Modelos Geostatísticos [1]

Paulo Justiniano Ribeiro Jr

Laboratório de Estatística e Geoinformação

Universidade Federal do Paraná e ESALQ/USP (Brasil)

em colaboração com

Peter J Diggle

Lancaster University (UK) e

Johns Hopkins University School of Public Health (USA)

52^a **RBRAS** e 12^o **SEAGRO, Santa Maria, RS, Jul 2007**

---

[1] *versão modificada de transparências do curso "Model Based Geostatistics" apresentado por PJD & PJRJr na IBC-2006, Montreal, CA e baseado no livro Diggle & Ribeiro Jr (2007), Springer.*

SESSION 1

Spatial statistics – an overview

Context for spatial data

- (geo)-referenced data
- GIS - Geographical Information Systems
- Spatial Analysis
- Spatial Statistics
  - discrete spatial variation
  - continuous spatial variation (geostatistical or point referenced data)
  - point process
  - and ... more complex structures

SESSION 2

Introduction and motivating examples

Geostatistics

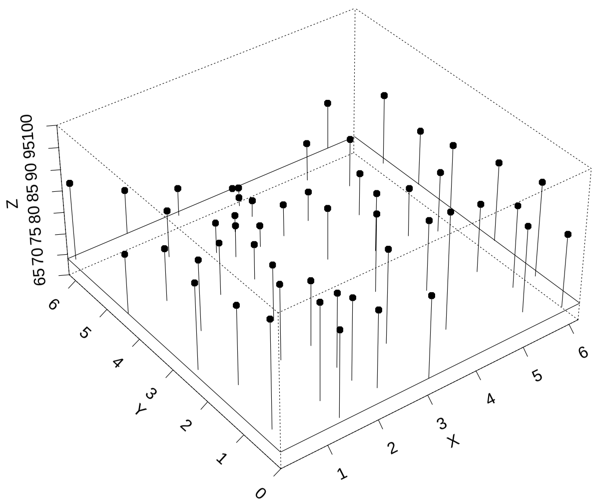
- traditionally, a self-contained methodology for spatial prediction, developed at École des Mines, Fontainebleau, France
- nowadays, that part of spatial statistics which is concerned with data obtained by spatially discrete sampling of a spatially continuous process

Motivating examples

In the following examples we should identify:

- **the structure of the available data**
- **the underlying process**
- **the scientific objectives**
- **the nature of the response variable(s) and potential co-variates**
- **combine elements/features for a possible statistical model**

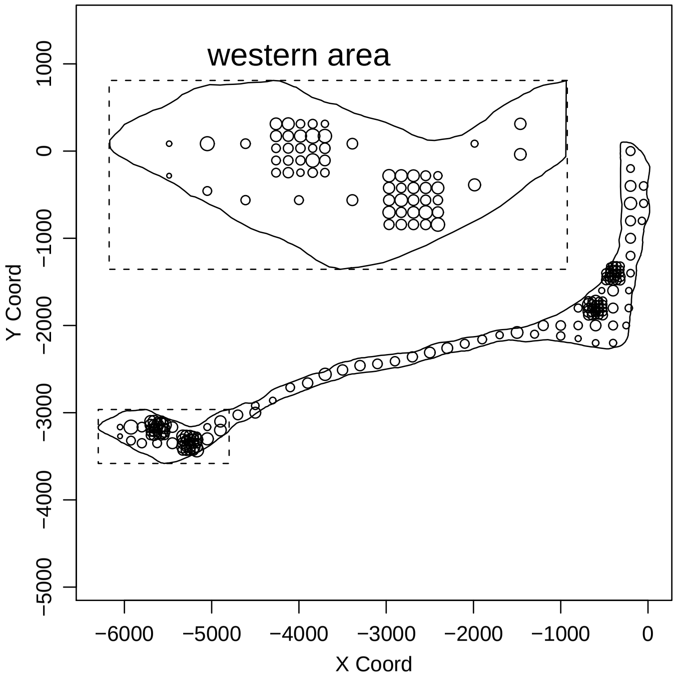
Example 1.1: Measured surface elevations



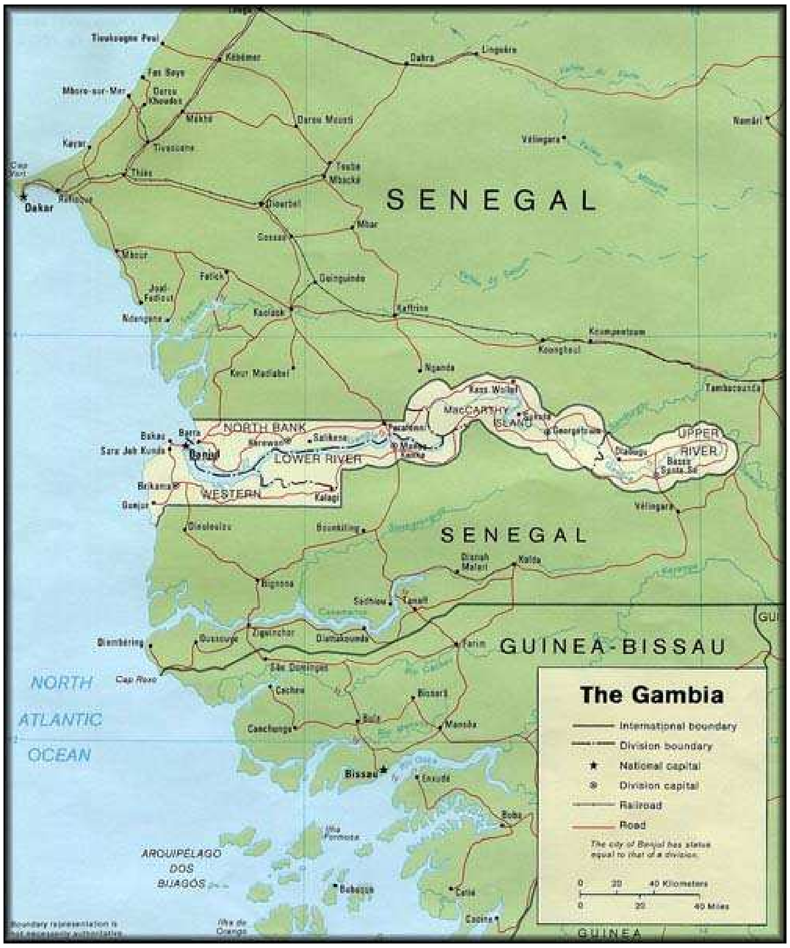
`require(geoR) ; data(elevation) ; ?elevation`

Potential distinction between $S(x)$ and $Y(x)$

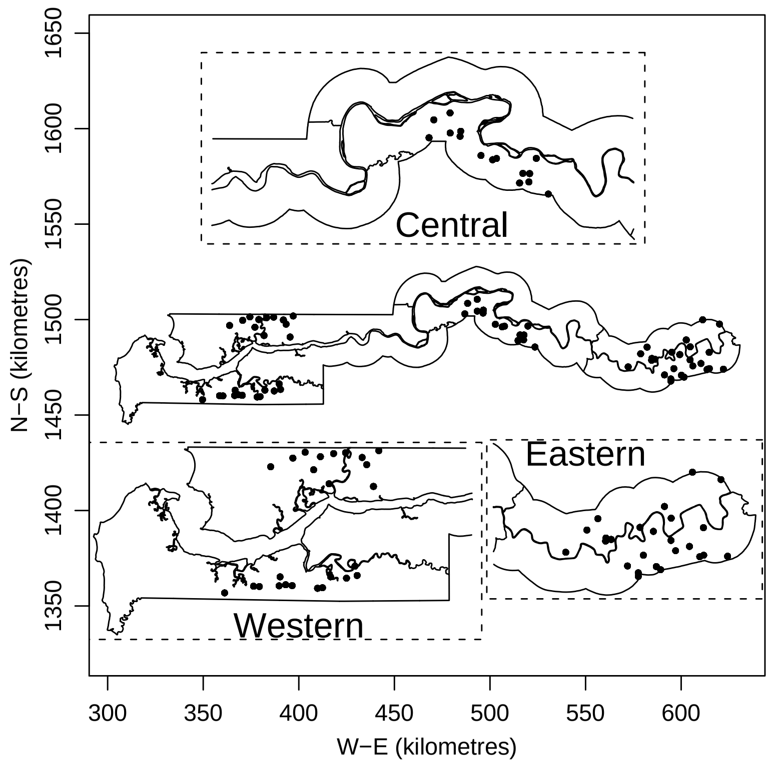
Example 1.2: Residual contamination from nuclear weapons testing



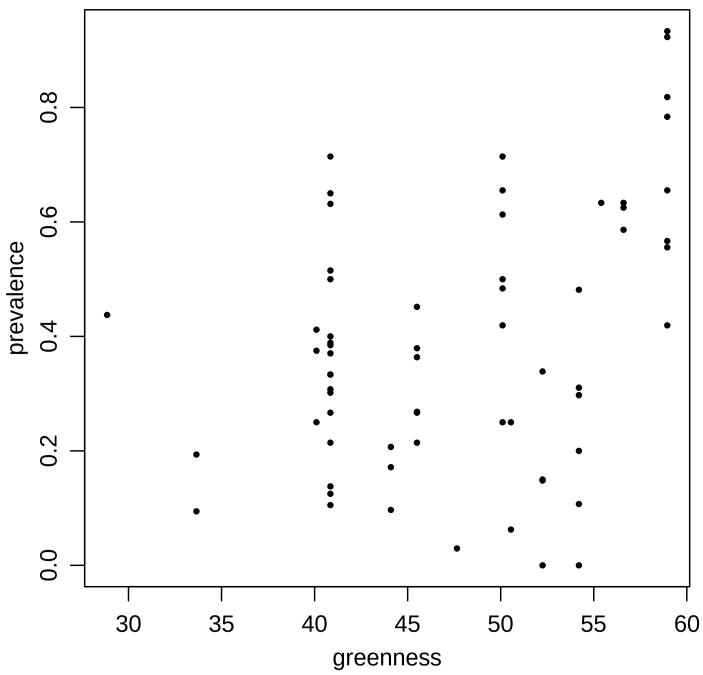

Sunset at Georgetown


SENEGAL
SENEGAL
GUINEA-BISSAU
NORTH ATLANTIC OCEAN
NORTH BANK
LOWER RIVER
WESTERN
MacCARTHY ISLAND
UPPER RIVER
Dakar
Banjul
Georgetown
Bissau
ARQUIPÉLAGO DOS BIJAGÓS
The Gambia
International boundary
Division boundary
National capital
Division capital
Railroad
Road
The city of Banjul has status equal to that of a division.
40 Kilometers
40 Miles
GUINEA

Example 1.3: Childhood malaria in Gambia

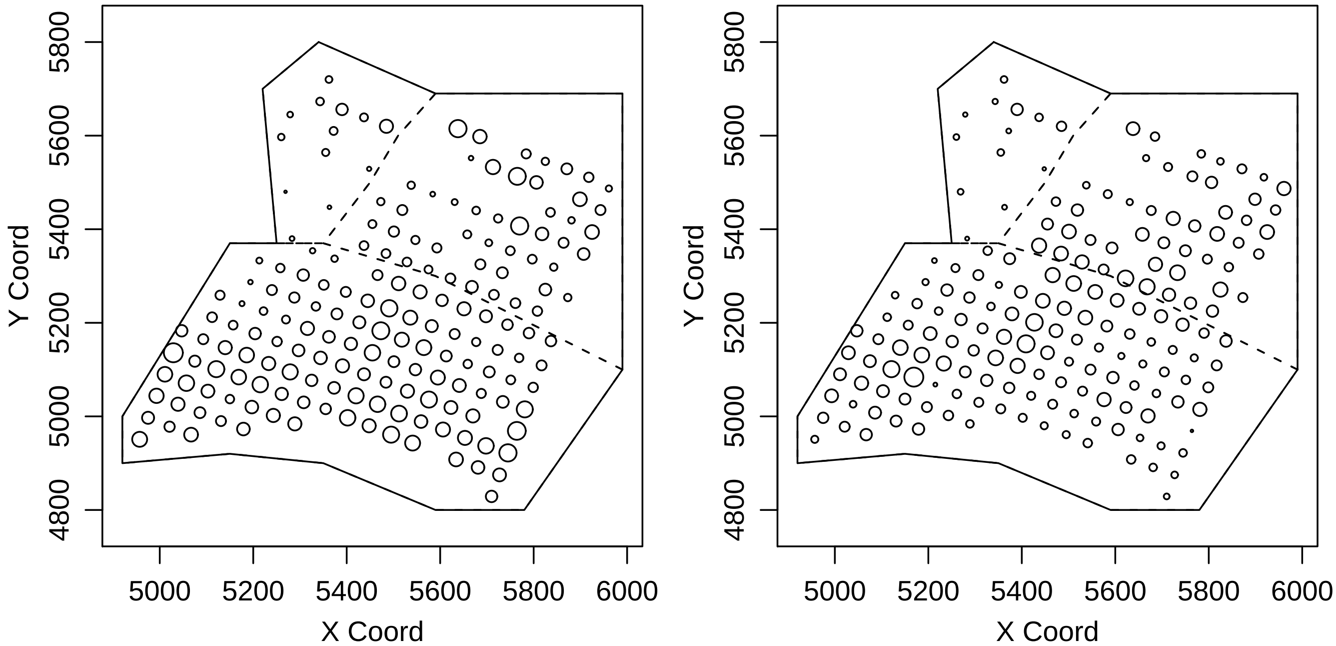


Example 1.3: continued



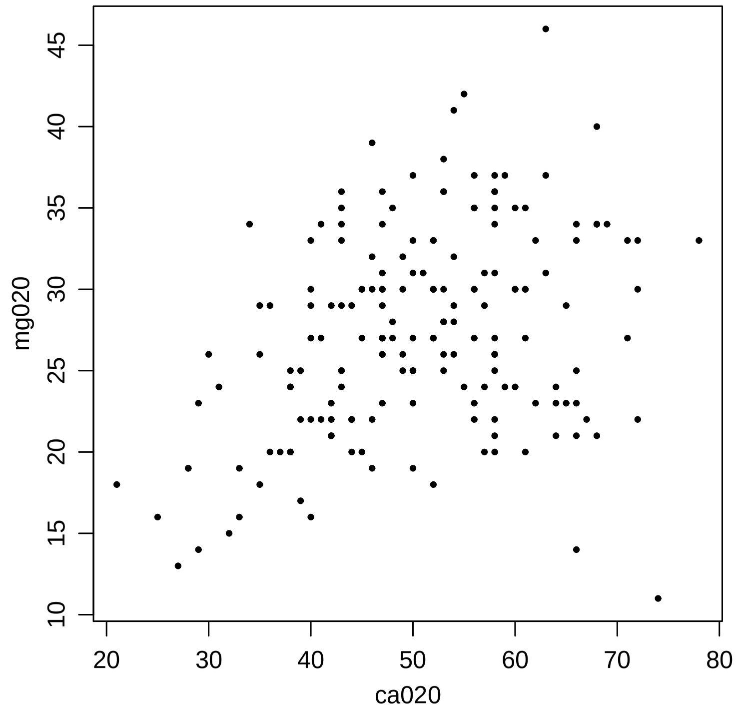
Correlation between prevalence and green-ness of vegetation?

Example 1.4: Soil data



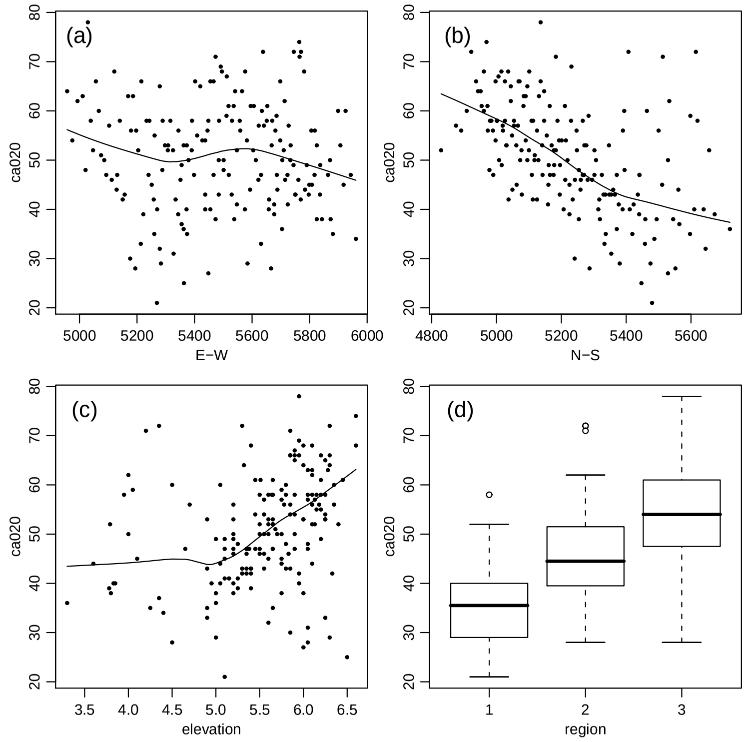
Ca (left-panel) and Mg (right-panel) concentrations

Example 1.4: Continued



Correlation between local Ca and Mg concentrations.

Example 1.4: Continued



Covariate relationships for Ca concentrations.

Terminology and Notation

- data format: (x_i, y_i)
- x_i fixed or stochastically independent of Y_i
- $S(x)$ is the assumed underlying stochastic process
- model specification: $[S, Y] = [S][Y|S]$

Note: $[\cdot]$ denotes *the distribution of*

Support

- x_i is in principle a point, but sometimes measurements are taken on (maybe small) portions
- revisiting the examples (e.g. elevation and rongelap) we can see contrating situations
- $S(x) = \int w(r)S^*(x-r)dr$
- smoothness of $w(s)$ contrains allowable forms for the correlation function
- support *vs* data from discrete spatial variation

Multivariate responses and covariates

- $Y(x_i)$ can be a vector of observable variable
- not necessarily measurements are taken at coincident locations
- data structure (x_i, y_i, d_i) can include covariates (potential explanatory variables)
- jargon: *external trend* and *trend surface* (coordinates or functions of them as covariates)
- distinction between multivariate responses and covariates is not aways sharp and pragmatically, it may depend on the objectives and/or availability of data
- revisiting examples

Sampling design

- uniform *vs* non-uniform
  - coverage of the area
  - estimation of spatial correlation
  - practical constraints
- preferential *vs* non-preferential
  - effects on inference
  - marked point process

Scientific objectives

- **Estimation: inference on model parameters**
- **Prediction: inference on the process ($S(x)$ or some functional of it)**
- **Hypotesis testing (typically not a main concern)**

Generalised linear models

- **GLM's and marginal and mixed models**
- **GLGM: Generalized linear geostatistical models**
- **ingredients:**
  1. **a Gaussian process $S(x)$, the *signal***
  2. **data generating mechanism given the signal**
  3. **relation to explanatory variables**

$$h(\mu_i) = S(x_i) + \sum_{k=1}^{p} \beta_k d_k(x_i)$$

- **Gaussian and other models - revisiting examples**

Model-based Geostatistics

- the application of general principles of statistical modelling and inference to geostatistical problems

- Example: kriging as minimum mean square error prediction under Gaussian modelling assumptions

Final remarks

- course (book) structure

- statistical (pre)-requisites:
  - exploratory analysis, regression, statistical modelling and inference.
  - likelihood and Bayesian inference.
  - computational methods, including MCMC.

