MODEL BASED GEOSTATISTICS

(Course Slides)

Presented by:
Paulo Justiniano Ribeiro Jr [1]

Departamento de Estatística
Universidade Federal do Paraná, Brasil

[1]*Correspondence Addrees:* Departamento de Estatística, Universidade Federal do Paraná, Cx. Postal 19.081, 81.531-990, Curitiba, PR, Brasil. E-mail: pj@est.ufpr.br

Acknowledgements

The course notes and slides are result of joint work with:

- **Professor Peter J. Diggle**
  Department of Mathematics and Statistics
  Lancaster University, UK
- **Ole Christensen**
  Center for Bioinformatik, Datalogisk Institut, Aarhus Universitet, Denmark
  formlely at: Department of Mathematics and Statistics, Lancaster University, UK
  and *Department of Mathematical Sciences, Aalborg University, Denmark*

PART I:

INTRODUCTION

1. **Basic Examples of Spatial Data**

2. **A Taxonomy for Spatial Statistics**

3. **Further Examples of Geostatistical Problems**

4. **Characteristic Features of Geostatistical Problems**

5. **Some History**

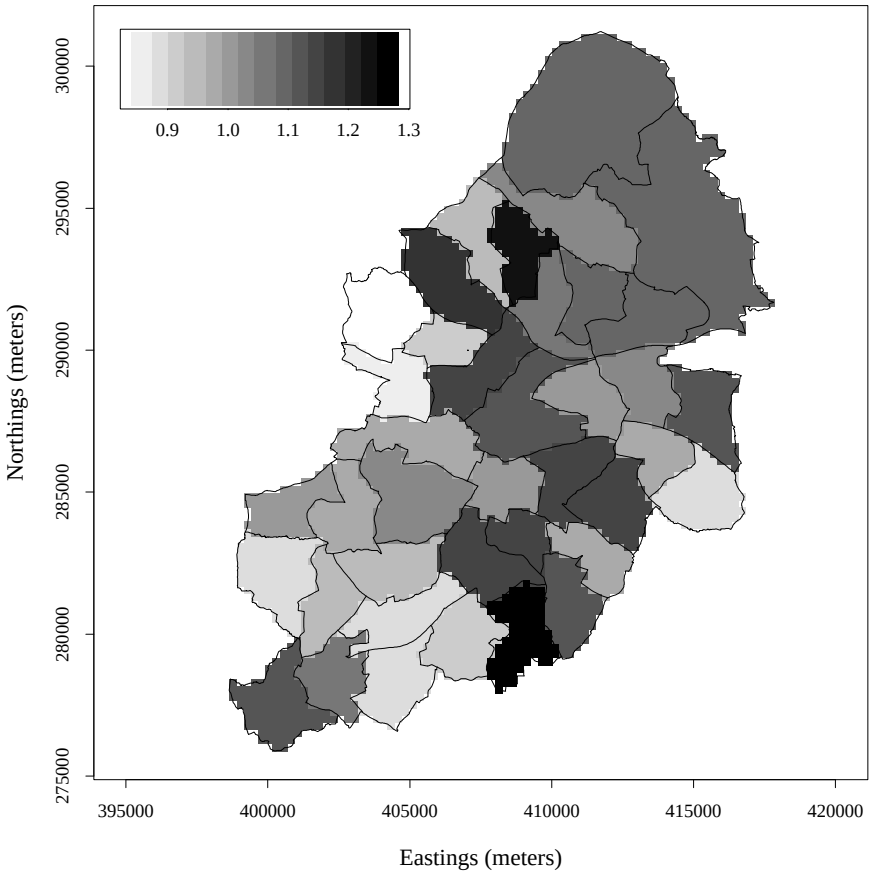
6. **Core Geostatistical Problems**

7. **Model Based Geostatistics**

1. **Basic Examples of Spatial Data**

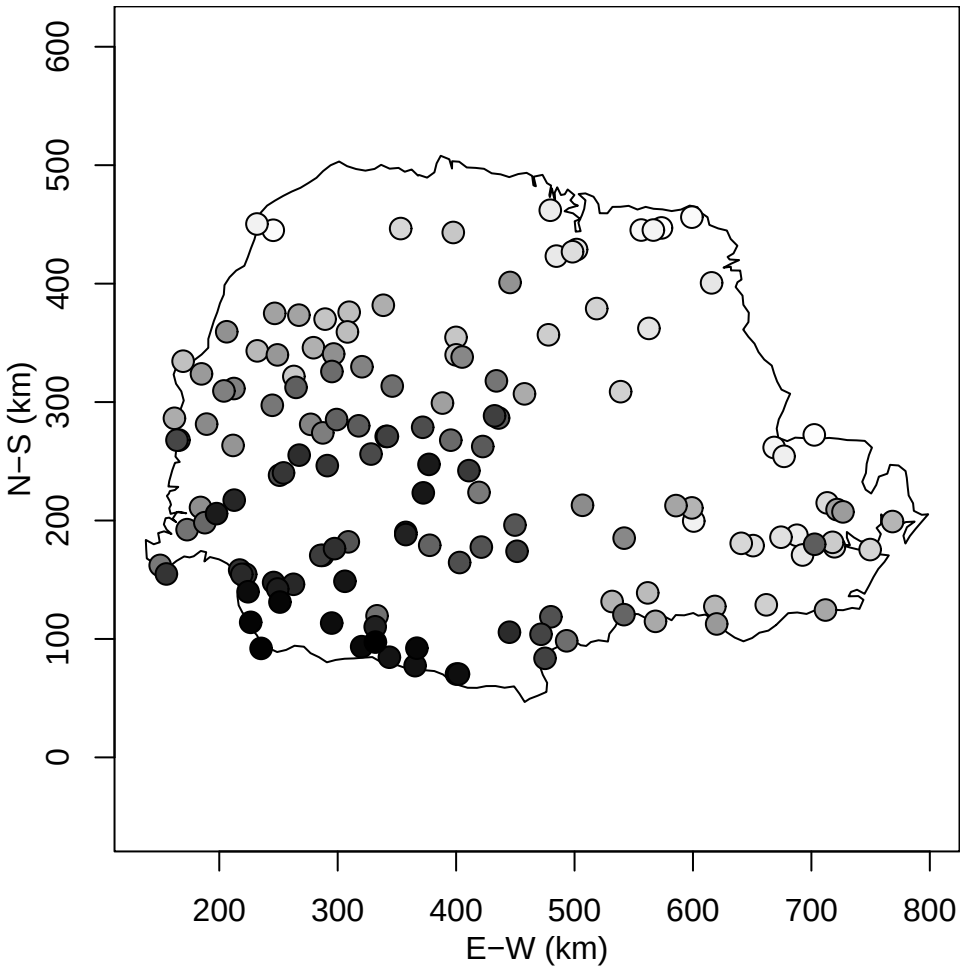
(a) **Cancer rates in administrative regions**

Grey-scale corresponds to estimated variation in relative risk of colorectal cancer in the 36 electoral wards of the city of Birmingham, UK.



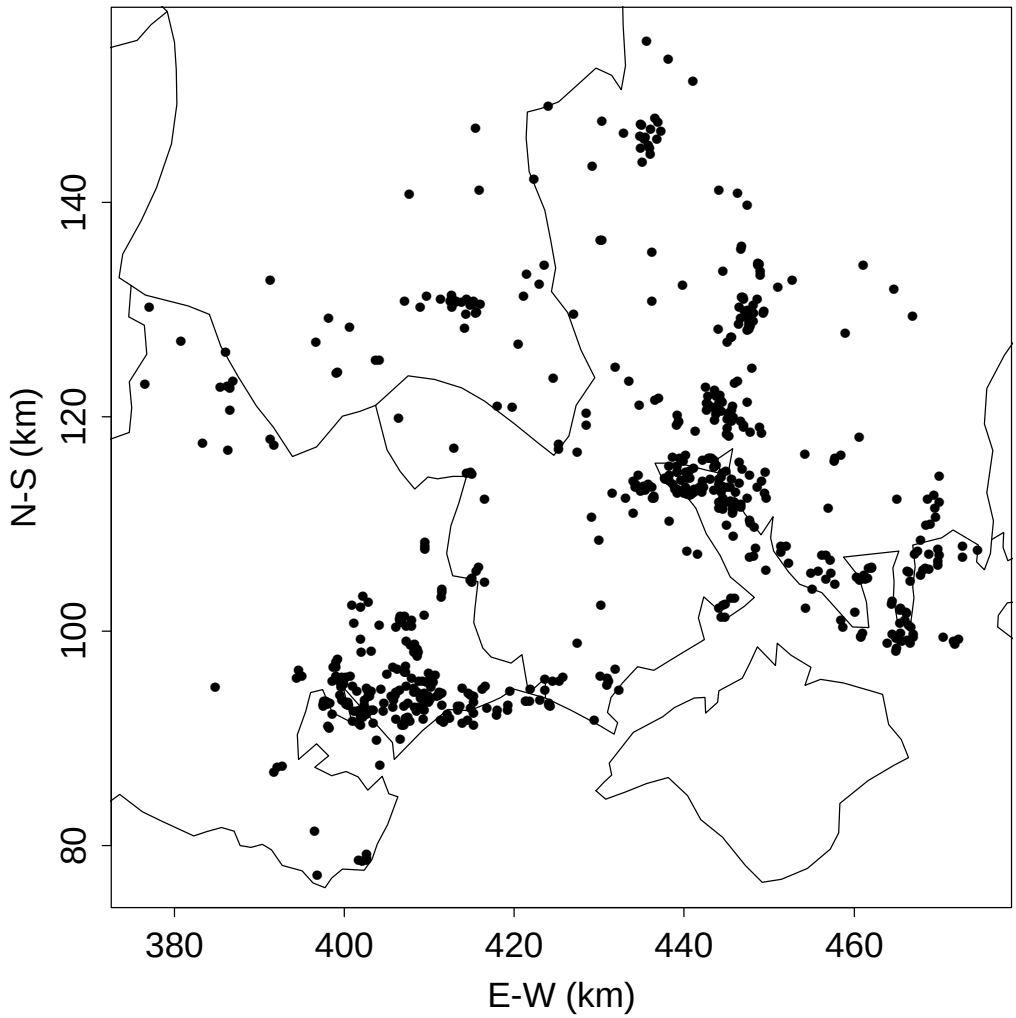
(b) **Rainfall in Paraná State, Brasil**

Rainfall measurements at 143 recording stations.
Average for the May-June period (dry season).



(c) Campylobacter cases in southern England

Residential locations of 651 cases of campylobacter reported over a one-year period in central southern England.



2. A Taxonomy of Spatial Statistics

(a) Discrete spatial variation

Basic structure. $Y_i : i = 1, ..., n$

- rarely arises naturally
- but often useful as a pragmatic strategy
- models typically defined indirectly from full conditionals, $[Y_i | Y_j, \forall j \neq i]$

(b) Continuous spatial variation

Basic structure. $Y(x) : x \in \mathbb{R}^2$

- data $(y_i, x_i) : i = 1, ..., n$, locations x_i may be:
  - non-stochastic (eg lattice to cover observation region A)
  - or stochastic, *but independent of the process* $Y(x)$

(c) Spatial point processes

Basic structure. Countable set of points $x_i \in \mathbb{R}^2$, generated stochastically.

- data sometimes converted to apparently discrete spatial variation by aggregation over sub-regions

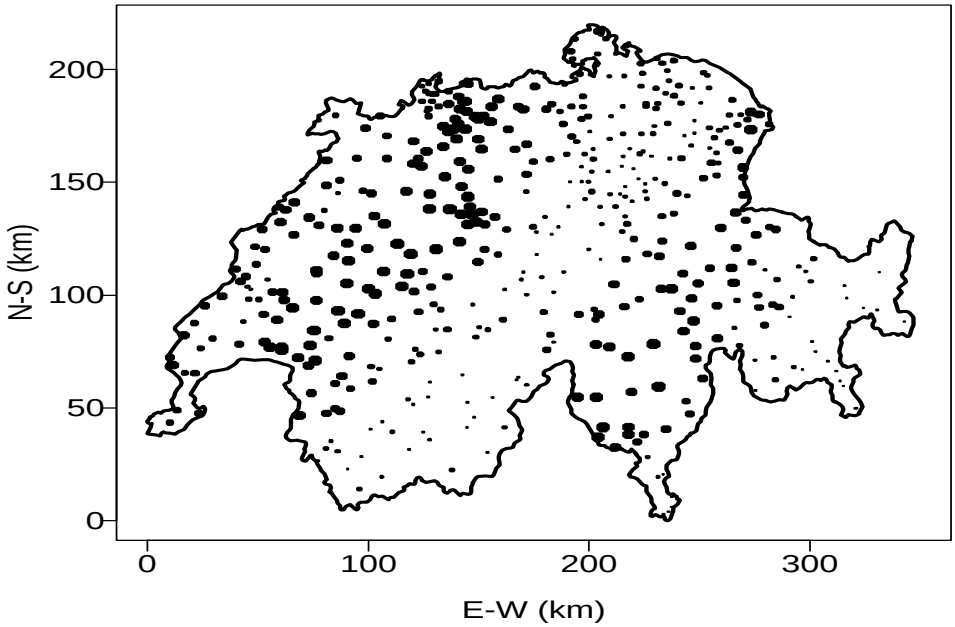
Spatial statistics is the collection of statistical methods in which spatial locations play an explicit role in the analysis of data.

Two strategic issues:

- don't confuse the *data-format* with the *underlying process*
- the choice of model may be influenced by the scientific objectives of the study – *analyse problems, not data*

3. Further Examples of Geostatistical Problems

(a) Swiss rainfall data

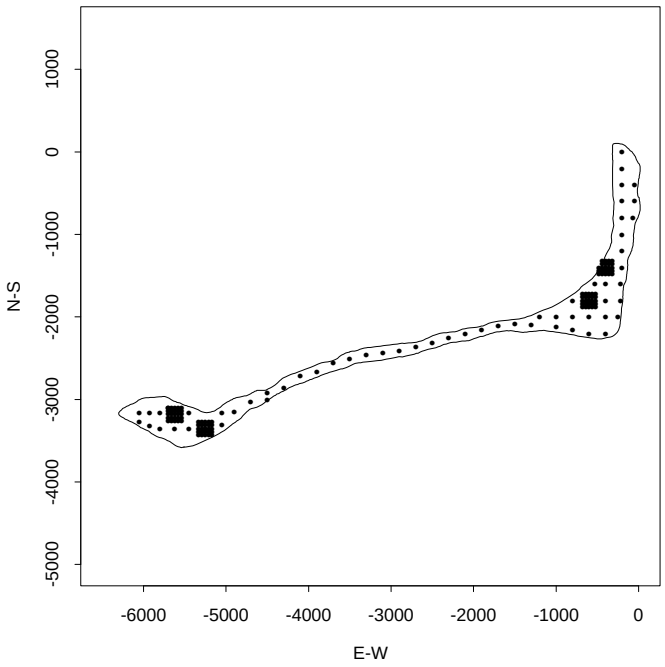


Locations shown as points with size proportional to the value of the observed rainfall.

- 467 locations in Switzerland
- daily rainfall measurements on 8th of May 1986
- data from:
  Spatial Interpolation Comparison 97
  (`ftp://ftp.geog.uwo.ca/SIC97/`).

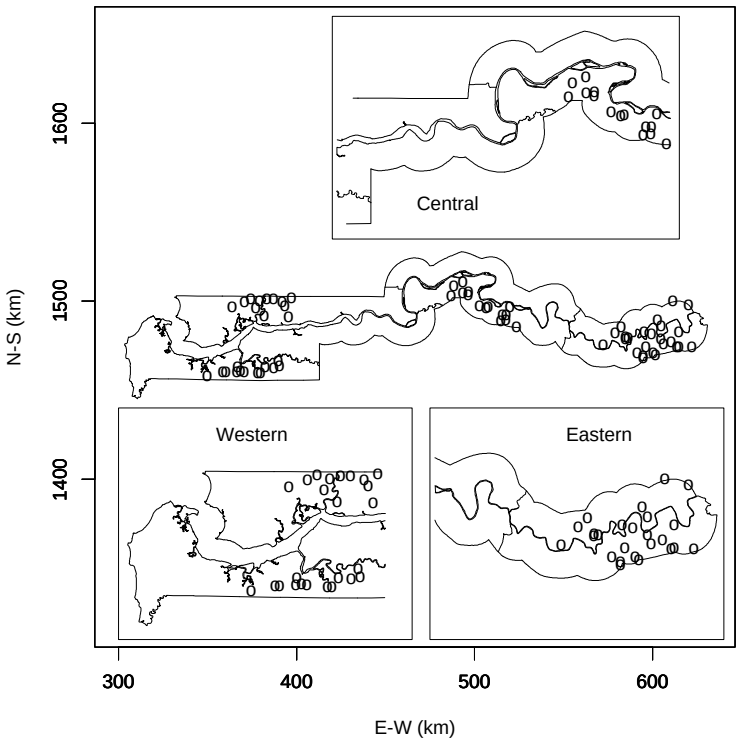
(b) **Rongelap Island**

- study of residual contamination, following nuclear weapons testing programme during 1950's
- island evacuated in 1985, is it now safe for re-settlement?
- survey yields noisy measurements Y_i of radioactive caesium concentrations
- initial grid of locations x_i at 200m spacing later
  supplemented by in-fill squares at 40m spacing.
- particular interest in maximum caesium concentration



(c) **Gambia malaria**

- survey of villages in Gambia
- in village i, data $Y_{ij} = 0/1$ denotes absence/presence of malarial parasites in blood sample from child j
- village-level covariates:
  - village locations
  - public health centre in village?
  - satellite-derived vegetation green-ness index
- child-level covariates:
  - age, sex, bed-net use
- interest in effects of covariates, and pattern of residual spatial variation



4. Characteristic Features of Geostatistical Problems

- data consist of **responses** Y_i associated with **locations** x_i
- in principle, Y could be determined from any location x within a continuous spatial region A
- it is reasonable to behave as if $\{Y(x) : x \in A\}$ is a stochastic process
- x_i is typically fixed. If the locations x_i are generated by a stochastic point process, it is reasonable to behave as if this point process is independent of the $Y(x)$ process
- scientific objectives include prediction of one or more functionals of a stochastic process $\{S(x) : x \in A\}$ which is dependent on the $Y(x)$ process.

5. **Some History**

- Origins in problems connected with estimation of ore reserves in mineral exploration/mining (Krige, 1951).

- Subsequent development largely independent of "mainstream" spatial statistics, initially by Matheron and colleagues at École des Mines, Fontainebleau.

- Parallel developments by Matérn (1946, 1960),
  Whittle (1954, 1962, 1963)

- Ripley (1981) re-casts kriging in terminology of stochastic process prediction

- Significant cross-fertilization during 1980's and 1990's (eg *variogram* is now a standard statistical tool for analysing correlated data in space and/or time).

- But still vigorous debate on practical issues:
  – prediction vs inference
  – role of explicit probability models

6. **Core Geostatistical Problems**

- **Design**
  - how many locations?
  - how many measurements?
  - spatial layout of the locations?
  - what to measure at each location?

- **Modelling**
  - probability model for the signal, $[S]$
  - conditional probability model for the measurements, $[Y|S]$

- **Estimation**
  - assign values to unknown model parameters
  - make inferences about (functions of) model parameters

- **Prediction**
  - evaluate $[T|Y]$, the conditional distribution of the target given the data

7. Model-Based Geostatistics

- declares explicit stochastic model
- apply general statistical principles for inference

Notation

$$(Y_i, x_i) : i = 1, ..., n$$

- $\{x_i : i = 1, ..., n\}$ is the **sampling design**
- $\{Y(x) : x \in A\}$ is the **measurement process**
- $\{S(x) : x \in A\}$ is the **signal process**
- $T = \mathcal{F}(S)$ is the **target for prediction**
- $[S, Y] = [S][Y|S]$ is the **geostatistical model**

Traditional geostatistics:

- avoids explicit references to the parametric specification of the model
- inference via variograms
  (Matheron's "estimating and choosing")
- complex variogram structures are often used
- concentrates on linear estimators
- specific methods/paradigms for:
  - point prediction (SK, OK, KTE, UK)
  - prediction of non-linear functionals (IK, DK)
  - predictive estimation (IK, DK)
  - simulations from the predictive distribution
    (SGSIM, SISIM, ...)
- the *kriging menu*

PART II:

SPATIAL PREDICTION AND GAUSSIAN MODELS

1. **The Gaussian Model**

2. **Specification of the Correlation Function**

3. **Stochastic Process Prediction**

4. **Linear Geostatistics**

5. **Prediction under the Gaussian Model**

6. **What does Kriging Actually do to the Data**

7. **Prediction of Functionals**

8. **Directional Effects**

9. **Non-stationary Gaussian Models**

1. The Gaussian Spatial Model

(a) $S(\cdot)$ is a stationary Gaussian process with

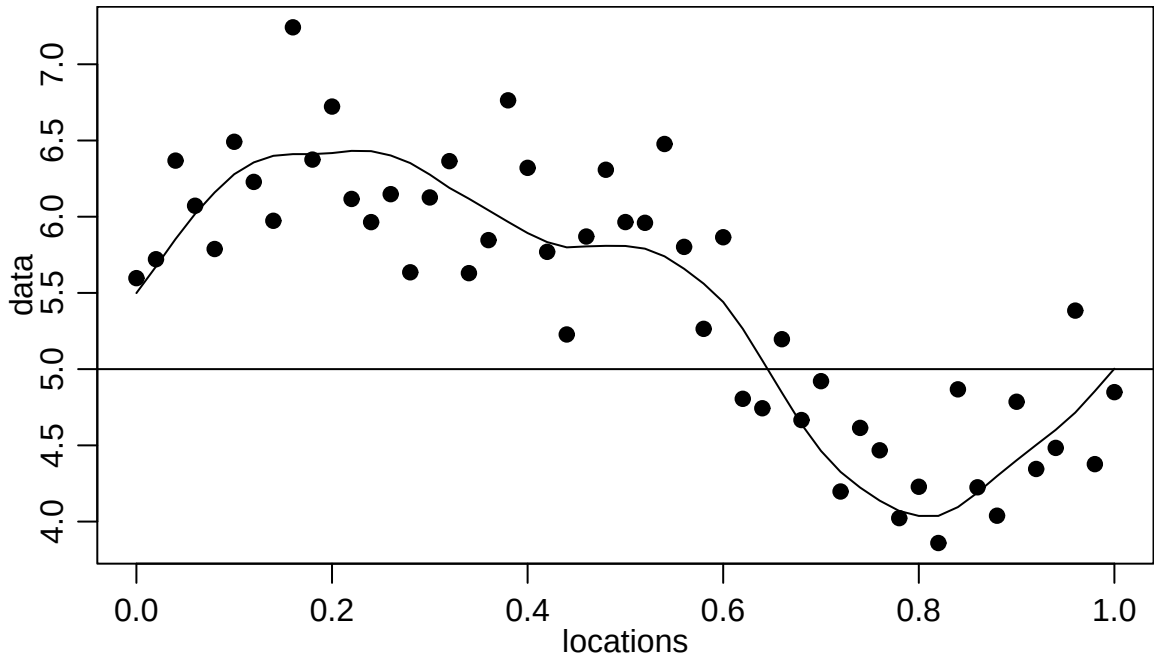
i. $\mathrm{E}[S(x)] = \mu$,

ii. $\mathrm{Var}\{S(x)\} = \sigma^2$

iii. $\rho(u) = \mathrm{Corr}\{S(x), S(x-u)\}$;

(b) the conditional distribution of Y_i given $S(\cdot)$ is Gaussian with mean $S(x_i)$ and variance τ^2;

(c) $Y_i : i = 1, ..., n$ are mutually independent, conditional on $S(\cdot)$.



simulated data in 1-D illustrating the elements of the model: data $Y(x_i)$ (dots), signal $S(x)$ (curve line) and mean μ. (horizontal line).

