Spatial Modelling

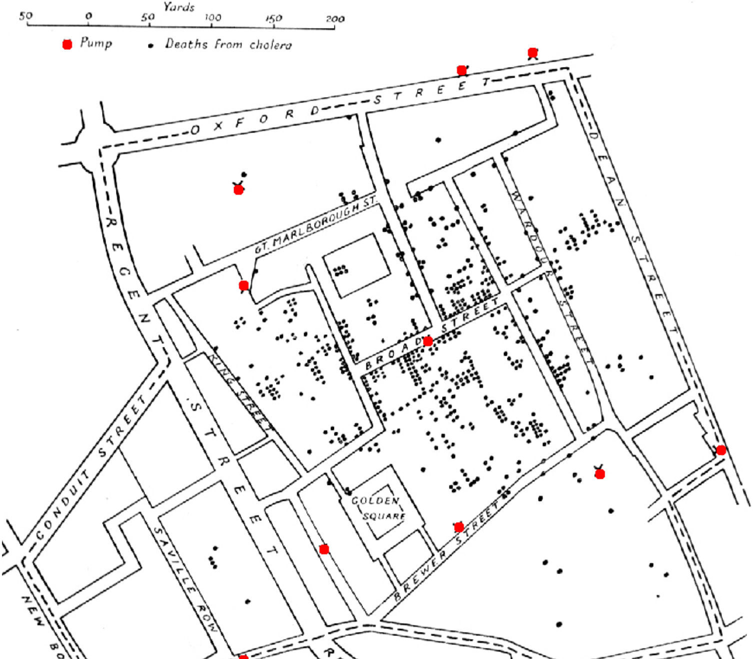
- Patrick Brown
- Cancer Care Ontario
- 620 University Ave, 12th floor.
- patrick.brown@cancercare.on.ca

With thanks to Paulo Ribeiro (Curitiba) and Chris Sherlock (Lancaster) for provision of course notes!

Motivating example

- John Snow and the Broad Street Pump
- Cholera outbreak in Soho, London, in 1854
- London had 1.5 million people and no sewage system
- The prevailing miasma theory said cholera was spread by bad air.
- Snow believed it was transmitted by contaminated water
- He talked to local residents, making note of where they lived

Locations of Cholera Victims



- Cholera cases are more likely to occur close to the Broad Street Pump
- Snow found microbes in the water
- He asked for the handle on the pump to be removed, which stopped the cholera epidemic

- Cholera cases are more likely to occur close to the Broad Street Pump
- Snow found microbes in the water
- He asked for the handle on the pump to be removed, which stopped the cholera epidemic
- Actually, the epidemic was declining before the handle was removed.

Spatial epidemiology

- Data are spatially referenced
- Two, or more, dimensions
- Is there a spatial pattern to disease incidence locations or rates?
- Can we quantify the spatial dependence?
- Is this a simple extension of time series analysis?

Time series analysis?

- A surprisingly complicated extension
- There is no natural ordering for spatial data
- In time series the present depends only on the past
- Y_t depends on Y_{t-1} (and Y_{t-2} and Y_{t-3}?)
- Continuous time $Y(t)$, $dY(t)/dt$, $d^2Y(t)$
- No such simplifications for spatial processes (or they're not as straightforward)
- Spatial models, compared to time series models, are typically
  - Simpler;
  - Computationally more demanding; and
  - limited in the size of dataset they can handle

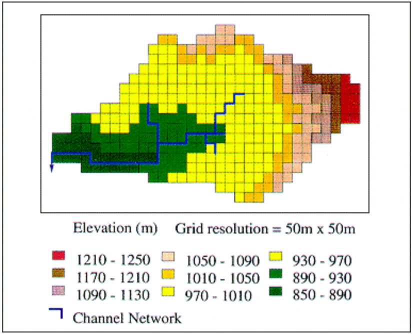
Spatial Statistics

- Geostatistical models
  - A surface which is defined everywhere on a region.
- Discrete spatial variation
  - A surface defined only at discrete points or regions (possibly irregular)
- Spatial point processes
  - Data consist of the locations of events

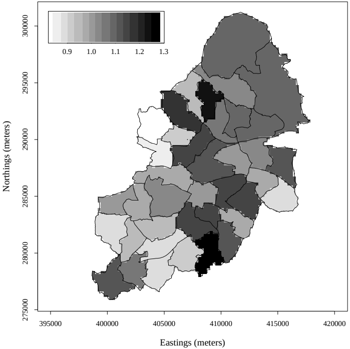
Discrete Processes

- Artificial lattice: pixel grid or census districts

Elevation of a drainage basin



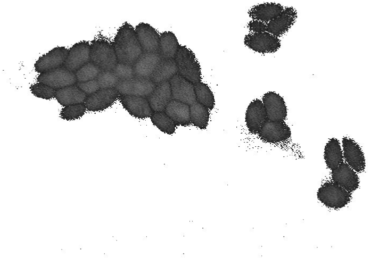
Cancer rates in Birmingham electoral wards