- computation (geoR and geoRglm)
  - R software (http://www.r-project.org)
  - geoR (http://www.leg.ufpr.br/geoR) and geoRglm (http://www.leg.ufpr.br/geoRglm) packages

Some computational resources

- **geoR package**:
  http://www.leg.ufpr.br/geoR
- **geoRglm package**:
  http://www.leg.ufpr.br/geoRglm
- **R-project**:
  http://www.R-project.org
- **CRAN spatial task view**:
  http://cran.r-project.org/src/contrib/Views/Spatial.html
- **AI-Geostats web-site**:
  http://www.ai-geostats.org

SESSION 3

An overview of model based geostatistics

Aims

- an overview of a *canonical* geostatistical analysis
- highlighting basic concepts, model features, results to be obtained
- steps of a typical data analysis
- using the *surface elevation data* as a running example

Design

- what and where to address questions of scientific interest
- *elevation data*: map the true surface
- how many: sample size
  - statistical criteria
  - but typically limited by pratical contraints: time, costs, operational issues, etc
- where: design locations
  - completely *random vs* completely *regular*
  - different motivations, need to compromise
  - *oportunistic* designs: concerns about preferential sampling and impact on inferences

Model formulation

- here ”unusually” before EDA (observational data)
- just a basic reference model antecipating issues for EDA
- *elevation data*: best guess of the true underlying surface from the available sparse data
- scientific reasoning: continuity and differentiability
- measurement process: distinction between $S(x)$ and $Y(x)$

A basic reference model

Gaussian geostatistics

The model:

- $[Y, S] = [S][Y|S]$
- Stationary Gaussian process $S(x) : x \in \mathbb{R}^2$
  - $\mathrm{E}[S(x)] = \mu$
  - $\mathrm{Cov}\{S(x), S(x')\} = \sigma^2 \rho(\|x - x'\|)$
- Mutually independent $Y_i|S(\cdot) \sim \mathrm{N}(S(x_i), \tau^2)$

Equivalent to:

$$Y(x) = S(x) + \epsilon$$

Correlation function

- core of the spatially continuous models
- $\rho(u)$ is positive definite
  (any $\sum_{i=1}^{m} a_i S(x_i)$ has a non-negative variance)
- here $u \geq 0$, symmetric
- typically assuming a parametric form for $\rho(\cdot)$
- The Mátern class

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

- stationarity assumption

Some possible extensions

- transformation of the response variable (Box-Cox)

$$Y^* = \begin{cases} (Y^{\lambda} - 1)/\lambda & : \lambda \neq 0 \\ \log Y & : \lambda = 0 \end{cases}$$

- non-constant mean model (covariates or trend surface)
- more general covariance functions
- non-stationary covariance structure

word of caution: decision on one will probably affect the other

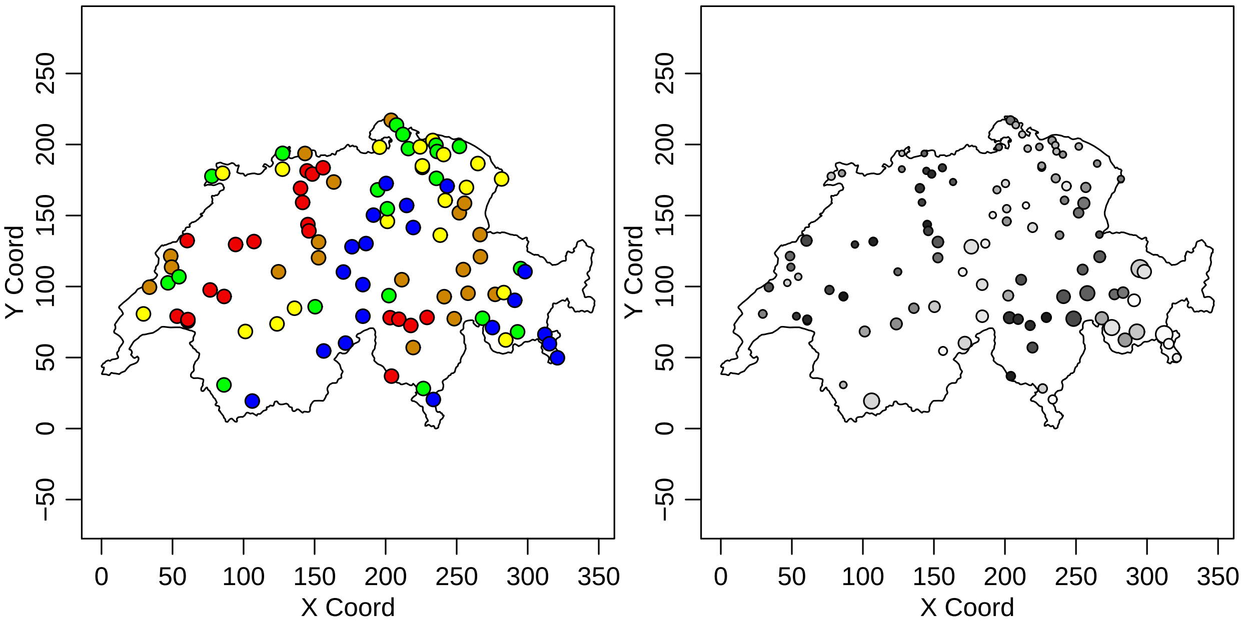
Exploratory data analysis

- **non-spatial *vs* spatial**
- **Non-spatial**
  - **outliers**
  - **non-normality**
  - **arbitrary mean model: choice of potential covariates**

Spatial EDA - some tools and issues

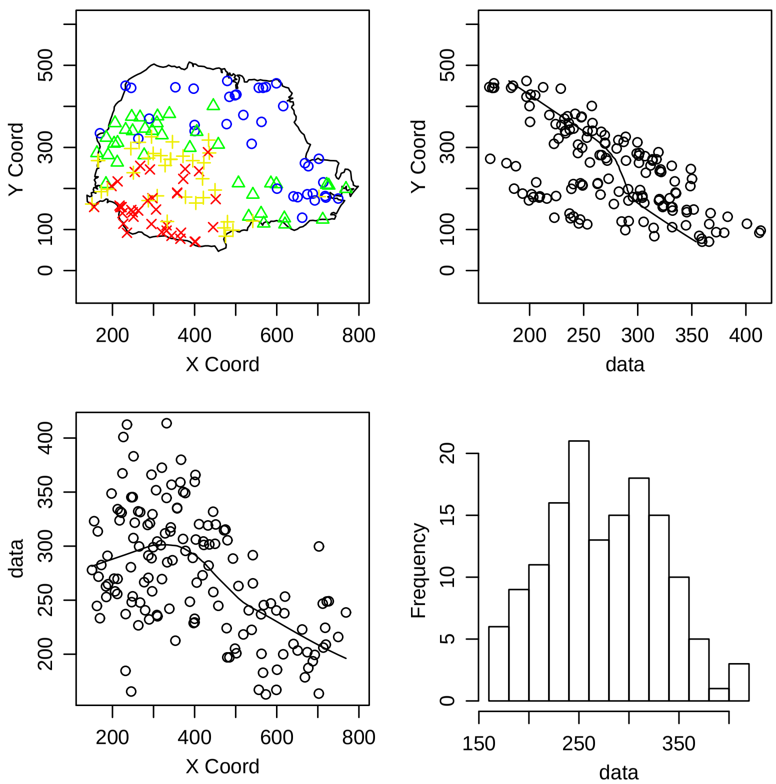
- **spatial outliers**
- **trend surfaces (scatterplots against covariates)**
- **other potential spatial covariates**
- **GIS tools**

Circle plot

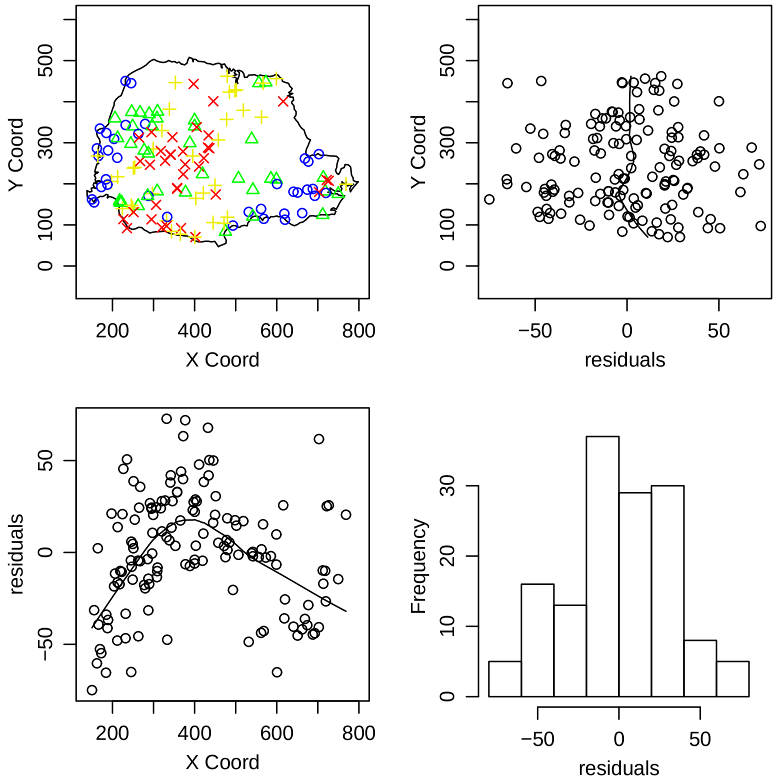


Two possible visualisations: left – data values divided in quintiles, right – gray shade proportional to data, circle sizes proportional to a covariate value (elevation).

A quick exploratory display



Residual plots



Comments

- spatially varying mean *vs* correlation in the response variables around the mean

- keep it simple!

- likelihood based methods for model choice

Variograms

- Theoretical variogram (for cte mean)

  $$2V(u) = \mathrm{Var}\{Y(x_i) - Y(x_j)\} = \mathrm{E}\{[Y(x_i) - Y(x_j)]^2\}$$

- Empirical (semi-)variogram: $\hat{V}(u)$

- biased for non-constant mean

- higher order polynomials *vs* spatial correlation

- Monte Carlo envelopes for empirical variograms

An aside: distinction between parameter estimation and spatial prediction

- assume a set of locations $x_i : i = 1, \ldots, n$ on a lattice covering the area

- interest: average level of pollution over the region

- consider the sample mean:

$$\bar{S} = n^{-1} \sum_{i=1}^{n} S_i$$

...

- within a parameter estimation problem
  - estimator of the constant mean parameter $\mu = \mathrm{E}[S(x)]$
  - precision given by the M.S.E. $\mathrm{E}\{[(\bar{S}-\mu)^2]\}$
  - $\mathrm{Var}[\bar{S}] = n^{-2}\sum_{i=1}^{n}\sum_{i=1}^{n}\mathrm{Cov}(S_i, S_j) \geq \sigma^2/n$

- within a prediction problem
  - predictor of the spatial average $S_A = |A|^{-1}\int_A S(x)dx$
  - precision given by the M.S.E. $\mathrm{E}[(\bar{S}-S_A)^2]$, S_A is r.v
  - precision (can even approach zero) given by

$$\begin{aligned}
\mathrm{E}[(\bar{S}-S_A)^2] &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(S_i, S_j) \\
&+ |A|^{-2}\int_A\int_A \mathrm{Cov}\{S(x), S(x')\}dxdx' \\
&- 2(n|A|)^{-1}\sum_{i=1}^{n}\int_A \mathrm{Cov}\{S(x), S(x_i)\}dx.
\end{aligned}$$

Inference

- parameter estimation: likelihood based methods (other approaches are also used)

- spatial prediction: *simple kriging*

$$\hat{S}(x) = \mu + \sum_{i=1}^{n} w_i(x)(y_i - \mu)$$

- straightforward extension for $\mu(x)$

- Parameter uncertainty?
  usually ignored in traditional geostatistics
  (plug-in prediction)

Summary(I): Notation

- $(Y_i, x_i) : i = 1, \ldots, n$ basic format for geostatistical data
- $\{x_i : i = 1, \ldots, n\}$ 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

Summary(II):A canonical geostatistical data analysis

Basic steps:

- exploratory data analysis
- model choice
- inference on the model parameters
- spatial prediction

Assumptions:

- stationarity (translation) global mean, variance and spatial correlation
- isotropy (rotation)
- Gaussianity

— demo 01 —

Summary(III):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

SESSION 4

Linear (Gaussian) geostatistical models

Opening remarks

- Gaussian stochastic process are widely used
- physical representation, behaviour and *tractability*
- underlying structure many geoestatistical methods
- benchmark for hierarquical models

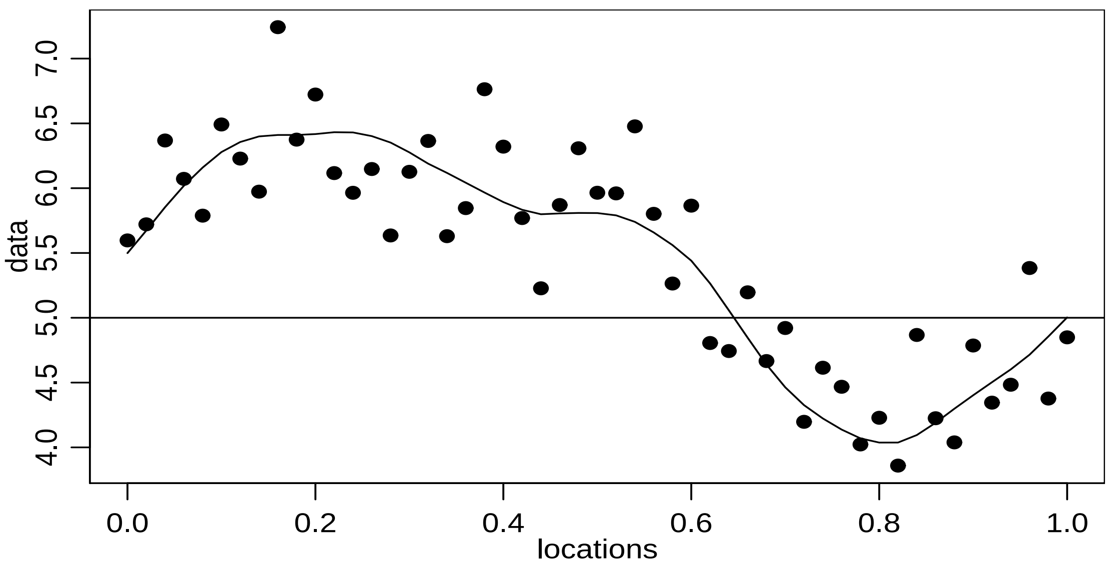
This section focus on characterization, properties and simulations of Gaussian models.

The reference model

- Equivalent model formulation for the Gaussian model

$$Y_i = \mu + S(x_i) + Z_i$$

- Schematic representation in 1-D:



Covariance function

- The assumed stationary Gaussian spatial process $S(x)$ is fully specified by:
  - the mean function $\mu = \mathrm{E}[S(x)]$
  - the covariance function $\mathrm{Cov}\{S(x), S(x')\} = \sigma^2 \rho(x, x')$
- A symmetric function $\mathrm{Cov}(\cdot)$ is positive definite if

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \,\mathrm{Cov}(x_i - x_j) \geq 0$$

  for all $a_i \in \mathbb{R}$, $x_i \in \mathbb{R}^d$ and $n \in \mathbb{N}$.

- a function $\mathrm{Cov}(\cdot) : \mathbb{R} \to \mathbb{R}$ is a valid covariance function iff $\mathrm{Cov}(\cdot)$ is positive definite
- geostatistics uses covariance functions to characterise spatial processes

Properties

1. $\mathrm{Cov}[Z(\mathrm{x}), Z(\mathrm{x}+0)] = \mathrm{Var}[Z(\mathrm{x})] = \mathrm{Cov}(0) \geq 0$
2. $\mathrm{Cov}(\mathrm{u}) = \mathrm{Cov}(\text{-}\mathrm{u})$
3. $\mathrm{Cov}(0) \geq | \mathrm{Cov}(\mathrm{u}) |$
4. $\mathrm{Cov}(\mathrm{u}) = \mathrm{Cov}[Z(\mathrm{s}), Z(\mathrm{x}+\mathrm{u})] = \mathrm{Cov}[Z(0), Z(\mathrm{u})]$
5. If $\mathrm{Cov}_j(\mathrm{u})$, $j = 1, 2, \ldots, k$, are valid cov. fc. then $\sum_{j=1}^{k} b_j \,\mathrm{Cov}_j(\mathrm{u})$ is valid for $b_j \geq 0 \forall j$
6. If $\mathrm{Cov}_j(\mathrm{u})$, $j = 1, 2, \ldots, k$, are valid cov. fc. then $\prod_{j=1}^{k} \mathrm{Cov}_j(\mathrm{u})$ is valid
7. If $\mathrm{Cov}(\mathrm{u})$ is valid in R^d, then is also valid in R^p , $p < d$

Smoothness

- A formal 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,

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

- $S(x)$ is *mean square differentiable* if there exists a process $S'(x)$ such that, for all x,

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

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

- 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 $u = 0$.

Spectral representation

Bochener Theorem (iff):

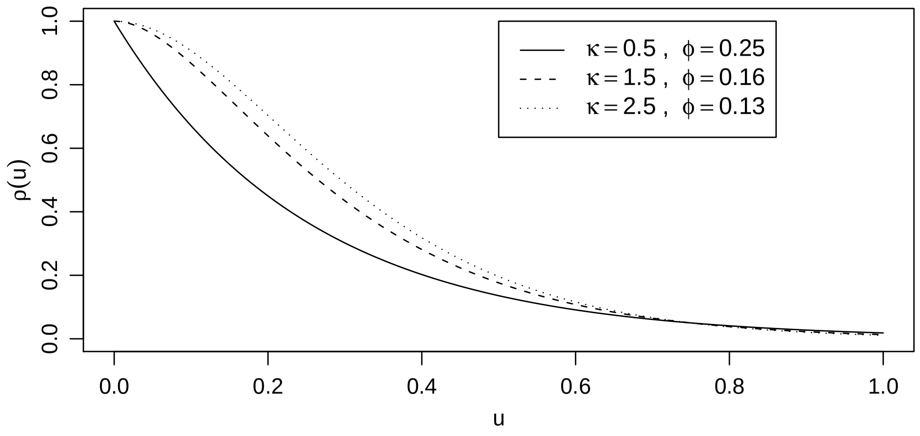
$$\mathrm{Cov}(u) = \int_{-\infty}^{+\infty} \exp\{iu\} s(w) dw$$

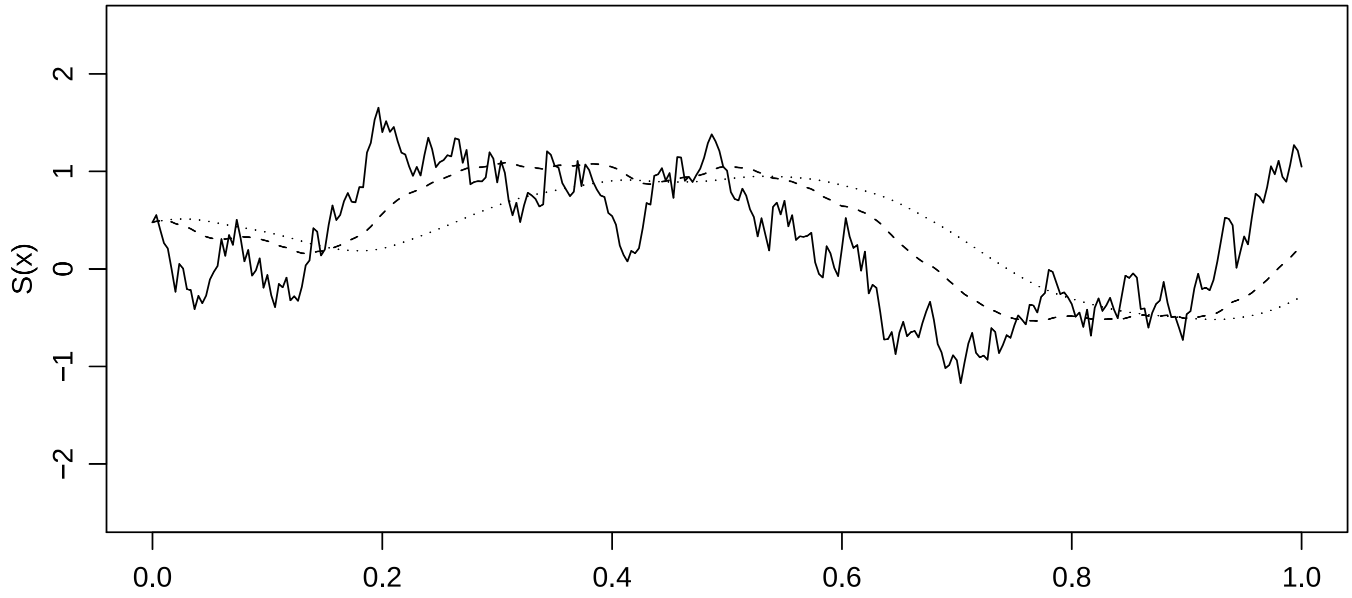
- $s(w)$ is the *spectral density function*
- $\mathrm{Cov}(u)$ and $s(w)$ form a Fourier pair (the latter can be expressed as a function of the former)
- provided an alternative way to estimate covariance structure from the data using *periodogram* – $\hat{s}(w)$
- provide ways for testinf valid covariance functions and/or to derive new ones

The Matérn family of correlation functions

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

- parameters $\kappa > 0$ (smoothness of $S(x)$) and $\phi > 0$ (extent of the spatial correlation)
- $K_{\kappa}(\cdot)$ denotes modified Bessel function of order κ
- for $\kappa = 0.5$, $\rho(u) = \exp\{-u/\phi\}$: exponential corr. fct.
- for $\kappa \rightarrow \infty$ $\rho(u) = \exp\{-(u/\phi)^2\}$: Gaussian corr. fct.
- κ and ϕ are not orthogonal
  - scale parameters ϕ are not comparable for different orders κ of the Matérn correlation function
  - reparametrisation: $\alpha = 2\phi\sqrt{\kappa}$
  - effects on parameter estimation





Notes

- many other are proposed in the literature
- correlation functions are typically, but not necessarily, decreasing functions
- models valid in d dimentions are valid for lower but not necessarily higher dimensions. Matérn is valid in 3-D.
- Matérn models are $\lceil \kappa - 1$ times differentiable. For the example, $\kappa = 0.5$, 1.5 and 2.5 correspond to processes mean square continuous, once and twice differentiable.
- Whittle (1954) proposed a special case with $\kappa = 1$
- for monotonic models, the pratical range is defined as the distance where the correlation is 0.05.
- we assume here punctual support for all data. For different supports (mis-aligned data) regularization is needed.