An Equivalent Formulation:

$$Y_i = S(x_i) + \epsilon_i : i = 1, ..., n.$$

where $\epsilon_i : i = 1, ..., n$ are mutually independent, identically distributed with $\epsilon_i \sim \mathrm{N}(0, \tau^2)$.

Then, the joint distribution of Y is multivariate Normal,

$$Y \sim \mathrm{MVN}(\mu \mathbf{1}, \sigma^2 R + \tau^2 I)$$

where:

$\mathbf{1}$ denotes an n-element vector of ones,

I is the $n \times n$ identity matrix

R is the $n \times n$ matrix with $(i, j)^{th}$ element $\rho(u_{ij})$ where
$u_{ij} = ||x_i - x_j||$, the Euclidean distance between x_i and x_j.

2. Specification of the Correlation Function

Differentiability of Gaussian processes

- A formal mathematical description of the smoothness of a spatial surface $S(x)$ is its degree of differentiability.
- A process $S(x)$ is *mean-square continuous* if, for all x, $\mathrm{E}[\{S(x+h) - S(x)\}^2] \to 0$ as $h \to 0$.
- $S(x)$ is *mean square differentiable* if there exists a process $S'(x)$ such that, for all x,

$$\mathrm{E}\left[\left\{\frac{S(x+h) - S(x)}{h} - S'(x)\right\}^2\right] \to 0 \text{ as } h \to 0$$

- the mean-square differentiability of $S(x)$ is directly linked to the differentiability of its covariance function

Theorem 3 Let $S(x)$ be a stationary Gaussian process with correlation function $\rho(u) : u \in \mathbb{R}$. Then:

- $S(x)$ is mean-square continuous iff $\rho(u)$ is continuous at $u = 0$;
- $S(x)$ is k times mean-square differentiable iff $\rho(u)$ is (at least) $2k$ times differentiable at $u = 0$.

(a) The Matérn family

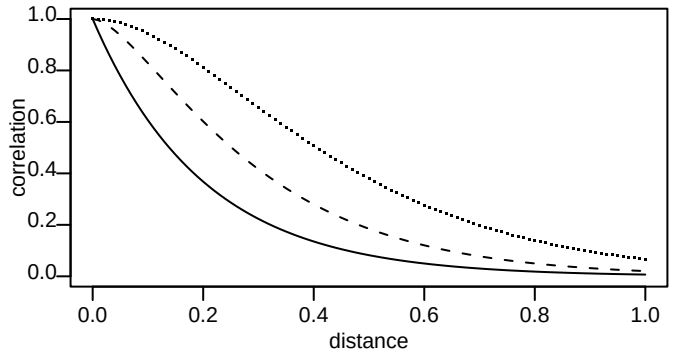
The correlation function is given by:

$$\rho(u) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^{\kappa}K_{\kappa}(u/\phi)$$

- κ and ϕ are parameters
- $K_{\kappa}(\cdot)$ denotes modified Bessel function of order κ
- valid for $\phi > 0$ and $\kappa > 0$.

- $\kappa = 0.5$: *exponential* correlation function
- $\kappa \to \infty$: *Gaussian* correlation function

- $S(x)$ is mean-square $\lceil \kappa - 1$ times differentiable



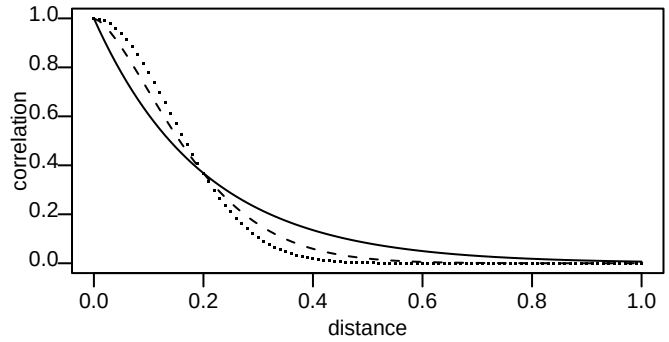
Three examples of the Matérn correlation function with $\phi = 0.2$ and $\kappa = 1$ (solid line), $\kappa = 1.5$ (dashed line) and $\kappa = 2$ (dotted line).

(b) The powered exponential family

$$\rho(u) = \exp\{-(u/\phi)^\kappa\}$$

- defined for $\phi > 0$ and $0 < \kappa \leq 2$
- ϕ and κ are parameters
- mean-square continuous (but non-differentiable) if $\kappa < 2$
- mean-square infinitely differentiable if $\kappa = 2$
- $\rho(u)$ very ill-conditioned when $\kappa = 2$
- $\kappa = 1$: *exponential* correlation function
- $\kappa = 2$: *Gaussian* correlation function

Conclusion: not as flexible as it looks

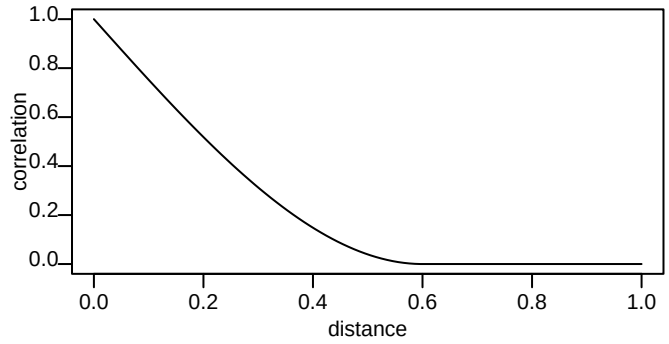


Three examples of the powered exponential correlation function with $\phi = 0.2$ and $\kappa = 1$ (solid line), $\kappa = 1.5$ (dashed line) and $\kappa = 2$ (dotted line).

(c) The spherical family

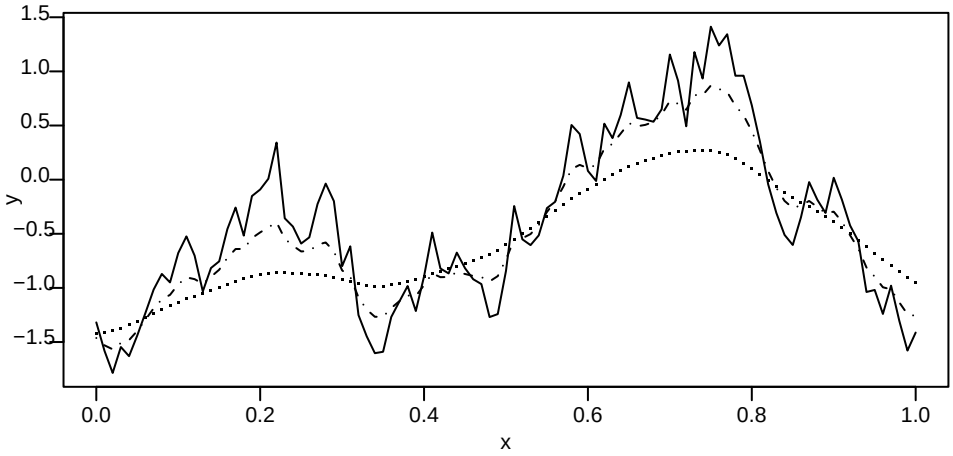
$$\rho(u;\phi) = \begin{cases} 1 - \frac{3}{2}(u/\phi) + \frac{1}{2}(u/\phi)^3 & : \ 0 \le u \le \phi \\ 0 & : \quad u > \phi \end{cases}$$

- $\phi > 0$ is parameter
- finite range
- non-differentiable at the origin

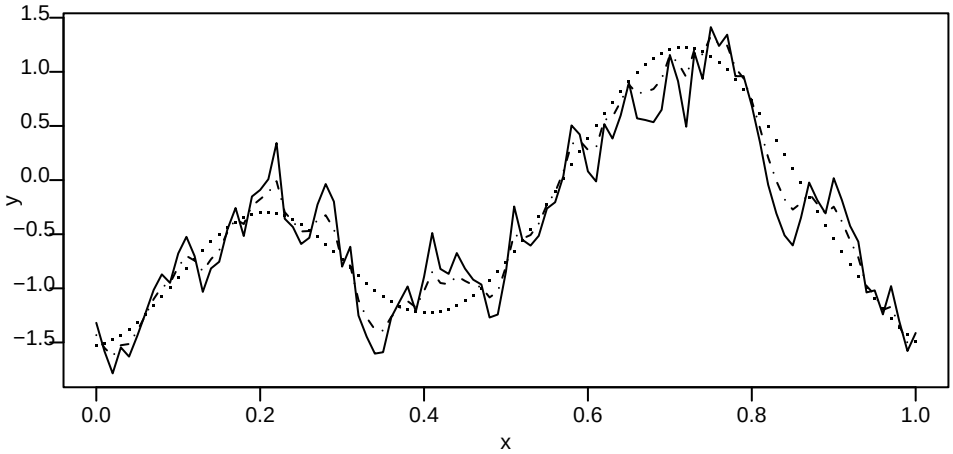


The spherical correlation function with $\phi = 0.6$.

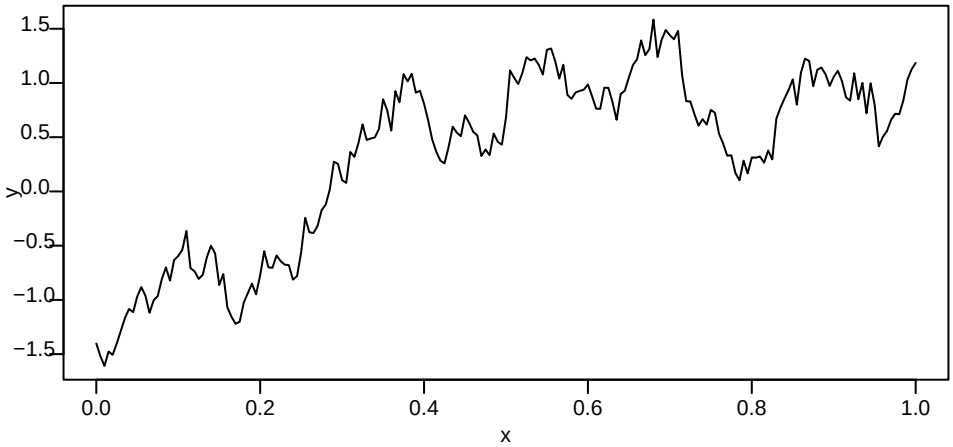
Comparable Simulations (same seed)



simulations with Matérn correlation functions with $\phi = 0.2$ and $\kappa = 0.5$ (solid line), $\kappa = 1$ (dashed line) and $\kappa = 2$ (dotted line).



simulations with powered exponential correlation function with $\phi = 0.2$ and $\kappa = 1$ (solid line), $\kappa = 1.5$ (dashed line) and $\kappa = 2$ (dotted line).



simulations with spherical correlation function ($\phi = 0.6$).

3. Stochastic Process Prediction

General results for prediction

goal: predict the realised value of a (scalar) r.v. T, using data y a realisation of a (vector) r.v. Y.

predictor: of T is any function of Y, $\hat{T} = t(Y)$

best choice: needs a criterion

MMSPE: the *best* predictor minimises

$$MSPE(\hat{T}) = \mathrm{E}[(T - \hat{T})^2]$$

Theorem 1.
The minimum mean square error predictor of T is

$$\hat{T} = \mathrm{E}(T|Y).$$

Theorem 2.
(a) The prediction mean square error of $\hat{T}$ is

$$\mathrm{E}[(T - \hat{T})^2] = \mathrm{E}_Y[\mathrm{Var}(T|Y)],$$

(the prediction variance is an estimate of the MSPE).
(b) $\mathrm{E}[(T - \hat{T})^2] \leq \mathrm{Var}(T)$, with equality if T and Y are independent random variables.

Comments

- We call $\hat{T}$ the *least squares predictor* for T, and $\mathrm{Var}(T|Y)$ its *prediction variance*

- $\mathrm{Var}(T) - \mathrm{Var}(T|Y)$ measures the contribution of the data (exploiting dependence between T and Y)

- point prediction, prediction variance are summaries

- complete answer is the distribution $[T|Y]$

- not transformation invariant:
  $\hat{T}$ the best predictor for T does NOT necessarily imply that $g(\hat{T})$ is the best predictor for $g(T)$.

4. Linear Gaussian Geostatistics

Suppose the **target** for prediction is $T = S(x)$

A **predictor** for T is a function $\hat{T} = \hat{T}(Y)$

The **mean square prediction error (MSPE)** is

$$MSPE(\hat{T}) = \mathrm{E}[(\hat{T} - T)^2]$$

The **the predictor** which minimises MSPE is

$$\hat{T} = \mathrm{E}[S(x)|Y]$$

Two approaches:

- **Model-based geostatistics:**
  - specify a probability model for $[Y, T]$
  - choose $\hat{T}$ to minimise $MSPE(\hat{T})$ amongst all functions $\hat{T}(Y)$
- **Traditional (linear) geostatistics:**
  - Assume that $\hat{T}$ is linear in Y, so that

$$\hat{T} = b_0(x) + \sum_{i=1}^{n} b_i(x) Y_i$$

  - Choose b_i to minimise $MSPE(\hat{T})$ within the class of linear predictors

Coincident results under Gaussian assumptions

5. Prediction Under The Gaussian Model

- assume that the target for prediction is $T = S(x)$
- $[T, Y]$ are jointly multivariate Gaussian.
- $\hat{T} = \mathrm{E}(T|Y)$, $\mathrm{Var}(T|Y)$ and $[T|Y]$ can be easily derived from a standard result:

Theorem 4. Let $X = (X_1, X_2)$ be jointly multivariate Gaussian, with mean vector $\mu = (\mu_1, \mu_2)$ and covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

ie $X \sim \mathrm{MVN}(\mu, \Sigma)$. Then, the conditional distribution of X_1 given X_2 is also multivariate Gaussian, $X_1|X_2 \sim \mathrm{MVN}(\mu_{1|2}, \Sigma_{1|2})$, where

$$\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$$

and

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

For the geostatistical model:

$[T, Y]$ is multivariate Gaussian with mean vector $\mu\mathbf{1}$ and variance matrix

$$\begin{bmatrix} \sigma^2 & \sigma^2\mathbf{r}' \\ \sigma^2\mathbf{r} & \tau^2 I + \sigma^2 R \end{bmatrix}$$

where r is a vector with elements $r_i = \rho(||x - x_i||) : i = 1, ..., n$.

Hence, using Theorem 4 with $X_1 = T$ and $X_2 = Y$, we find that the minimum mean square error predictor for $T = S(x)$ is

$$\hat{T} = \mu + \sigma^2\mathbf{r}'(\tau^2 I + \sigma^2 R)^{-1}(Y - \mu\mathbf{1}) \qquad (1)$$

with prediction variance

$$\mathrm{Var}(T|Y) = \sigma^2 - \sigma^2\mathbf{r}'(\tau^2 I + \sigma^2 R)^{-1}\sigma^2\mathbf{r}. \qquad (2)$$

Notes

1. Because the conditional variance does not depend on Y, the prediction mean square error is equal to the prediction variance.

2. Equality of prediction mean square error and prediction variance is a special property of the multivariate Gaussian distribution, not a general result.

3. In conventional geostatistical terminology, construction of the surface $\hat{S}(x)$, where $\hat{T} = \hat{S}(x)$ is given by (1), is called *simple kriging*. This name is a reference to D.G. Krige, who pioneered the use of statistical methods in the South African mining industry (Krige, 1951).

6. What Does Kriging Actually Do to the Data?

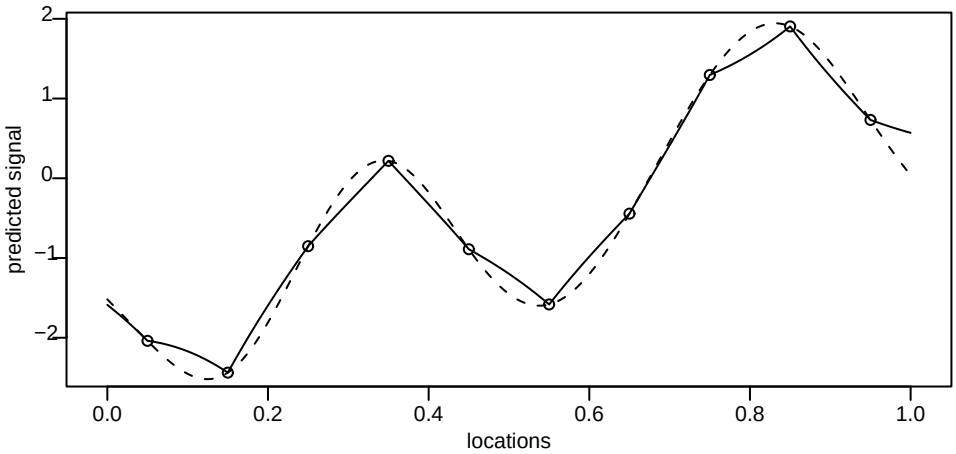
The minimum mean square error predictor for $S(x)$ is given by

$$\begin{aligned} \hat{T} = \hat{S}(x) &= \mu + \sum_{i=1}^{n} w_i(x)(Y_i - \mu) \\ &= \{1 - \sum_{i=1}^{n} w_i(x)\}\mu + \sum_{i=1}^{n} w_i(x) Y_i \end{aligned}$$

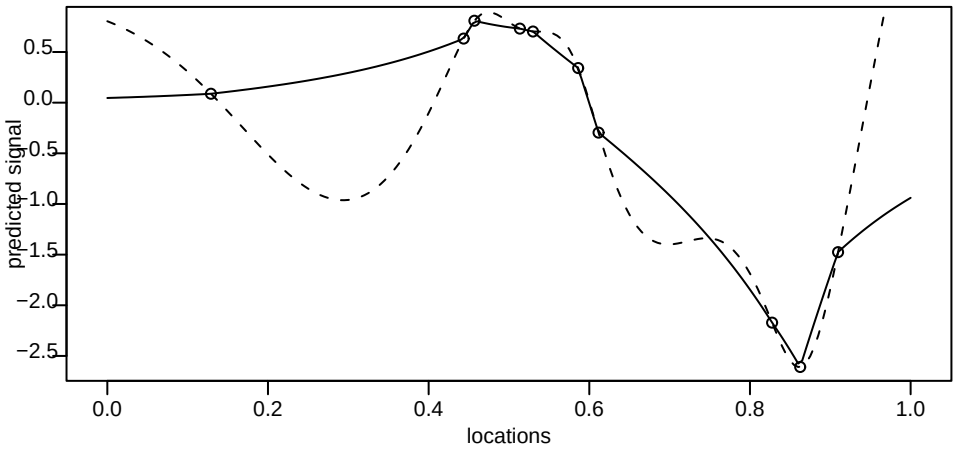
- the predictor $\hat{S}(x)$ compromises between its unconditional mean μ and the observed data Y
- the nature of the compromise depends on the target location x, the data-locations x_i and the values of the model parameters.
- call the $w_i(x)$ the *prediction weights*.

6.1 **Effects on predictions**

(a) **Varying the correlation function**

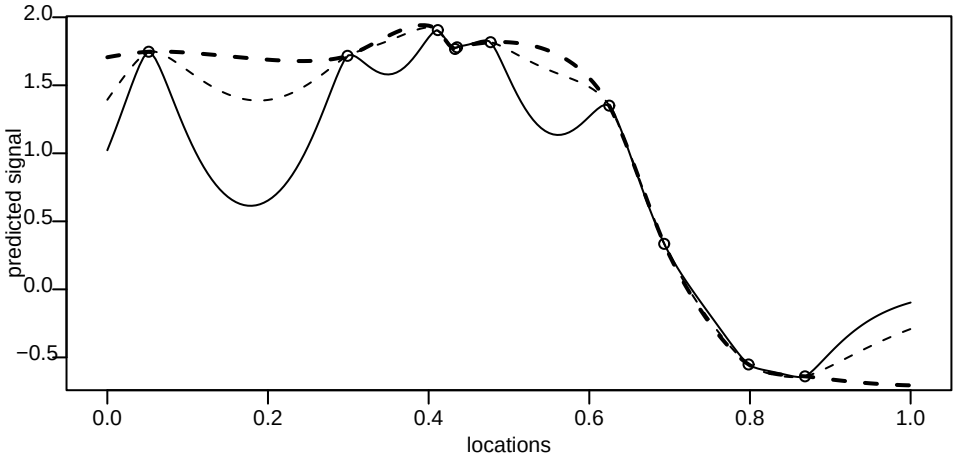


Predictions from 10 equally spaced data-points using exponential (solid line) or Matérn of order 2 (dashed line) correlation functions.



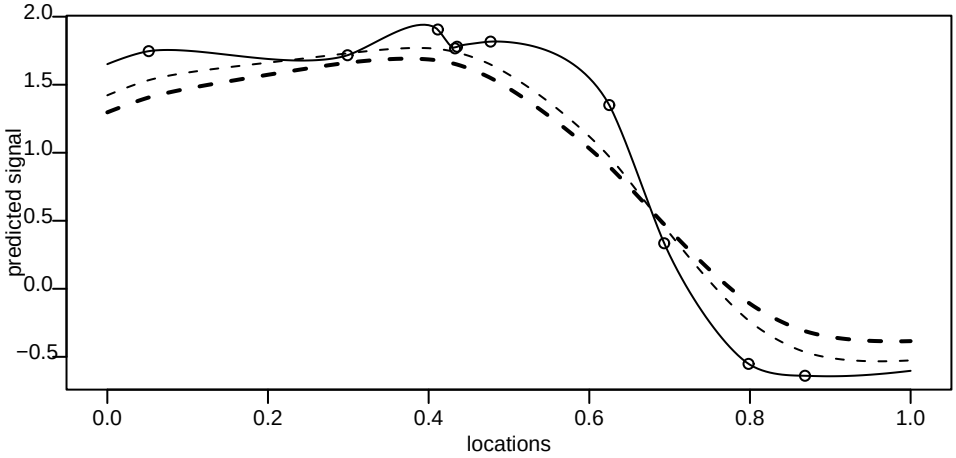
Predictions from 10 randomly spaced data-points using exponential (solid line) or Matérn of order 2 (dashed line) correlation functions.

(b) **Varying the correlation parameter**

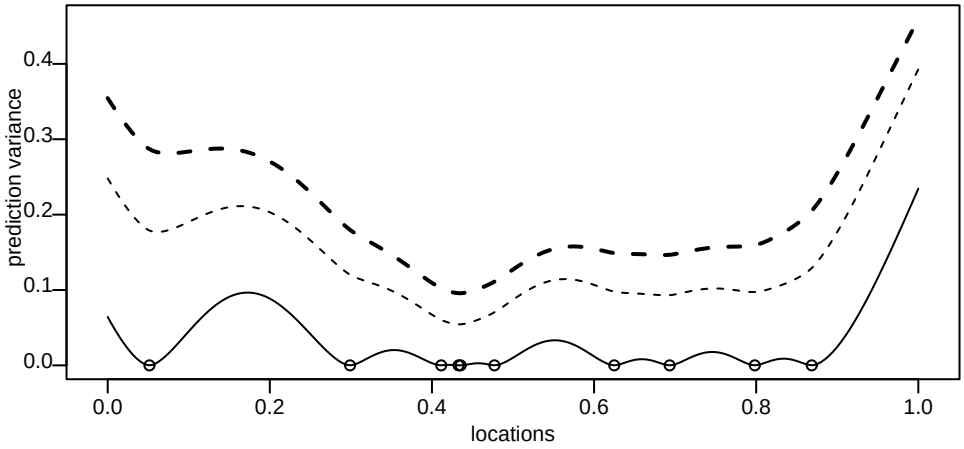


Predictions from 10 randomly spaced data-points using the Matérn ($\kappa = 2$) correlation function and different values of ϕ: 0.05 (solid line), 0.1 (dashed line) and 0.5 (thick dashed line).