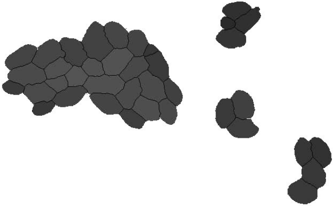
Lattice processes

- Natural lattices: school district boundaries or cell wall boundaries

Cell images



Discrete spatial process

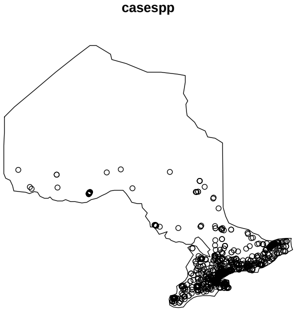


Colours represent cell density

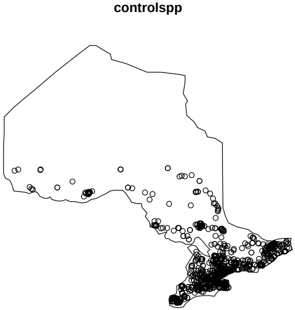
Point processes

- Lung cancer in Ontario

Cases



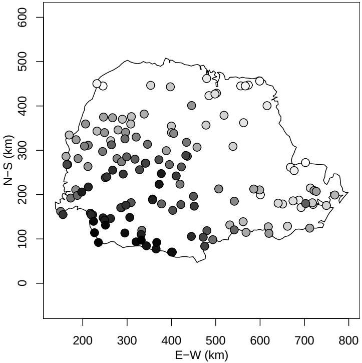
Controls



- Is there spatial variation in cancer incidence?
  - More cases near a specific location such as a power plant?
  - Cases tending to cluster near other cases?

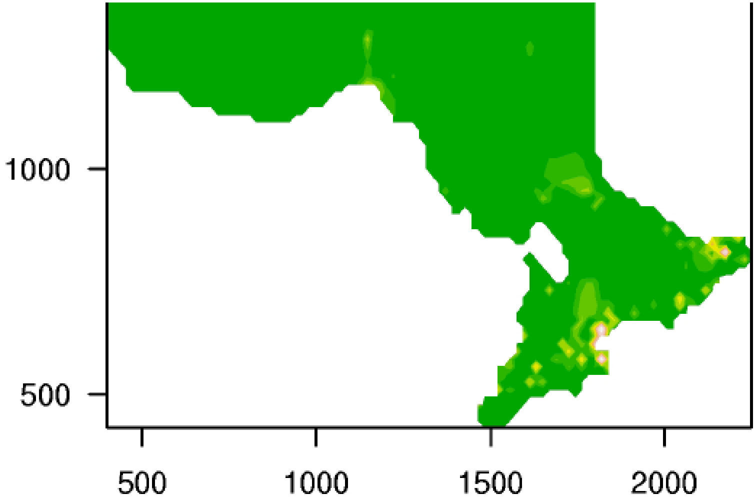
Geostatistics

- Rainfall in Parana state, Brazil
- Exists everywhere, but is only evaluated at a few points



Geostatistics

- Intensity of lung cancer cases in Ontario
- Unobserved, estimated by modelling



This Course

- Geostatistics — 6 weeks
- Point Processes — 3 weeks
- Discrete spatial variation — 1 week
- Markov Random fields ???
- Spatio-temporal models ??

Labs

- 1 hour in 256 McCaul, following lectures
- Using R and WinBUGS
- analyses of spatial datasets
- you're encouraged to work on your own computers.

Assessment

- One small project 20%
- One larger project with a presentation 40%
- Exam 40 %

Books

Main books:

- Diggle and Ribeiro (2006) Model-based Geostatistics `amazon.com/dp/0387329072`
- Diggle (2003) Statistical Analysis of Spatial Point Patterns `amazon.com/dp/0340740701`

Other books:

- Moeler and Wagerpetersen “Statistical inference and simulation for spatial point processes” `www.myilibrary.com/browse/open.asp?ID=19973` for a more technical treatment of point process and model fitting
- Rue and Held “Gaussian Markov Random Fields” `amazon.com/dp/1584884320`, if we do Markov random fields.
- See also *http://www.ai-geostats.org*

Contents

1. **INTRODUCTION**
2. **BASIC GEOSTATISTICAL MODEL**
3. **EXPLORATORY VARIOGRAM ANALYSIS**
4. **PARAMETER ESTIMATION FOR GAUSSIAN MODELS**
5. **SPATIAL PREDICTION FOR GAUSSIAN MODELS**
6. **GENERALISED LINEAR GEOSTATISTICAL MODELS**
7. **FURTHER TOPICS, HISTORY AND PHILOSOPHY**

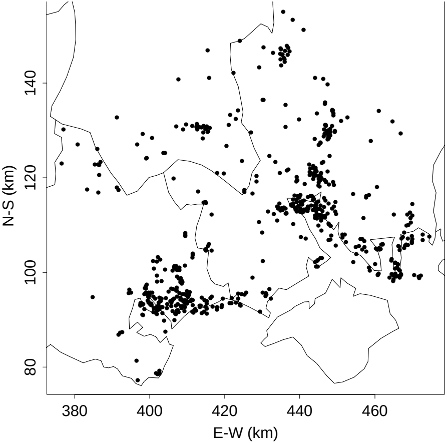
PART I: INTRODUCTION

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. **Core Geostatistical Problems**

