(\mathbf{s}))$$

$$C_\epsilon(\mathbf{s} - \mathbf{u}) = \exp(C_\delta(\mathbf{s} - \mathbf{u})) - 1$$

with $C_\delta(\mathbf{s} - \mathbf{u})$ a standard covariance function

Second-order Structure

- Latent process:

$$E\{\Lambda(\mathbf{s})\} = \mu(\mathbf{s}) \quad , \quad \text{cov}\{\Lambda(\mathbf{s}), \Lambda(\mathbf{u})\} = \mu(\mathbf{s})\mu(\mathbf{u})C_\epsilon(\mathbf{s}-\mathbf{u})$$

- Data:

$$\begin{aligned} \text{E}\{Z_i\} &= t_i\mu_i \\ \text{cov}\{Z_i, Z_j\} &= t_it_j\mu_i\mu_jC_\epsilon(\mathbf{s}_i-\mathbf{s}_j), \qquad i \neq j \\ \frac{1}{2}\text{var}\{Z_i - Z_j\} &= t_it_j\mu_i\mu_j\gamma_\epsilon(\mathbf{s}_i-\mathbf{s}_j) + \frac{1}{2}\Big(t_i\mu_i + t_j\mu_j + \sigma_\epsilon^2[t_i\mu_i - t_j\mu_j]^2\Big) \end{aligned}$$

with $\mu_i = \mu(\mathbf{s}_i)$ and $\sigma_\epsilon^2 = C_\epsilon(\mathbf{0})$

Residuals

From trend estimates compute 'residuals' in the form of ratios

$$R_i = \frac{Z_i}{t_i \widehat{\mu}_i}, \qquad i = 1, \ldots, n$$

Treating trend estimates as known

$$\mathsf{E}\{R_i\} \approx 1 \quad , \quad \mathsf{var}\{R_i\} \approx \sigma_\epsilon^2 + \frac{1}{t_i \mu_i}$$

and for any $i \neq j$

$$\frac{1}{2}\mathsf{var}\{R_i - R_j\} \approx \gamma_\epsilon(\mathbf{s}_i - \mathbf{s}_j) + \frac{1}{2}\Big(\frac{t_i \mu_i + t_j \mu_j}{t_i t_j \mu_i \mu_j}\Big)$$

Prediction of Latent Process

The Poisson kriging predictor of $\Lambda(\mathbf{s}_0)$ based on the residuals is the one that minimizes

$$\text{MSPE}(\hat{\Lambda}(\mathbf{s}_0)) = E\{(\Lambda(\mathbf{s}_0) - \hat{\Lambda}(\mathbf{s}_0))^2\}$$

over the class of linear unbiased predictors

$$\hat{\Lambda}(\mathbf{s}_0) = \mu(\mathbf{s}_0) \sum_{i=1}^{n} \lambda_i(\mathbf{s}_0) R_i$$

that are (approximately) unbiased

$$\sum_{i=1}^{n} \lambda_i(\mathbf{s}_0) = 1$$

Prediction of Latent Process (cont.)

- The optimal coefficients (weights) $\boldsymbol{\lambda}(\mathbf{s}_0) = (\lambda_1(\mathbf{s}_0), \ldots, \lambda_n(\mathbf{s}_0))$ are obtained as the solution of the linear system of equations