(c) **Varying the noise-to-signal ratio**



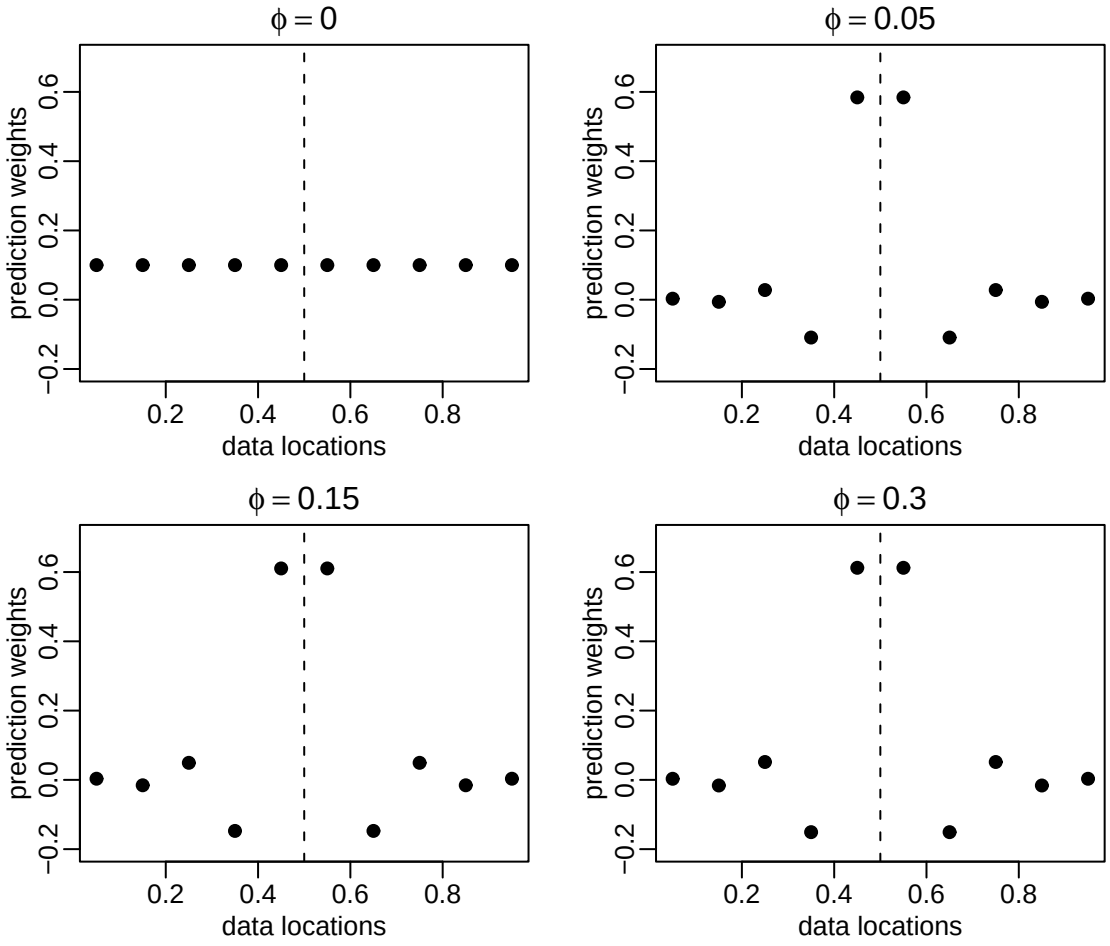
Predictions from 10 randomly spaced data-points using the Matérn correlation function and different values of τ^2: 0 (solid line), 0.25 (dashed line) and 0.5 (thick dashed line).



Prediction variances from 10 randomly spaced data-points using the Matérn correlation function and different values of τ^2: 0 (solid line), 0.25 (dashed line) and 0.5 (thick dashed line).

6.2 Effects on kriging weights

(a) The prediction weights: varying ϕ



Prediction weights for 10 equally spaced data-points with target location $x = 0.50$.

i. **varying parameter** $\phi = 0, 0.05, 0.15, 0.30$

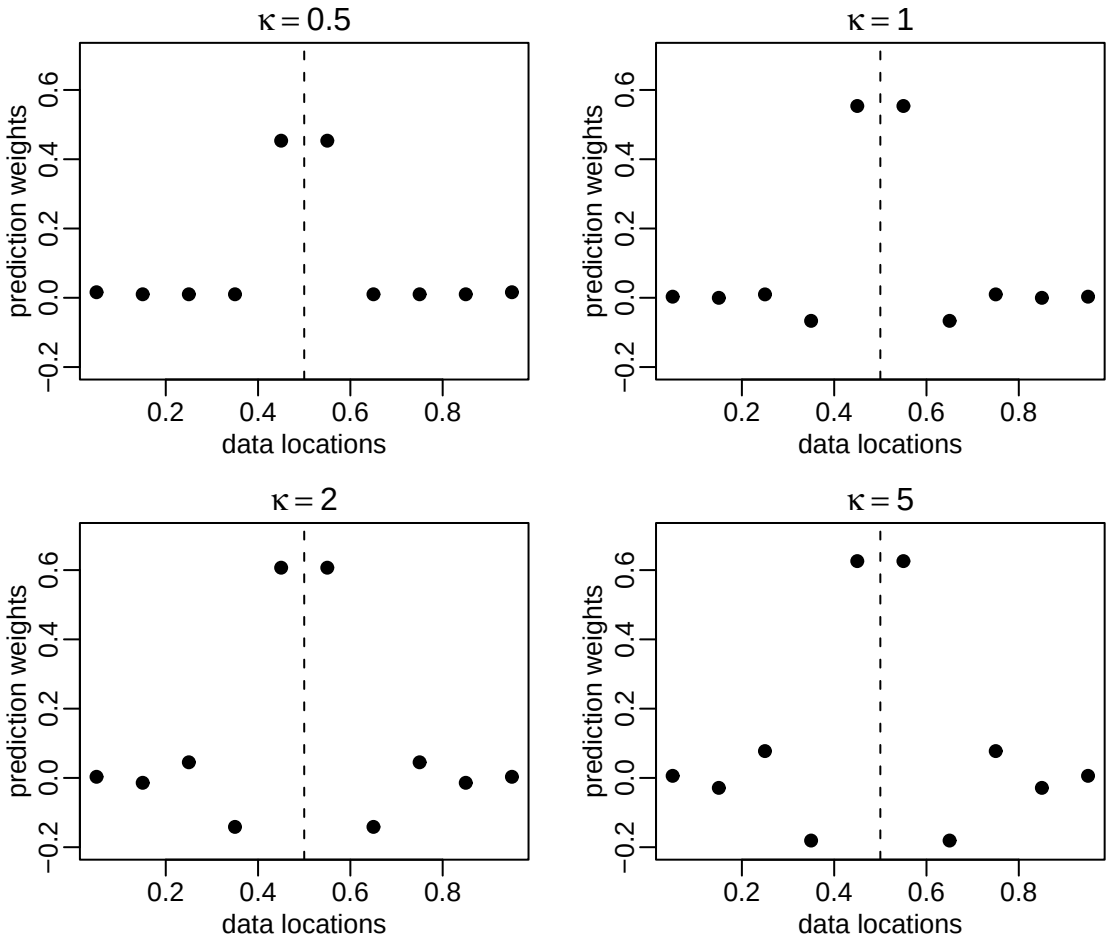
ii. locations: equally spaced $x_i = -0.05 + 0.1i$: $i = 1, ..., 10$

iii. prediction location: $x = 0.50$

iv. correlation function: Matérn with $\kappa = 2$

v. nugget: $\tau^2 = 0$

(b) **The prediction weights: varying** κ



Prediction weights for 10 equally spaced data-points with target location $x = 0.50$.

i. **varying parameter** $\kappa = 0.5, 1, 2, 5$

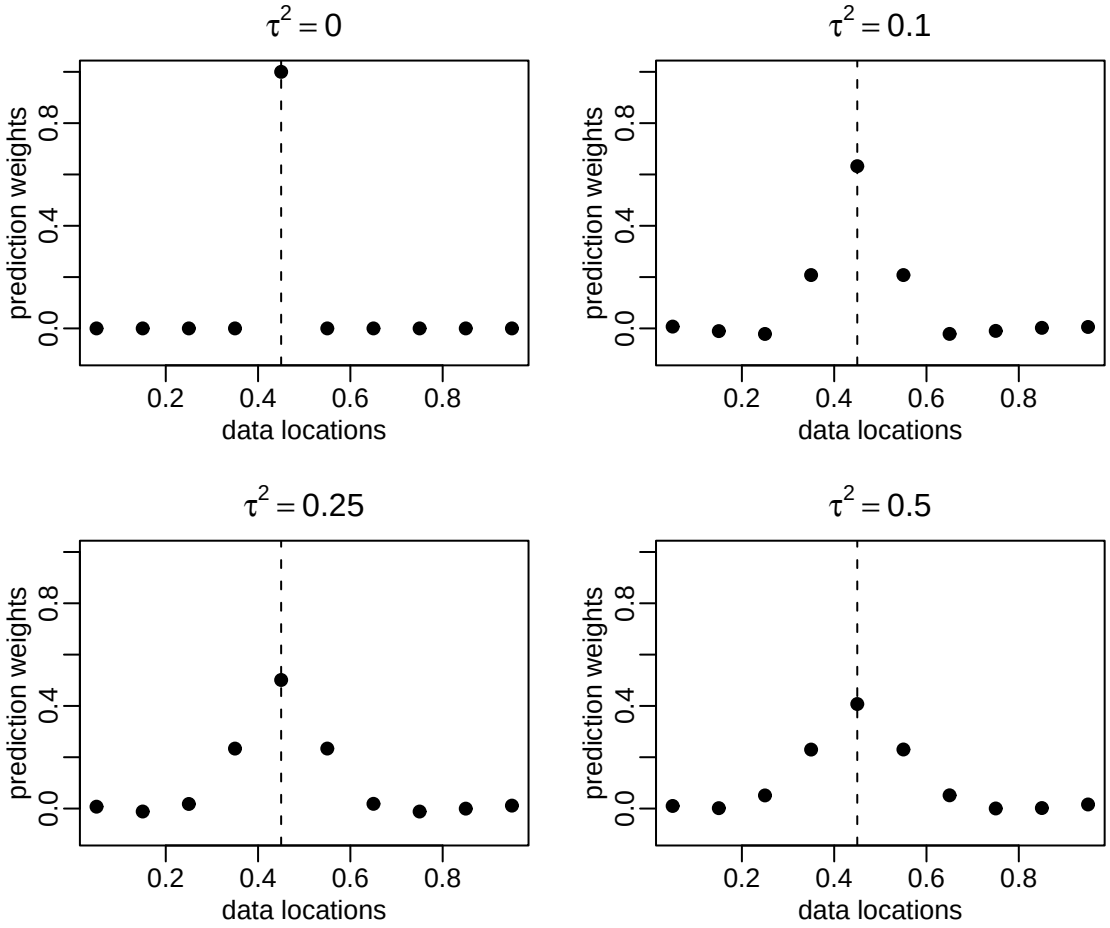
ii. locations: equally spaced $x_i = -0.05 + 0.1i$: $i = 1, ..., 10$

iii. prediction location: $x = 0.50$

iv. correlation function: Matérn with $\phi = 0.1$

v. Nugget: $\tau^2 = 0$

(c) The prediction weights: varying τ^2



Prediction weights for 10 equally spaced data-points with target location $x = 0.45$.

i. **varying parameter** $\tau^2 = 0, 0.1, 0.25, 0.5$

ii. locations: equally spaced $x_i = -0.05 + 0.1i$: $i = 1, ..., 10$

iii. prediction location: $x = 0.45$

iv. correlation function: Matérn with $\kappa = 2$ and $\phi = 0.1$

7. Prediction of Functionals

Let T be any *linear* functional of S,

$$T = \int_A w(x)S(x)dx$$

for some prescribed weighting function $w(x)$.

Under the Gaussian model:

- $[T, Y]$ is multivariate Gaussian;
- $[T|Y]$ is univariate Gaussian;
- the conditional mean and variance are:

$$\mathrm{E}[T|Y] = \int_A w(x)\mathrm{E}[S(x)|Y]dx$$

$$\mathrm{Var}[T|Y] = \int_A \int_A w(x)w(x')\mathrm{Cov}\{S(x), S(x')\}dxdx'$$

Note in particular that

$$\hat{T} = \int_A w(x)\hat{S}(x)dx$$

In words:

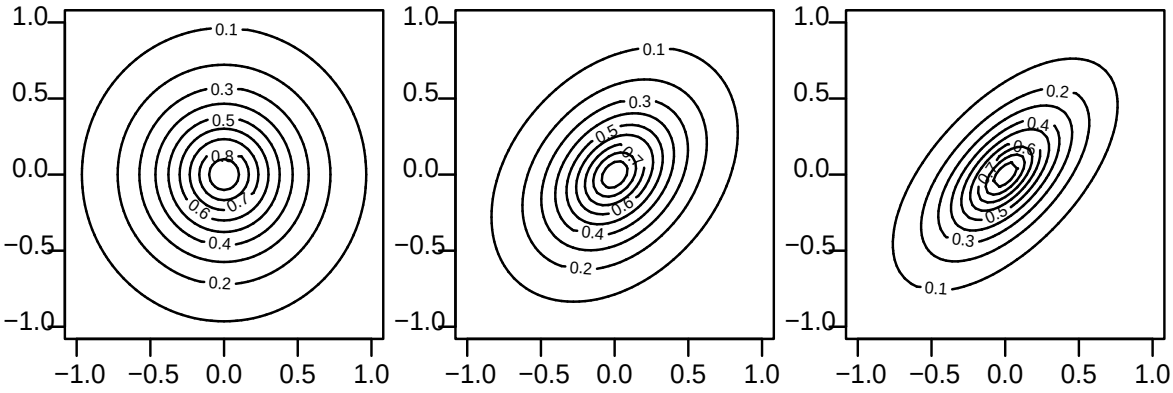
- given a predicted surface $\hat{S}(x)$, it is legitimate simply to calculate any linear property of this surface and to use the result as the predictor for the corresponding linear property of the true surface $S(x)$

- it is *NOT* legitimate to do this for prediction of non-linear properties

- for example, the maximum of $\hat{S}(x)$ is a very bad predictor for the maximum of $S(x)$ (this problem will be addressed later)

8. **Directional Effects**

- Environmental conditions can induce directional effects (wind, soil formation, etc).
- As a consequence the spatial correlation may vary with the direction.

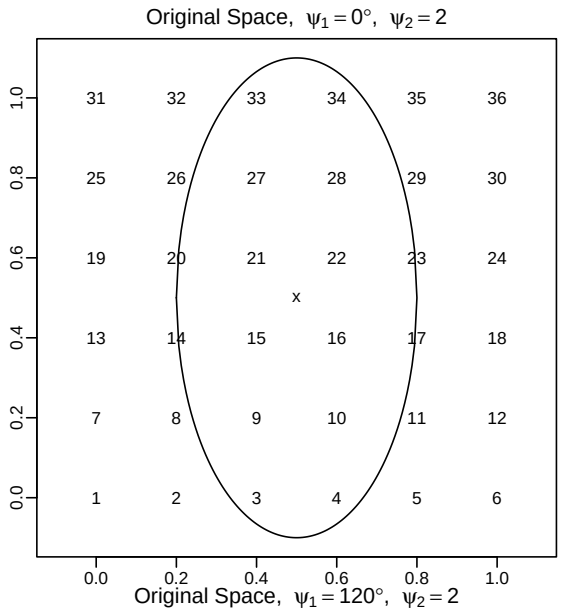


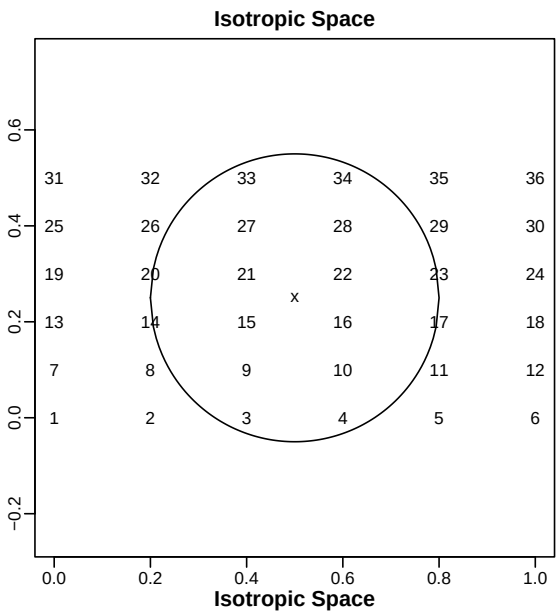
correlation contours for a isotropic model (left) and two anisotropic models (center and right).

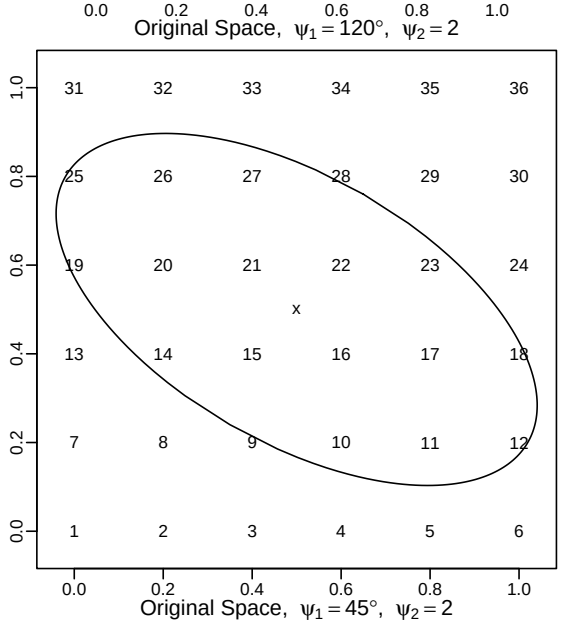
- a possible approach: *geometric anisotropy*.
- two more parameters: the *anisotropy angle* ψ_A and the *anisotropy ratio* ψ_R.
- rotation and stretching of the original coordinates:

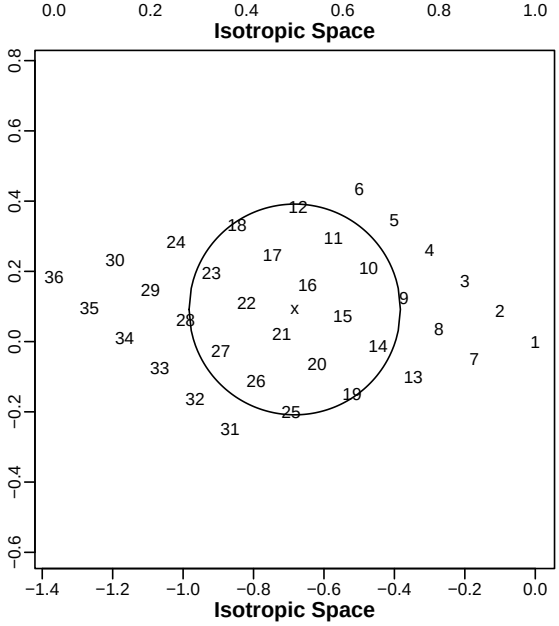
$$(x_1\prime, x_2\prime) = (x_1, x_2) \begin{pmatrix} \cos(\psi_A) & -\sin(\psi_A) \\ \sin(\psi_A) & \cos(\psi_A) \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\psi_R} \end{pmatrix}$$

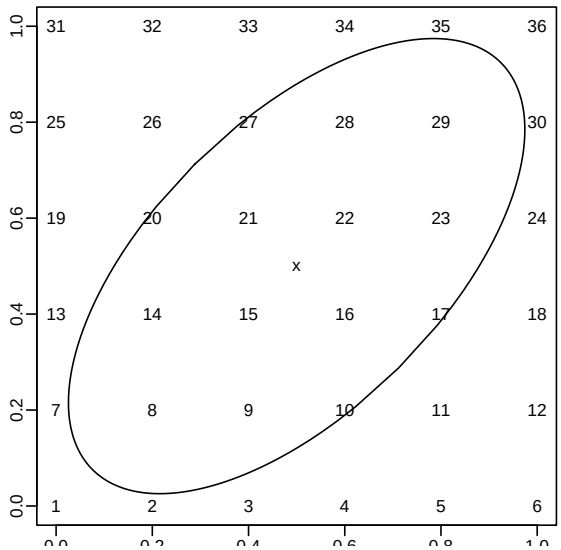
Geometric Anisotropy Correction

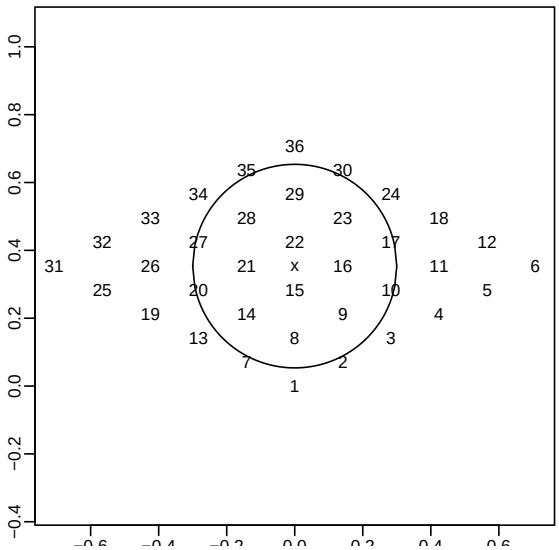












9. Non-Stationary Gaussian Models

Stationarity is a convenient working assumption, which can be relaxed in various ways.

- **Functional relationship between mean and variance?**
  Can sometimes be resolved by a transformation of the data.

- **Non-constant mean?**
  Replace constant μ by

$$\mu(x) = F\beta = \sum_{j=1}^{k} \beta_j f_j(x)$$

  for measured covariates $f_j(x)$ (or non-linear versions).

  Note: sometimes called **universal kriging** or **kriging with external trend**.

- **Non-stationary random variation?**

  Intrinsic variation a weaker hypothesis than stationarity (process has stationary increments, cf random walk model in time series), widely used as default model for discrete spatial variation (Besag, York and Molié, 1991).

  Spatial deformation methods (Sampson and Guttorp, 1992) seek to achieve stationarity by transformation of the geographical space, x.

- as always, need to balance increased flexibility of general modelling assumptions against over-modelling of sparse data, leading to poor identifiability of model parameters.

PART III:

PARAMETRIC ESTIMATION

1. **Second-Moment Properties**

2. **Variogram Analysis**

3. **Likelihood Inference**

4. **Plug-in Prediction**

5. **Gaussian Transformed Models**

6. **A Case Study**

7. **Anisotropic Models**

8. **Model Validation**

1. Second-Moment Properties

- the **variogram** of a process $Y(x)$ is the function

$$V(x, x') = \frac{1}{2}\mathrm{Var}\{Y(x) - Y(x')\}$$

- for the spatial Gaussian model, with $u = ||x - x'||$,

$$V(u) = \tau^2 + \sigma^2\{1 - \rho_0(u)\}$$

- basic structural parameters of the spatial Gaussian model are:
  - *the nugget variance*: τ^2
  - *the sill*: $\sigma^2 = \mathrm{Var}\{S(x)\}$
  - *the total sill*: $\tau^2 + \sigma^2 = \mathrm{Var}\{Y(x)\}$
  - *the range*: ϕ, such $\rho_0(u) = \rho(u/\phi)$
- practical implications:
  - any reasonable version of the (linear) spatial Gaussian model has at least three parameters
  - but you need a lot of data (or contextual knowledge) to justify estimating more than three parameters
- the **Matérn** family uses a fourth parameter to determine the differentiability of $S(x)$

Paradigmas for parameter estimation

- **Ad hoc (variogram based) methods**
  - compute the empirical variogram
  - fit a theoretical covariance model

- **Likelihood-based methods**
  - typically under Gaussian assumptions
  - Optimal under stated assumptions, but computationally expensive and may lack robustness?

- **Bayesian implementation**,
  combining estimation with prediction, becoming more widely accepted (amongst statisticians!)