Basic Examples of Spatial Data

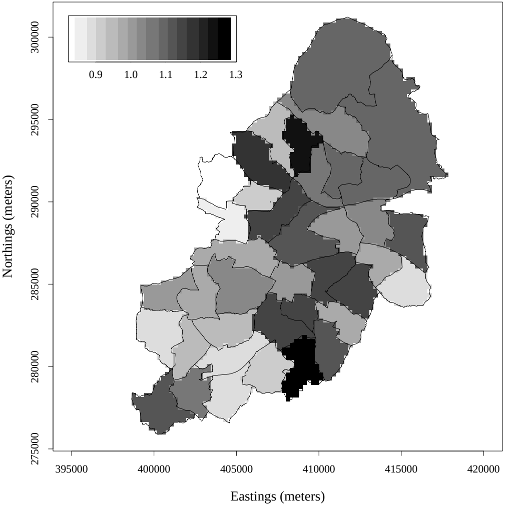
- **Campylobacter cases in southern England**

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



Cancer rates in administrative regions

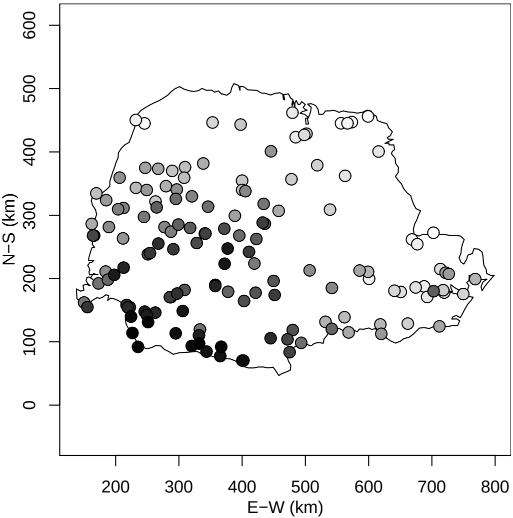
Grey-scale corresponds to estimated variation in relative risk of colorectal cancer in the 36 electoral wards of the city of Birmingham, UK.



Rainfall in Paraná State, Brasil

Rainfall measurements at 143 recording stations.

Average for the May-June period (dry season).



A Taxonomy of Spatial Statistics

1. **Spatial point processes**

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

   e.g. cases of campylobacter.

2. **Discrete spatial variation**

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

   e.g. number of cancer cases in an electoral region.

   - rarely arises naturally
   - but often useful as a pragmatic strategy

3. **Continuous spatial variation**

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

   Data $(y_i, \mathbf{x}_i) : i = 1, ..., n$

   e.g. rainfall measurements at locations x_i.

   Locations may be:

   - non-stochastic (eg lattice to cover observation region A)
   - or stochastic, *but independent of the process* $Y(\mathbf{x})$

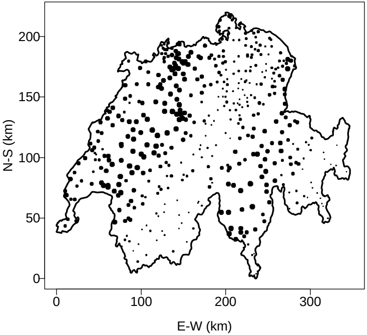
Spatial statistics is the collection of statistical methods in which spatial locations play an explicit role in the analysis of data.

Geostatistics is that part of spatial statistics concerned with data obtained by spatially discrete sampling of a spatially continuous process.

Don't confuse the *data-format* with the *underlying process*

Further Examples of Geostatistical Problems

Swiss rainfall data

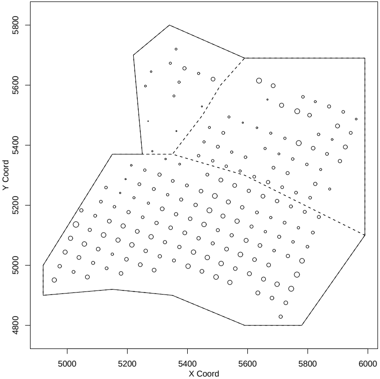


- Locations shown as points with size proportional to the value of the observed rainfall.
- 467 locations in Switzerland
- daily rainfall measurements on 8th of May 1986

data from: *Spatial Interpolation Comparison 97*
`http://www.ai−geostats.org/resources/famous_geostats_data.htm`

Calcium and magnesium contents in a soil

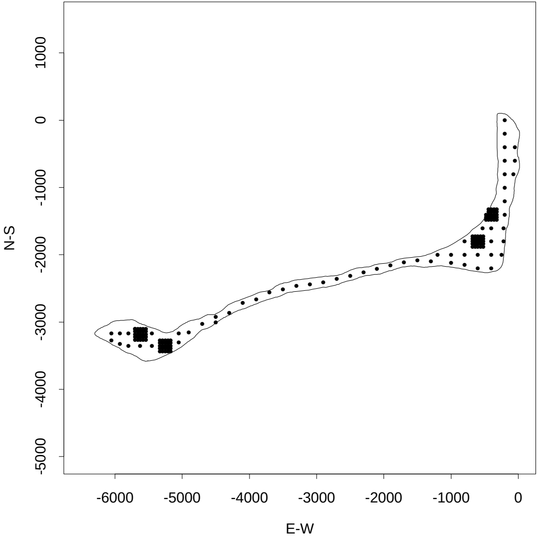
178 measurements of Calcium and Magnesium contents taken on the $0 - 20cm$ (and $20 - 40cm$) soil layers



- fertility maps
- assess effects of soil regime and elevation
- joint model for Ca and Mg

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?