Properties of the process (revisited)

- Strict stationarity
- Weak (second-order, covariance) stationarity
- isotropy

Variogram representations

- *intrinsic* stationarity (intrinsic random functions, Matheron,1973)
- validity (Gneiting, Sasvári and Schlather, 2001)

Simulating from the model

- For a finite set of locations x, $S(x)$ is multivariate Gaussian.
- A "standard" way for obtaining (unconditional) simulations of $S(x)$ is:
  - define the locations
  - define values for model parameters
  - compute Σ using the correlation function
  - obtain $\Sigma^{1/2}$, e.g. by Cholesky factorization of singular value decomposition
  - obtain simulations $S = \Sigma^{1/2} Z$ where Z is a vector of normal scores.

Simulating from the model (cont.)

- **Large simulations are often need in practice and require other methods, e.g.:**
  - **Wood and Chan (1994) – fast fourier transforms**
  - **Rue and Tjelmeland (2002) – approximation by Markov Gaussian Random Fields**
    Gibbs scheme using approximated sparse $(n-1) \times (n-1)$ full conditionals (GMRFlib)
  - **Schlather (2001) – package *RandomFields* : implements a diversity of methods (circulant embedding, turning bands, etc)**

— demo 02 —

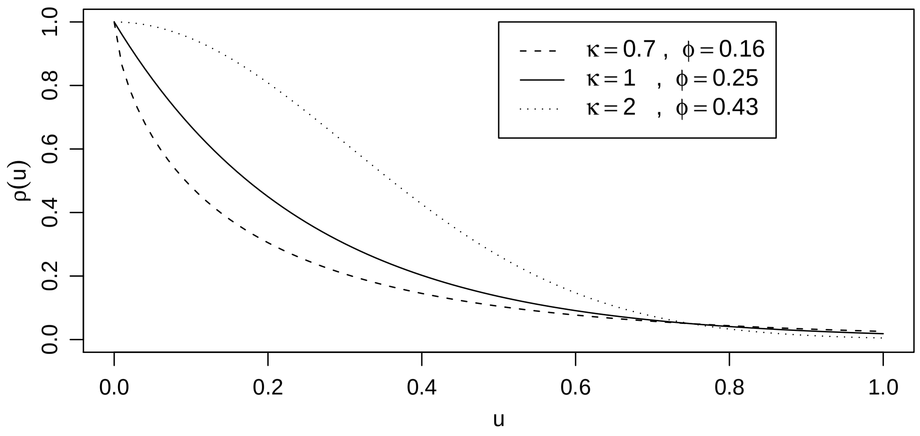
SESSION 5

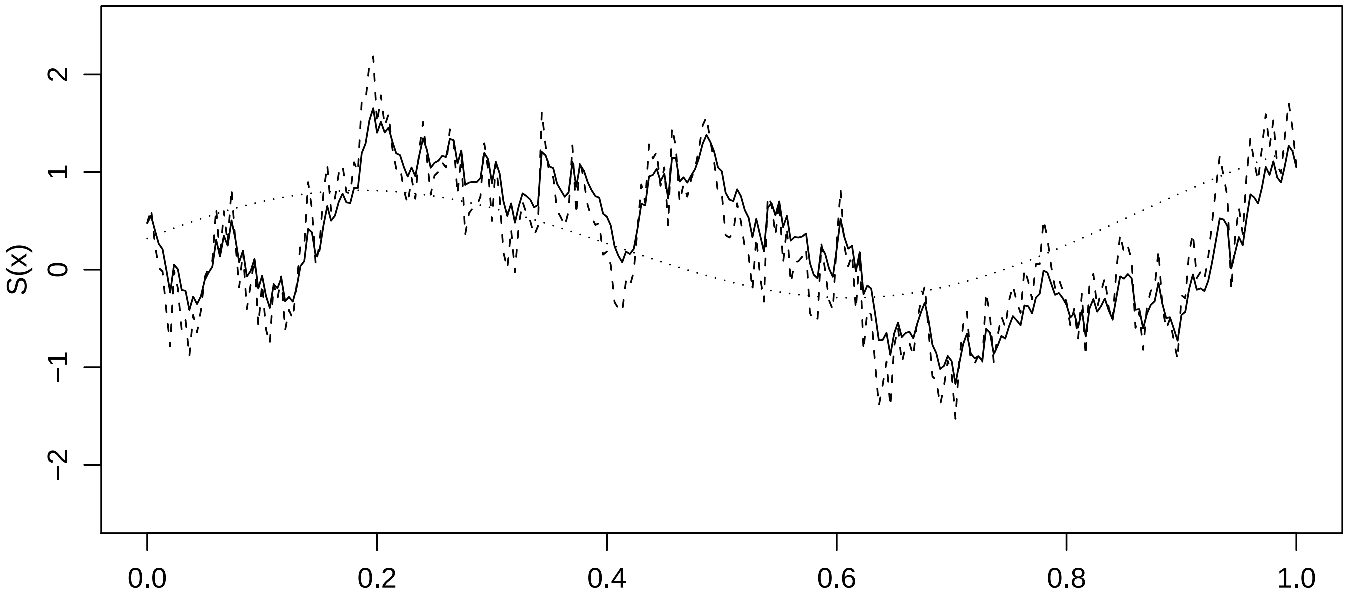
Linear (Gaussian) geostatistical models (cont.)

Other families (I): powered exponential

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

- **scale parameter ϕ and shape parameter κ**
- **non-orthogonal parameters**
- $0 < \kappa \leq 2$
- **non-differentiable for $\kappa < 2$ e infinitely dif. for $\kappa = 2$**
- **asymptotically behaviour (pratical range)**





Other families (II): spherical model

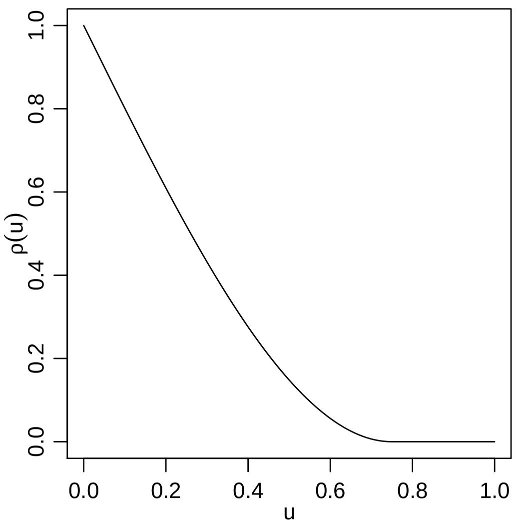
$$\rho(u) = \begin{cases} [1 - 1.5(u/\phi) + 0.5(u/\phi)^3] & \text{for} \quad 0 \le u \le \phi \\ 0 & \text{for} \quad u > \phi \end{cases}$$

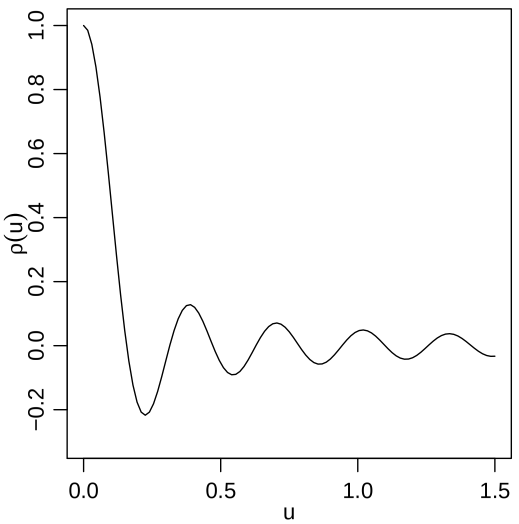
- finite range ϕ
- non- differentiable at origin
- only once differentiable at $u = \phi$
- potential difficulties for MLE
- overlapping volume between two spheres

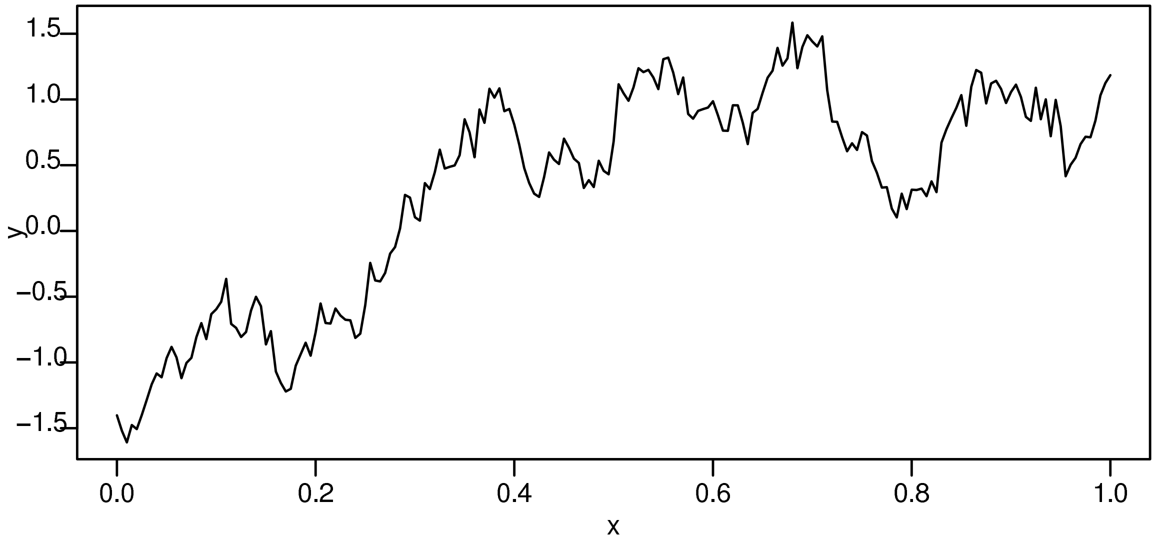
Other families (III): wave model

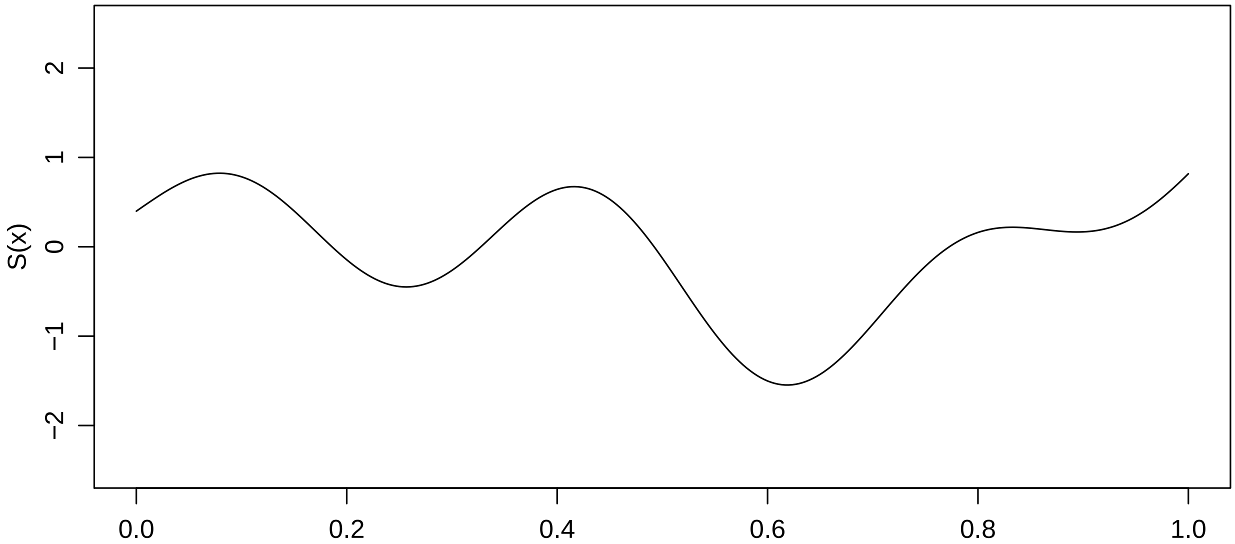
$$\rho(u) = (u/\phi)^{-1}\sin(u/\phi)$$

- non-monotone
- oscilatory behaviour reflected in realisations









The *nugget* effect

- **discontinuity at the origin**
- **interpretations**
  - $\mathrm{Var}[\boldsymbol{Y}|\boldsymbol{S}]$
  - **measurement error ($\boldsymbol{Y}$)**
  - **micro-scale variation ($\boldsymbol{S}$)**
  - **combination of both**
- **importance for sampling design**
- **usually indistinguishable**
- **except repeated measurements at coincident locations**
- **impact on predictions and their variance**

Spatial trends

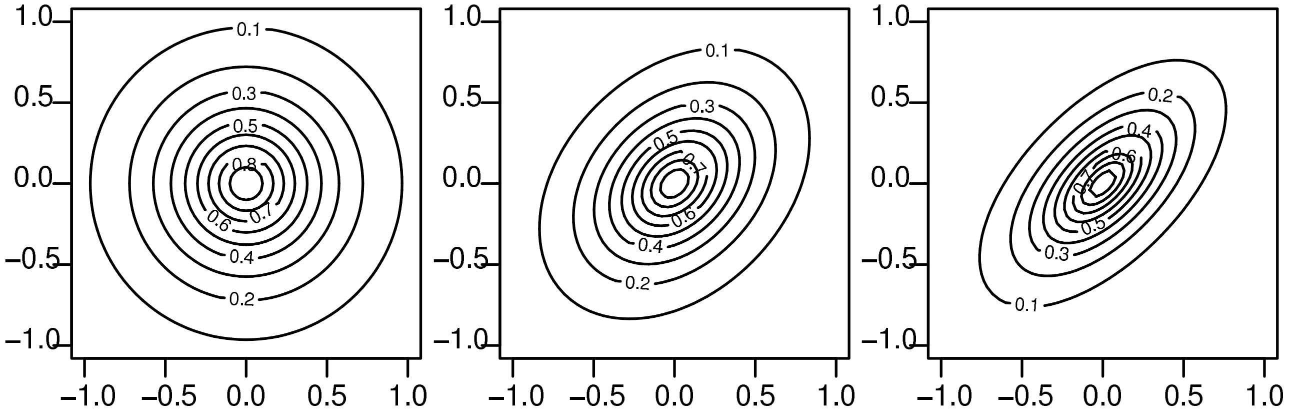
- term reffers to a variable mean function $\mu(x)$
- trend surface and covariates
- deterministic *vs* stochastic: interpretation of the process
- exploratory analysis: possible non-linear relations

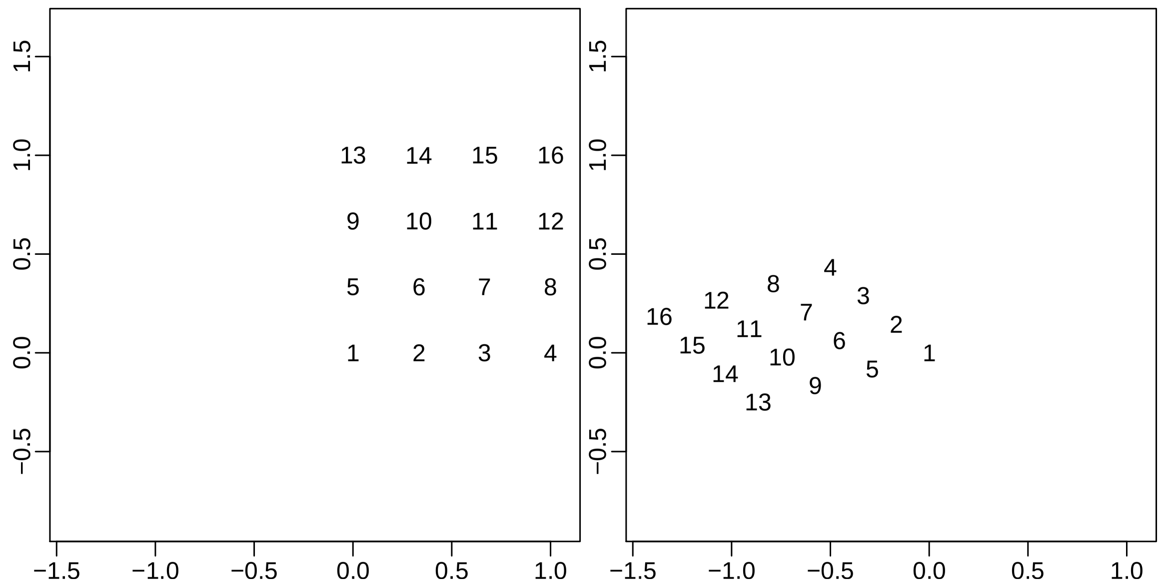
Directional effects

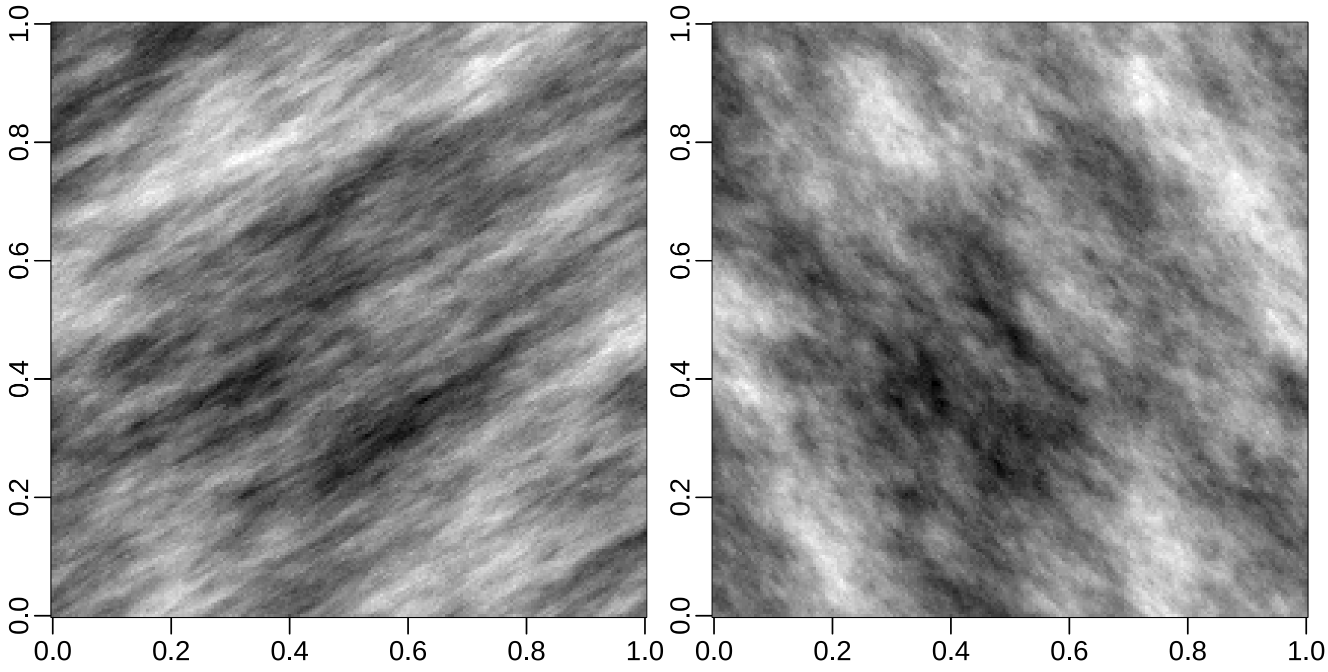
- environmental conditions wind, flow, soil formation, etc) can induce directional effects
- non-invariant properties of the cov. function under *rotation*
- simplest model: geometric anisotropy
- new coordinates by rotation and stretching of the original coordinates:

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

- add two parameters to the covariance function
- (ψ_A, ψ_R) anisotropy angle and ratio parameters







Realisations of a geometrically anisotropic Gaussian spatial processes whose principal axis runs diagonally across the square region with anisotropy parameters $(\pi/3, 4)$ for the left-hand panel and $(3\pi/4, 2)$ for the right-hand panel.