2. Variogram Analysis

- The variogram is defined by

$$V(x, x') = \frac{1}{2}\mathrm{Var}\{Y(x) - Y(x')\}$$

- if $Y(x)$ is stationary,

$$V(x, x') = V(u) = \frac{1}{2}\mathrm{E}[\{Y(x) - Y(x')\}^2]$$

  where $u = ||x - x'||$

- suggests an empirical estimate of $V(u)$:

$$\hat{V}(u) = \mathrm{average}\{[y(x_i) - y(x_j)]^2\}$$

  where each average is taken over all pairs $[y(x_i), y(x_j)]$ such that $||x_i - x_j|| \approx u$

- for a process with non-constant mean (covariates), trend-removal can be used as follows:
  - define $r_i = Y_i - \hat{\mu}(x_i)$
  - define $\hat{V}(u) = \mathrm{average}\{(r_i - r_j)^2\}$, where each average is taken over all pairs (r_i, r_j)

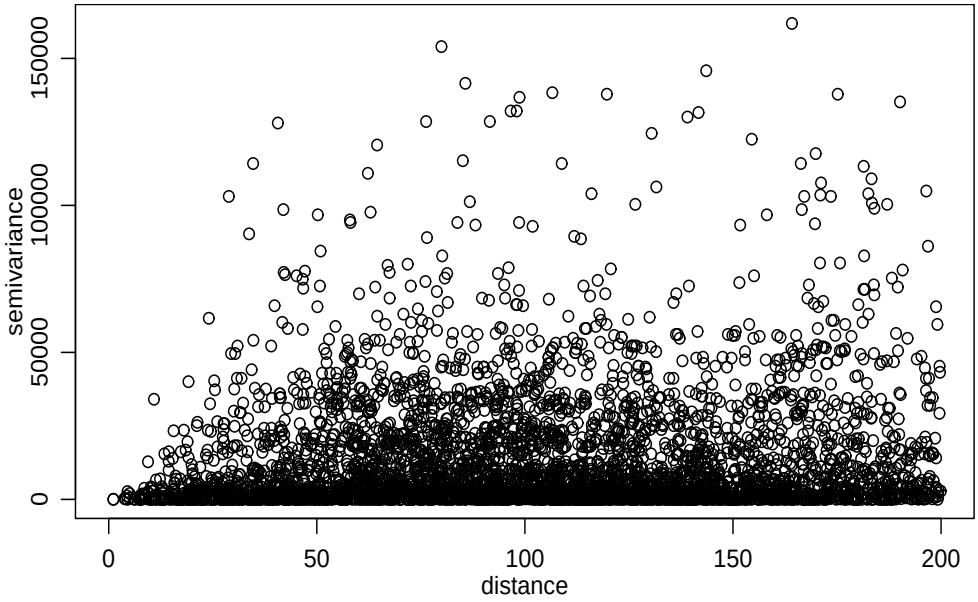
(a) The variogram cloud

- define the quantities

$$\begin{aligned} r_i &= Y_i - \hat{\mu}(x_i) \\ u_{ij} &= ||x_i - x_j|| \\ v_{ij} &= \frac{(r_i - r_j)^2}{2} \end{aligned}$$

- the **variogram cloud** is a scatterplot of the points (u_{ij}, v_{ij})

Example: Swiss rainfall data

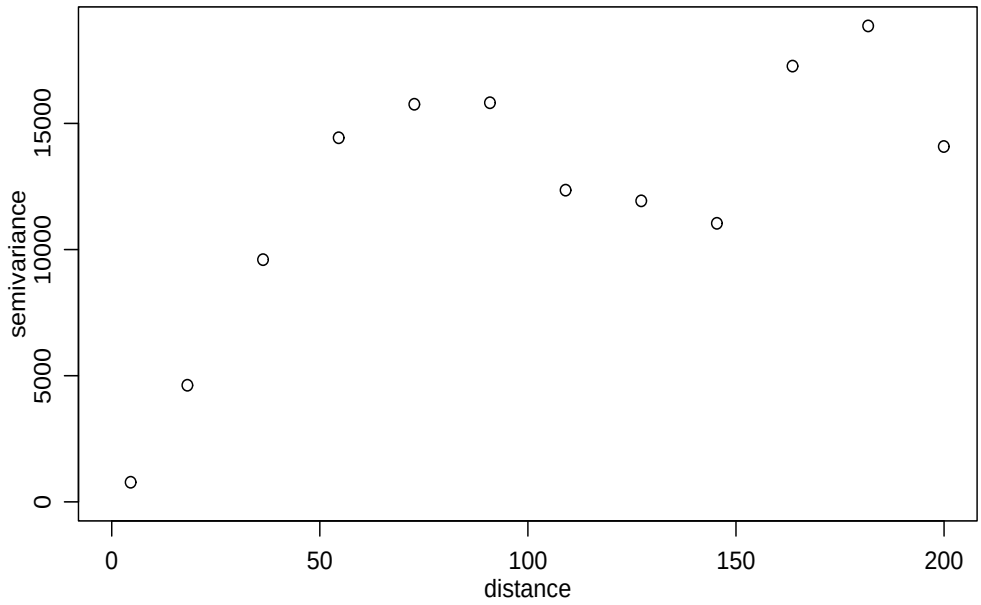


- under the spatial Gaussian model:
  - $v_{ij} \sim V(u_{ij})\chi^2_1$
  - the v_{ij} are correlated
- the variogram cloud is therefore unstable, both pointwise and in its overall shape

(b) **The empirical variogram**

- derived from the variogram cloud by **averaging within bins**: $u-h/2 \leq u_{ij} < u+h/2$
- forms k bins, each averaging over n_k pairs
- removes the first objection to the variogram cloud, but not the second
- is sensitive to mis-specification of $\mu(x)$

Example: Swiss rainfall data.



Empirical binned variogram

(c) The fitted variogram

Estimate $\tilde{\theta}$ to minimise a particular criterium

eg, the weighted least squares (Cressie, 1993)

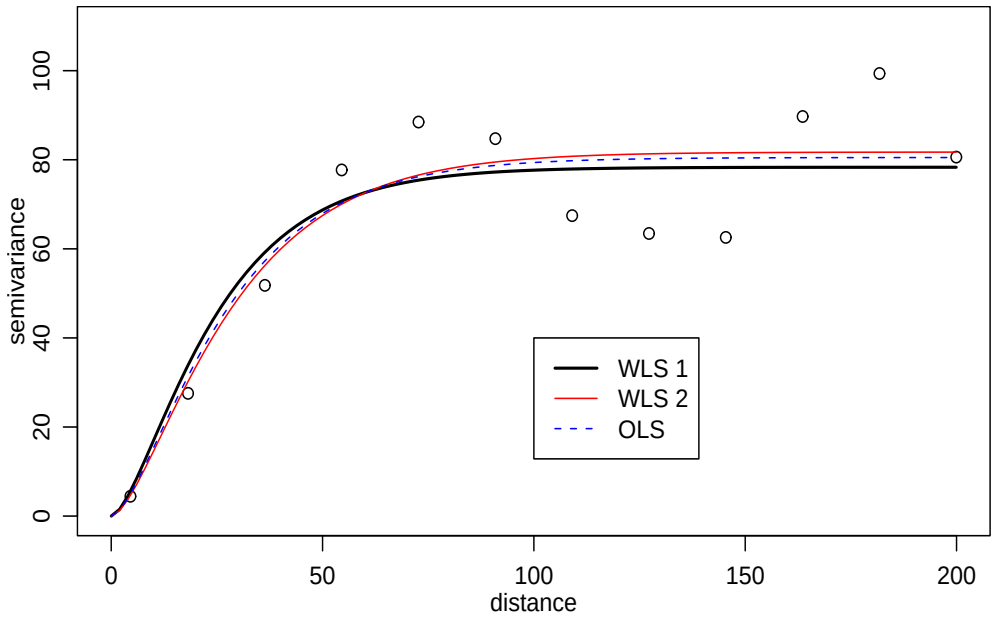
$$S(\theta) = \sum_k n_k \{[\bar{V}_k - V(u_k;\theta)]/V(u_{ij};\theta)\}^2$$

where $\bar{V}_k$ is average of n_k variogram ordinates v_{ij}.

Other criteria: OLS, WLS with weights given n_k only, ...

- Corresponds to biased estimating equation for θ, although still widely used in practice.
- Potentially misleading because of inherent correlations amongst successive $\bar{V}_k$

Example: Swiss rainfall data

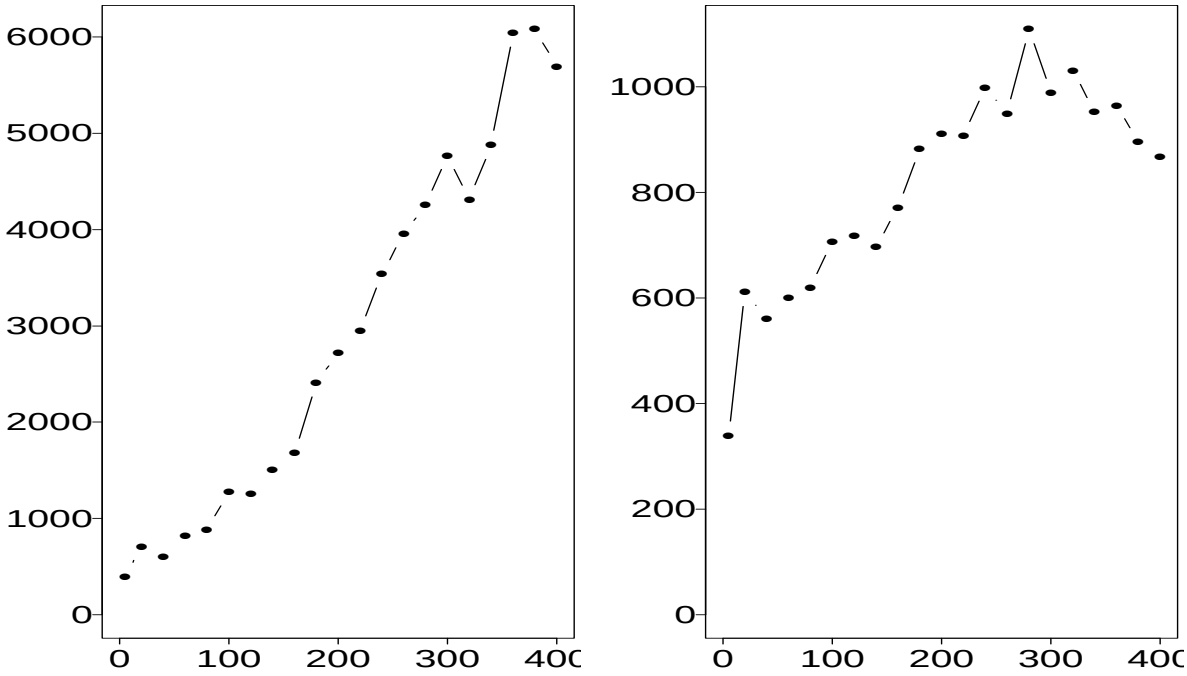


Empirical variogram with WLS with n_k only (thick line), Cressie's WLS (full line) and OLS (dashed line)

Further comments on empirical variograms

(a) Variograms of raw data and residuals can be very different

Example: Paraná rainfall data.

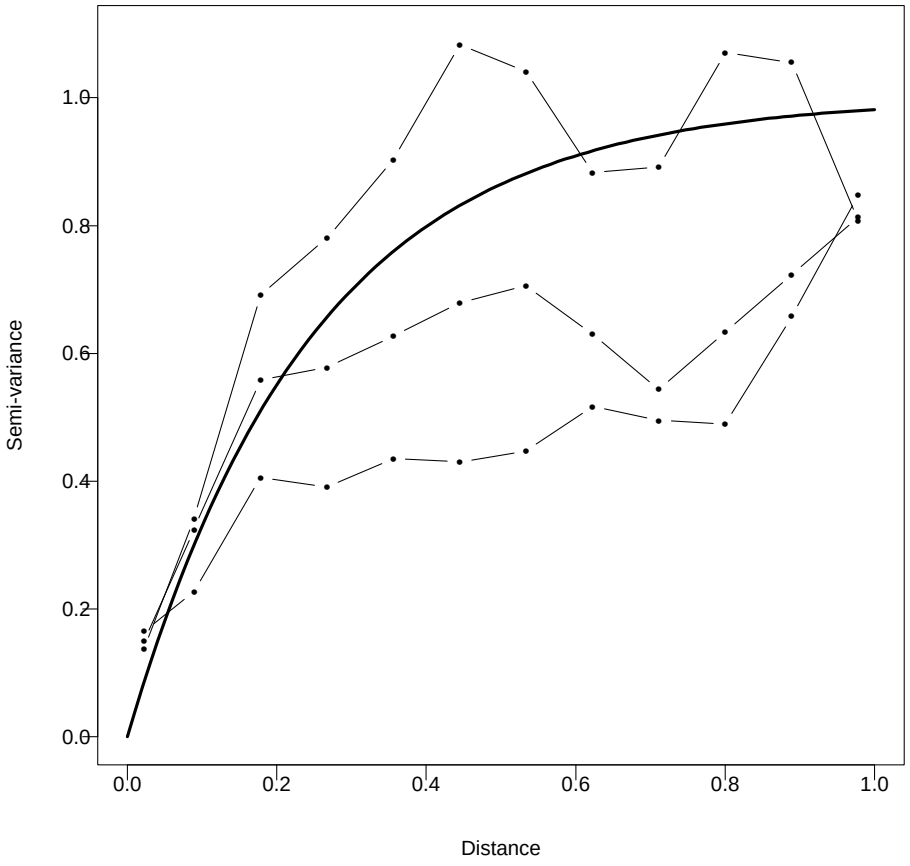


empirical variograms of raw data (left-hand panel) and of residuals after linear regression on latitude, longitude and altitude (right-hand panel).

- variogram of raw data includes variation due to
  large-scale geographical trend

- variogram of residuals eliminates this source
  of variation

(b) How unstable are empirical variograms?



- thick solid line shows true underlying variogram
- fine lines show empirical variograms from three
  independent simulations of the same model
- high autocorrelations amongst $\hat{V}(u)$ for successive
  values of u imparts misleading smoothness

3. Likelihood Inference

(a) **Maximum likelihood estimation (ML)** The Gaussian model is given by:

$$Y_i|S \sim \mathrm{N}(S(x_i), \tau^2)$$

- $S(x_i) = \mu(x_i) + S_c(x_i)$
- $S_c(\cdot)$ is a zero mean stationary Gaussian process with covariance parameters (σ^2, ϕ, κ),
- $\mu(x_i) = F\beta = \sum_{j=1}^{k} f_k(x_i)\beta_k$, where $f_k(x_i)$ is a vector of covariates at location x_i

Which allows us to write:

$$Y_i = \mu(x_i) + S_c(x_i) + \epsilon_i : i = 1, ..., n$$

Then

$$Y \sim \mathrm{MVN}(F\beta, \sigma^2 R + \tau^2 I)$$

and the likelihood function is

$$\begin{aligned} L(\beta, \tau, \sigma, \phi, \kappa) \propto \quad & -0.5\{\log|(\sigma^2 R + \tau^2 I)| + \\ & (y - F\beta)'(\sigma^2 R + \tau^2 I)^{-1}(y - F\beta)\}. \end{aligned}$$

which maximisation yields the maximum likelihood estimates of the model parameters.

Maximum likelihood estimation (cont.)

Computational details:

- reparametrise $\nu^2 = \frac{\tau^2}{\sigma^2}$ and denote: $\sigma^2 V = \sigma^2(R + \nu^2 I)$
- the log-likelihood function is maximised for

$$\hat{\beta}(V) = (F'V^{-1}F)^{-1}F'V^{-1}y$$

$$\hat{\sigma}^2 = n^{-1}(y - F\hat{\beta})'V^{-1}(y - F\hat{\beta})$$

- then substituting (β, σ^2) by $(\hat{\beta}, \hat{\sigma^2})$ in 3 the maximisation reduces to

$$L(\tau_r, \phi, \kappa) \propto -0.5\{n \log|\hat{\sigma^2}| + \log|(R + \nu^2 I)|\}$$

- For the Matérn correlation function we suggest to take κ in a discrete set $\{0.5, 1, 2, 3, ..., N\}$

(b) Restricted Maximum Likelihood Estimation (REML)

The REML principle is as follows:

- under the assumed model for $\mathrm{E}[Y] = F\beta$, transform the data linearly to $Y^* = AY$ such that the distribution of Y^* does not depend on β;
- estimate $\theta = (\nu^2, \sigma^2, \phi, \kappa)$ by maximum likelihood applied to the transformed data Y^*

We can always find a suitable matrix A without knowing the true values of β or θ, for example

$$A = I - F(F'F)^{-1}F'$$

The REML estimators for θ can be computed by maximising

$$\begin{aligned} L^*(\theta) \;\propto\; -\frac{1}{2}\Big\{&\log|\sigma^2 V| + \log|F'\{\sigma^2 V\}^{-1}F| \\ &+(y - F\tilde{\beta})'\{\sigma^2 V\}^{-1}(y - F\tilde{\beta}))\Big\}, \end{aligned}$$

where $\tilde{\beta} = \hat{\beta}(\theta)$.

Comments on REML

- introduced in the context of variance components estimation in designed experiments (Patterson and Thompson, 1971)
- leads to less biased estimators in small samples
- $L^*(\theta)$ depends on F, and therefore on a correct specification of the model for $\mu(x)$,
- $L^*(\theta)$ does not depend on the choice of A.
- Given the model for $\mu(\cdot)$, the method enjoys the same objectivity as does maximum likelihood estimation.
- widely recommended for geostatistical models.
- REML is more sensitive than ML to misspecification of the model for $\mu(x)$.
- for designed experiments the mean $\mu(x)$ is usually well specified
- however in the spatial setting there is no sharp distinction between $\mu(x)$ and $S_c(x)$.

Profile likelihood

Variability of the parameter estimators can be inspected by looking at the log-likelihood surface defined by (3).

Surface reflects the information contained in the data about the model parameters.

Dimension of the log-likelihood surface does not allow direct inspection.

Computing profile likelihoods

- Suppose a model with parameters (α, ψ)
- denote its likelihood by $L(\alpha, \psi)$
- To inspect the likelihood for α replace the nuisance parameters ψ by their ML estimators $\hat{\psi}(\alpha)$, for each value of α
- This gives the profile likelihood:

$$L_p(\alpha) = L(\alpha, \hat{\psi}(\alpha)) = \max_{\psi}(L(\alpha, \psi)).$$

4. Plug-In Prediction – Kriging

Quite often the interest is to predict

- the realised value of the process $S(\cdot)$ at a point,
- or the average of $S(\cdot)$ over a region,

$$T = |B|^{-1} \int_B S(x)dx$$

where $|B|$ denotes the area of the region B.

For the Gaussian model we've seen that the minimum MSPE predictor for $T = S(x)$ is

$$\hat{T} = \mu + \sigma^2 \mathbf{r}'(\tau^2 I + \sigma^2 R)^{-1}(Y - \mu \mathbf{1})$$

with prediction variance

$$\mathrm{Var}(T|Y) = \sigma^2 - \sigma^2 \mathbf{r}'(\tau^2 I + \sigma^2 R)^{-1}\sigma^2 \mathbf{r}$$

where the only unknowns are the model parameters.

The **plug-in prediction** consists of replacing the true parameters by their estimates.

Comments

- ML estimates and simple kriging
- REML estimates and ordinary kriging
- The *ad-hoc* prediction
  (a) Estimate β by OLS, $\tilde{\beta} = (F'F)^{-1}F'Y$, and construct residuals $Z = Y - F\tilde{\beta}$.
  (b) Calculate the empirical variogram of Z and use it for model formulation and parameter estimation.
  (c) Re-estimate β by GLS and use the fitted model for prediction.
- the role of empirical variograms:
  – diagnostics (model-based approach)
  – inferential tool (ad-hoc approach)
- The conventional approach to kriging (M-B or ad-hoc) is to plug-in estimated parameter values and proceed *as if the estimates were the truth*.
  This approach:
  – usually gives good point predictions when predicting $T = S(x)$
  – but often under-estimates prediction variance
  – and can produce poor results when predicting other targets T

5. Gaussian Transformed Models

The Gaussian model might be clearly inappropriate for variables with asymmetric distributions.

Some flexibility with an extra parameter λ defining a Box-Cox transformation.

Terminology: *Gaussian-transformed model*.

The model is defined as follows:

- assume a variable $Y^* \sim MVN(F\beta, \sigma^2 V)$
- the data, denoted $y = (y_1, ..., y_n)$, are generated by a transformation of the linear Gaussian model $Y = h_\lambda^{-1}(Y^*)$ such that:

$$Y_i^* = h_\lambda(Y) = \begin{cases} \frac{(y_i)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y_i) & \text{if } \lambda = 0 \end{cases}$$

The log-likelihood is:

$$\begin{aligned} \ell(\beta, \theta, \lambda) = & -\frac{1}{2}\{\log|\sigma^2 V| \\ & \quad +(h_\lambda(y) - F\beta)'\{\sigma^2 V\}^{-1}(h_\lambda(y) - F\beta)\} \\ & + \sum_{i=1}^{n} \log\left((y_i)^{\lambda-1}\right) \end{aligned}$$

Notes:

- $\lambda = 0$ corresponds to a particular case: log-Gaussian model

- Inference strategies:

(a) λ as a random parameter (De Oliveira, Kedem and Short, 1997). Prediction averages over a range of models.

(b) An alternative approach is to estimate λ and then fix it equal to the estimate when performing prediction (Christensen, Diggle and Ribeiro, 2000).

i. find the *best* transformation maximising the profile likelihood for λ

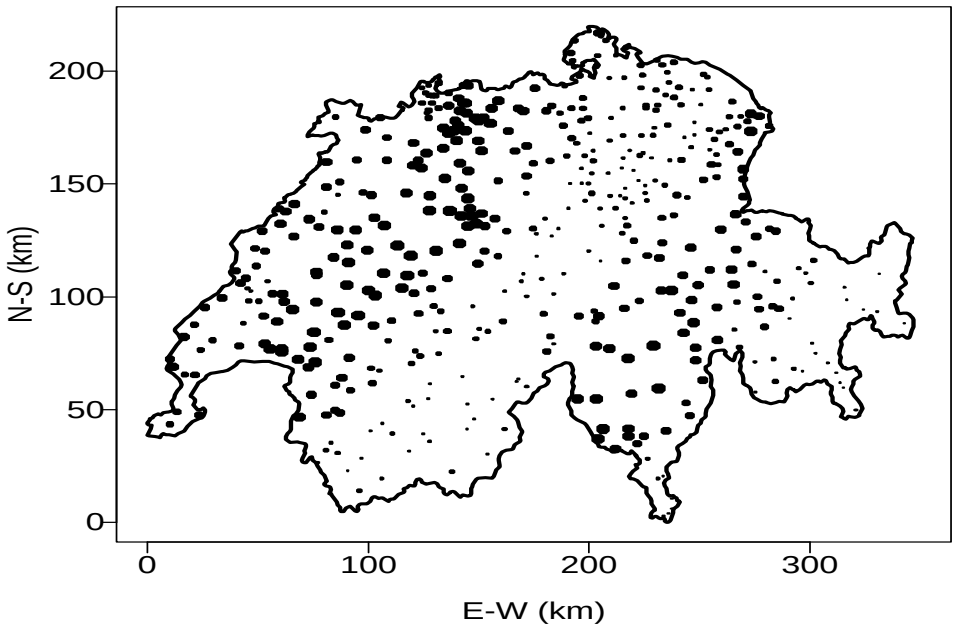
ii. fix the transformation, transform data

iii. inferences on transformed scale

iv. back-transform results

Difficulties with negative values and back-transformation.

6. **A Case Study: Swiss rainfall data**



Locations of the data points with points size proportional to the value of the observed data. Distances are in kilometres.