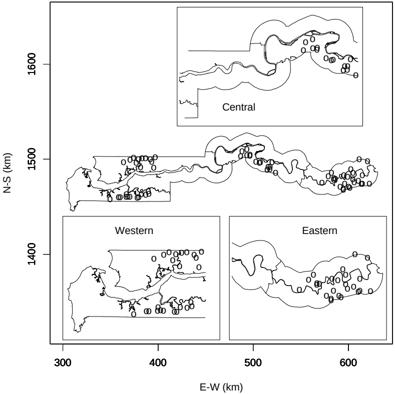


- survey yields noisy measurements Y_i of radioactive caesium concentrations
- initial grid of locations x_i at 200m spacing later supplemented

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
- interest in effects of covariates, and pattern of residual spatial variation

- village-level covariates:
  - village locations
  - public health centre in village?
  - satellite-derived vegetation green-ness index
- child-level covariates:
  - age, sex, bed-net use



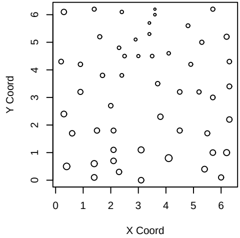
Characteristic Features of Geostatistical Data

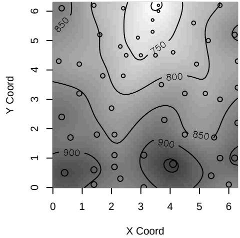
- data consist of **responses** $Y_i := Y(\mathbf{x}_i)$ associated with **locations** $\mathbf{x}_i$
- in principle, Y could be determined from any location $\mathbf{x}$ within a continuous spatial region A
- it is reasonable to behave as if $\{Y(\mathbf{x}) : \mathbf{x} \in A\}$ is a stochastic process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

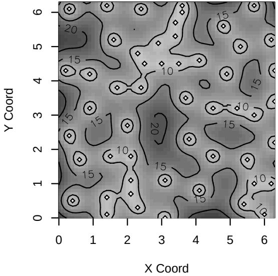
Core Geostatistical Problems

- **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

A basic example: elevation data







Raw data; kriging (with raw data overlaid); and kriging standard errors.

PART II: BASIC GEOSTATISTICAL MODEL

1. **Notation**
2. **The Signal Process**
3. **The Measurement Process**
4. **The Correlation Function**
5. **Model Extensions (1)**

Model-Based Geostatistics

Basic model

response = mean effect + signal + noise

Notation

- $\{\mathbf{x}_i : i = 1, ..., n\}$ is the **sampling design**
- μ or $\mu_i := \mu(\mathbf{x}_i)$ is the **trend** or **mean effect**
- $\{Y(\mathbf{x}) : \mathbf{x} \in A\}$ is the **measurement process**
- $Y_i := Y(\mathbf{x}_i)$ a.k.a. the **response**
- $\{S(\mathbf{x}) : \mathbf{x} \in A\}$ is the **signal process**
- $T = \mathcal{F}(S)$ is the **target for prediction**
- $[S, Y] = [S][Y|S]$ is the **geostatistical model**

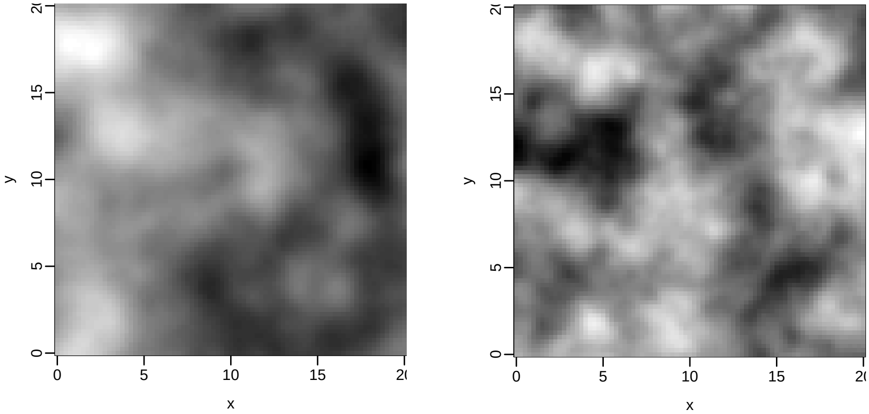
Data consist of pairs $(y_i, \mathbf{x}_i) : i = 1, ..., n$, possibly with covariates measured at each $\mathbf{x}_i$.

The Signal Process

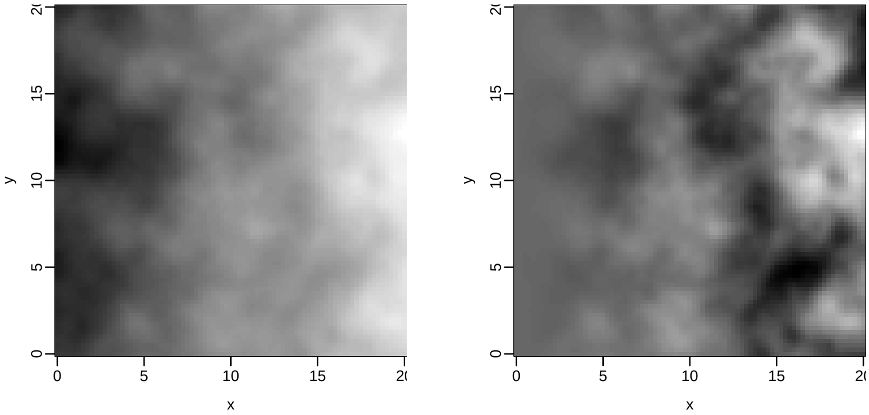
Model the signal process $S(\mathbf{x})$ as a **Gaussian random field** (GRF), also known as a Gaussian process. Initially assume it is **stationary** and **isotropic**

- A **stationary** process is one whose probability distribution is invariant under translation.
- An **isotropic** process is one whose probability distribution is invariant under rotation.