Non-stationary models

- Stationarity is a convenient working assumption, which can be relaxed in various ways.
  - Functional relationship between mean and variance: sometimes handled by a data transformation
  - Non-constant mean: replace constant μ by

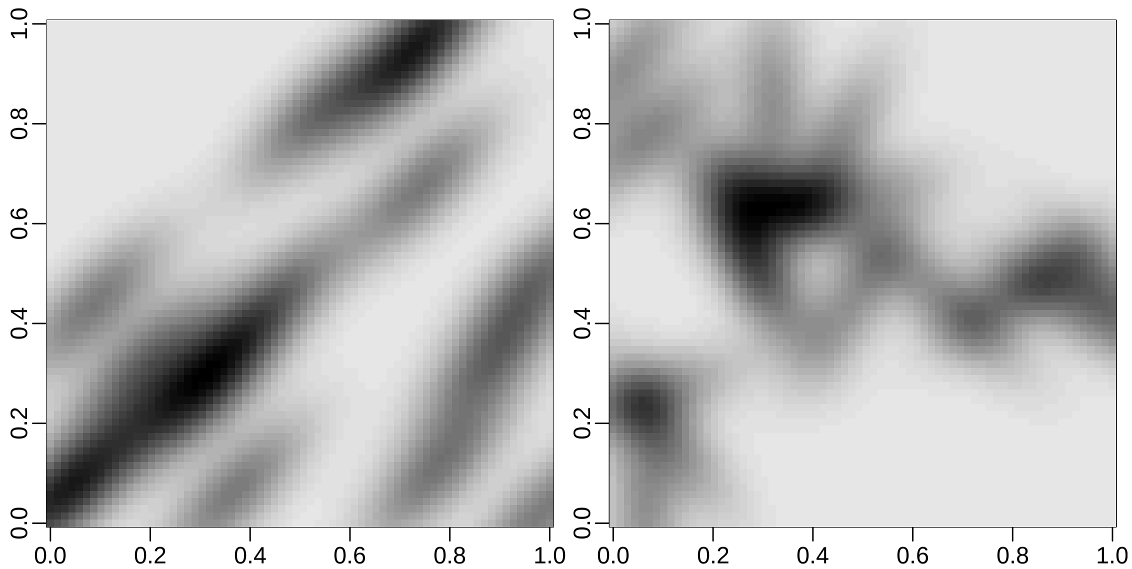
$$\mu(x) = F\beta = \sum_{j=1}^{k} \beta_j f_j(x)$$

Non-stationary models (cont.)

- 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 complex transformations of the geographical space, x.
  - spatial convolutions (Higdon 1998, 2002; Fuentes e Smith)
  - low-rank models (Hastie, 1996)
  - non-Euclidean distances (Rathburn, 1998)
  - locally directional effects

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

An illustration



Globally (left) and locally (right) directional models

Other topics

- transformed Gaussian models
- non-Gaussian (GLM) models (Gotway & Stroup, 1997 ; Diggle, Tawn & Moyeed, 1998)
- unconditional and conditional simulations
- decomposing the error term Z ("nugget effect"): Z = short scale variation + measurement error
- multivariate models
  - second order properties
  - constructions

Constructing multivariate models

One example: A common-component model

- assume independent processes $S_0^*(\cdot)$, $S_1^*(\cdot)$ and $S_2^*(\cdot)$
- Define a bivariate process $S(\cdot) = \{S_1(\cdot), S_2(\cdot)\}$
- $S_j(x) = S_0^*(x) + S_j^*(x) : j = 1, 2.$
- $S(\cdot)$ is a valid bivariate process with covariance structure

$$\mathrm{Cov}\{S_j(x), S_{j'}(x-u)\} = \mathrm{Cov}_0(u) + I(j = j')\,\mathrm{Cov}_j(u)$$

- for different units it requires an additional scaling parameters so that $S_{0j}^*(x) = \sigma_{0j} R(x)$ where $R(x)$ has unit variance.

More general contructions are presented by Chilès and Delfiner, 1999; Gelfand, Schmidt, Banerjee & Sirmans (2004), Schmidt & Gelfand (2003)

SESSION 6

Parameter Estimation

Opening remarks

- The canonical problem is spatial prediction of the form

$$\hat{S}(x) = \mu(x) + \sum_{i=1}^{n} w_i(x)(y_i - \mu(x))$$

- The prediction problem can be tackled by adopting some criteria (e.g. minimise MSPE)

$$MSPE(\hat{T}) = \mathrm{E}\left[(T - \hat{T})^2\right] \text{ e.g. above } T = S(x)$$

- However this requires knowledge about model parameters

- infer first and second-moment properties of the process from the available data

First moment properties (trend estimation)

- The OLS estimator

$$\tilde{\beta} = (D'D)^{-1}D'Y$$

  is unbiased irrespective the covariance structure (assuming the model is correct)

- A more efficient GLS estimator:

$$\hat{\beta} = (D'V^{-1}D)^{-1}D'V^{-1}Y$$

  - unbiased
  - smaller variance
  - MLE
  - requires knowledge about covariance parameters

- for non-cte mean, OLS residuals can inform about covariance structure

$$R = Y - D\tilde{\beta}$$

- strategies: two stages (which can be interactive) or joint estimation

Second order properties

Under the assumed model, for $u = ||x_i - x_j||$

- **Variances and covariances:**

$$\mathrm{Var}\left[Y(x)\right] = \tau^2 + \sigma^2 \quad \mathrm{Cov}\left[Y(x_i), Y(x_j)\right] = \sigma^2 \rho(||u||)$$

- **The (theoretical) variogram**

$$\mathrm{V}(x_i, x_j) = \mathrm{V}(u) = \frac{1}{2} \mathrm{Var}\left\{S(x_i) - S(x_j)\right\}$$

- **Under stationarity**

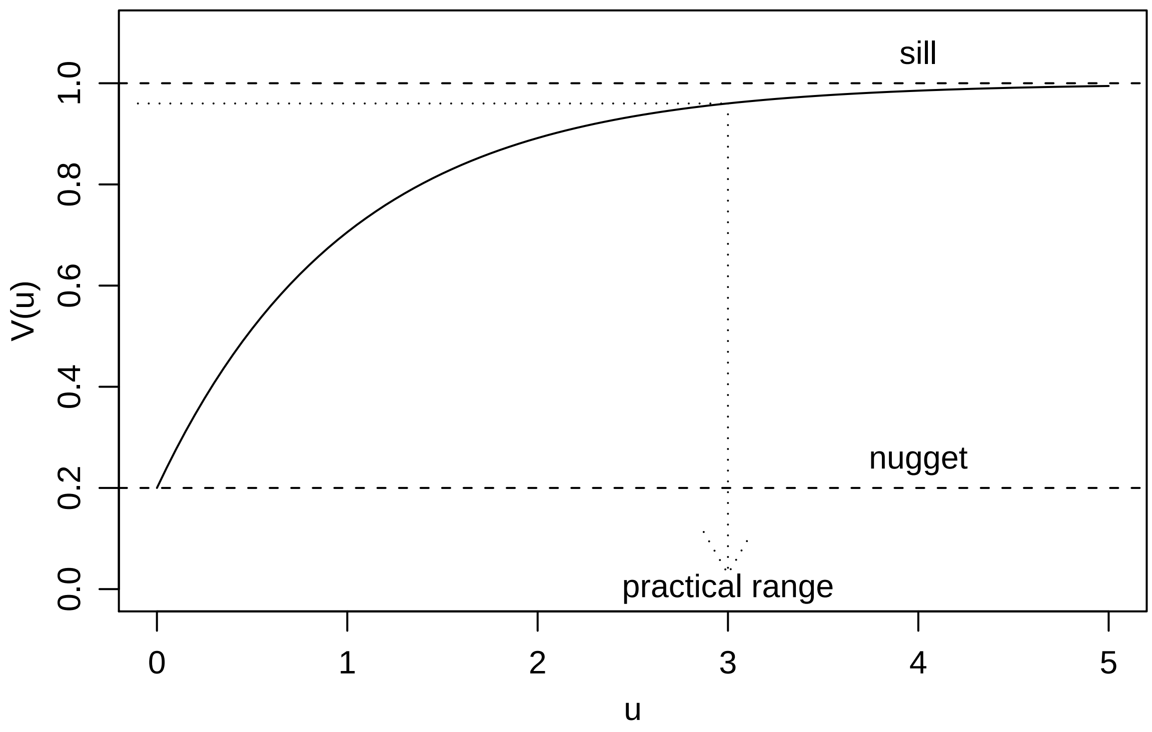
$$V(u) = \tau^2 + \sigma^2\{1 - \rho(u)\}$$

- **So, the theoretical variogram is a function which sumarises all the properties of the process**

Terminology

- *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** $\rho_0(u) = \rho(u/\phi)$
- *the pratical range*: u_0, **such**
  - $\rho(u) = 0$ **(finite range correlation models)**
  - $\rho(u) = 0.95\sigma^2$ **(correlation functions approaching zero asymptotically)**
  - **or, in terms of variogram** $\mathrm{V}(u) = \tau^2 + 0.95\sigma^2$
  - ***this is just a practical convention!***

Schematic representation



Paradigms for parameter estimation

- Ad hoc (variogram based) methods
  - compute an empirical variogram
  - fit a theoretical covariance model
- Likelihood-based methods
  - typically under Gaussian assumptions
  - more generally needs MCMC or approximations
  - Optimal under stated assumptions, robustness issues
  - full likelihood not feasible for large data-sets
  - variations on the likelihood function (*pseudo-likelihoods*)
- Bayesian implementation, combines estimation and prediction

Empirical variograms

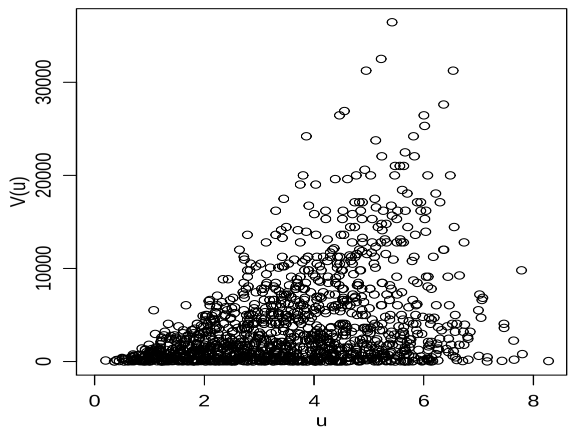
- The theoretical variogram suggests an empirical estimate of $V(u)$:

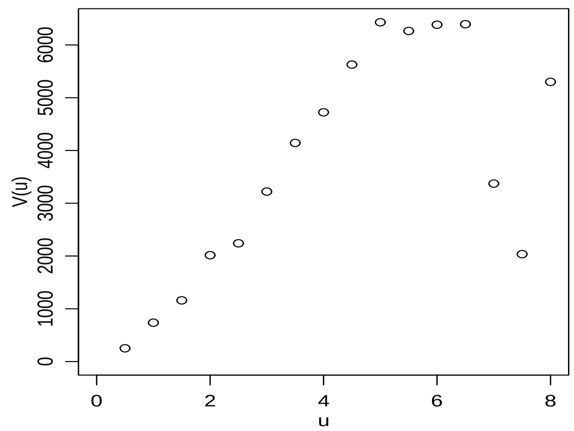
$$\hat{V}(u_{ij}) = \text{average}\{0.5[y(x_i) - y(x_j)]^2\} = \text{average}\{v_{ij}\}$$

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

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

- the empirical variogram is derived from the variogram cloud by averaging within bins: $u - h/2 \leq u_{ij} < u + h/2$





- *sample variogram* ordinates V_k; $(k-1)h < u_{ij} < kh$
- convention $u_k = (k - 0.5)h$ (mid-point of the interval)
- may adopt distinct h_k
- excludes zero from the smallest bin (deliberate)
- typically limited at a distance $u < u_{max}$

Variations on empirical variograms

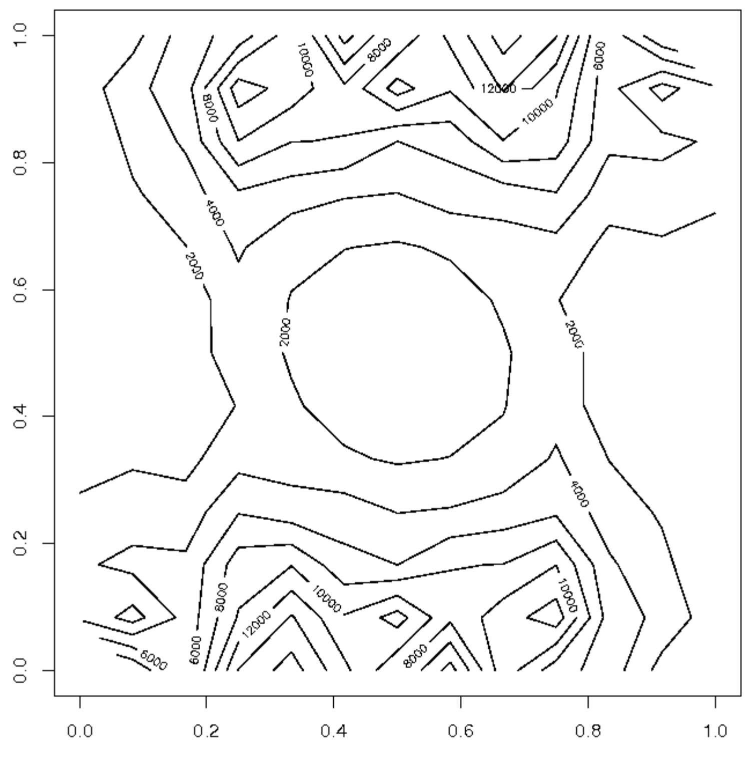
- for a process with non-constant mean (covariates) replace $y(x_i)$ by residuals $r(x_i) = y(x_i) - \hat{\mu}(x_i)$ from a trend removal

- usage of kernel or spline smoothers, however notice $\frac{1}{2}n(n-1)$ points are not independent

- may not be worth the trouble (bandwidth issues, etc) considering exploratory purposes

- a diversity of alternative estimators is available

Exploring directional effects



Difficulties with empirical variograms

- $v_{ij} \sim V(u_{ij})\chi_1^2$

- the v_{ij} are correlated

- the variogram cloud is therefore unstable, both pointwise and in its overall shape

- binning removes the first objection to the variogram cloud, but not the second

- is sensitive to mis-specification of $\mu(x)$

Variogram model fitting

- fitting a typically non-linear variogram function (as e.g. the Matérn) to the empirical variogram provides a way to estimate the models parameters.

- e.g. a weighted least squares criteria minimises

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

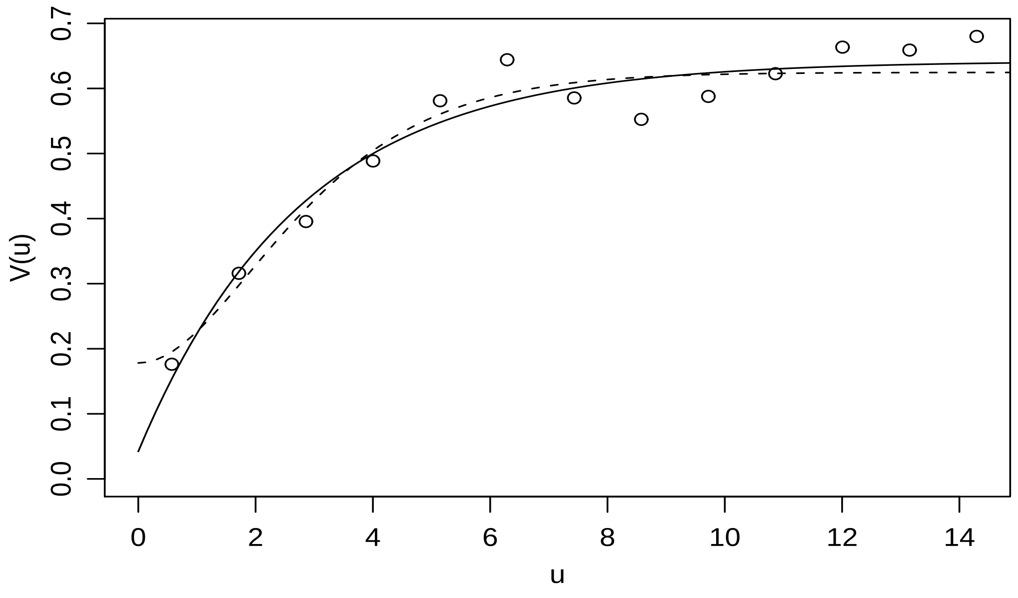
  where θ denotes the vector of covariance parameters and $\bar{V}_k$ is average of n_k variogram ordinates v_{ij}.

- in practice u is usually limited to a certain distance

- variations includes:
  – fitting models to the variogram cloud
  – other estimators for the empirical variogram
  – different proposals for weights
  – ...