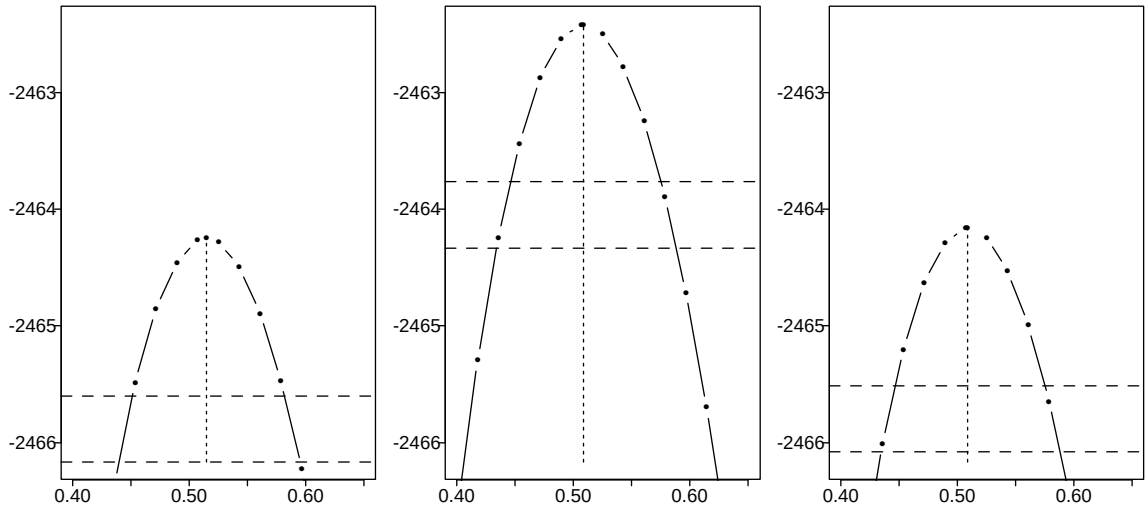
- 467 locations in Switzerland
- daily rainfall measurements on 8th of May 1986
- The data values are integers where the unit of measurement is $1/10$ mm
- 5 locations where the value is equal to zero.

Swiss rainfall data (cont.)

Estimating the transformation parameter and using the Matérn model for the correlation function.

κ	$\hat{\lambda}$	$\log \hat{L}$
0.5	0.514	-2464.246
1	0.508	-2462.413
2	0.508	-2464.160

Maximum likelihood estimate $\hat{\lambda}$ and the corresponding value of the log-likelihood function $\log \hat{L}$ for different values of the Matérn parameter κ.



Profile likelihoods for λ. Left: $\kappa = 0.5$, middle: $\kappa = 1$, right: $\kappa = 2$. The two lines correspond to 90% and 95% percent quantiles for a $\frac{1}{2}\chi^2(1)$-distribution.

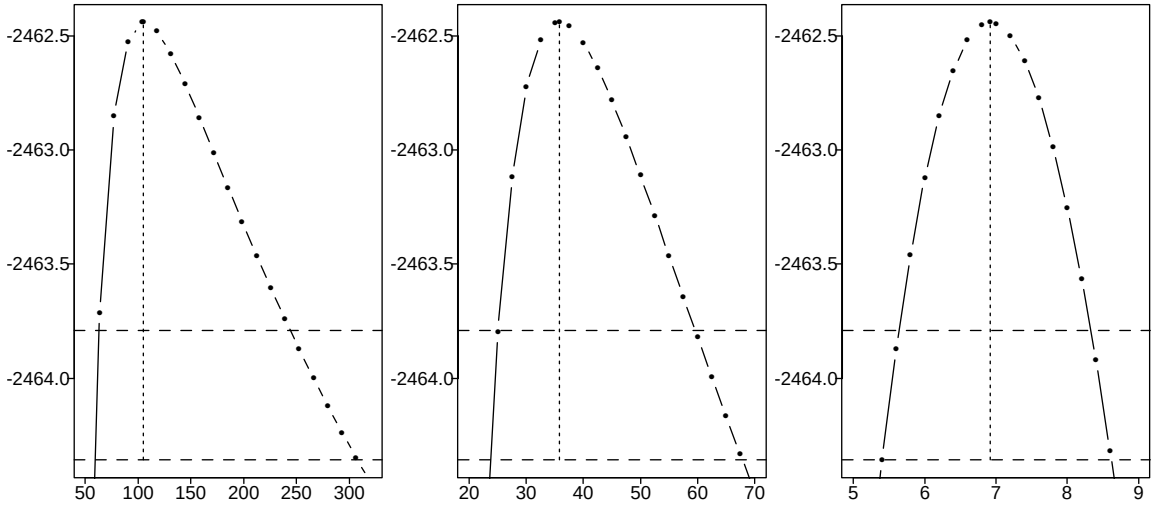
Neither untransformed nor log-transformed are indicated.

Swiss rainfall data (cont.)

Estimates for the model with $\lambda = 0.5$

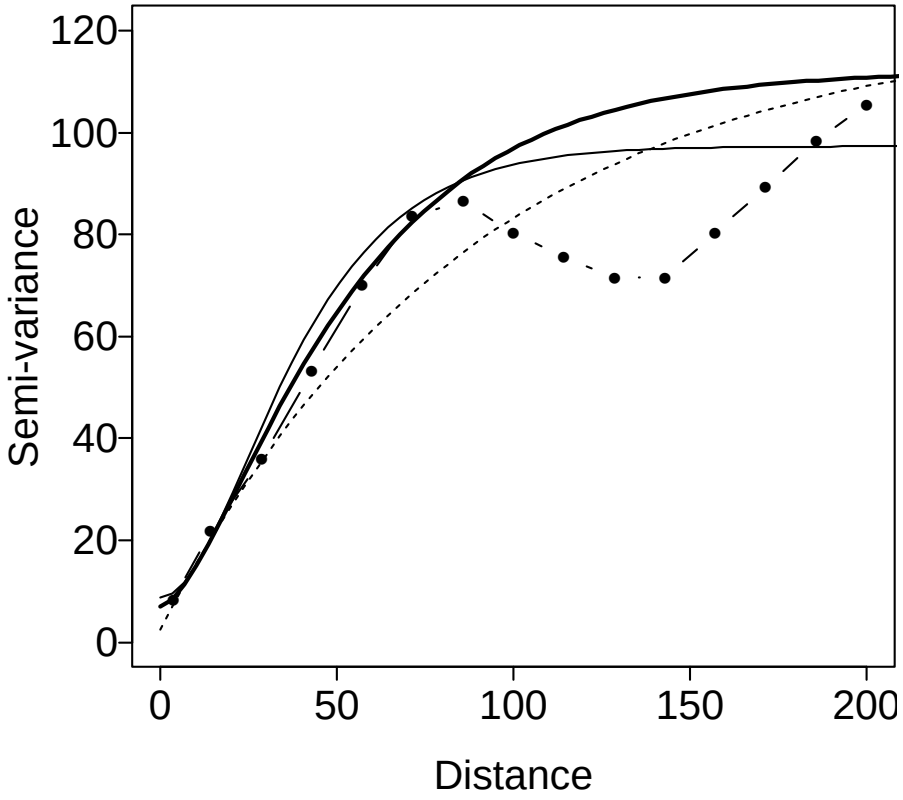
κ	$\hat{\beta}$	$\hat{\sigma}^2$	$\hat{\phi}$	$\hat{\tau}^2$	$\log \hat{L}$
0.5	18.36	118.82	87.97	2.48	-2464.315
1	20.13	105.06	35.79	6.92	-2462.438
2	21.36	88.58	17.73	8.72	-2464.185

Maximum likelihood estimates $\hat{\beta}$, $\hat{\phi}$, $\hat{\sigma}$, $\hat{\tau}$ and the corresponding value of the likelihood function $\log \hat{L}$ for different values of the Matérn parameter κ, for $\lambda = 0.5$



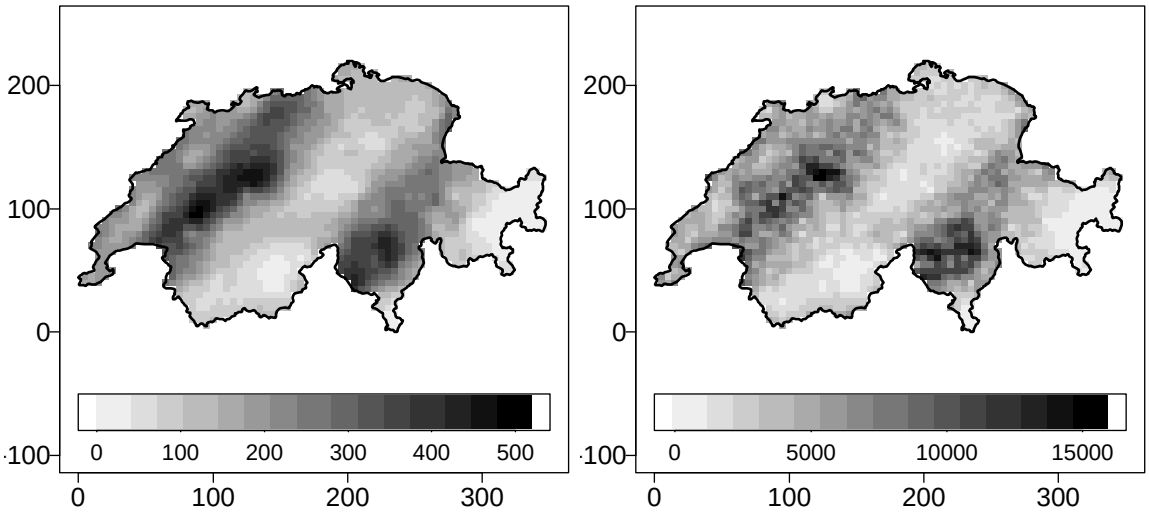
Profile likelihood for covariance parameters with $\kappa = 1$ and $\lambda = 0.5$. Left: σ^2, middle: ϕ, right: τ^2.

Swiss rainfall data (cont.)



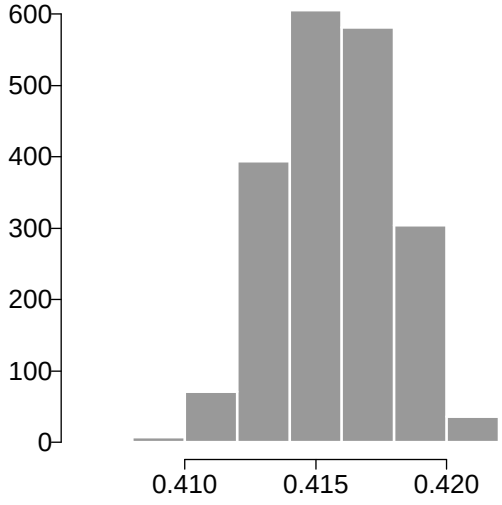
Estimated semivariances for square-root transformed data (dot-dashed line), compared with the theoretical semivariogram model, with parameters equal to the maximum likelihood estimates. $\kappa = 0.5$ (dashed line), $\kappa = 1$ (thick solid line), $\kappa = 2$ (thin solid line).

Swiss rainfall data (cont.)



Maps with predictions (left) and prediction variances (right).

Prediction of the percentage of the area where $Y(x) \geq 200$: $\tilde{A}_{200}$ is 0.4157



Samples from the predictive distribution of $\tilde{A}_{200}$.

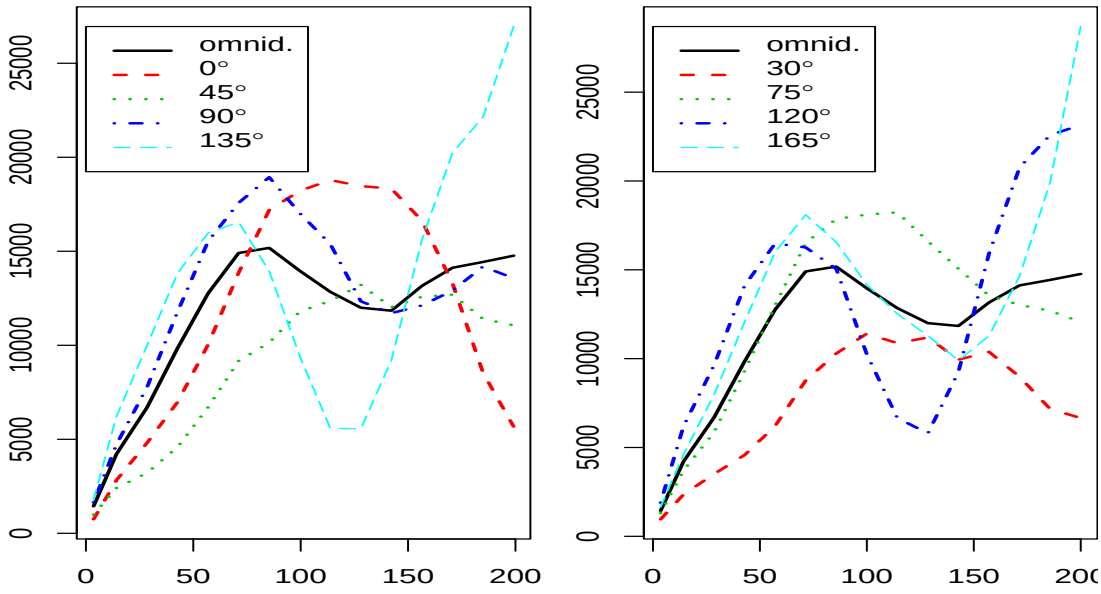
7. Estimating Anisotropic Models

Likelihood based methods

- Two more parameters for the correlation function
- Increases the dimension of the numerical minimisation problem
- In practice a lot of data might be needed

Variogram based methods

- Compute variograms for different directions
- Tolerance angles, in particular for irregularly distributed data
- Fit variogram model using directional variograms



Directional variograms for the Swiss rainfall data.

8. Model Validation and Comparison:

Using validation data

Data divided in two groups: data for model fitting and data for validation

Frequently in practice data are scarce and to expensive to be left out

"Leaving-one-out"

- One by one, for each datum:

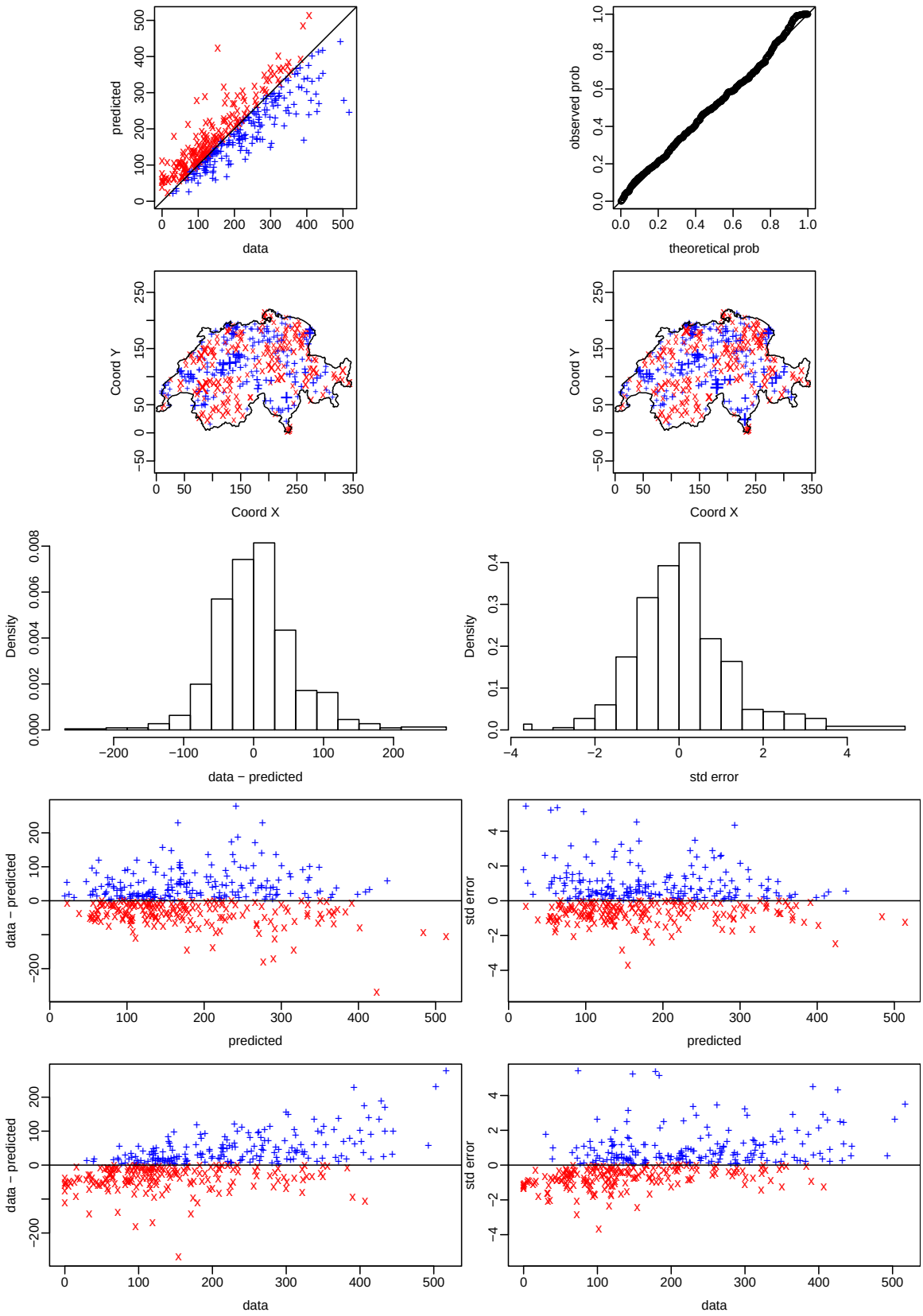
  (a) remove the datum from the data-set

  (b) (re-estimate model parameters)

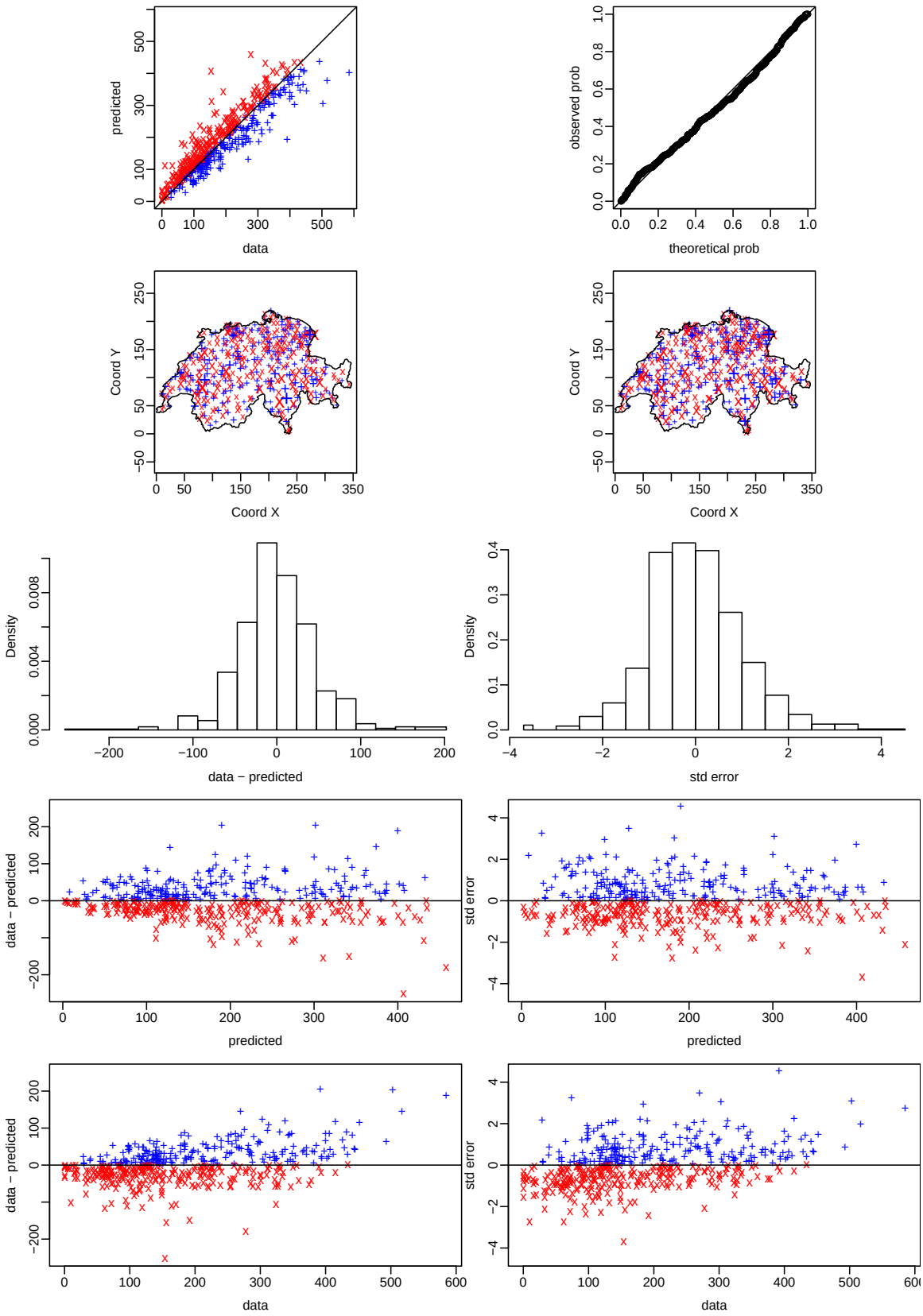
  (c) predict at the datum location

- Compare original data with predicted values.

100 fitting data + 367 validation data



all data – “leaving-one-out”



PART IV:

BAYESIAN INFERENCE FOR THE GAUSSIAN MODEL

1. **Basic Concepts**

2. **Bayesian Analysis of the Gaussian Model**

3. **A Case Study: Swiss Rainfall data**

1. Bayesian Analysis - Basic Concepts

Bayesian inference deals with parameter uncertainty by treating parameters as random variables, and expressing inferences about parameters in terms of their conditional distributions, given all observed data.

For inference about model parameters, the full model specification now should include the model parameters:

$$[Y, \theta] = [\theta][Y|\theta]$$

Bayes' Theorem allows us to calculate:

$$[Y, \theta] = [Y|\theta][\theta] = [Y][\theta|Y]$$

Thus,

$$[\theta|Y] = [Y|\theta][\theta]/[Y]$$

is the *posterior distribution* where

$$[Y] = \int [Y|\theta][\theta]d\theta.$$

The Bayesian paradigm:

(a) **Model**

- the full model specification consists of $[Y, \theta] = [Y|\theta][\theta]$.
- formulate a model for the observable variable Y.
- this model defines $[Y|\theta]$ (and hence an expression for the log-likelihood $\ell(\theta; Y)$)

(b) **Prior**

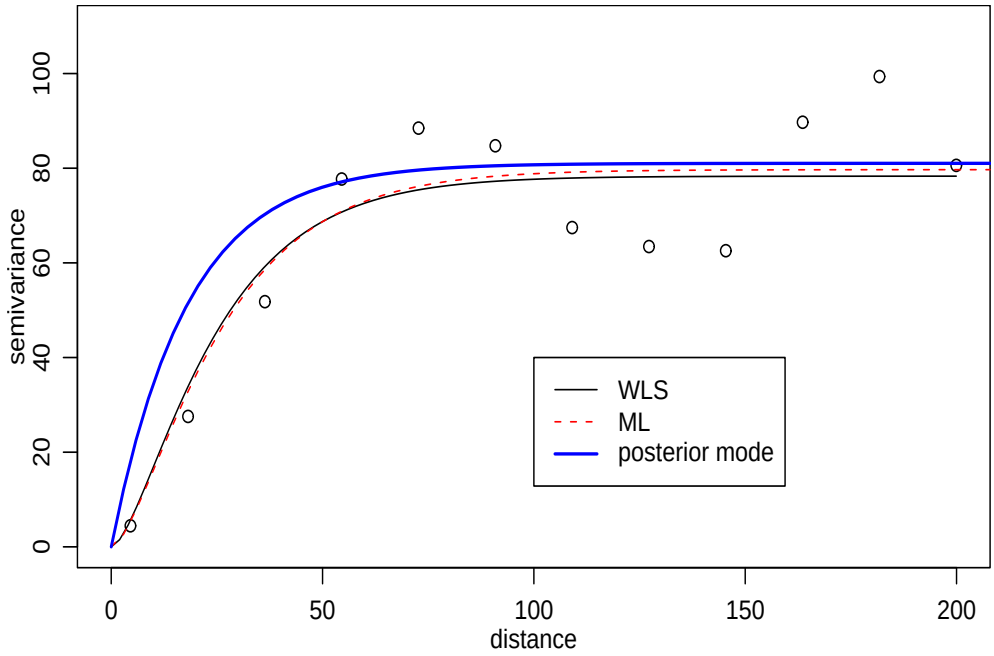
- before we observe Y, the marginal $[\theta]$ expresses our uncertainty about θ
- call $[\theta]$ *prior distribution* for θ

(c) **Posterior**

- having observed Y, it is no longer an unknown (randomly varying) quantity
- therefore revise uncertainty about θ by conditioning on the observed value of Y
- call $[\theta|Y]$ *posterior distribution* for θ, and use it to make inferential statements

NOTE: the likelihood function occupies a central role in both classical and Bayesian inference

Example: Swiss rainfall data



Fitted variograms (100 data), using three different methods of estimation: (a) curve-fitting (thin line); (b) ML (dashed line); (c) posterior mode (thick line).