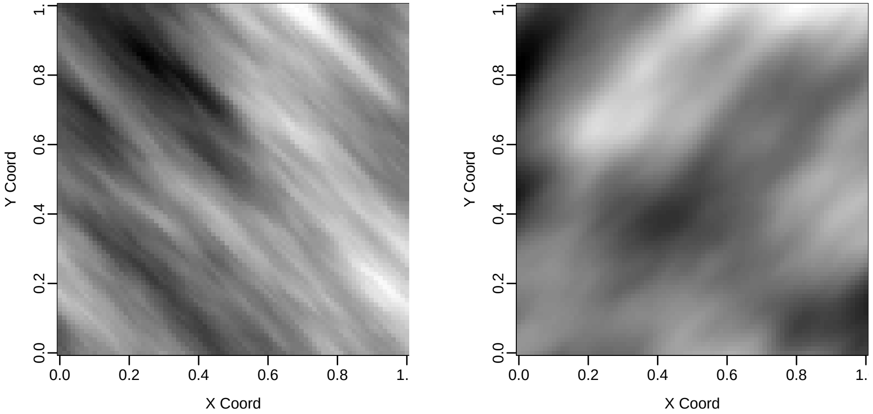
Stationary, Isotropic



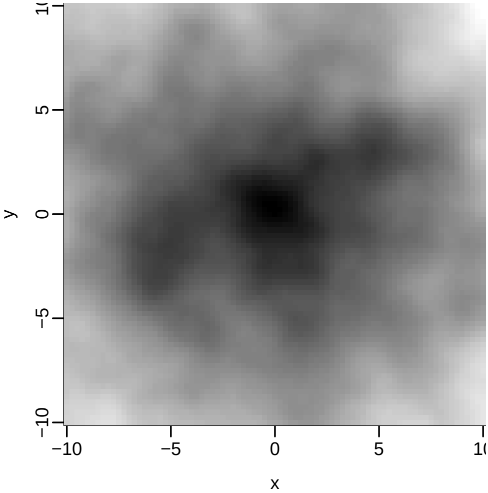
Non-stationary



Stationary, Anisotropic



Isotropic, Non-stationary



Distribution

If $S(\mathbf{x})$ is a stationary isotropic Gaussian process (SGP) then for any set of points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in A$

$$\left[\begin{array}{c} S(\mathbf{x}_1) \\ . \\ S(\mathbf{x}_n) \end{array} \right] \sim N(m\mathbf{1}, \sigma^2\mathbf{R})$$

where $\mathbf{1}$ is a vector of ones and

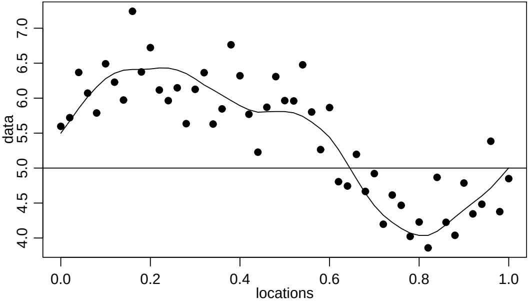
$$R_{ij} = \mathrm{Corr}\left[S(\mathbf{x}_i), S(\mathbf{x}_j)\right] = \rho(||\mathbf{x}_i - \mathbf{x}_j||)$$

for some function $\rho(\cdot)$. Without loss of generality we will always take $m = 0$. Clearly at any one point $\mathbf{x}$

$$S(\mathbf{x}) \sim N(0, \sigma^2)$$

The Measurement Process

1. the conditional distribution of $Y(\mathbf{x}_i)$ given $S(\mathbf{x}_i)$ is $Y_i | s(\mathbf{x}_i) \ \sim \ N(\mu + s(\mathbf{x}_i), \tau^2)$;
2. $Y_i : i = 1, ..., n$ are mutually independent, conditional on $S(\cdot)$.



Simulated data in 1-D illustrating the elements of the model: data $Y(\mathbf{x}_i)$ (●), signal $S(\mathbf{x})$ ($\frown$) and mean μ (—).

An Equivalent Formulation:

$$Y_i = \mu + S(\mathbf{x}_i) + \epsilon_i : i = 1, ..., n.$$

where $S(\mathbf{x})$ has mean 0, and $\epsilon_i : i = 1, ..., n$ are mutually independent, identically distributed with $\epsilon_i \sim \mathrm{N}(0, \tau^2)$.
The joint distribution of $\mathbf{Y}$ is multivariate Normal,

$$\mathbf{Y} := \begin{bmatrix} Y(\mathbf{x}_1) \\ . \\ Y(\mathbf{x}_n) \end{bmatrix} \sim \mathrm{N}(\mu \mathbf{1}, \sigma^2 \mathbf{R} + \tau^2 \mathbf{I})$$

where:
$\mathbf{1}$ is a vector of 1's
$\mathbf{I}$ is the $n \times n$ identity matrix
$\mathbf{R}$ is the $n \times n$ matrix with $(i,j)^{th}$ element $\rho(u_{ij})$ where
$u_{ij} := ||\mathbf{x}_i - \mathbf{x}_j||$, the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$.