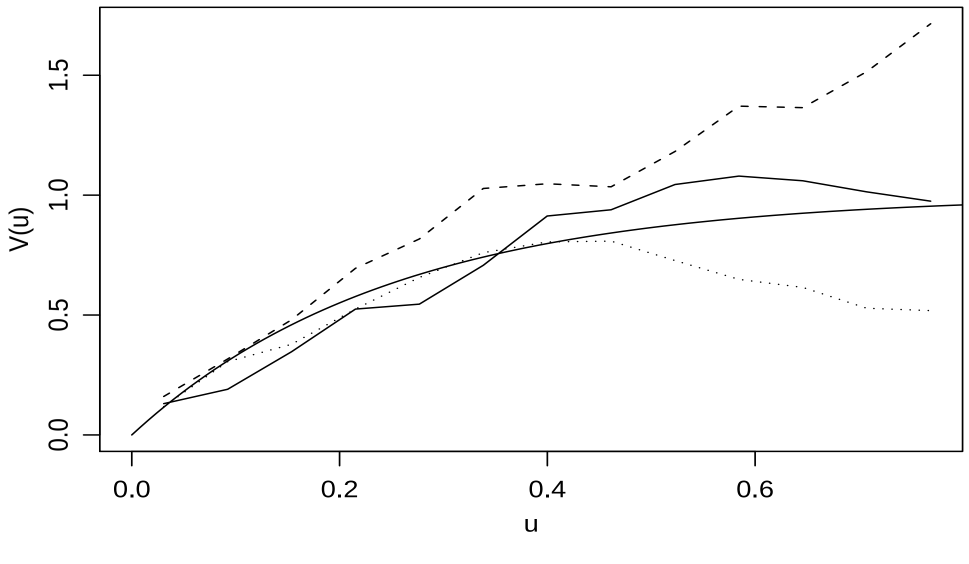
Comments on variograms - I

- equally good fits for different ”extrapolations” at origin



Comments on variograms - II

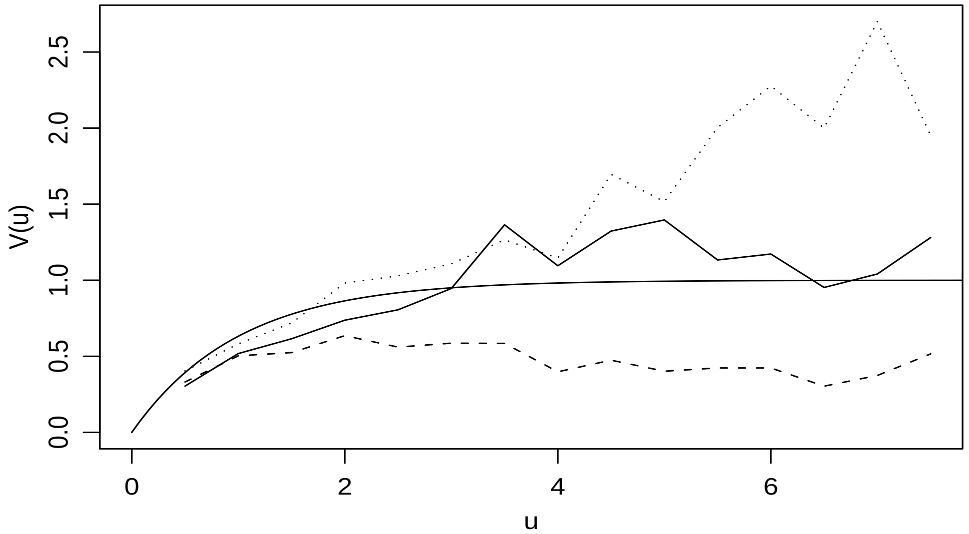
- correlation between variogram points points



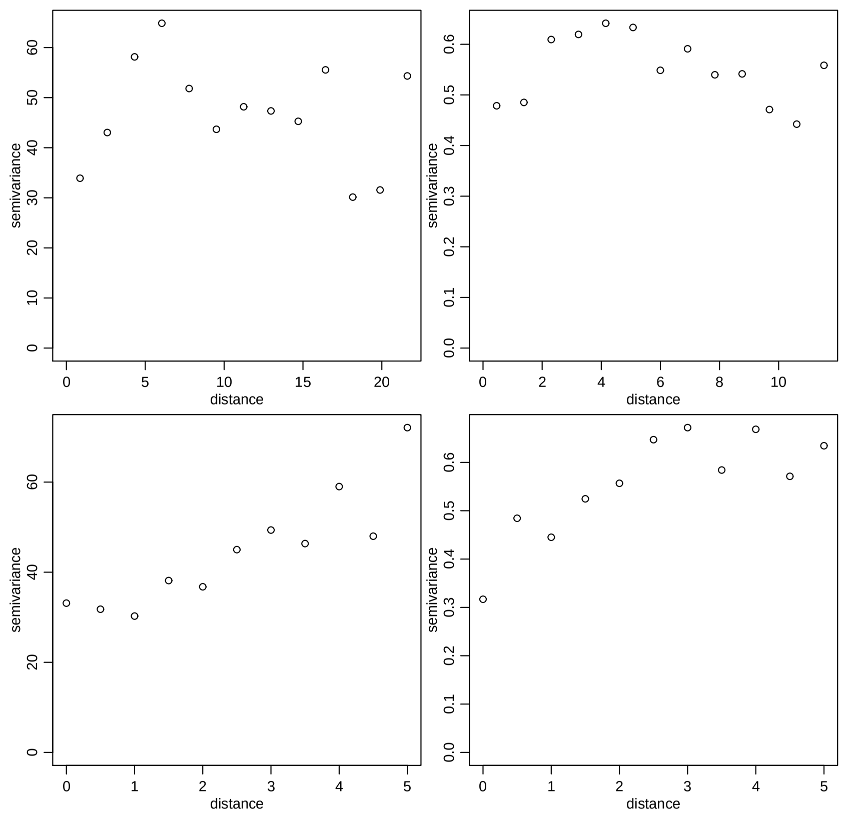
Empirical variograms for three simulations from the same model.

Comments on variograms - III

- sensitivity to the specification of the mean
- solid smooth line: true model, dotted: empirical variogram, solide: empirical variogram from true residuals, dashed: empirical variogram from estimated residuals.



Comments on variograms - IV



Computing variograms

— demo on variograms —

Parameter estimation: maximum likelihood

For the basic geostatistical model

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

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.

Or more generally for

$$\begin{aligned} S(x_i) &= \mu(x_i) + S_c(x_i) \\ \mu(x_i) &= D\beta = \sum_{j=1}^{k} f_k(x_i)\beta_k \end{aligned}$$

where $d_k(x_i)$ is a vector of covariates at location x_i

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

The likelihood function is

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

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

$$\begin{aligned} \hat{\beta}(V) &= (D'V^{-1}D)^{-1}D'V^{-1}y \\ \hat{\sigma}^2(V) &= n^{-1}(y - D\hat{\beta})'V^{-1}(y - D\hat{\beta}) \end{aligned}$$

- concentrated likelihood: substitute (β, σ^2) by $(\hat{\beta}, \hat{\sigma^2})$ and the maximisation reduces to

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

Some technical issues

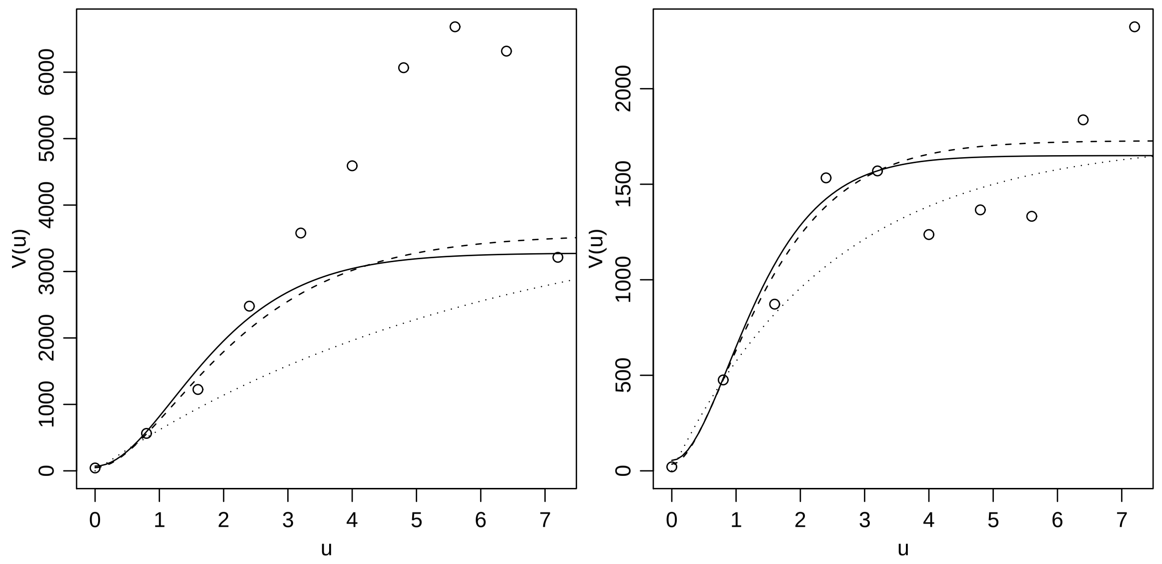
- poor quadratic approximations, unreliable Hessian matrices
- identifiability issues for more than tow parameters in the correlation function
- for models such as *Mat'ern* and *powered exponential* ϕ and κ are not orthogonal
- For the Matérn correlation function we suggest to take κ in a discrete set $\{0.5, 1, 2, 3, \ldots, N\}$ ("profiling")
- other possible approach is reparametrization such as replacing ϕ by $\alpha = 2\sqrt{\kappa}\phi$ (Handcock and Wallis)
- stability: e.g. Zhang's comments on σ^2/ϕ
- reparametrisations and asymptotics, e.g. $\theta_1 = \log(\sigma^2/\phi^{2\kappa})$ and $\theta_2 = \log(\phi^{2\kappa})$

Note: variations on the likelihood

- we strongly favor likelihood based methods.
- examining profile likelihoods can be reavealing on model identifiability and parameter uncertainty.
- restricted maximum likelihood is widely recommended leading to less biased estimators but is sensitive to mis-specification of the mean model. In spatial models distinction between $\mu(x)$ and $S(x)$ is not sharp.
- composite likelihood uses independent contributions for the likelihood function for each pair of points.
- approximate likelihoods are useful for large data-sets.
- Markov Random Fields can be used to approximate geo-statistical models.
- ...

Example: Surface elevation data

model with constant mean							
model	$\hat\mu$			$\hat\sigma^2$	$\hat\phi$	$\hat\tau^2$	logL
$\kappa = 0.5$	863.71			4087.6	6.12	0	−244.6
$\kappa = 1.5$	848.32			3510.1	1.2	48.16	−242.1
$\kappa = 2.5$	844.63			3206.9	0.74	70.82	−242.33
model with linear trend							
model	$\hat\beta_0$	$\hat\beta_1$	$\hat\beta_2$	$\hat\sigma^2$	$\hat\phi$	$\hat\tau^2$	logL
$\kappa = 0.5$	919.1	−5.58	−15.52	1731.8	2.49	0	−242.71
$\kappa = 1.5$	912.49	−4.99	−16.46	1693.1	0.81	34.9	−240.08
$\kappa = 2.5$	912.14	−4.81	−17.11	1595.1	0.54	54.72	−239.75



SESSION 7

Geostatistical spatial prediction (*kriging*)

Prediction – general results

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, $\hat{T} = t(Y)$

a criterion – MMSPE: the *best* predictor minimises

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

The MMSEP of T is $\hat{T} = \mathrm{E}(T|Y)$

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 MSPE($\hat{T}$)).

$\mathrm{E}[(T - \hat{T})^2] \leq \mathrm{Var}(T)$, with equality if T and Y are independent random variables.

Prediction – general results (cont.)

- 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 $[T|Y]$ (analytically or a sample from it)

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

Prediction – Linear Gaussian model

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

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

- $[S(x), Y]$ are jointly multivariate Gaussian. with mean vector $\mu 1$ and variance matrix

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

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

- $\hat{T} = \mathrm{E}[S(x)|Y] = \mu + \sigma^2 \mathrm{r}'(\tau^2 I + \sigma^2 R)^{-1}(Y - \mu 1) \qquad (1)$

- $\mathrm{Var}[S(x)|Y] = \sigma^2 - \sigma^2 \mathrm{r}'(\tau^2 I + \sigma^2 R)^{-1} \sigma^2 \mathrm{r}$

Prediction – Linear Gaussian model (cont.)

- for the Gaussian model $\hat{T}$ is linear in Y, so that

$$\hat{T} = w_0(x) + \sum_{i=1}^{n} w_i(x) Y_i$$

- equivalent to a least squares problem to find w_i which minimise $MSPE(\hat{T})$ within the class of linear predictors.
- Because the conditional variance does not depend on Y, the prediction MSE is equal to the prediction variance.
- Equality of prediction MSE and prediction variance is a special property of the multivariate Gaussian distribution, not a general result.

Prediction – Linear Gaussian model (cont.)

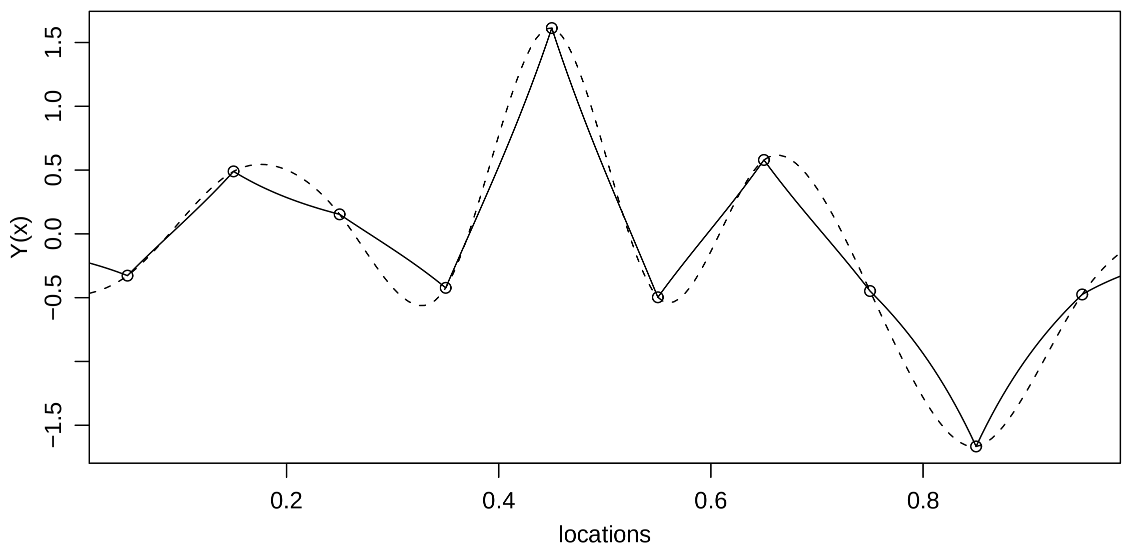
- Construction of the surface $\hat{S}(x)$, where $\hat{T} = \hat{S}(x)$ is given by (1), is called simple kriging.
- Assumes known model parameters.
- This name is a reference to D.G. Krige, who pioneered the use of statistical methods in the South African mining industry (Krige, 1951).

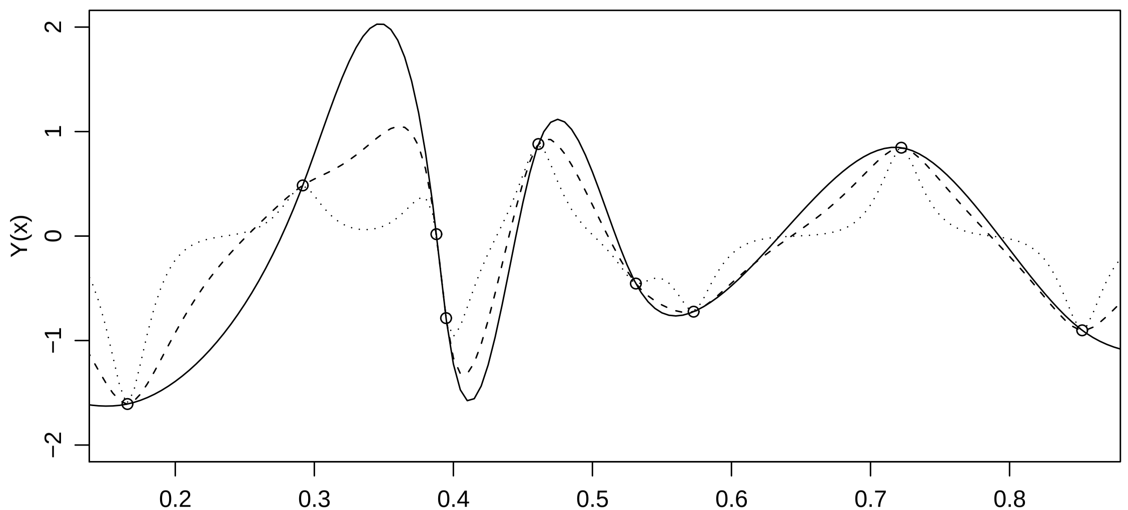
Features of spatial prediction

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

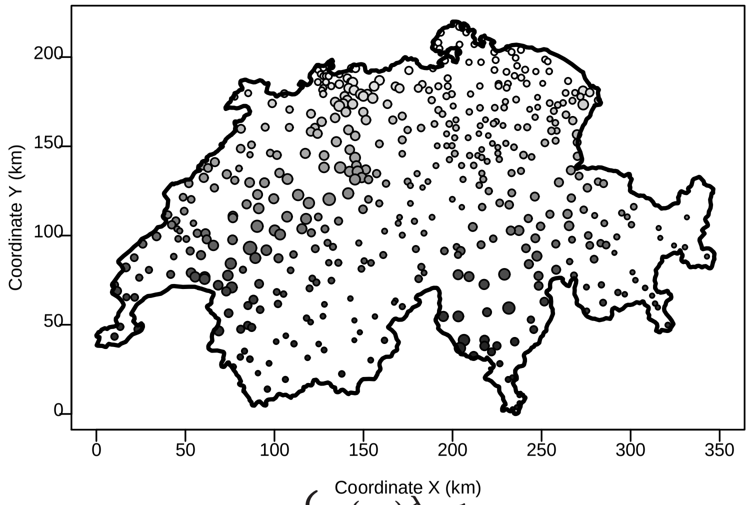
$$\begin{aligned}\hat{T} = \hat{S}(x) &= \mu + \sum_{i=1}^{n} w_i(x)(Y_i - \mu) \\ &= \{1 - \sum_{i=1}^{n} w_i(x)\}\mu + \sum_{i=1}^{n} w_i(x) Y_i\end{aligned}$$

- shows 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,
- $w_i(x)$ are the prediction weights.





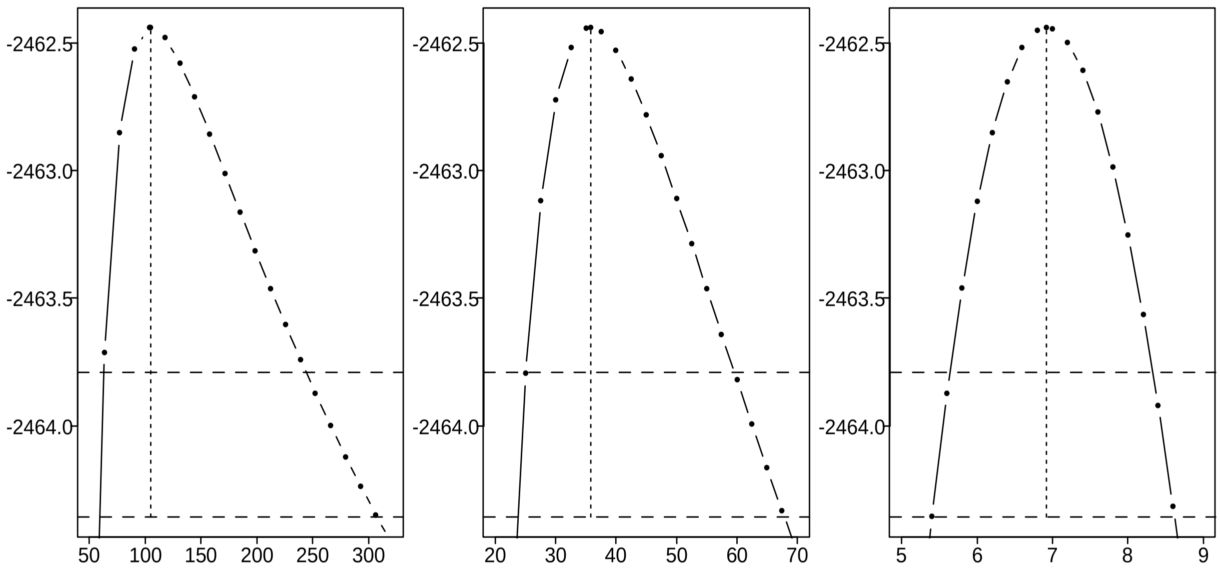
Swiss rainfall data – trans-Gaussian model

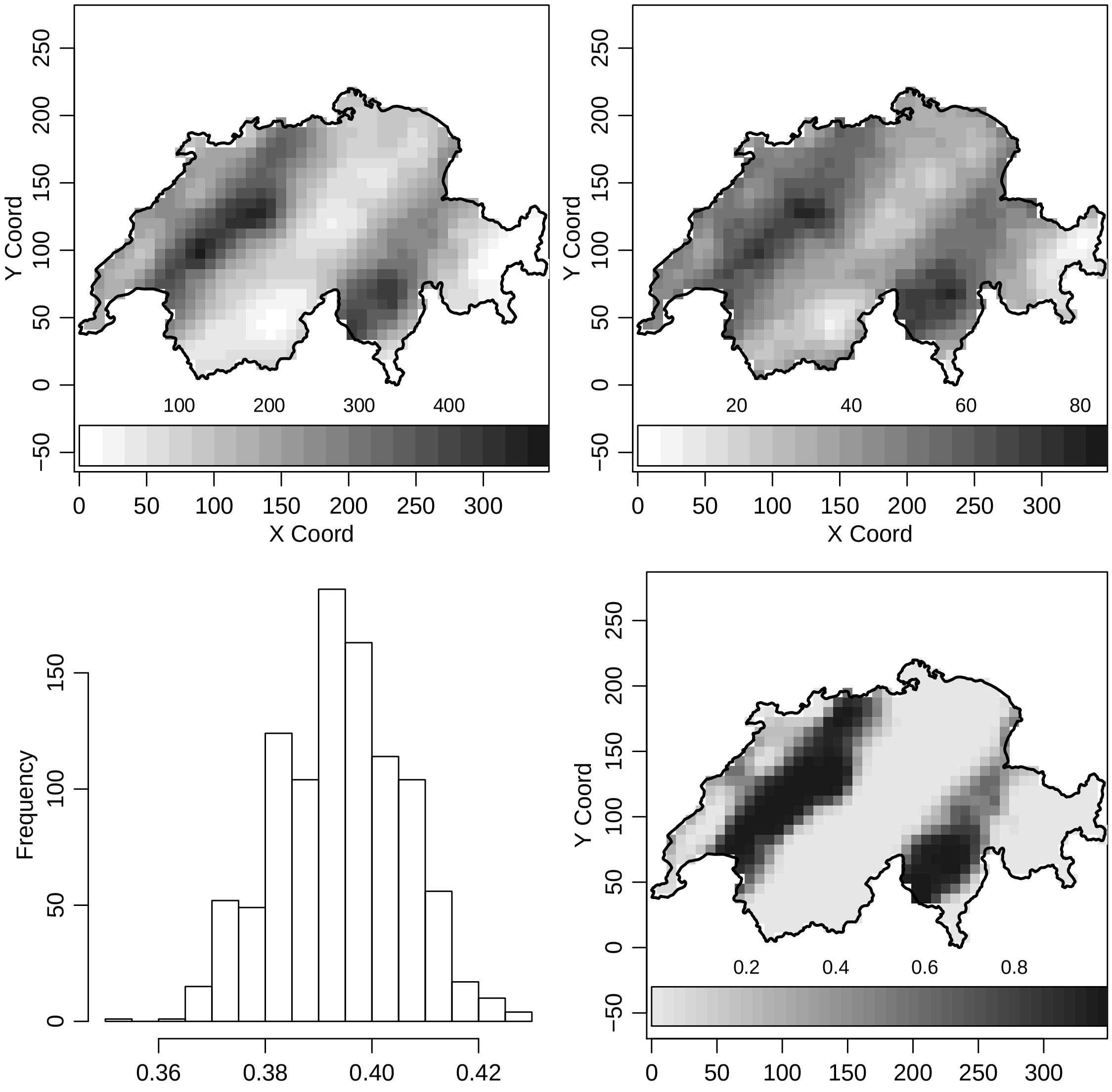


$$Y_i^* = h_\lambda(Y) = \begin{cases} \frac{(y_i)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y_i) & \text{if } \lambda = 0 \end{cases}$$

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

κ	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\phi}$	$\hat{\tau}^2$	$\log \hat{L}$
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





SESSION 8

Bayesian Inference

Bayesian Basics

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.