Prediction

Because Bayesian inference treats θ as a random variable, it makes no formal distinction between parameter estimation problems and prediction problems, and thereby provides a natural means of allowing for parameter uncertainty in predictive inference.

The general idea for prediction is to formulate a model for

$$[Y, T, \theta] = [Y, T|\theta][\theta]$$

and make inferences based on the conditional distribution

$$\begin{aligned}[T|Y] &= \int [T, \theta|Y] d\theta \\ &= \int [\theta|Y][T|Y, \theta] d\theta\end{aligned}$$

Comparing plug-in and Bayesian

- the plug-in prediction corresponds to inferences about $[T|Y, \hat{\theta}]$
- Bayesian prediction is a weighted average of plug-in predictions, with different plug-in values of θ weighted according to their conditional probabilities given the observed data.

Bayesian prediction is usually more cautious than plug-in prediction, or in other words:

- allowance for parameter uncertainty usually results in wider prediction intervals

Notes:

(a) Until recently, the need to evaluate the integral which defines $[Y]$ represented a major obstacle to practical application.

(b) Development of Markov Chain Monte Carlo (MCMC) methods has transformed the situation.

(c) BUT, for geostatistical problems, reliable implementation of MCMC is not straightforward. Geostatistical models don't have a natural Markovian structure for the algorithms work well.

(d) in particular for the Gaussian model other algorithms can be implemented.

2. Results for the Gaussian Model

Uncertainty only in the mean parameter

Assume for now that only the mean parameter β is regarded as random with (conjugate) prior:

$$\beta \sim \mathrm{N}\left(m_\beta \,;\, \sigma^2 V_\beta\right)$$

The posterior is given by

$$\begin{aligned}
[\beta|Y] \;&\sim\; \mathrm{N}((V_\beta^{-1} + F'R^{-1}F)^{-1}(V_\beta^{-1}m_\beta + F'R^{-1}y) \,; \\
&\qquad \sigma^2\,(V_\beta^{-1} + F'R^{-1}F)^{-1}) \\
&\sim\; \mathrm{N}\left(\hat{\beta} \,;\, \sigma^2\, V_{\hat{\beta}}\right)
\end{aligned}$$

The predictive distribution is

$$p(S^*|Y, \sigma^2, \phi) \;=\; \int p(S^*|Y, \beta, \sigma^2, \phi)\, p(\beta|Y, \sigma^2, \phi)\, d\beta.$$

with mean and variance given by

$$\begin{aligned}
E[S^*|Y] \;&=\; (F_0 - r'V^{-1}F)(V_\beta^{-1} + F'V^{-1}F)^{-1}V_\beta^{-1}m_\beta + \\
&\qquad \left[r'V^{-1} + (F_0 - r'V^{-1}F)(V_\beta^{-1} + F'V^{-1}F)^{-1}F'V^{-1}\right] Y \\
\mathrm{Var}[S^*|Y] \;&=\; \sigma^2 \left[V_0 - r'V^{-1}r + \right. \\
&\qquad \left. (F_0 - r'V^{-1}F)(V_\beta^{-1} + F'V^{-1}F)^{-1}(F_0 - r'V^{-1}F)'\right].
\end{aligned}$$

The predictive variance has three interpretable components: a priori variance, the reduction due to the data and the uncertainty in the mean.

$V_\beta \to \infty$ corresponds to universal (or ordinary) kriging.

Uncertainty for all model parameters

Assume (w.l.g.) a model without measurement error and the prior $p(\beta, \sigma^2, \phi) \propto \frac{1}{\sigma^2} p(\phi)$.

The posterior distribution:

$$p(\beta, \sigma^2, \phi|y) \;=\; p(\beta, \sigma^2|y, \phi)\; p(\phi|y)$$

$$pr(\phi|y) \;\propto\; pr(\phi)\; |V_{\hat{\beta}}|^{\frac{1}{2}}\; |R_y|^{-\frac{1}{2}}\; (S^2)^{-\frac{n-p}{2}}.$$

Algorithm 1:

(a) Discretise the distribution $[\phi|y]$, i.e. choose a range of values for ϕ which is sensible for the particular application, and assign a discrete uniform prior for ϕ on a set of values spanning the chosen range.

(b) Compute the posterior probabilities on this discrete support set, defining a discrete posterior distribution with probability mass function $\tilde{pr}(\phi|y)$, say.

(c) Sample a value of ϕ from the discrete distribution $\tilde{pr}(\phi|y)$.

(d) Attach the sampled value of ϕ to the distribution $[\beta, \sigma^2|y, \phi]$ and sample from this distribution.

(e) Repeat steps (3) and (4) as many times as required; the resulting sample of triplets (β, σ^2, ϕ) is a sample from the joint posterior distribution.

The predictive distribution is given by:

$$\begin{aligned} p(S^*|Y) &= \int\int\int p(S^*, \beta, \sigma^2, \phi|Y)\, d\beta\, d\sigma^2\, d\phi \\ &= \int\int\int p\big(s^*, \beta, \sigma^2|y, \phi\big)\ d\beta\, d\sigma^2\, pr(\phi|y)\ d\phi \\ &= \int p(S^*|Y, \phi)\ p(\phi|y)\, d\phi. \end{aligned}$$

To sample from this distribution:

Algorithm 2:

(a) Discretise $[\phi|Y]$, as in Algorithm 1.

(b) Compute the posterior probabilities on the discrete support set. Denote the resulting distribution $\tilde{pr}(\phi|y)$.

(c) Sample a value of ϕ from $\tilde{pr}(\phi|y)$.

(d) Attach the sampled value of ϕ to $[s^*|y, \phi]$ and sample from it obtaining realisations of the predictive distribution.

(e) Repeat steps (3) and (4) as many times as required to generate a sample from the required predictive distribution.

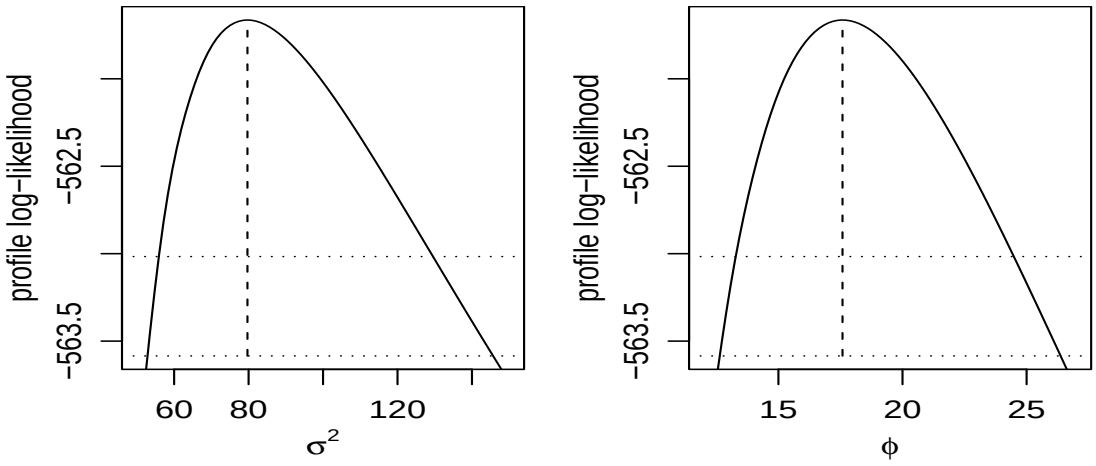
Note:

(a) The algorithms are of the same kind to treat τ and/or κ as unknown parameters.

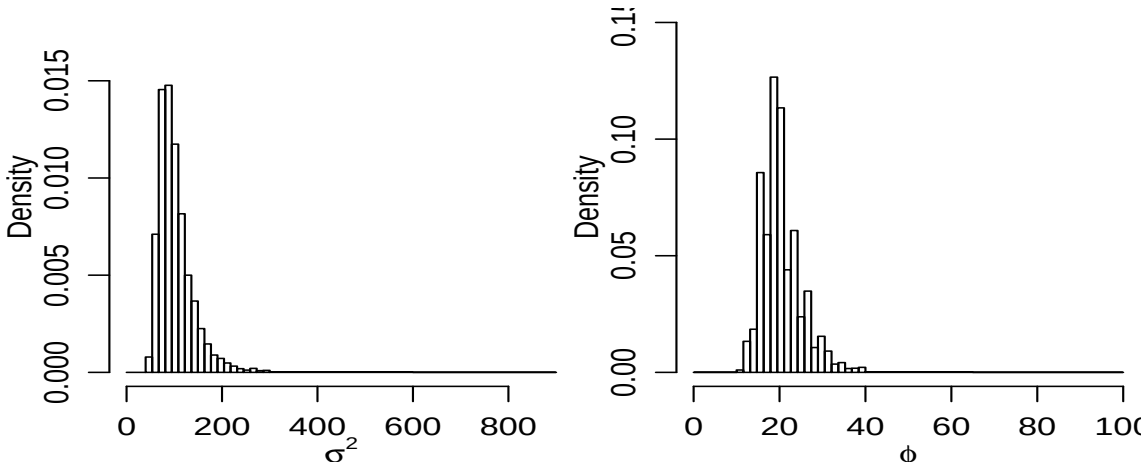
(b) We specify a discrete prior distribution on a multi-dimensional grid of values.

(c) This implies extra computational load (but no new principles)

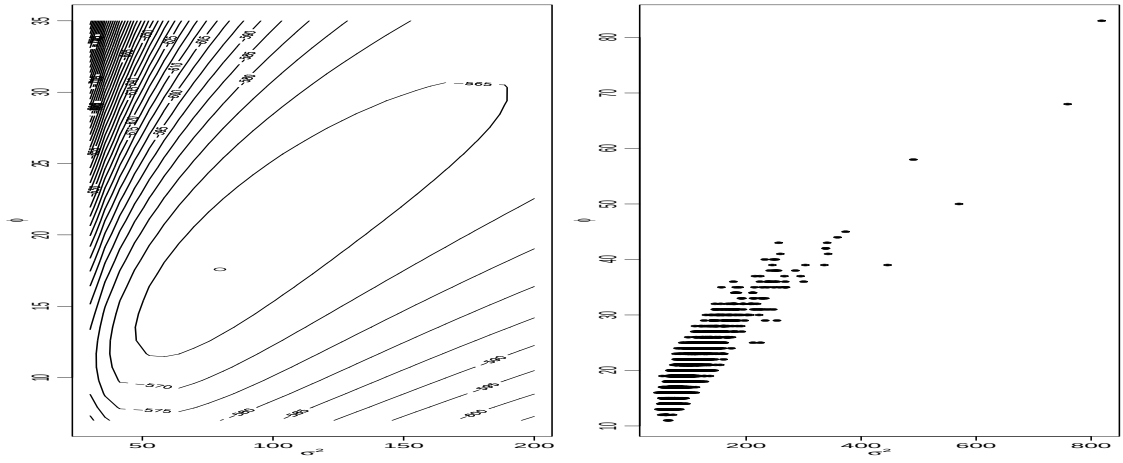
3. A Case Study: Swiss rainfall, 100 data



Profile likelihoods for covariance parameters: σ^2; ϕ

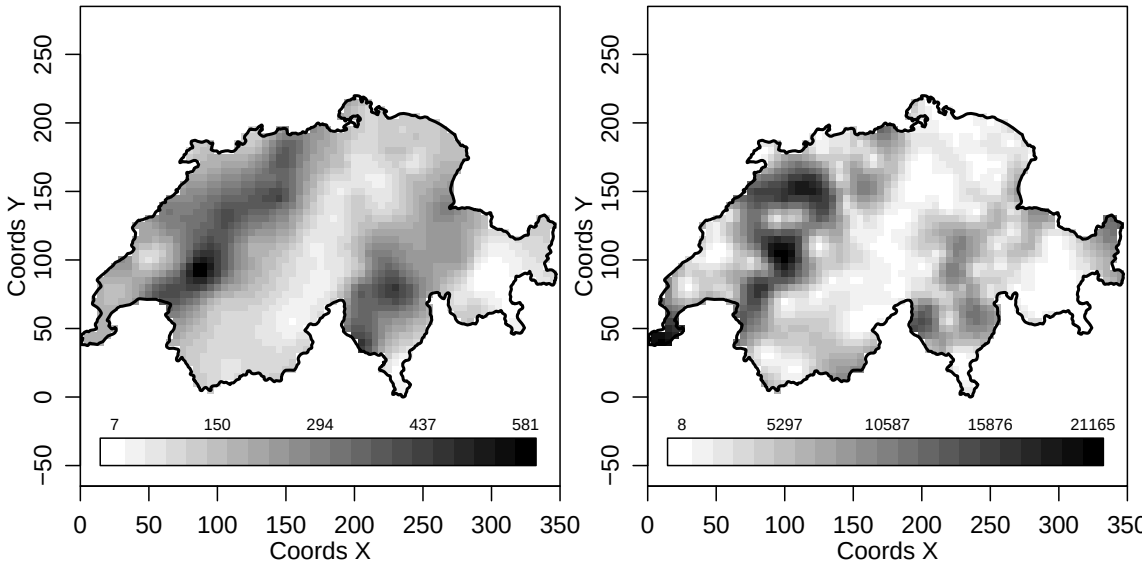


Posterior distributions for covariance parameters: σ^2; ϕ

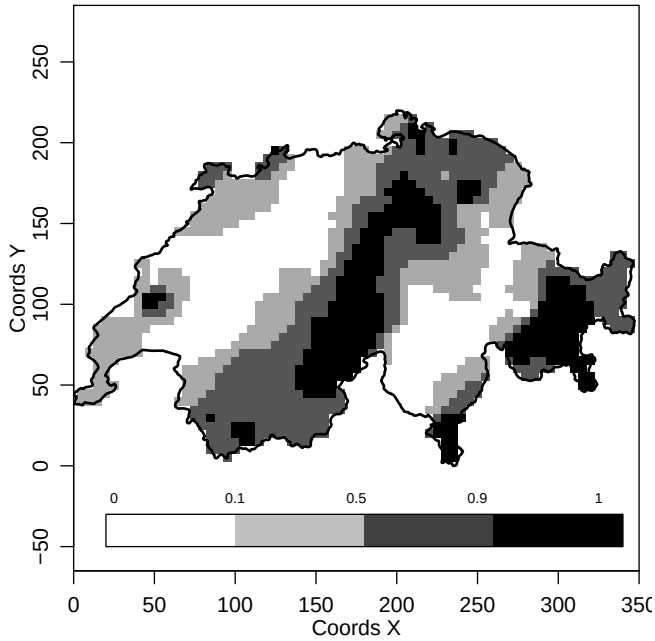


2D profile log-likelihood (left) and samples from posterior distributions (right) for parameters σ^2 and ϕ

Swiss rainfall: prediction results

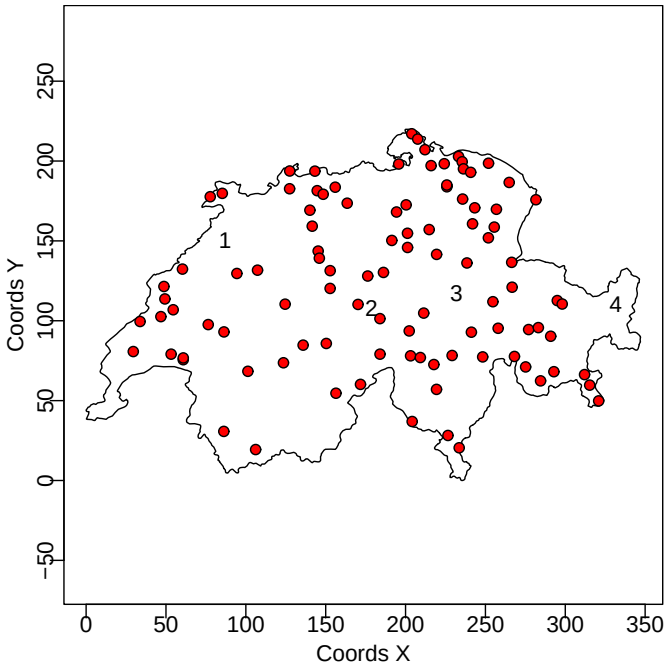


Predicted signal surfaces and associated measures of precision for the rainfall data: (a) posterior mean; (b) posterior variance

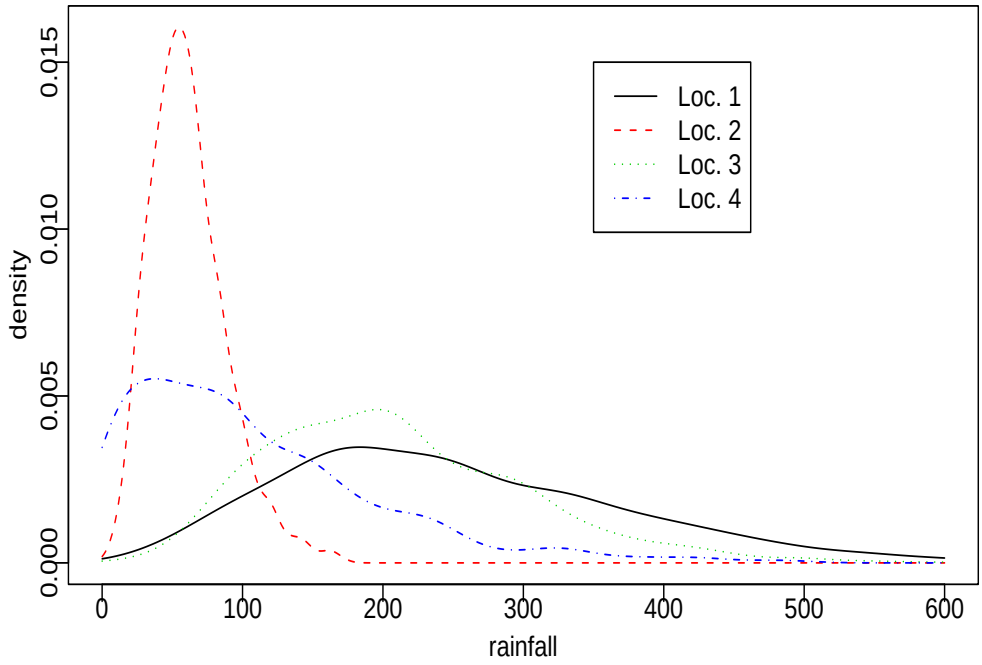


Posterior probability contours for levels 0.10, 0.50 and 0.90 for the random set $T = \{x : S(x) < 150\}$

Swiss rainfall: prediction results (cont.)



Recording stations and selected prediction locations (1 to 4)



Bayesian predictive distributions for average rainfall at selected locations.

Comments

- nugget effect
- other covariates (altitude, temperature, etc)
- data “peculiarities”
- (variogram) estimation: REML *vs* Bayesian
- priors
- Bayesian implementation algorithm
- spatial-temporal modelling
- vector random fields
- *ad hoc* **vs(?)** *model-based*

PART V:

GENERALIZED LINEAR SPATIAL MODELS

1. **Generalized linear mixed models**

2. **Inference for the generalized linear geostatistical model**

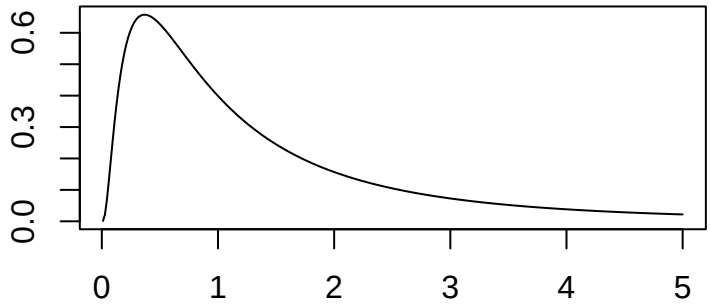
3. **Application of MCMC to Generalized Linear Prediction**

4. **Case-study: Rongelap Island**

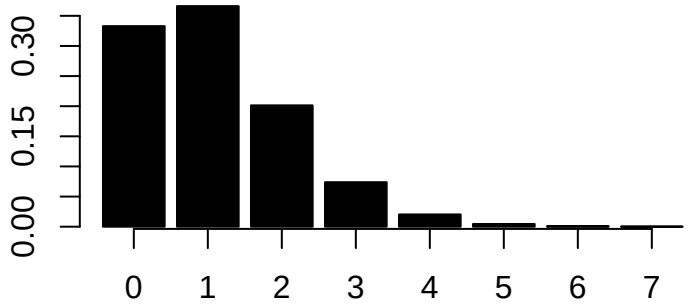
5. **Case-study: Gambia Malaria**

Non-Gaussian data

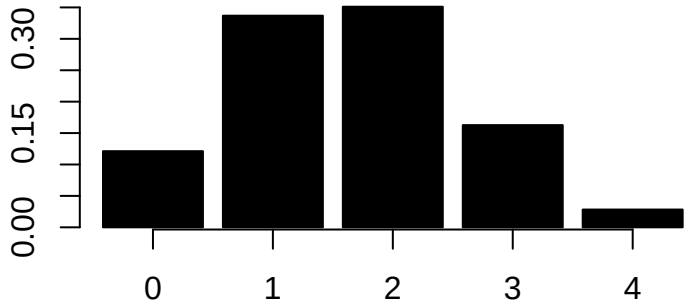
Positive data:



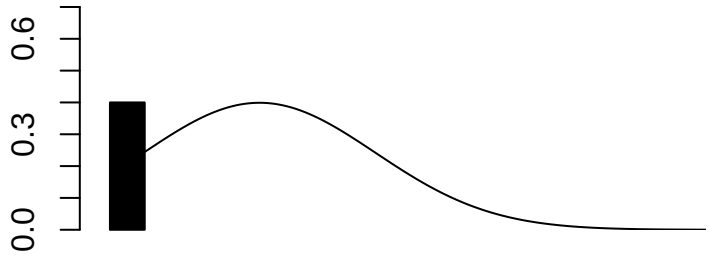
Count data:



Binomial data:



Positive data with zeros:



Towards a model specification

The linear model

$$Y = F\beta + \varepsilon$$

can be written as:

$$Y_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \sum_{j=1}^{k} f_{ij}\beta_j$$

and generalized in two ways

$$Y_i \sim Q(\mu, ...)$$
$$h(\mu_i) = \sum_{j=1}^{k} f_{ij}\beta_j,$$

where Q is distribution in the exponential family and $h(\cdot)$: known link function

- Generalized Linear Models (GLM)

- no longer requires: normality, homocedasticity, continuous scale

- linear model: particular case

Generalized Linear Mixed Models

Classical generalized linear model has

- $Y_i : i = 1, ..., n$
  mutually independent, with $\mu_i = \mathrm{E}[Y_i]$
- $h(\mu_i) = \sum_{j=1}^{k} f_{ij}\beta_j$,
  for known link function $h(\cdot)$

Generalized linear mixed model has

- $Y_i : i = 1, ..., n$
  mutually independent, with $\mu_i = \mathrm{E}[Y_i]$, conditional on realised values of a set of latent random variables U_i
- $h(\mu_i) = U_i + \sum_{j=1}^{k} f_{ij}\beta_j$,
  for known link function $h(\cdot)$

Generalized linear geostatistical model has

- $Y_i : i = 1, ..., n$
  mutually independent, with $\mu_i = \mathrm{E}[Y_i]$, conditional on realised values of a set of latent random variables U_i
- $h(\mu_i) = U_i + \sum_{j=1}^{p} f_{ij}\beta_j$,
  for known link function $h(\cdot)$
- $U_i = S(x_i)$
  where $\{S(x) : x \in \mathbb{R}^2\}$ is a spatial stochastic process

Examples

$x_1, \ldots, x_n$ locations with observations

Poisson-log

- $[Y(x_i) \mid S(x_i)]$ is Poisson with density

$$f(z; \mu) = \exp(-\mu)\mu^z/z! \;\; z = 0, 1, 2, \ldots$$

- link: $E[Y(x_i) \mid S(x_i)] = \mu_i = \exp(S(x_i))$

Binomial-logit

- $[Y(x_i) \mid S(x_i)]$ is binomial with density

$$f(z; \mu) = \binom{r}{z}(\mu/r)^z(1 - \mu/r)^{r-\mu} \;\; z = 0, 1, \ldots, r$$

- link: $\mu_i = \mathrm{E}[Y(x_i) \mid S(x_i)]$, $S(x_i) = \log(\mu_i/(r - \mu_i))$

Likelihood function

$$L(\theta) = \int_{\mathbb{R}^n} \prod_i^n f(y_i; h^{-1}(s_i)) f(s \mid \theta) ds_1, \ldots, ds_n$$

High-dimensional integral !!!