Do exercise 1a

The Correlation Function

Positive definiteness

- The variance of some linear combination $a_1 S(\mathbf{x}_1) + \cdots + a_n S(\mathbf{x}_n)$ is

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \mathrm{Cov}\left[S(\mathbf{x}_i), S(\mathbf{x}_j)\right] = \sigma^2 \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j R_{ij}$$

- This must be positive for all possible $a_i \in \Re$ (or possibly zero).
- Not all candidate correlation functions posses this property.

Positive Definite Matrices

- A is positive definite if $x'Ax > 0$ for all x.
- A necessary and sufficient condition for positive definiteness is for all the Eigenvalues of A to be positive.
- Variance matrices must be positive definite, since if $Y \sim N(0, \Sigma)$ then for a vector a then $\mathrm{Var}\left[a'Y\right] = a'\Sigma a$.
- For $a'Y$ to have positive variance for all a, then Σ must be positive defininte.

Positive Definite Functions

- $f(x)$ is a function of $x \in \Re^n$.
- for any set of points $x_1 \ldots x_m$
- create matrix $A_{ij} = f(x_i - x_j)$
- If A is always positive definite, then f is a positive definite function
- for f to be a spatial variance function it must be positive definite
- Otherwise we could find a set of points $u_1 \ldots u_m$ and a vector b such that

  $$\text{Var}\left[b' \begin{pmatrix} Y(u_1) \\ \vdots \\ Y(u_m) \end{pmatrix} \right]$$

  is negative.

Characteristics of positive definite functions

- Non-negative, and monotone decreasing.
- **Bochner's Theorem** states that all p d functions have positive Fourier transforms.
- The Exponential function and the Gaussian density are positive definite.
- The positive definite constraint leads us to use a small number parametric families for covariance functions.

Differentiability of Gaussian processes

- A formal mathematical description of the smoothness of a spatial surface $S(\mathbf{x})$ is its degree of differentiability.
- $S(\mathbf{x})$ is *mean-square continuous* if, for all $\mathbf{x}$,

$$\mathbb{E}\left[\{S(\mathbf{x}+\mathbf{h})-S(\mathbf{x})\}^2\right] \rightarrow 0 \text{ as } ||h|| \rightarrow 0$$

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

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

- the mean-square differentiability of $S(\mathbf{x})$ is directly linked to the differentiability of its covariance function $\rho(u)$.

Theorem Let $S(\mathbf{x})$ be a SGP with correlation function $\rho(u) : u \in \mathbb{R}$. Then:

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

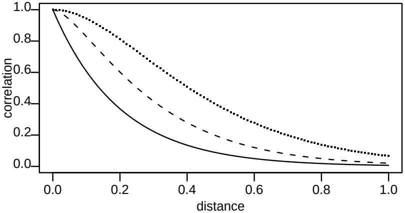
The Matérn family

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

$S(\mathbf{x})$ is mean-square m times differentiable if and only if $\kappa > m$



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).

Simulating the Matérn

```
library(RandomFields)

x <- y <- seq(0, 20, by=1/4)

f <- GaussRF(x=x, y=y, model="whittlematern", grid=TRUE,
             param=c(mean=0, variance=1, nugget=0,
               scale=1, alpha=2))

image(x, y, f, col=topo.colors(100))
```

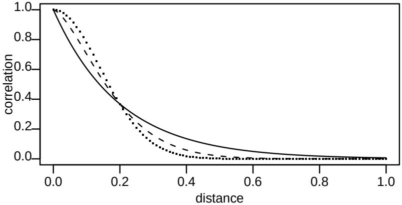
The “alpha” parameter is the roughness parameter κ in our notation.

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

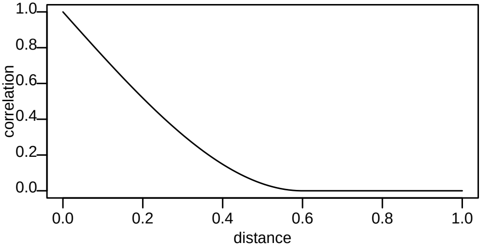


Three examples of the powered exponential correlation function with $\phi = 0.2$ and $\kappa = 1$ (solid line), $\kappa = 1.5$ (dashed line) and $\kappa = 2$ (dotted line).

The spherical family

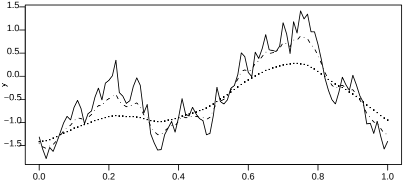
$$\rho(u;\phi) = \begin{cases} 1 - \frac{3}{2}(u/\phi) + \frac{1}{2}(u/\phi)^3 & : \quad 0 \leq u \leq \phi \\ 0 & : \quad u > \phi \end{cases}$$

- $\phi > 0$ is parameter
- finite range
- non-differentiable at the origin
- Has strange properties in the frequency domain which makes estimation unstable.
- Despite the problems it's very widely used.

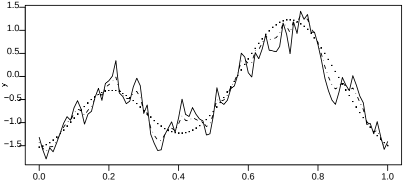


The spherical correlation function with $\phi = 0.6$.