- model specification includes model parameters:

$$[\boldsymbol{Y}, \boldsymbol{\theta}] = [\boldsymbol{\theta}][\boldsymbol{Y}|\boldsymbol{\theta}]$$

- inference using Bayes' Theorem:

$$[\boldsymbol{Y}, \boldsymbol{\theta}] = [\boldsymbol{Y}|\boldsymbol{\theta}][\boldsymbol{\theta}] = [\boldsymbol{Y}][\boldsymbol{\theta}|\boldsymbol{Y}]$$

- to derive the posterior distribution

$$[\boldsymbol{\theta}|\boldsymbol{Y}] = [\boldsymbol{Y}|\boldsymbol{\theta}][\boldsymbol{\theta}]/[\boldsymbol{Y}] \propto [\boldsymbol{Y}|\boldsymbol{\theta}][\boldsymbol{\theta}]$$

- The prior distribution $[\boldsymbol{\theta}]$ express the uncertainty about the model parameters

- The posterior distribution $[\boldsymbol{\theta}|\boldsymbol{Y}]$ express the *revised* uncertainty after observing $\boldsymbol{Y}$
- conjugacy is achieved in particular models where convenient choices of $[\boldsymbol{\theta}]$ produces $[\boldsymbol{\theta}|\boldsymbol{Y}]$ within the same family
- more generally $[\boldsymbol{\theta}|\boldsymbol{Y}]$ may be an unknown and $[\boldsymbol{Y}] = \int [\boldsymbol{Y}|\boldsymbol{\theta}][\boldsymbol{\theta}]d\boldsymbol{\theta}$ may need to be evaluated numerically.
- probability statements and estimates are based on the posterior density obtained through

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{\ell(\boldsymbol{\theta}; \boldsymbol{y})\pi(\boldsymbol{\theta})}{\int \ell(\boldsymbol{\theta}; \boldsymbol{y})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

  are usually expressed as summary statistics (mean, median, mode) and/or Bayesian credibility intervals
- credible intervals are not uniquely defined (e.g. quantile based, highest density interval, etc)

Prediction

For Bayesian prediction expand the Bayes' theorem to include the prediction target, allowing for uncertainty on model parameters to be accounted for.

- and for prediction

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

- derive the predictive distribution

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

- can be interpreted as a weighted prediction over possible values of $[\theta|Y]$
- in general, as data becomes more abundant $[\theta|Y]$ concentrates around $\hat{\theta}$

Bayesian inference for the geostatistical model

Bayesian inference for the geostatistcal model expands the previous results acknowledging for Y and S as specified by the adopted model.

- model specification:

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

- inference using Bayes' Theorem:

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

- to derive the posterior distribution

$$[\theta|Y] = \int [\theta, S|Y] dS = \int \frac{[Y|S, \theta][S|\theta][\theta]}{[Y]} dS$$

- where $[\boldsymbol{Y}] = \int\int [\boldsymbol{Y}|\boldsymbol{\theta}][\boldsymbol{S}|\boldsymbol{\theta}][\boldsymbol{\theta}]d\boldsymbol{S}d\boldsymbol{\theta}$ is typically difficult to evaluate

- For prediction

$$[\boldsymbol{Y}, \boldsymbol{T}, \boldsymbol{S}, \boldsymbol{\theta}] = [\boldsymbol{Y}, \boldsymbol{T}|\boldsymbol{S}, \boldsymbol{\theta}][\boldsymbol{S}|\boldsymbol{\theta}][\boldsymbol{\theta}]$$

- derive the predictive distribution

$$[\boldsymbol{T}|\boldsymbol{Y}] = \int\int [\boldsymbol{T}, \boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{Y}]d\boldsymbol{S}d\boldsymbol{\theta} = \int\int [\boldsymbol{T}|\boldsymbol{Y}, \boldsymbol{S}, \boldsymbol{\theta}][\boldsymbol{S}, \boldsymbol{\theta}|\boldsymbol{Y}]d\boldsymbol{S}d\boldsymbol{\theta}$$

- and explore the conditional independence structure of the model to simplify the calculations

Notes I

- likelihood function occupies a central role in both classical and Bayesian inference

- plug-in prediction corresponds to inferences about $[\boldsymbol{T}|\boldsymbol{Y}, \hat{\theta}]$

- Bayesian prediction is a weighted average of plug-in predictions, with different plug-in values of $\boldsymbol{\theta}$ weighted according to their conditional probabilities given the observed data.

- Bayesian prediction is usually more cautious than plug-in prediction.
  Allowance for parameter uncertainty usually results in wider prediction intervals

Notes II

1. The need to evaluate the integral which defines $[Y]$ represented a major obstacle to practical application,

2. development of Markov Chain Monte Carlo (MCMC) methods has transformed the situation.

3. 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.

4. in particular for the Gaussian model other algorithms can be implemented.

Results for the Gaussian models - I

- fixing covariance parameters and assuming a (conjugate) prior for β
$$\beta \sim \mathrm{N}\left(m_\beta \; ; \; \sigma^2 V_\beta\right)$$

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

- and 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

$$\begin{aligned}
E[S^*|Y] &= (D_0 - r'V^{-1}D)(V_\beta^{-1} + D'V^{-1}D)^{-1}V_\beta^{-1}m_\beta + \\
&\left[r'V^{-1} + (D_0 - r'V^{-1}D)(V_\beta^{-1} + D'V^{-1}D)^{-1}D'V^{-1}\right] Y \\
\mathrm{Var}[S^*|Y] &= \sigma^2 \left[V_0 - r'V^{-1}r + \right. \\
&\left. (D_0 - r'V^{-1}D)(V_\beta^{-1} + D'V^{-1}D)^{-1}(D_0 - r'V^{-1}D)'\right].
\end{aligned}$$

- predicted mean balances between prior and weighted average of the data
- The predictive variance has three interpretable components: a priori variance, the reduction due to the data and the uncertainty in the mean.
- $V_\beta \to \infty$ results can be related to REML and universal (or ordinary) kriging.

Results for the Gaussian models - II

- fixing correlation parameters and assuming a (conjugate) prior for $[\beta, \sigma^2] \sim N\chi^2_{ScI}\left(m_b, V_b, n_\sigma, S^2_\sigma\right)$ given by:

$$[\beta|\sigma^2] \sim \mathrm{N}\left(m_\beta\ ;\ \sigma^2 V_\beta\right) \text{ and } [\sigma^2] \sim \chi^2_{ScI}(n_\sigma, S^2_\sigma)$$

- The posterior is $[\beta, \sigma^2|y, \phi] \sim N\chi^2_{ScI}\left(\tilde{\beta}, V_{\tilde{\beta}}, n_\sigma + n, S^2\right)$

$$\tilde{\beta} = V_{\tilde{\beta}}(V_b^{-1}m_b + D'R^{-1}y)$$

$$V_{\tilde{\beta}} = (V_b^{-1} + D'R^{-1}D)^{-1}$$

$$S^2 = \frac{n_\sigma S^2_\sigma + m_b' V_b^{-1} m_b + y'R^{-1}y - \tilde{\beta}' V_{\tilde{\beta}}^{-1}\tilde{\beta}}{n_\sigma + n}$$

- The predictive distribution $[S^*|y] \sim t_{n_\sigma+n}\left(\mu^*, S^2\Sigma^*\right)$
- with mean and variance given by

$$\begin{aligned} \mathrm{E}[S^*|y] &= \mu^*, \\ \mathrm{Var}[S^*|y] &= \frac{n_\sigma + n}{n_\sigma + n - 2}\, S^2\Sigma^*, \end{aligned}$$

$$\begin{aligned} \mu^* &= (D^* - r'V^{-1}D)V_{\tilde\beta}V_b^{-1}m_b \\ &\quad + \left[r'V^{-1} + (D^* - r'V^{-1}D)V_{\tilde\beta}D'V^{-1}\right]y, \\ \Sigma^* &= V^0 - r'V^{-1}r + (D^* - r'V^{-1}D)(V_b^{-1} + V_{\hat\beta}^{-1})^{-1}(D^* - r'V^{-1}D)'. \end{aligned}$$

- valid if $\tau^2 = 0$
- for $\tau^2 > 0$, $\nu^2 = \tau^2/\sigma^2$ can regarded as a correlation parameter

Results for the Gaussian models - III

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

- The posterior distribution for the parameters is:

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

- where $p(\beta, \sigma^2|y, \phi)$ can be obtained analytically and

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

- analogous results for more general prior:

$$[\beta|\sigma^2, \phi] \sim N\left(m_b, \sigma^2 V_b\right) \quad \text{and} \quad [\sigma^2|\phi] \sim \chi^2_{ScI}\left(n_\sigma, S^2_\sigma\right),$$

- choice of prior for ϕ can be critical. (Berger, De Oliveira & Sansó, 2001)

Algorithm 1:

1. 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.

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

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

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

5. 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:

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

Algorithm 2:

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

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

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

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

5. Repeat steps (3) and (4) to generate a sample from the required predictive distribution.

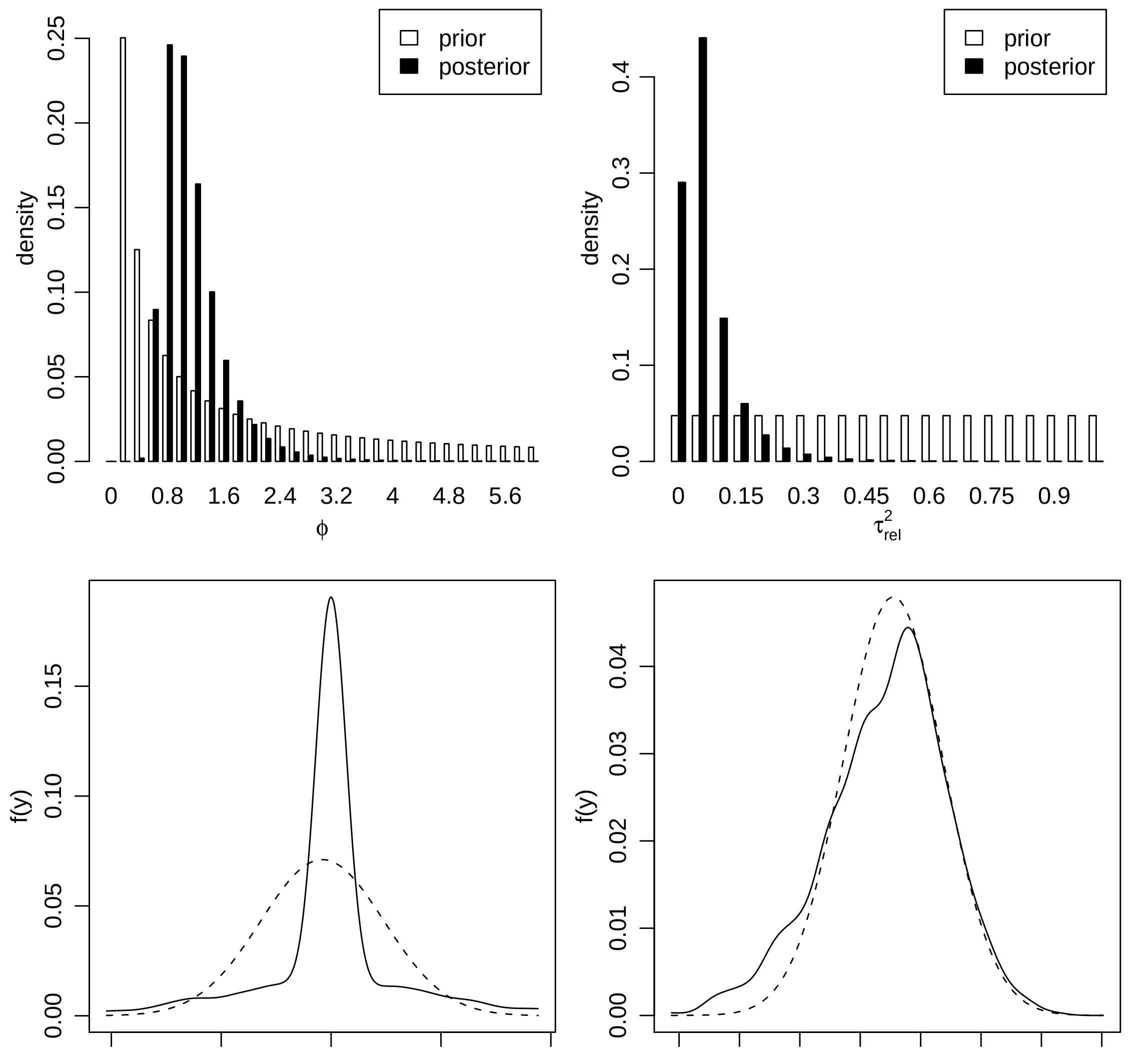
Notes

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

2. We specify a discrete prior distribution on a multi-dimensional grid of values.

3. This implies extra computational load (but no new principles)

Elevation data



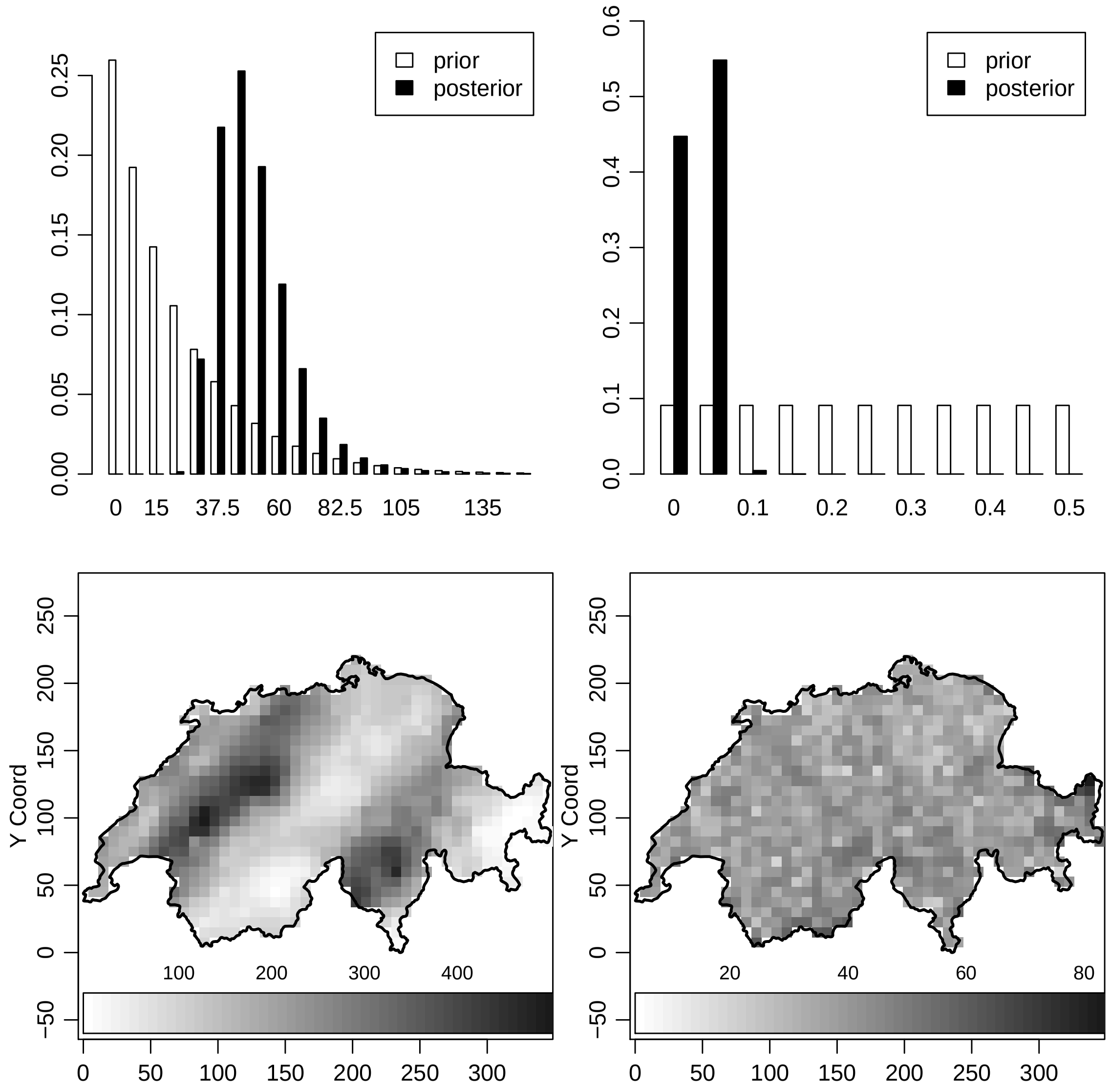
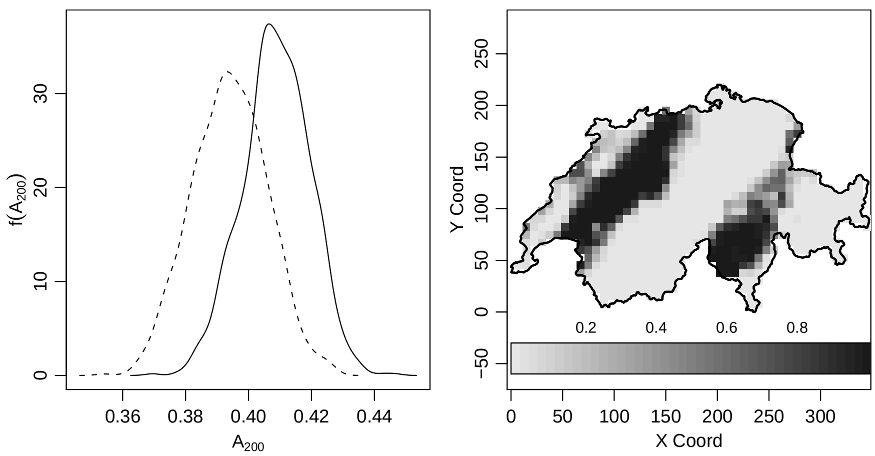


Table 1: Swiss rainfall data: posterior means and 95% central quantile-based credible intervals for the model parameters.

parameter	estimate	95% interval
β	**144.35**	**[53.08 , 224.28]**
σ^2	**13662.15**	**[8713.18 , 27116.35]**
ϕ	**49.97**	**[30 , 82.5]**
ν^2	**0.03**	**[0 , 0.05]**



SESSION 9

Generalised linear geostatistical models

Generalized linear geostatistical model

- Preserving the assumption of a zero mean, stationary Gaussian process $S(\cdot)$,
- our basic model can be generalized replacing the assumption of mutually independent $Y_i|S(\cdot) \sim \mathrm{N}(S(x), \tau^2)$ by assuming $Y_i|S(\cdot)$ are mutually independent within the class of generalized linear models (GLM)
- with a link function $h(\mu_i) = \sum_{j=1}^{p} d_{ij}\beta_j + S(x_i)$
- this defines a generalized linear mixed model (GLMM) with correlated random effects
- which provides a way to adapt classical GLM for geostatistical applications.