Inference For The Generalized Linear Geostatistical Model

- likelihood evaluation involves high-dimensional numerical integration
- approximate methods (eg Breslow and Clayton, 1993) are of uncertain accuracy
- MCMC is feasible, although not routine.

Application of MCMC to Generalized Linear Prediction

- Ingredients

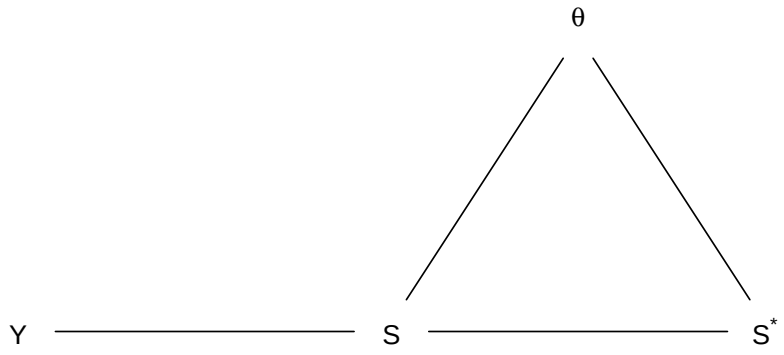
  · Prior distributions for regression parameters β and covariance parameters θ

  · Data: $Y = (Y_1, ..., Y_n)$

  · $S = (S(x_1), ..., S(x_n))$

  · $S^* =$ all other $S(x)$

- Conditional independence structure



- Use output from chain to construct posterior probability statements about $[T|Y]$, where $T = \mathcal{F}(S^*)$

Case-study: Rongelap Island

This case-study illustrates a model-based geostatistical analysis combining:

- a Poisson log-linear model for the sampling distribution of the observations, conditional on a latent
  Gaussian process which represents spatial variation
  in the level of contamination

- Bayesian prediction of non-linear functionals of the latent process

- MCMC implementation

Details are in Diggle, Moyeed and Tawn (1998).

Radiological survey of Rongelap Island

- **Rongelap Island**
  - approximately 2500 miles south-west of Hawaii
  - contaminated by nuclear weapons testing during 1950's
  - evacuated in 1985
  - now safe for re-settlement?
- **The statistical problem**
  - field-survey of ^{137}Cs measurements
  - estimate spatial variation in ^{137}Cs radioactivity
  - compare with agreed safe limits

Poisson Model for Rongelap Data

- Basic measurements are nett counts Y_i over time-intervals t_i at locations x_i $(i = 1, ..., n)$

- Suggests following model:

  · $S(x) : x \in R^2$ stationary Gaussian process (local radioactivity)

  · $Y_i | \{S(\cdot)\} \sim$ Poisson(μ_i)

  · $\mu_i = t_i \lambda(x_i) = t_i \exp\{S(x_i)\}$.

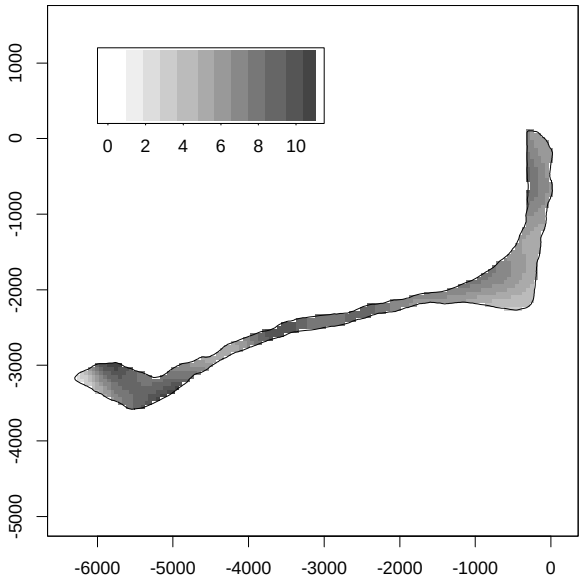
- Aims:

  · predict $\lambda(x)$ over whole island

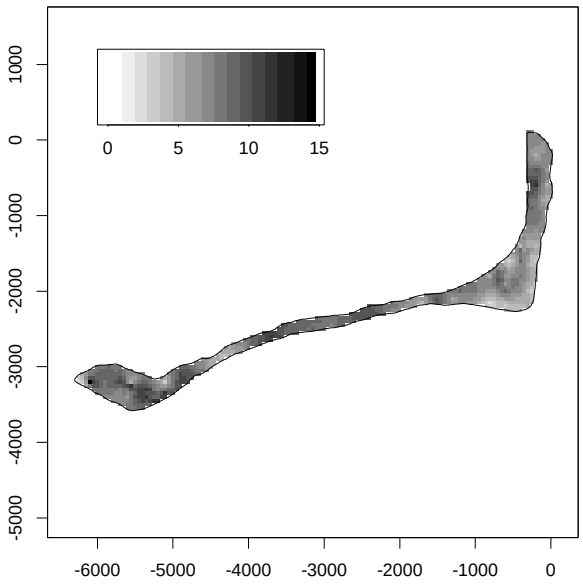
  · max $\lambda(x)$

  · arg(max $\lambda(x)$)

Predicted radioactivity surface using log-Gaussian kriging



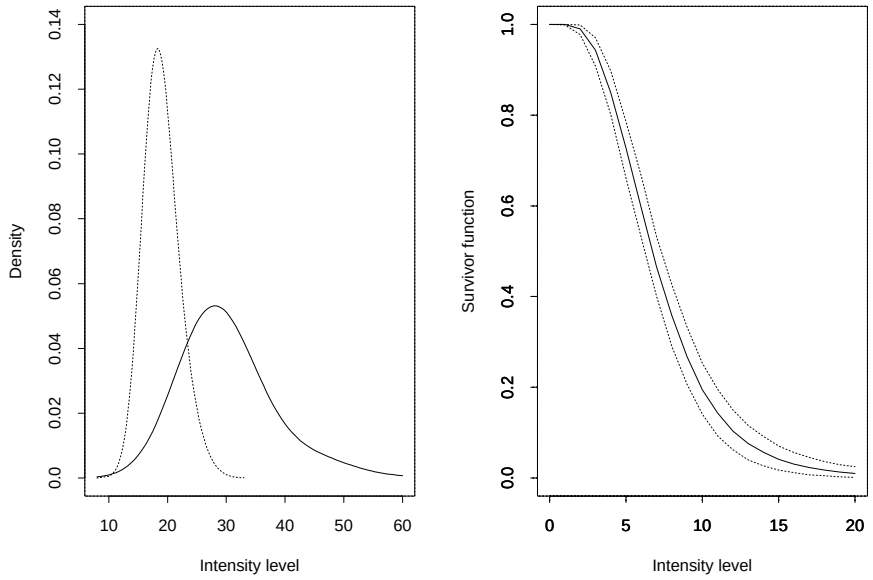
Predicted radioactivity surface using Poisson log-linear model with latent Gaussian process



- The two maps above show the difference between:
  - log-Gaussian kriging of observed counts per unit time
  - log-linear analysis of observed counts
- the principal visual difference is in the extent of spatial smoothing of the data, which in turn stems from the different treatments of the nugget variance

Bayesian prediction of non-linear functionals of the radioactivity surface

The left-hand panel shows the predictive distribution of maximum radioactivity, contrasting the effects of allowing for (solid line) or ignoring (dotted line) parameter uncertainty; the right-hand panel shows 95% pointwise credible intervals for the proportion of the island over which radioactivity exceeds a given threshold.



- The two panels of the above diagram illustrate Bayesian prediction of non-linear functionals of the latent
  Gaussian process in the Poisson log-linear model

- the left-hand panel contrasts posterior distributions of the maximum radioactivity based on:
  (i) the fully Bayesian analysis incorporating the effects of parameter uncertainty in addition to uncertainty in the latent process (solid line)
  (ii) fixing the model parameters at their estimated values, ie allowing for uncertainty only in the latent process

- the right-hand panel gives posterior estimates with 95% point-wise credible intervals for the proportion of the island over which radioactivity exceeds a given threshold (dotted line).

Case-study: Gambia malaria

- In this example, the spatial variation is of secondary scientific importance.

- The primary scientific interest is to describe how the prevalence of malarial parasites depends on explanatory variables measured:
  - on villages
  - on individual children

- There is a particular scientific interest in whether a vegetation index derived from satellite data is a useful predictor of malaria prevalence, as this would help health workers to decide how to make best use of scarce resources.

Data-structure

- 2039 children in 65 villages
- test each child for presence/absence of malaria parasites

Covariate information at child level:

- age (days)
- sex (F/M)
- use of mosquito net (none, untreated, treated)

Covariate information at village level:

- location
- vegetation index, from satellite data
- presence/absence of public health centre

Logistic regression model

Logistic model for presence/absence in each child:

- $Y_{ij} = 0/1$ for absence/presence of malaria parasites in jth child in ith village
- $f_{ij} =$ child-specific covariates
- $w_i =$ village-specific covariate
- $\text{logit}P(Y_{ij} = 1|S(\cdot)) = f'_{ij}\beta_1 + {w_i}'\beta_2 + S(x_i)$

Is it reasonable to assume conditionally independent infections within same village?

If not, we might wish to extend the model to allow for non-spatial extra-binomial variation:

- $U_i \sim \text{N}(0, \nu^2)$
- $\text{logit}P(Y_{ij} = 1|S(\cdot), U) = f'_{ij}\beta_1 + {w_i}'\beta_2 + U_i + S(x_i)$

Exploratory analysis

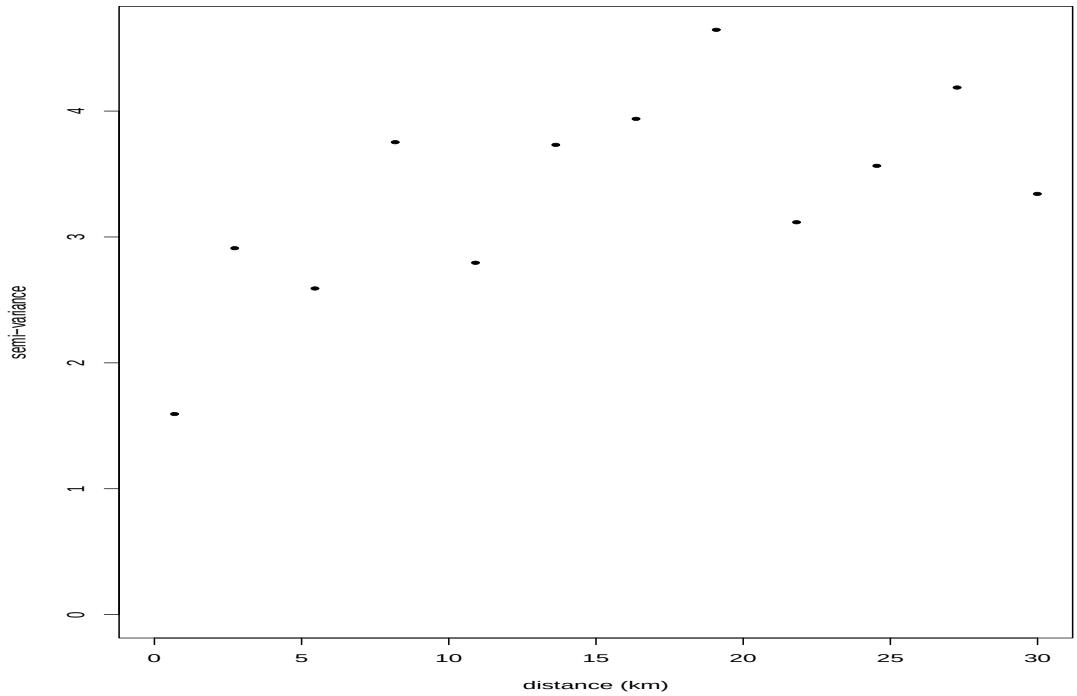
- fit standard logistic linear model, ignoring $S(x)$ and/or U
- compute for each village:

  $N_i = \sum_{j=1}^{n_i} Y_{ij}$

  $\mu_i = \sum_{j=1}^{n_i} \hat{P}_{ij}$

  $\sigma_i^2 = \sum_{j=1}^{n_i} \hat{P}_{ij}(1 - \hat{P}_{ij})$
- compute village-residuals, $r_i = (N_i - \mu_i)/\sigma_i$
- apply conventional geostatistics to derived data r_i
- variogram indicates residual spatial structure

Variogram of residuals



Model-based geostatistical analysis

α = intercept term in linear predictor

β_1 = regression coefficient for age

β_2 = regression coefficient for bed-net use

β_3 = regression coefficient for treated bed-net

β_4 = regression coefficient for green-ness index

β_5 = regression coefficient for presence of public health centre in village

ν^2 = variance of non-spatial random effects U_i

σ^2 = variance of spatial process $S(x)$

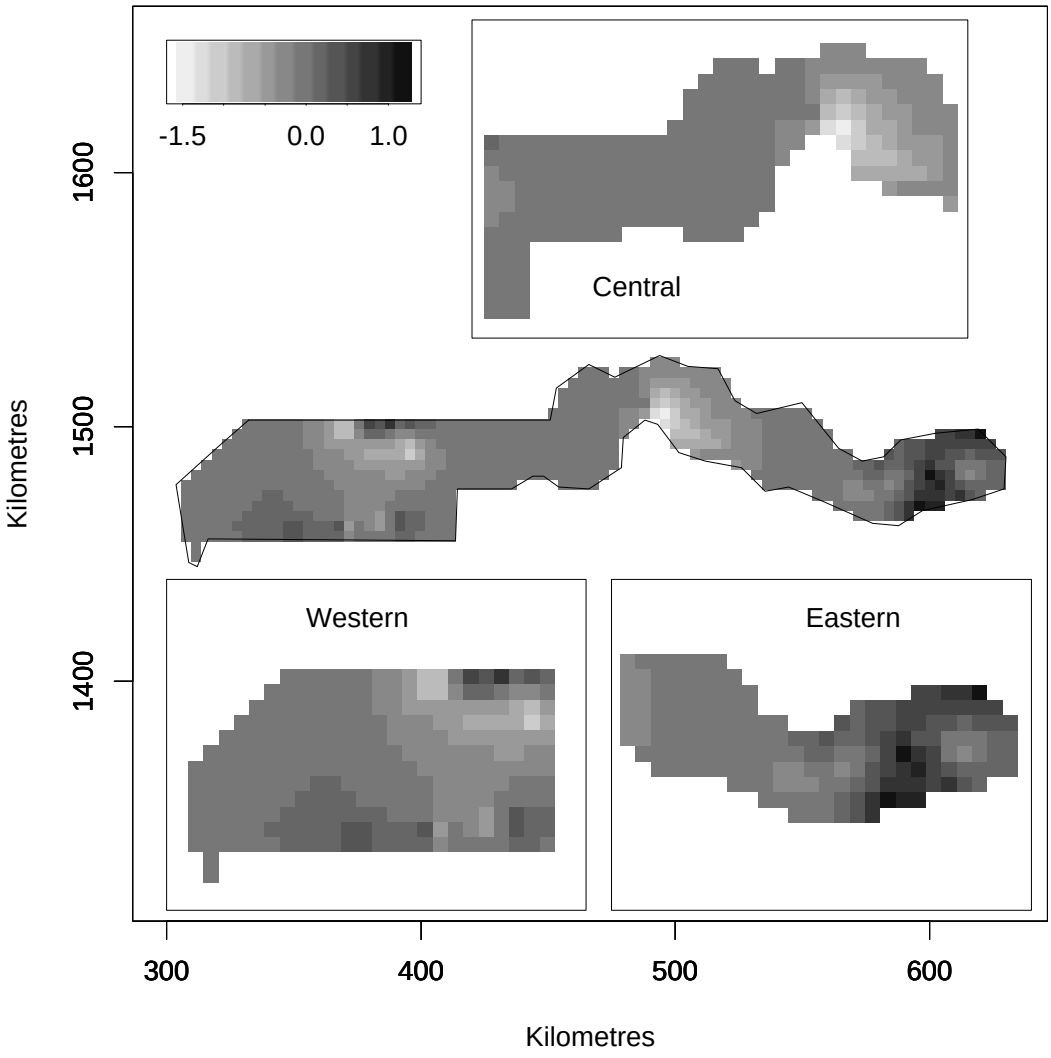
ϕ = rate of decay of spatial correlation with distance

κ = shape parameter for Matérn correlation function

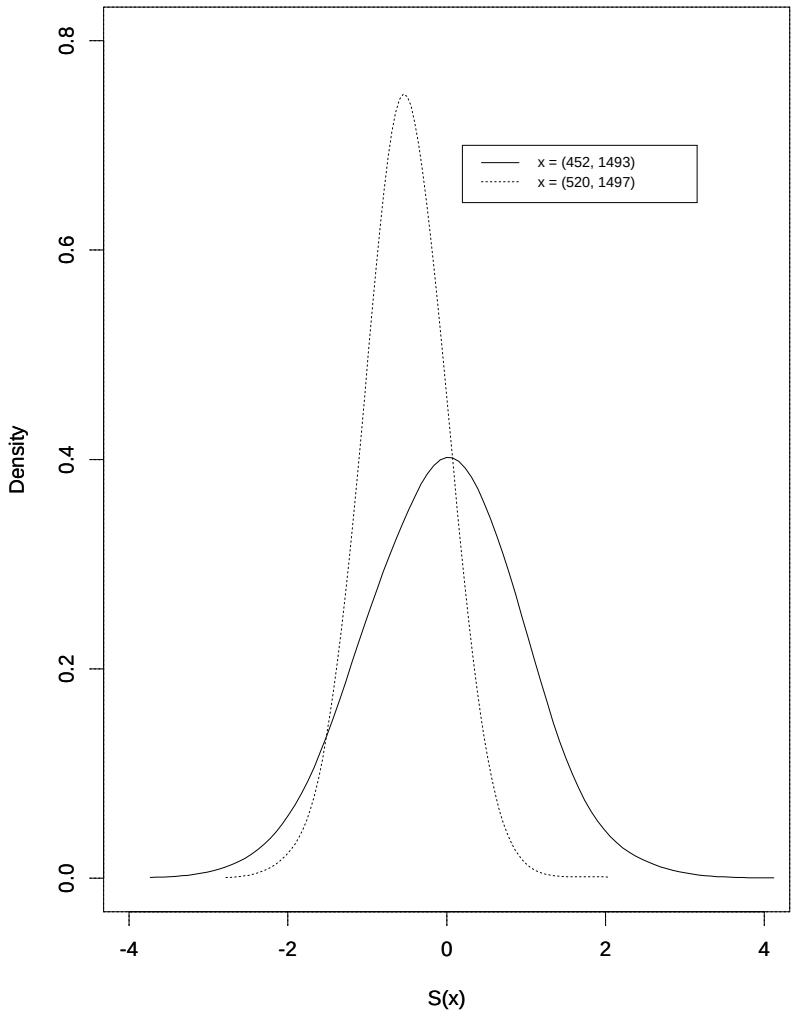
Param.	2.5% Qt.	97.5% Qt.	Mean	Median
α	-4.232073	1.114734	-1.664353	-1.696228
β_1	0.000442	0.000918	0.000677	0.000676
β_2	-0.684407	-0.083811	-0.383750	-0.385772
β_3	-0.778149	0.054543	-0.355655	-0.355632
β_4	-0.039706	0.071505	0.018833	0.020079
β_5	-0.791741	0.180737	-0.324738	-0.322760
ν^2	0.000002	0.515847	0.117876	0.018630
σ^2	0.240826	1.662284	0.793031	0.740790
ϕ	1.242164	53.351207	11.653717	7.032258
κ	0.150735	1.955524	0.935064	0.830548

- note concentration of posterior for ν^2 close to zero

Map of the predicted surface $\hat{S}(x)$ (posterior mean)

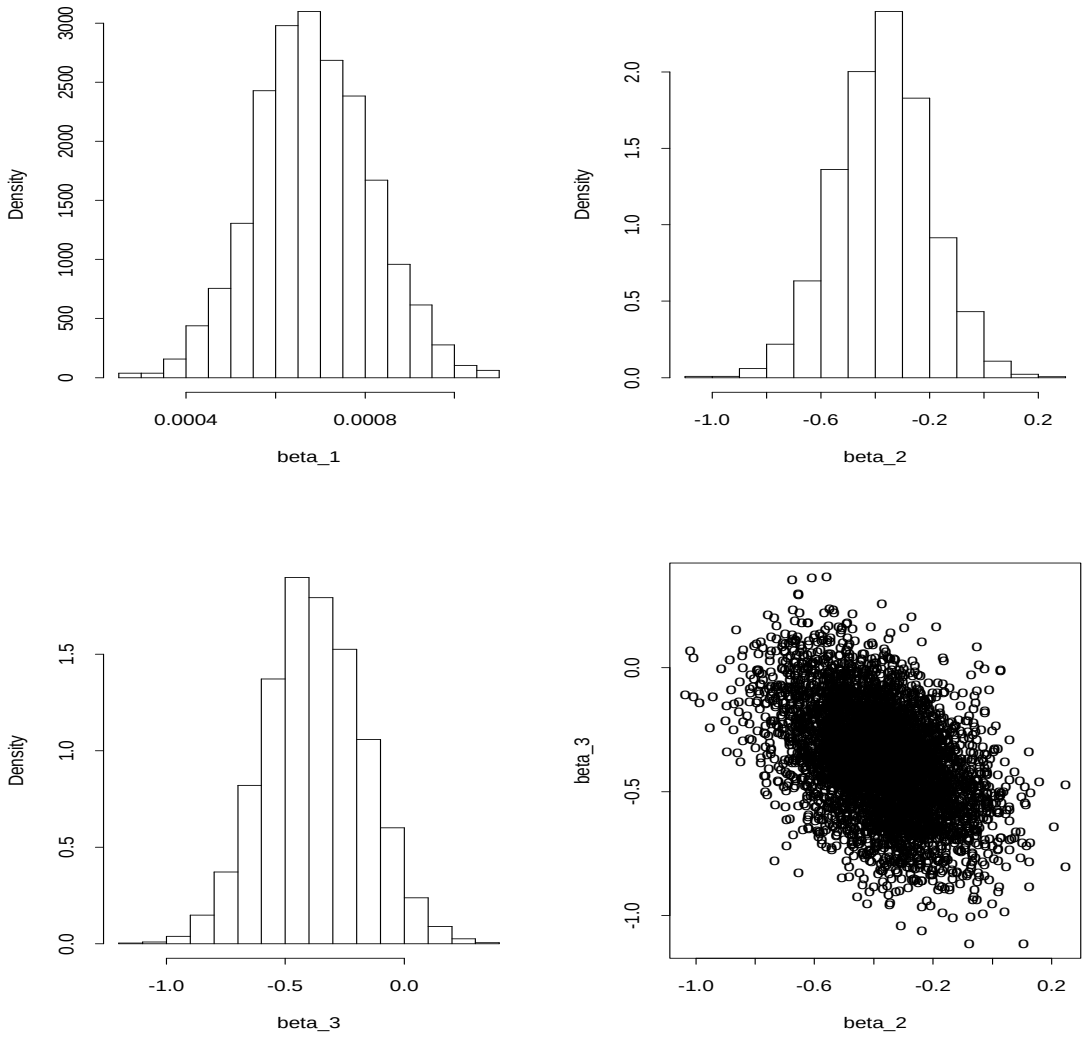


Posterior density estimates for $S(x)$ at two selected locations.