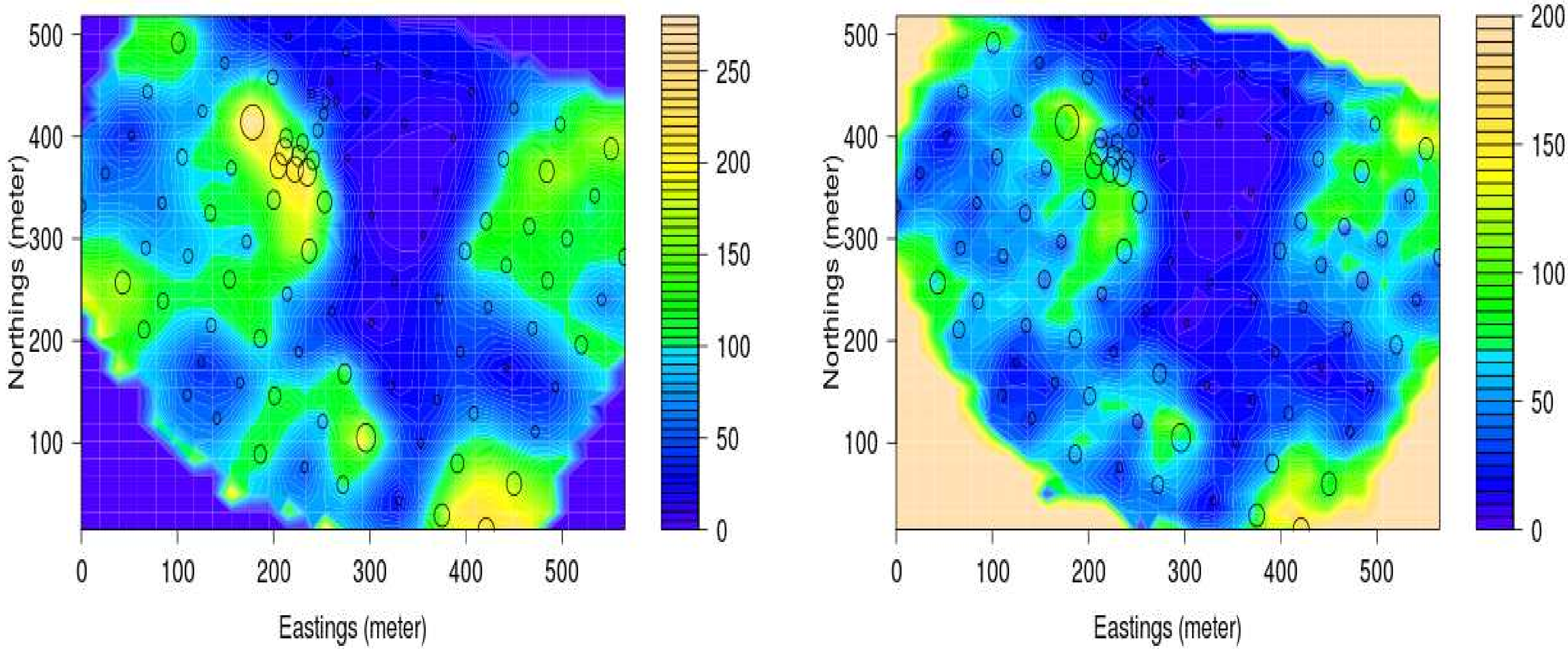
$$\begin{cases} \frac{\lambda_j}{t_j \mu_j} + \sum_{i=1}^{n} \lambda_i C_\epsilon(\mathbf{s}_i - \mathbf{s}_j) - m_0 = C_\epsilon(\mathbf{s}_j - \mathbf{s}_0); & \text{for } j = 1, \ldots, n \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases}$$

- An uncertainty measure is

$$\begin{aligned} \widehat{\sigma}^2(\mathbf{s}_0) &= \text{MSPE}(\widehat{\Lambda}^K(\mathbf{s}_0)) \\ &= \mu^2(\mathbf{s}_0)\left(\sigma_\epsilon^2 - \sum_{i=1}^{n} \lambda_i C_\epsilon(\mathbf{s}_i - \mathbf{s}_0) + m_0\right) \end{aligned}$$

- Poisson kriging predictor has the same drawbacks of the (regular) kriging predictor, plus a new one

Predictive Inference from Weed Data



Maps of $\widehat{\Lambda}^K(\mathbf{s}_0)$ (left) and $\widehat{\sigma}(\mathbf{s}_0)$ (right)

THANKS FOR YOUR ATTENTION

victor.deoliveira@utsa.edu
http://faculty.business.utsa.edu/vdeolive