GLGM

- usually just a single realisation is available, in contrast with GLMM for longitudinal data analysis
- The GLM approach is most appealing when follows an natural sampling mecanism such as Poisson model for counts and logist-linear models for binary/binomial responses
- in principle transformed models can be considered for skewed distributions
- variograms for such processes can be obtained although providing a less obvious summary statistics
- empirical variograms of GLM residuals can be used for exploratory analysis

An example: a Poisson model

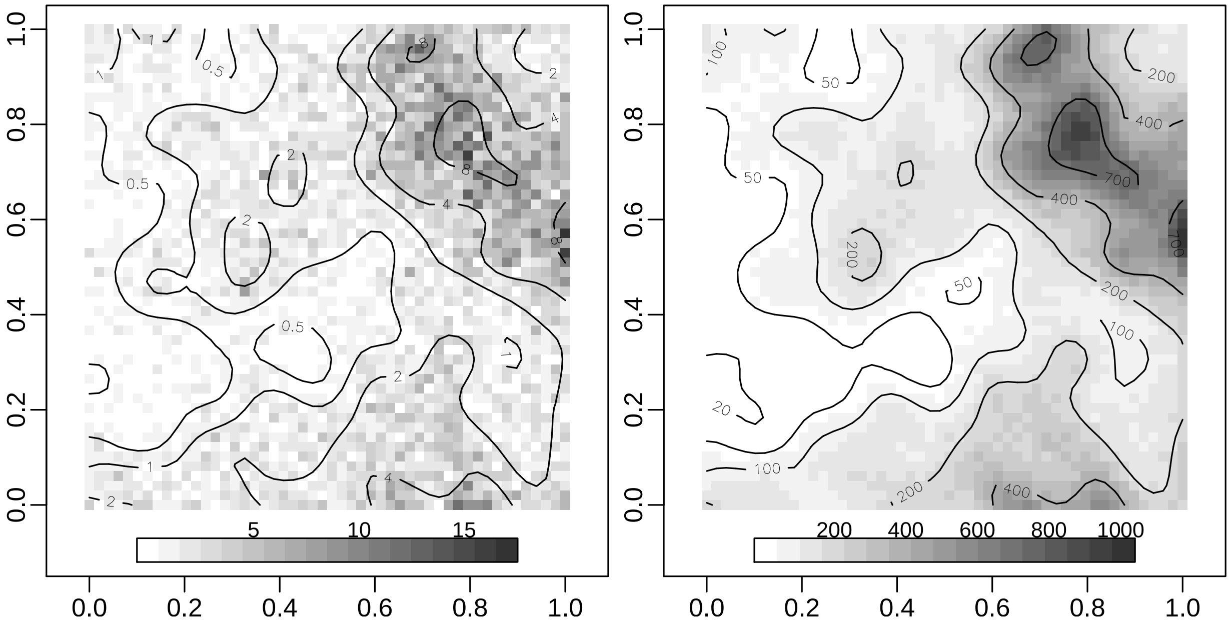
- $[Y(x_i) \mid S(x_i)]$ is Poisson with density

$$f(y_i; \zeta_i) = \exp(-\zeta_i)\zeta_i^{y_i}/y_i! \quad y_i = 0, 1, 2, \ldots$$

- link: $E[Y(x_i) \mid S(x_i)] = \zeta_i = h(\mu_i) = h(\mu + S(x_i))$
- log-link $h(\cdot) = \exp(\cdot)$
- more generaly the models can be expanded allowing for covariates and/or uncorrelated random effects

$$h(\mu_i) = \sum_{j=1}^{p} d_{ij}\beta_j + S(x_i) + Z_i$$

which, differently from Gaussian models, distinguish between the terms of the nugget effect: Poisson variation accounts for the anologue of measurement error and spatially uncorrelated component to the short scale variation



Simulations from the Poisson model; grey-scale shading represents the data values on a regular grid of sampling locations and contours represents the conditional expectation surface, with $\mu = 0.5$ on the left panel and $\mu = 5$ on the right panel.

Another example: a Binomial logistic model

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

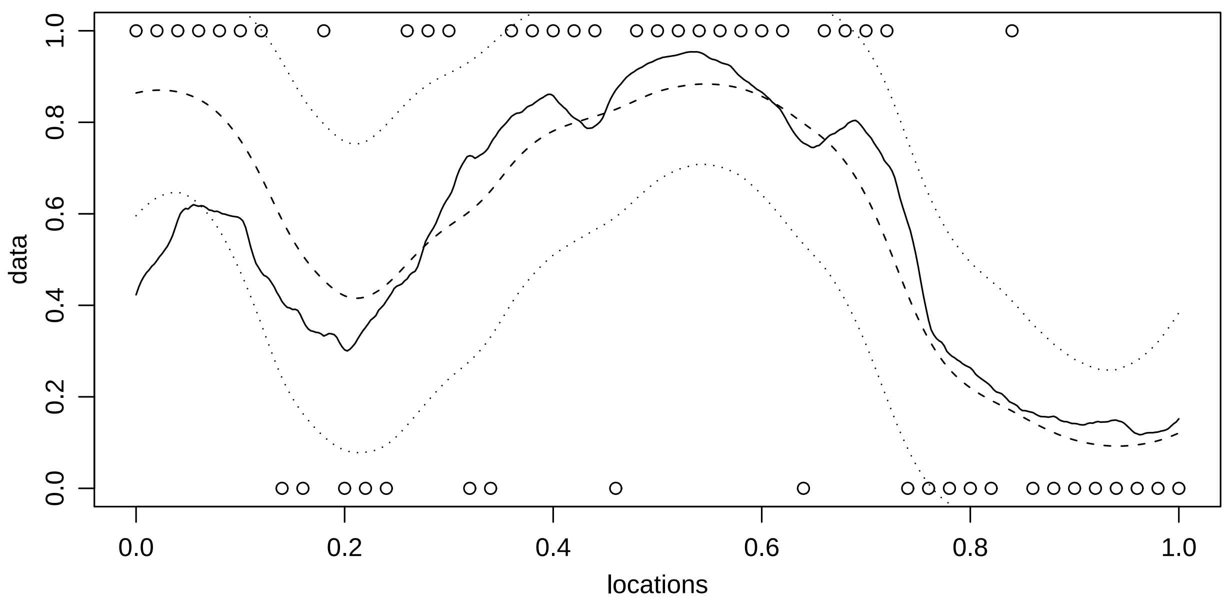
$$f(y_i; \zeta_i) = \binom{n_i}{y_i} \zeta_i^{y_i} (1-\zeta_i)^{(n_i - y_i)} \quad y_i = 0, 1, \ldots, n_i$$

- logistic link: $E[Y(x_i) \mid S(x_i)] = n_i \zeta_i = \frac{\exp\{\mu_i\}}{1+\exp\{\mu_i\}}$
- mean: $\mu_i = \mu + S(x_i)$
- again can be expanded as

$$h(\mu_i) = \sum_{j=1}^{p} d_{ij}\beta_j + S(x_i) + Z_i$$

- typically more informative with larger values of n_i

An simulated example from binary model



- in this example the binary sequence is not much informative on $S(x)$
- wide intervals compared to the prior mean of $p(x)$

Inference

- Likelihood function

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

- Involves a high-dimensional (numerical) integration
- MCMC algorithms can exploit the conditional independence scructure of the model

$$\begin{tikzcd}
 & \theta \arrow[dl, dash] \arrow[dr, dash] & & \beta \arrow[dl, dash] \arrow[dr, dash] & \\
S^* \arrow[rr, dash] & & S \arrow[rr, dash] & & Y
\end{tikzcd}$$

Prediction with known parameters

- 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 reduce Monte Carlo error by
  - calculating $\mathrm{E}[T(S^*)|s(j)]$ directly
  - estimate $\mathrm{E}[T(S^*)|y]$ by $\frac{1}{m}\sum_{j=0}^{m} \mathrm{E}[T(S^*)|s(j)]$

MCMC for conditional simulation

- Let $S = D'\beta + \Sigma^{1/2}\Gamma$, $\Gamma \sim N_n(0, I)$.
- 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 $\xi(\gamma) = \gamma + \frac{h}{2}\nabla \log f(\gamma \mid y)$.
- E.g for the Poisson-log Spatial model:
  $\nabla \log f(\gamma|y) = -\gamma + (\Sigma^{1/2})'(y - \exp(s))$ where $s = \Sigma^{1/2}\gamma$.
- Expression generalises to other generalised linear spatial models.
- MCMC output $\gamma_1, \ldots, \gamma_m$. Multiply by $\Sigma^{1/2}$ and obtain: $s(1), \ldots, s(m)$ from $[S|y]$.

MCMC for Bayesian inference

Posterior:

- 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)

Predictive:

- 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)

Comments

- Marginalisation w.r.t β and σ^2 is possible using conjugate priors
- Discrete prior for ϕ is an advantage (reduced computing time).
- thinning: not to store a large sample of high-dimensional quantities.
- similar algorithms for MCMC maximum likelihood estimation

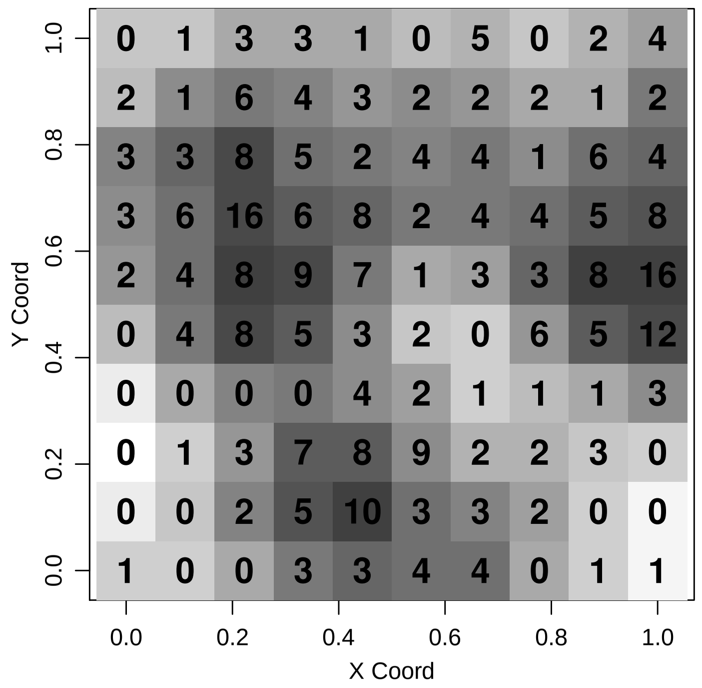
Example code from geoRglm

— demo 03 —

SESSION 10

Case studies on generalised linear geostatistical models

A simulated Poisson data



R code for simulation

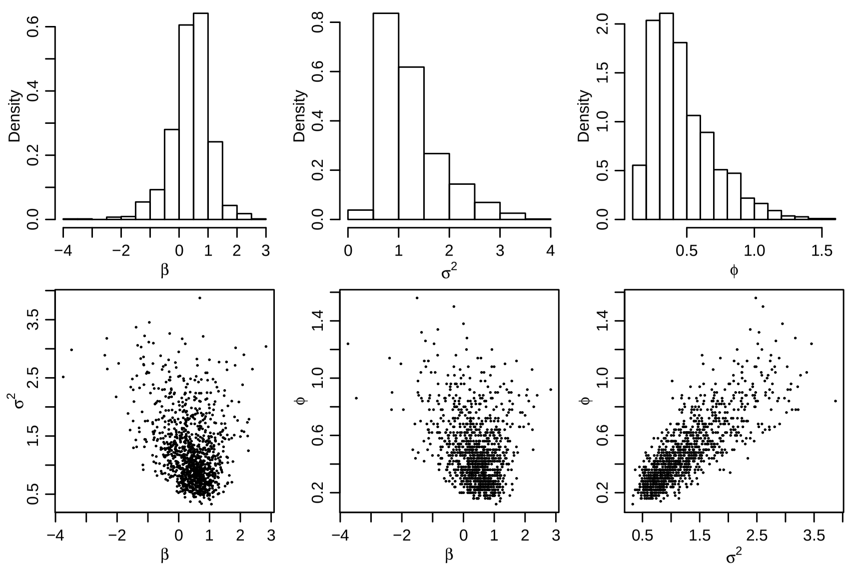
```
## setting the seed
> set.seed(371)
## defining the data locations on a grid
> cp <- expand.grid(seq(0, 1, l = 10), seq(0, 1, l = 10))
## simulating from the S process
> s <- grf(grid = cp, cov.pars = c(2, 0.2), cov.model = "mat",
+     kappa = 1.5)
## visualising the S process
> image(s, col = gray(seq(1, 0.25, l = 21)))
## inverse link function
> lambda <- exp(0.5 + s$data)
## simulating the data
> y <- rpois(length(s$data), lambda = lambda)
## visualising the data
> text(cp[, 1], cp[, 2], y, cex = 1.5, font = 2)
```

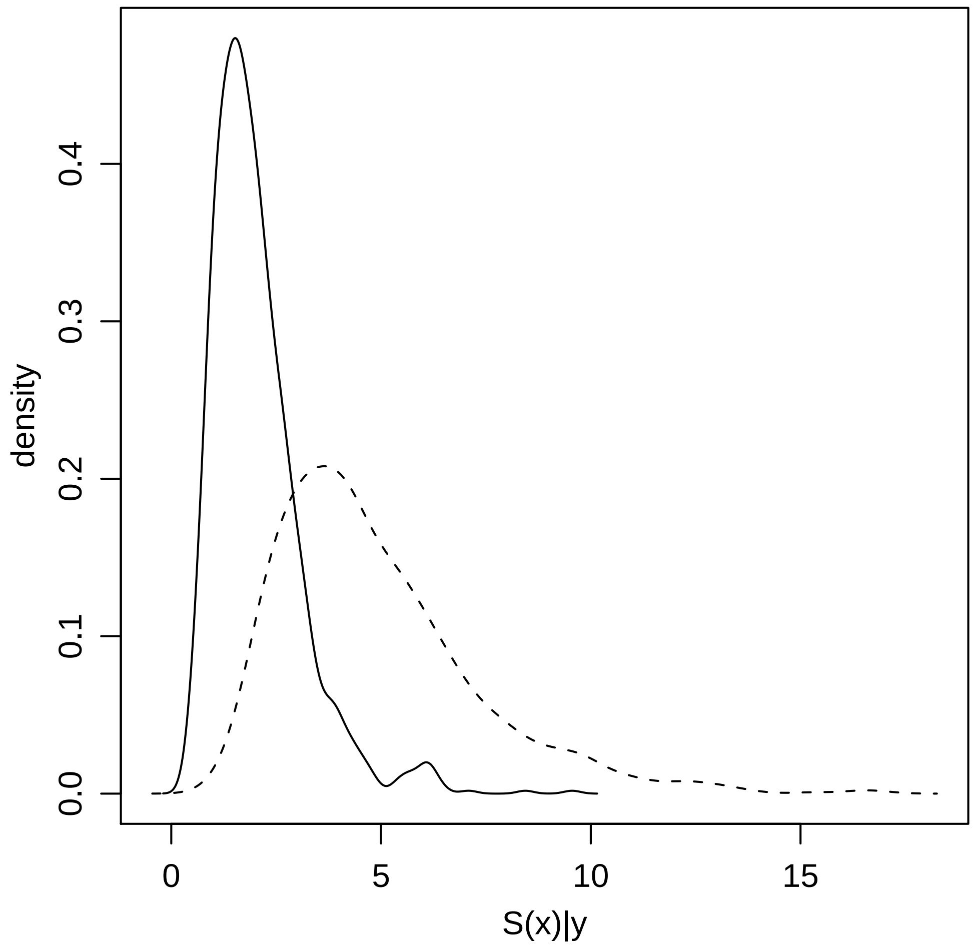
R code for the data analysis

```
set.seed(371)
## calibracao do algoritmo MCMC
MCc <- mcmc.control(S.scale=0.025, phi.sc=0.1, n.iter=110000,
                    burn.in=10000, thin=100, phi.start=0.2)
## especificacao de priors
PGC <- prior.glm.control(phi.prior="exponential", phi=0.2,
                   phi.discrete=seq(0,2,by=0.02),tausq.rel=0)
## opo de saida
OC <- output.glm.control(sim.pred=T)
## escolhendo 2 localizacoes para predicao
locs <- cbind(c(0.75, 0.15), c(0.25, 0.5))
##
pkb <- pois.krige.bayes(dt, loc=locs, prior=PGC, mcmc=MCc, out=OC)
```

Summaries of the posterior for the simulated Poisson data: posterior means and 95% central quantile-based intervals.

parameters	true values	posterior mean	95% interval
β	**0.5**	**0.4**	**[0.08 , 1.58]**
σ^2	**2.0**	**1.24**	**[0.8 , 2.76]**
ϕ	**0.2**	**0.48**	**[0.3 , 1.05]**





Rongelap Island

— see other set of slides —

The Gambia malaria

— see other set of slides —

SESSION 11

More on generalised linear geostatistical models

Covariance functions and variograms

- **In non-Gaussian settings, the variogram is a less natural summary statistic but can still be useful as a diagnostic tool**

- **for GLGM the model with constant mean:**

$$\mathrm{E}\left[Y(x_i)|S(x_i)\right] = \mu_i = g(\alpha + S_i) \quad v_i = v(\mu_i)$$

$$\begin{aligned}
\gamma_Y(u) &= \mathrm{E}[\frac{1}{2}(Y_i - Y_j)^2] \\
&= \frac{1}{2}\mathrm{E}_S[\mathrm{E}_Y[(Y_i - Y_j)^2|S(\cdot)]] \\
&= \frac{1}{2}\left(\mathrm{E}_S[\{g(\alpha + S_i) - g(\alpha + S_j)\}^2] + 2\mathrm{E}_S[v(g(\alpha + S_i))]\right) \\
&\approx g'(\alpha)^2\gamma_S(u) + \bar{\tau}^2
\end{aligned}$$

- **the variogram on the Y-scale is approximately proportional to the variogram of $S(\cdot)$ plus an intercept**

- **the intercept represents an average nugget effect induced by the variance of the error distribution of the model**

- **however it relies on a linear approximation to the inverse link function**

- **it may be inadequate for diagnostic analysis since the essence of the generalized linear model family is its explicit incorporation of a non-linear relationship between Y and $S(x)$.**

- **The exact variogram depends on higher moments of $S(\cdot)$**

- **explicit results are available only in special cases.**

Spatial survival analysis

- specified through hazard function $h(t) = f(t)/\{1 - F(t)\}$,
- $h(t)\delta t$ is the conditional probability event will occour in the interval $(t, t + \delta t)$, given it has not occour until time t
- *proportional hazards model* with $\lambda_0(t)$, an unspecified baseline hazard function

$$h_i(t) = \lambda_0(t) \exp(d_i' \beta)$$

- $h_i(t)/h_j(t)$ does not change over time
- alternativelly, fully specified models are proposed
- *frailty* corresponds to *random effects* can be introduced by

$$h_i(t) = \lambda_0(t) \exp(z_i' \beta + U_i) = \lambda_0(t) W_i \exp(d_i' \beta)$$

- e.g. *log-Gaussian frailty model* and *gamma frailty model*
- replacing U_i by $S(x_i)$ introduces ***spatial frailties*** (Li & Ryan, 2002; Banerjee, Wall & Carlin, 2003)
- $\mathrm{E}[S(x)] = -0.5 \,\mathrm{Var}[S(x)]$ preserves interpretation of $\exp\{S(x)\}$ as a frailty process
- other possible approaches, e.g. Henderson, Shimakura and Gorst (2002) extends the gamma-frailty model

Geostatistical models for point process

- Two possible connections between point process and geostatistics:

    1. measurement process replaced by a point process
    2. choice of data locations for $Y(x_i)$

Cox point processes

Definition:

A Cox process is a point process in which there is an unobserved, non-negative-valued stochastic process $S = \{S(x) : x \in \mathbb{R}^2\}$ such that, conditional on S, the observed point process is an inhomogeneous Poisson process with spatially varying intensity $S(x)$.