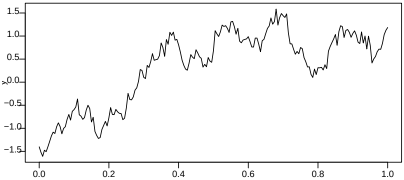
Comparable Simulations (same seed)



Matérn
$\phi = 0.2$ and $\kappa = 0.5$ (—),
$\kappa = 1$ (- - -) and $\kappa = 2$ (. . .).



Powered exponential
$\phi = 0.2$ and $\kappa = 1$ (—),
$\kappa = 1.5$ (- - -) and $\kappa = 2$ (. . .).



Spherical
($\phi = 0.6$).

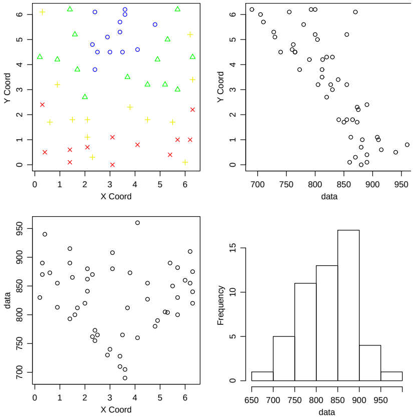
Model Extensions (1)

Removal of trends

Re-examine the elevation data; there is evidence for a linear (quadratic?) trend with co-ordinates.

In general replace constant μ with

$$\mu_i = \mu(\mathbf{x}_i) = \mathbf{f}(\mathbf{x}_i)'\beta = \sum_{j=1}^{k} \beta_j f_j(\mathbf{x}_i)$$



So that

$$Y_i = \mu_i + S(\mathbf{x}_i) + \epsilon_i : i = 1, ..., n.$$

where $\mathbb{E}\left[S(\mathbf{x})\right] = 0$; $S(\mathbf{x})$ remains stationary and isotropic.

Covariates

- $f_1(\mathbf{x}) = \mathbf{1}$ allows for an overall mean
- To incorporate a linear trend in the elevation data, $f_2(\mathbf{x})$ and $f_3(\mathbf{x})$ would be the x and y coords of $\mathbf{x}$.
- In general $f_j(\mathbf{x}_i)$ is any measured covariate at $\mathbf{x}_i$ (or function of it).

NB although the linear trend is only obvious for the y-coord for the elevation data, in general we would fit similar trend effects to both coordinates so as to be independent of the particular axis directions.
Q: how many more parameters would be required for a quadratic trend?

PART III: Exploratory Variogram Analysis

1. **The Variogram**
2. **The Empirical Variogram**
3. **Monte Carlo Variogram Envelope**

NB: Assumption that non-spatial exploratory analysis has already been performed.

The Variogram

The variogram is defined by

$$V(\mathbf{x}, \mathbf{x}') := \frac{1}{2}\mathrm{Var}\left[Y(\mathbf{x}) - Y(\mathbf{x}')\right]$$

Let S be an isotropic SGP with

$$\mathbb{E}\left[S(\mathbf{x})\right] = 0 \quad , \quad \mathrm{Var}\left[S(\mathbf{x})\right] = \sigma^2$$

and correlation function $\rho(u)$. Let the response be

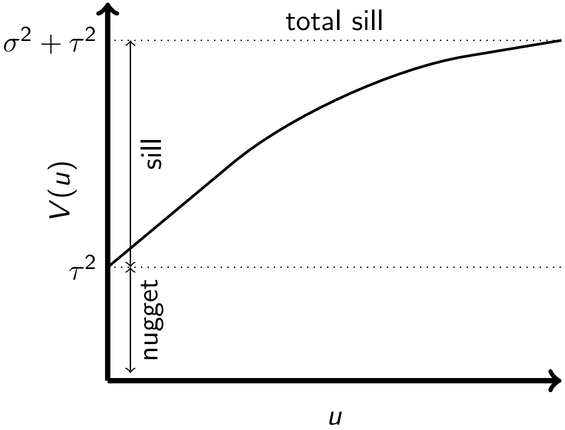
$$Y(\mathbf{x}_i) = \mu_i + S(\mathbf{x}_i) + \epsilon_i$$

where ϵ_i is i.i.d. Gaussian noise $\epsilon_i \sim N(0, \tau^2)$.
Then, writing $u = ||\mathbf{x} - \mathbf{x}'||$, the variogram of Y is

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

For proof see handout.

Interpreting the Variogram



- *the nugget variance*: τ^2
- *the sill*: $\sigma^2 = \mathsf{Var}\left[S(\mathbf{x})\right]$
- *the total sill*: $\tau^2 + \sigma^2 = \mathsf{Var}\left[Y(\mathbf{x})\right]$
- *the range*: ϕ, such that $\rho(u; \phi) = \rho(u/\phi; 1)$

Variogram v Correlation

- Why not just use the correlation function?
- The Variogram is defined for non-stationary processes
- The Variogram is easier to estimate for irregular data

The Empirical Variogram

- if $Y(\mathbf{x})$ is stationary and isotropic,

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

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

$$\hat{V}(u) = \mathrm{average}\{[y(\mathbf{x}_i) - y(\mathbf{x}_j)]^2\}$$

  where each average is taken over all pairs $[y(\mathbf{x}_i), y(\mathbf{x}_j)]$ such that $||\mathbf{x}_i - \mathbf{x}_j|| \approx u$