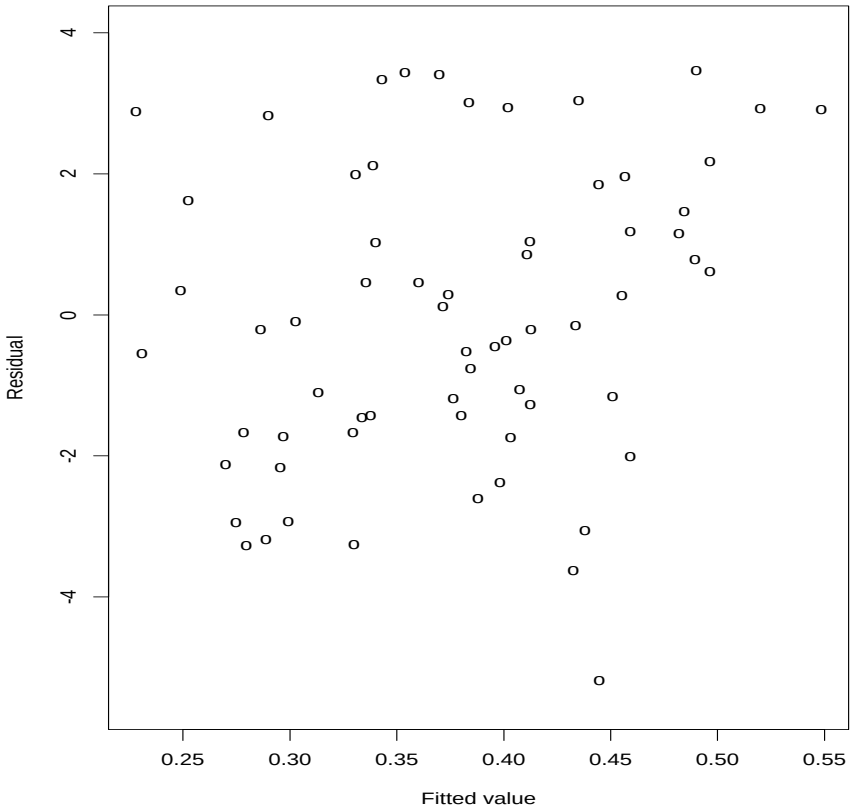
- solid curve – remote location (452, 1493),
- dashed curve – location (520, 1497), close to observed sites in central region.

Empirical posterior distributions for regression parameters



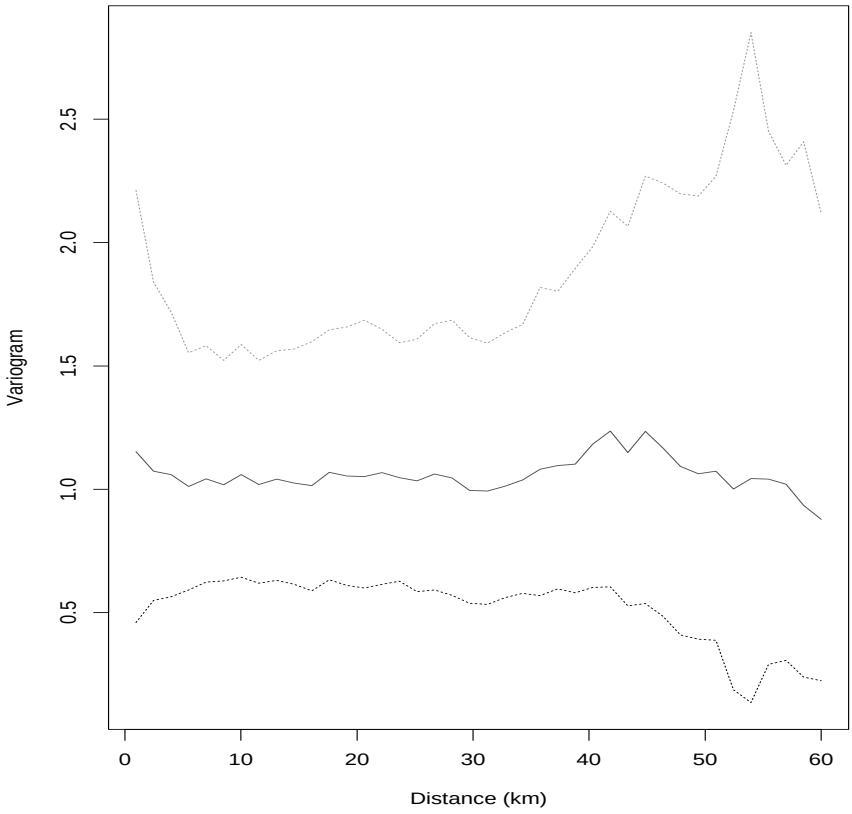
- $\beta_1 =$ effect of age
- $\beta_2 =$ effect of untreated bed-nets
- $\beta_3 =$ additional effect of treated bed-nets

Goodness-of-fit for Gambia malaria model



Village-level residuals against fitted values.

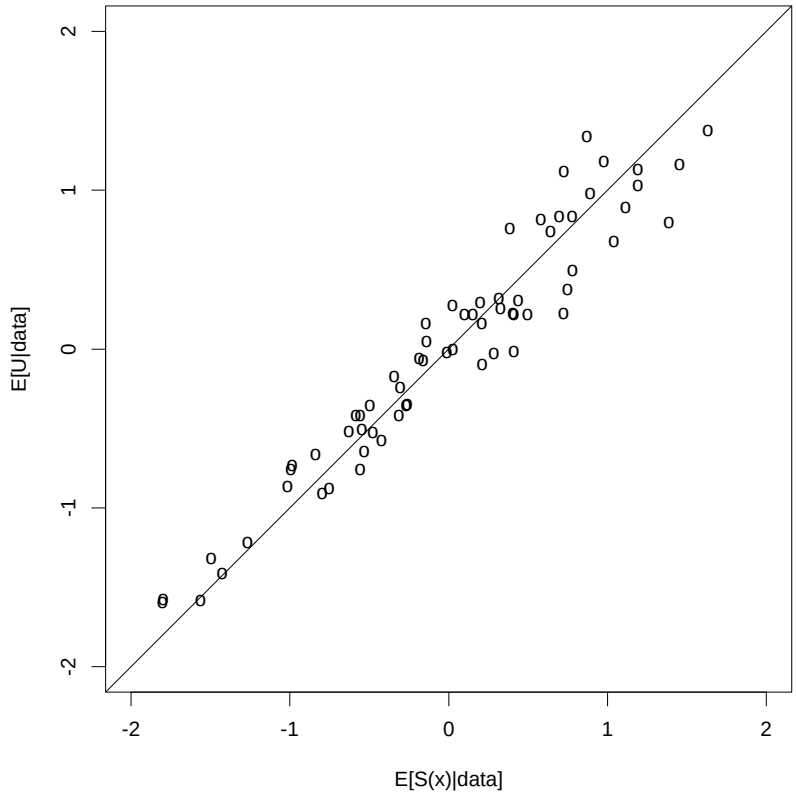
- $r_{ij} = (Y_{ij} - \hat{p}_{ij})/\sqrt{\{\hat{p}_{ij}(1 - \hat{p}_{ij})\}}$
- $r_i = \sum r_{ij}/\sqrt{n_i}$
- intended to check adequacy of model for p_{ij}



Standardised residual empirical variogram plot (village-level data and pointwise 95% posterior intervals constructed from simulated realisations of fitted model).

- $r_{ij} = (Y_{ij} - \hat{p}^*_{ij})/\sqrt{\{\hat{p}^*_{ij}(1 - \hat{p}^*_{ij})\}}$
- $r_i = \sum r_{ij}/\sqrt{n_i}$
- $\text{logit}p^*_{ij} = \hat{\alpha} + f'_{ij}\hat{\beta} + \hat{S}(x_i)$
- intended to check adequacy of model for $S(x)$

Is a geostatistical model necessary?



Plot of estimated posterior means of random effects $\hat{U}_i$ from non-spatial GLMM against estimated posterior means $\hat{S}(x_i)$ at observed locations in geostatistical model.

- high correlation represents strong empirical evidence of spatial dependence
- but explicit modelling of spatial dependence has small effect on inferences about regression parameters

Generalised linear Spatial Models

Diggle, Tawn and Moyeed (1998).

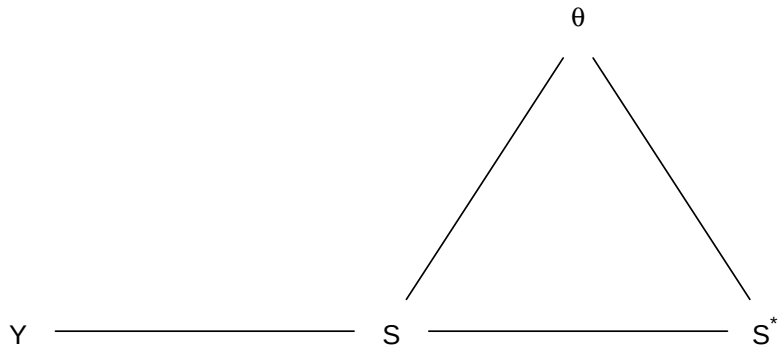
Error distribution $[Y(x) \mid S(x)]$

Link function $\mathrm{E}[Y(x) \mid S(x)] = h^{-1}(S(x))$

Gaussian random field $\{S(x) : x \in A\}$

- $\mathrm{E}[S(x)] = d(x)^T\beta$
- $\mathrm{Cov}\,(S(x), S(x')) = \sigma^2\rho(\|x - x'\|; \phi) + \tau^2\mathbf{1}_{\{x=x'\}}$.

Structure:



- Data : $Y = (Y(x_1), ..., Y(x_n))$
- $S = (S(x_1), ..., S(x_n))$
- S^* = some other $S(x)$
- θ parameters

Prediction with known parameters

From figure, prediction is separated into three steps.

- Simulate $s(1), \ldots, s(m)$ from $[S|y]$ (using MCMC).

- Simulate $s^*(j)$ from $[S^*|s(j)]$, $j = 1, \ldots, m$ (multivariate Gaussian)

- Approximate $\mathrm{E}[T(S^*)|y]$ by

$$\frac{1}{m} \sum_{j=0}^{m} T(s^*(j))$$

If possible: instead calculate $\mathrm{E}[T(S^*)|s(j)]$, $j = 1, \ldots, m$ directly, and estimate $\mathrm{E}[T(S^*)|y]$ by

$$\frac{1}{m} \sum_{j=0}^{m} \mathrm{E}[T(S^*)|s(j)]$$

Advantage: smaller Monte Carlo error

MCMC for conditional simulation

Let $S = D^{\mathrm{T}}\beta + \Sigma^{1/2}\Gamma$, $\Gamma \sim N_n(0, I)$.

Conditional density of $[\Gamma\,|Y = y]$

$$f(\gamma|y) \propto f(y|\gamma)f(\gamma)$$

Langevin-Hastings algorithm

Proposal: γ' from a $N_n(\xi(\gamma), hI)$ where $\xi(\gamma) = \gamma + \frac{h}{2}\nabla \log f(\gamma \mid y)$.

Poisson-log Spatial model:

$\nabla \log f(\gamma|y) = -\gamma + (\Sigma^{1/2})^{\mathrm{T}}(y - \exp(s))$ where $s = \Sigma^{1/2}\gamma$.

Expression generalises to other generalised linear spatial models.

MCMCM output $\gamma_1, \ldots, \gamma_m$. Multiply by $\Sigma^{1/2}$ and obtain: $s(1), \ldots, s(m)$ from $[S|y]$.

Random walk Metropolis algorithm

- Current state: γ.
- Proposal: γ' from a $N_n(\gamma, hI)$.
- Accept: γ' with probability

$$a(\gamma, \gamma') = 1 \wedge \frac{\pi(\gamma')}{\pi(\gamma)}.$$

Langevin-Hastings algorithm

- Current state: γ.
- Proposal: γ' from a $N_n(\xi(\gamma), hI)$ where $\xi(\gamma) = \gamma + \frac{h}{2}\nabla \log \pi(\gamma)$.
- Accept: γ' with probability

$$a(\gamma, \gamma') = 1 \wedge \frac{\pi(\gamma')q(\gamma', \gamma)}{\pi(\gamma)q(\gamma, \gamma')}.$$

Models considered here:

$$\nabla \log f(\gamma|y) = -\gamma + (\Sigma^{1/2})^{\mathrm{T}} \left\{ (y_i - g^{-1}(s_i)) \frac{g_c(g^{-1}(s_i))}{g(g^{-1}(s_i))} \right\}_{i=1}^{n}$$

.

where $s = D\beta + \Sigma^{1/2}\gamma$.

Geometric Ergodicity

Convergence rate to equilibrium is geometrically fast.

Geometrical ergodicity $\Rightarrow$ Central Limit Theorem:

$$\sqrt{m}\left(\frac{1}{m}\sum_{j=1}^{m}\psi(\Gamma(j)) - \mathrm{E}[\psi(\Gamma)|y]\right) \widetilde{\rightarrow} N(0, \sigma_{\psi}^2).$$

when $\psi(\cdot)^2 \leq V(\cdot)$

Random walk Metropolis:
Conditions (Jarner and Hansen, 2000):

$$\limsup_{\mu \to m_1, m_2} \frac{1}{g'(\mu)}\left|\frac{1}{\mu - y_i} + \frac{g_c''(\mu)}{g_c'(\mu)} - \frac{g''(\mu)}{g'(\mu)}\right| < \infty,$$

$$\limsup_{\mu \to m_1, m_2} \frac{y_i - \mu}{g(\mu)}\frac{g_c'(\mu)}{g'(\mu)} \leq 0,$$

imply geometric ergodicity, with $V(\gamma) = f(\gamma|y)^{-1/2}$.

Langevin-Hastings:(Robert and Tweedie,1996)
Poisson-log model ($\nabla(\gamma) = -\gamma + Q^{\mathrm{T}}(y - \exp(s))$) not geom. ergodic.

Truncated Langevin-Hastings

$$\nabla(\gamma)^{\mathbf{trunc}} = -\gamma + Q^{\mathrm{T}}R(\gamma)$$

where $R(\gamma)$ is a bounded function. Geometrically ergodic with $V(\gamma) = \exp(t\|\gamma\|)$.

MCMC for Bayesian inference

Parameters in underlying Gaussian model:

$S \sim MVN(D^{\mathrm{T}}\beta, \sigma^2\kappa(\phi))$.

$S = D^{\mathrm{T}}\beta + \sigma\kappa(\phi)^{1/2}\Gamma$, $\Gamma \sim N_n(0, I)$.

Metropolis-within-Gibbs algorithm:

- Update Γ from $[\Gamma|y, \beta, \log(\sigma), \log(\phi)]$ (Langevin-Hasting described earlier)
- Update β from $[\beta|\Gamma, \log(\sigma), \log(\phi)]$ (RW-Metropolis)
- Update $\log(\sigma)$ from $[\log(\sigma)|\Gamma, \beta, \log(\phi)]$ (RW-Metropolis)
- Update $\log(\phi)$ from $[\log(\phi)|\Gamma, \beta, \log(\sigma)]$ (RW-Metropolis)

Prediction:

- Simulate $(s(j), \beta(j), \sigma^2(j), \phi(j))$, $j = 1, \ldots, m$ (using MCMC)
- Simulate $s^*(j)$ from $[S^*|s(j), \beta(j), \sigma^2(j), \phi(j)]$, $j = 1, \ldots, m$ (multivariate Gaussian)

Discrete prior for ϕ is an advantage (reduced computing time).

Thinning not to store a large sample of high-dimensional quantities.

Marginalisation Integrating out β and σ^2.

Conjugate priors:

$$\beta \mid \sigma^2 \sim \mathrm{N}\left(m_b \,;\, \sigma^2 V_b\right)$$

$$1/\sigma^2 \sim \frac{1}{S_c^2}\chi^2\left(n_c\right)$$

[scaled-inverse-$\chi^2(n_c, S_c^2)$]

Marginalisation

- Posterior distribution is $f(s|y) = f(y|s)f_I(s)$, where f_I is the marginalised density for S.
- Marginalisation makes $[S^*|S]$ a multivariate-t distribution.

Procedure:

- Simulate $s(1), \ldots, s(m)$ from $[S|y]$ (using MCMC).
- Simulate $s^*(j)$ from $[S^*|s(j)]$, $j = 1, \ldots, m$ (multivariate-t distribution)
- Approximate $\mathrm{E}[T(S^*)|y]$ by

$$\frac{1}{m}\sum_{j=0}^{m} T(s^*(j))$$

Parameter estimation for generalized linear spatial model

- Methods of moments (empirical variogram/covariogram)
- Approximate methods: pseudo-likelihood (accuracy unknown).

Monte Carlo approximation to likelihood

Geyer and Thompson, 1992 ; Geyer, 1994.

Define $\tilde{f}(y) = f(y \mid s)\tilde{f}(s)$ for some density $\tilde{f}(s)$.

$$L(\theta) = \int_{\mathbb{R}^n} f(y \mid s) f(s \mid \theta) ds \propto \int_{\mathbb{R}^n} \frac{f(y \mid s) f(s \mid \theta)}{f(y \mid s)\tilde{f}(s)} \tilde{f}(s \mid y) d$$

$$= \tilde{\mathrm{E}}[\frac{f(S \mid \theta)}{\tilde{f}(S)} \mid y$$

where $\tilde{\mathrm{E}}[\cdot \mid y]$ is expectation w.r.t $\tilde{f}(s \mid y)$.

- How to chose $\tilde{f}(s)$?
  - Could chose $\tilde{f}(s) = f(s \mid \theta_0)$ for some θ_0.
  - Instead we use $\tilde{f}(s) = f_I(s \mid \psi_0)$ for some ψ_0, where $\theta = (\beta, \sigma^2, \psi)$, and $f_I(s \mid \psi_0)$ is the marginalised density $f_I(s \mid \psi_0) = \int\int f_I(s \mid \beta, \sigma^2, \psi_0)\pi(\beta, \sigma^2) d\beta d\sigma^2$ (using conjugate prior on (β, σ^2)).
- Very useful as exploratory tool; exploring other correlation functions, anisotropy, etc.

Priors

$S \sim N_n(D\beta, \sigma^2 R(\phi))$

Non-informative priors:

- $\pi(\sigma^2) \propto 1/\sigma^2$ commonly used, but gives improper posterior !
- $\pi(\phi)$ must be proper.
- Thesis gives conditions for proper posterior.

Conjugate priors:

$$\beta \mid \sigma^2 \sim \mathrm{N}\left(m_b \,;\, \sigma^2 V_b\right)$$

$$1/\sigma^2 \sim \frac{1}{S_\sigma^2}\chi^2\left(n_\sigma\right)$$

[scaled-inverse-$\chi^2(n_\sigma, S_\sigma^2)$]

Marginalisation

- Conditional distribution becomes $f(s|y) \propto f(y|s)f_I(s)$,
  where f_I is the marginalised density of S.
- Marginalisation makes $[S^*|S]$ a multivariate-t distribution.

Computationally advantageous in MCMC-algorithm.

Conjugate prior for ϕ ?

PART VI:

Further topics and conclusions

1. **Multivariate methods**

2. **Non-linear methods**

3. **Space-time models**

4. **Marked point processes**

5. **Closing remarks**

Multivariate methods

- within linear Gaussian setting, extension to multivariate data is straightforward in principle

- but specification of a useful class of default models for cross-covariance structure is not straightforward

- detailed implementation should be problem-specific: for example,
  - (Y_1, Y_2) qualitatively different, and measured at a common set of locations?
  - Y_2 a low-cost surrogate for Y_1 of primary interest?

Non-linear methods

- non-linear mean response $\mu(x)$ with Gaussian errors – straightforward in principle

  Example: plume model for dispersal of pollutant

- intrinsically non-linear systems specified by system of stochastic differential equations – a challenging problem

  Example: soil conductivity/transmissivity

Space-time models

- Emerging space-time data-sets are **big**, and present severe computational challenges.

- Specific models are best defined in context.

- Some examples:

  1. Calibration of radar reflectance against ground-truth rainfall intensity (Brown, Diggle, Lord and Young, 2001).
     (a) $Y_{it} : i = 1, ..., n$ – ground-truth log-rainfall intensity at small number of sites x_i
     (b) $U(x, t) : x \in A$ – log-radar reflectance measured effectively continuously over a study region A
     (c) Empirical model,

     $$Y_{it} = \alpha + B(x_i, t)U(x_i, t)$$

     where $B(x, t)$ is continuous space-time Gaussian process
     (d) Spatial prediction,

     $$\hat{Y}(x, t) = \hat{\alpha} + \hat{B}(x, t)U(x, t)$$

2. On-line disease surveillance (Brix and Diggle, 2001)

   (a) Data give population density $\lambda_0(x)$ (approximately), plus locations of daily incident cases

   (b) Model space-time point process of incident cases as Cox process:

   – Poisson process with intensity

$$\lambda(x,t) = \lambda_0(x)\exp\{\alpha + Z(x,t)\}$$

   – Space-time Gaussian process $Z(x,t)$ models variation in disease risk

   – Interested in early detection of **changes in the risk surface**,

$$\lambda(x,t)/\lambda(x,t-1) = \exp\{Z(x,t)-Z(x,t-1)\}$$

Marked point processes

Definition: a joint probability model for a stochastic **point process** P, and an associated set of random variables, or **marks**, M

Different possible structural assumptions:

- $[P, M] = [P][M]$
  The **random field model** – often assumed implicitly in geostatistical work.

- $[P, M] = [P|M][M]$
  Preferential sampling – sampling locations are determined by partial knowledge of the underlying
  mark process
  Example. deliberate siting of air pollution monitors in badly polluted areas.

- $[P, M] = [P][M|P]$
  Often appropriate when the mark process is only
  defined at the sampling locations.
  Example. Presence/absence of disease amongst individual members of a population.

Implications of ignoring violations of the random field model are not well understood.

Closing remarks

- all models are wrong, but some models are useful
- whatever model is adopted, inferential procedures which respect general statistical principles are likely to out-perform *ad hoc* procedures
- ignoring parameter uncertainty can seriously prejudice nominal prediction intervals
- the Bayesian paradigm gives a workable integration of parameter estimation and stochastic process prediction, but results can be sensitive to joint prior specifications.
- the best models are developed by statisticians and subject-matter scientists working in collaboration.