- fits into the general geostatistical framework
- derived as limiting form of a geostatistical model as $\delta \rightarrow 0$ for locations on lattice-spacing δ
- *log-Gaussian Cox process* is a tractable form of Cox process (e.g. Möller, Syversveen and Waagepetersen,1998; Brix & Diggle, 2001)
- inference generally requires computationally intensive Monte Carlo methods, whose implementation involves careful tuning

- moment-based method provides an analogue of the variogram, for exploratory analysis and *preliminary* estimation of model parameters

Cox point processes

- intensity surface $\Lambda(x) = \exp\{S(x)\}$
- has mean and variance $\lambda = \exp\{\mu + 0.5\gamma(0)\}$
- also represents the expected number of points per unit area in the Cox process, and $\phi(u) = \exp\{\gamma(u)\} - 1$.
- $K(s)$: reduced second moment measure of a stationary point process
- $\lambda K(s)$: expected number of further points within distance s of an arbitrary point of the process
- For the log-Gaussian Cox process

$$K(s) = \pi s^2 + 2\pi\lambda^{-2} \int_0^s \phi(u) u du$$

- A non-parametric estimator:

$$\hat{K}(s) = \frac{|A|}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} w_{ij}^{-1} I(u_{ij} \leq s)$$

- w_{ij} allows for *edge correction*
- *preliminary estimates* of model parameters can then be obtained by minimising a measure of the discrepancy between theoretical and empirical K-functions

Geostatistics and marked point processes

locations X signal S measurements Y

- Usually write geostatistical model as

$$[S, Y] = [S][Y|S]$$

- What if X is stochastic? Usual implicit assumption is

$$[X, S, Y] = [X][S][Y|S],$$

hence can ignore $[X]$ for inference about $[S, Y]$.

- Resulting likelihood:

$$L(\theta) = \int [S][Y|S]dS$$

Marked point processes

locations X marks Y

- X is a point process
- Y need only be defined at points of X
- natural factorisation of $[X, Y]$?

Example 1. Spatial distribution of disease

$$\begin{array}{ll} X: & \text{population at risk} \\ Y: & \text{case or non-case} \end{array}$$

- Natural factorisation is $[X, Y] = [X][Y|X]$
- Usual scientific focus is $[Y|X]$
- Hence, can ignore $[X]$

Example 2. Growth of natural forests

X : location of tree
Y : size of tree

- Two candidate models:
  - competitive interactions $\Rightarrow [X, Y] = [X][Y|X]$
  - environmental heterogeneity $\Rightarrow [X, Y] = [Y][X|Y]$?
- focus of scientific interest?

Preferential sampling

locations X signal S measurements Y

- Conventional model:

$$[X, S, Y] = [S][X][Y|S] \quad (1)$$

- Preferential sampling model:

$$[X, S, Y] = [S][X|S][Y|S, X] \quad (2)$$

- Key point for inference: even if $[Y|S, X]$ in (2) and $[Y|S]$ in (1) are algebraically the same, the term $[X|S]$ in (1) cannot be ignored for inference about $[S, Y]$, because of the shared dependence on the unobserved process S

A model for preferential sampling

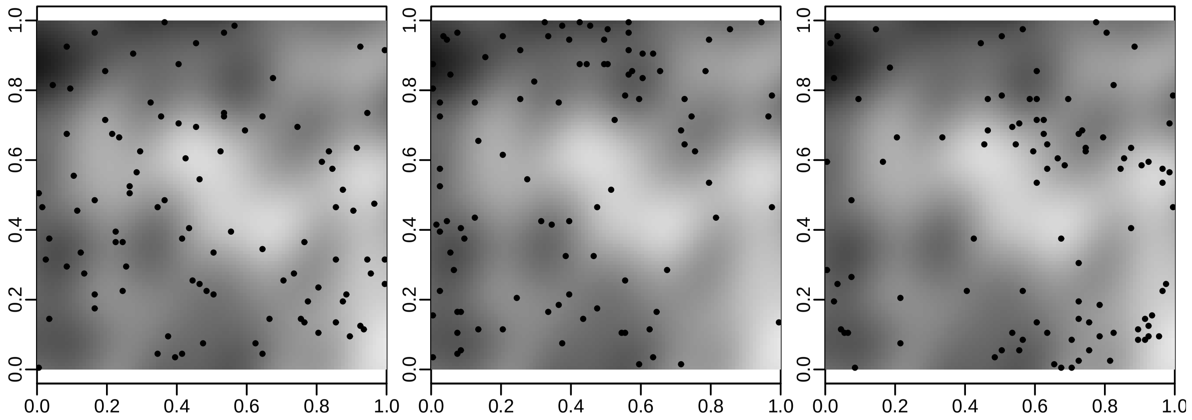
$$[X, S, Y] = [S][X|S][Y|S, X]$$

- $[S] = \mathrm{SGP}(0, \sigma^2, \rho)$ (stationary Gaussian process)
- $[X|S]$ = inhomogenous Poisson process with intensity

$$\lambda(x) = \exp\{\alpha + \beta S(x)\}$$

- $[Y|S, X] = \mathrm{N}\{\mu + S(x), \tau^2\}$ (independent Gaussian)

Simulation of preferential sampling model



Locations (dots) and underlying signal process (grey-scale):

- left-hand panel: uniform non-preferential
- centre-panel: clustered preferential
- right-hand panel: clustered non-preferential

Likelihood inference

$$[X, S, Y] = [S][X|S][Y|S, X]$$

- data are X and Y, hence likelihood is

$$L(\theta) = \int [X, S, Y]dS = \mathrm{E}_S\left[[X|S][Y|S, X]\right]$$

- evaluate expectation by Monte Carlo,

$$L_{MC}(\theta) = m^{-1}\sum_{j=1}^{m}[X|S_j][Y|S_j, X]$$

- use anti-thetic pairs $S_{2j} = -S_{2j-1}$

Practical solutions to weak identifiability

1. Strong Bayesian priors (if you can believe them)

2. Explanatory variables as surrogate for S

3. Two-stage sampling

SESSION 12

Geostatistical design

Geostatistical design

- **Retrospective**

  Add to, or delete from, an existing set of measurement locations $x_i \in A : i = 1, \ldots, n$.

- **Prospective**

  Choose optimal positions for a new set of measurement locations $x_i \in A : i = 1, \ldots, n$.

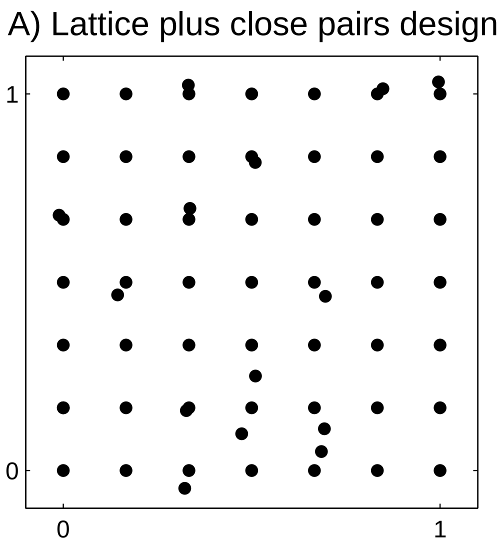
Naïve design folklore

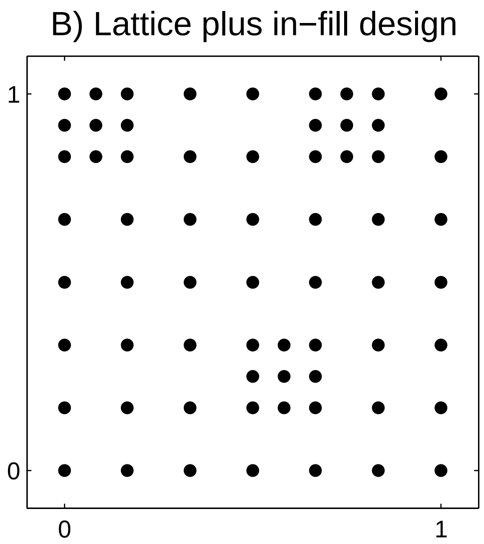
- **Spatial correlation decreases with increasing distance.**
- **Therefore, close pairs of points are wasteful.**
- **Therefore, spatially regular designs are a good thing.**

Less naïve design folklore

- **Spatial correlation decreases with increasing distance.**
- **Therefore, close pairs of points are wasteful if you know the correct model.**
- **But in practice, at best, you need to estimate unknown model parameters.**
- **And to estimate model parameters, you need your design to include a wide range of inter-point distances.**
- **Therefore, spatially regular designs should be tempered by the inclusion of some close pairs of points.**

Examples of compromise designs





Designs for parameter estimation

Comparison of random and square lattice designs, each with $n = 100$ *sample locations, with respect to three design criteria: spatial maximum of mean square prediction error* $M(x)$*; spatial average of mean square prediction error* $M(x)$*; scaled mean square error,* $100 \times MSE(T)$*, for* $T = \int S(x)dx$*. The simulation model is a stationary Gaussian process with parameters* $\mu = 0$*,* $\sigma^2 + \tau^2 = 1$*, correlation function* $\rho(u) = \exp(-u/\phi)$ *and nugget variance* τ^2*. The tabulated figures are averages of each design criterion over* $N = 500$ *replicate simulations.*

Model parameters		max $M(x)$		average $M(x)$		$MSE(T)$	
		Random	Lattice	Random	Lattice	Random	Lattice
$\tau^2 = 0$	$\phi = 0.05$	9.28	8.20	0.77	0.71	0.53	0.40
	$\phi = 0.15$	5.41	3.61	0.40	0.30	0.49	0.18
	$\phi = 0.25$	3.67	2.17	0.26	0.19	0.34	0.10
$\tau^2 = 0.1$	$\phi = 0.05$	9.57	8.53	0.81	0.76	0.54	0.41
	$\phi = 0.15$	6.22	4.59	0.50	0.41	0.56	0.28
	$\phi = 0.25$	4.44	3.34	0.37	0.30	0.47	0.22
$\tau^2 = 0.3$	$\phi = 0.05$	10.10	9.62	0.88	0.86	0.51	0.40
	$\phi = 0.15$	7.45	6.63	0.65	0.60	0.68	0.43
	$\phi = 0.25$	6.23	5.70	0.55	0.51	0.58	0.38

A Bayesian design criterion

Assume goal is prediction of $S(x)$ for all $x \in A$.

$$[S|Y] = \int [S|Y, \theta][\theta|Y] d\theta$$

For retrospective design, minimise

$$\bar{v} = \int_A \mathrm{Var}\{S(x)|Y\} dx$$

For prospective design, minimise

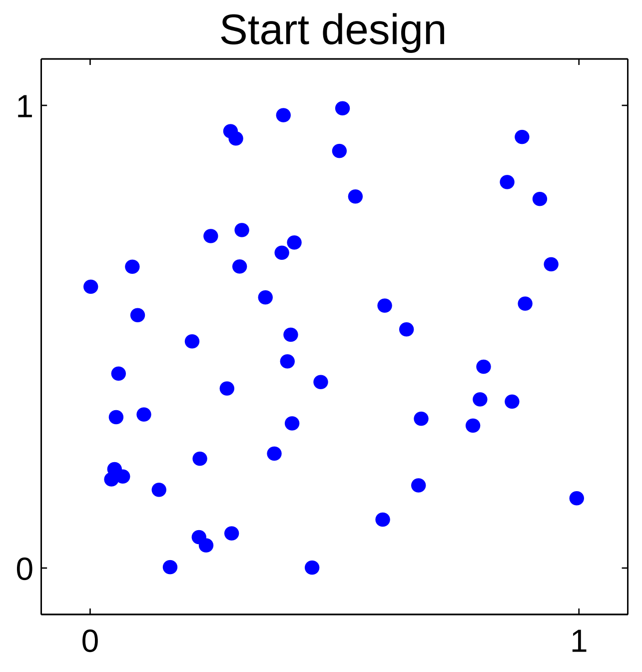
$$\mathrm{E}(\bar{v}) = \int_y \int_A \mathrm{Var}\{S(x)|y\} f(y) dy$$

where $f(y)$ corresponds to

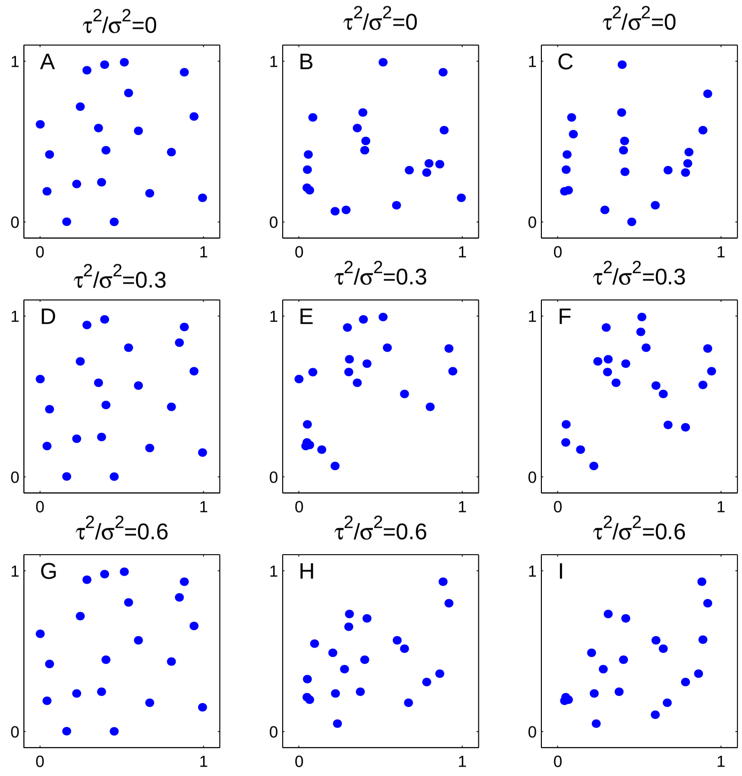
$$[Y] = \int [Y|\theta][\theta] d\theta$$

Results

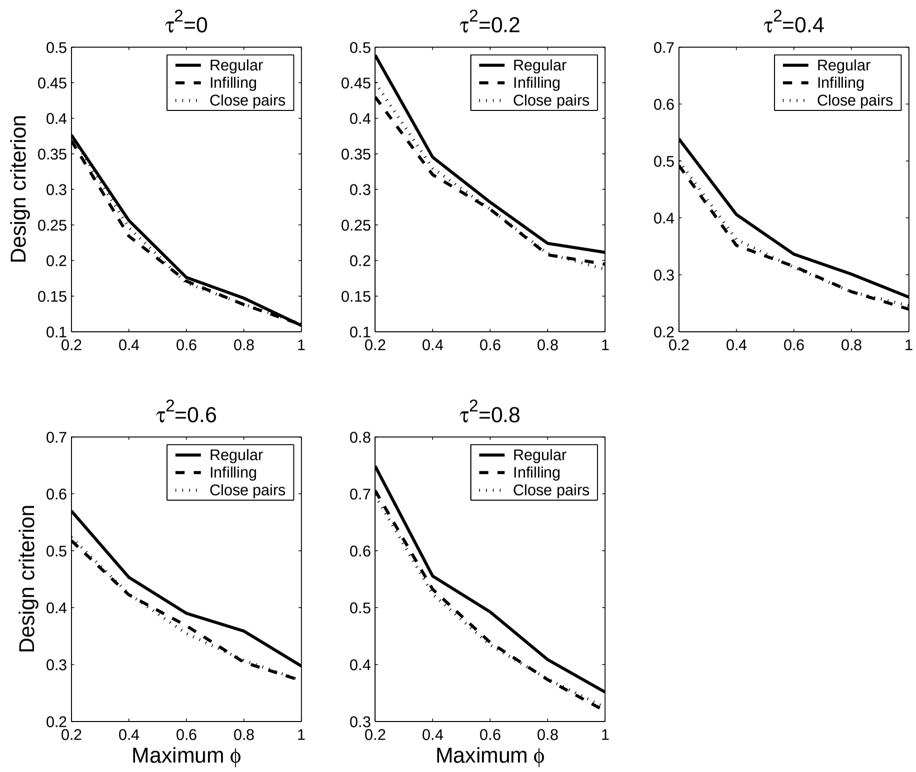
Retrospective: deletion of points from a monitoring network



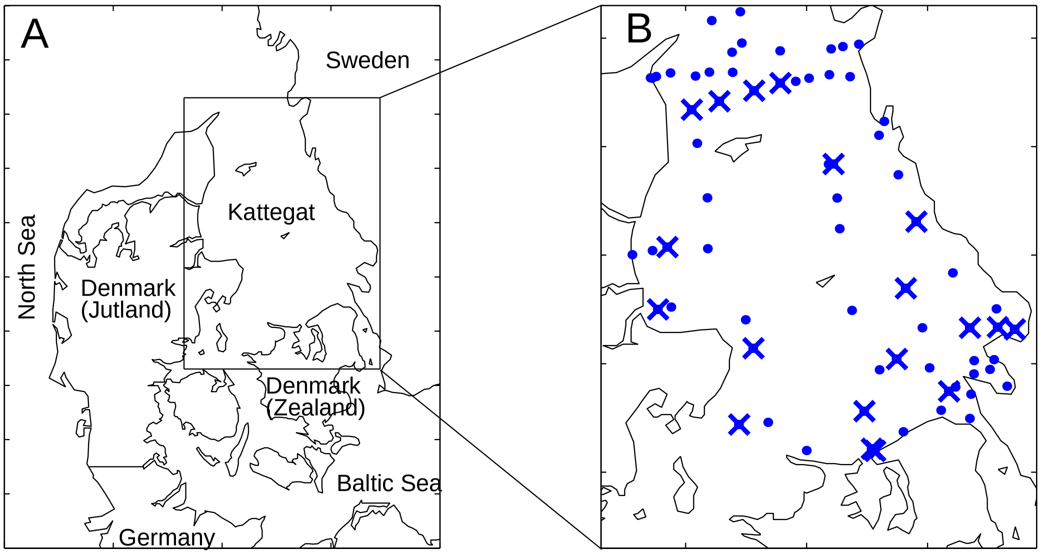
Selected final designs



Prospective: regular lattice vs compromise designs



Monitoring salinity in the Kattegat basin



Solid dots are locations deleted for reduced design.

Further remarks on geostatistical design

1. Conceptually more complex problems include:
   (a) design when some sub-areas are more interesting than others;
   (b) design for best prediction of non-linear functionals of $S(\cdot)$;
   (c) multi-stage designs (see next session).
2. Theoretically optimal designs may not be realistic (eg Loa loa photo).
3. Goal here is NOT optimal design, but to suggest constructions for good, general-purpose designs.

Closing remarks

- There is *nothing special* about geostatistics.
- Parameter uncertainty can have a material impact on prediction.
- Bayesian paradigm deals naturally with parameter uncertainty.
- Implementation through MCMC is not wholly satisfactory:
  – sensitivity to priors?
  – convergence of algorithms?
  – routine implementation on large data-sets?

- **Model-based approach clarifies distinctions between:**
  - **the substantive problem;**
  - **formulation of an appropriate model;**
  - **inference within the chosen model;**
  - **diagnostic checking and re-formulation.**
- **Analyse problems, not data:**
  - **what is the scientific question?**
  - **what data will best allow us to answer the question?**
  - **what is a reasonable model to impose on the data?**
  - **inference: avoid** *ad hoc* **methods if possible**
  - **fit, reflect, re-formulate as necessary**
  - **answer the question.**