- for a process with non-constant mean (covariates), trends may be removed as follows:
  - let $\hat{\beta}$ be the OLS estimate
  - and $\hat{\mu}(\mathbf{x}_i) = \mathbf{f}(\mathbf{x}_i)'\hat{\beta}$
  - define $r_i := Y_i - \hat{\mu}(\mathbf{x}_i)$
  - define $\hat{V}(u) = \mathrm{average}\{(r_i - r_j)^2\}$,
    where each average is taken over all pairs (r_i, r_j)

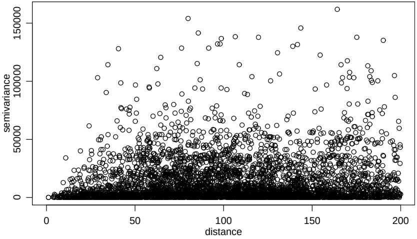
The variogram cloud

- define the quantities

$$\begin{aligned} r_i &= y_i - \hat{\mu}(\mathbf{x}_i) \\ u_{ij} &= ||\mathbf{x}_i - \mathbf{x}_j|| \\ v_{ij} &= \frac{(r_i - r_j)^2}{2} \end{aligned}$$

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

Example: Swiss rainfall data

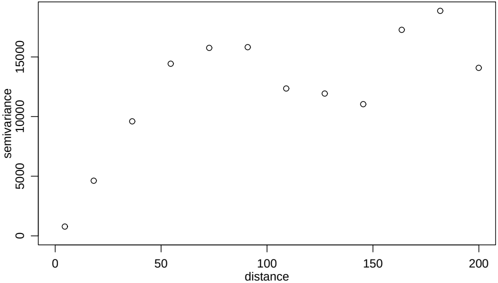


- under the spatial Gaussian model:
  - $V_{ij} \sim V(u_{ij})\chi_1^2$
  - the v_{ij} are correlated
- the variogram cloud is therefore unstable, both pointwise and in its overall shape

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 $\mu(\mathbf{x})$

Example: Swiss rainfall data.



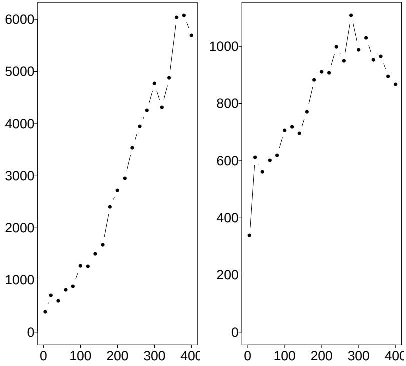
Empirical binned variogram

Do exercise 3a (lecturer), b/c (class)

Variograms of raw data and residuals can be very different

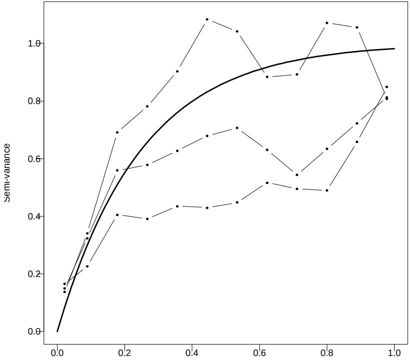
Example: Paraná rainfall data.

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

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

Monte Carlo Variogram Envelope

A simple test for spatial correlation.

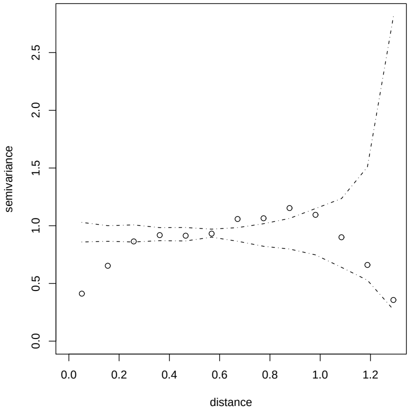
- $\mathbf{H}_0$: there is no spatial correlation.
- Under H_0 the relative spatial positions of the data (or residuals) are irrelevant
- Under H_0 the data may be permuted and the resulting empirical variogram should arise from the same underlying distribution of variograms as the original.

The Algorithm

Repeat k times

1. randomly permute the data
2. calculate the empirical variogram for this permutation

For each u use the lowest and highest (or 5^{th} and 95^{th} percentiles) of the simulated $V(u)$'s as envelopes (under H_0) for the true value $V(u)$.



Variogram and envelope for simulated data with $\mu = 0, \sigma^2 = 1, \tau^2 = 0.25$ and $\phi = 0.3$.

PART IV: PARAMETER ESTIMATION FOR GAUSSIAN MODELS

1. **Maximum Likelihood Estimation**
2. **Model Extensions (2) - Box-Cox**
3. **A Case Study: Swiss rainfall data**
4. **Model Extensions (3) - Anisotropy**
5. **Not Covered Here**
6. **Bayesian estimation of parameters**

Maximum Likelihood Estimation

The model

$$Y(\mathbf{x}_i) = \mu_i + S(\mathbf{x}_i) + \epsilon_i$$

- **mean** $\mu_i = \mu(\mathbf{x}_i) = \sum_{j=1}^{k} \beta_j f_j(\mathbf{x}_i)$ i.e. $\boldsymbol{\mu} = \mathbf{F}\boldsymbol{\beta}$
- **SGP** $S(\mathbf{x})$ with $\mathbb{E}\left[S(\mathbf{x})\right] = 0$, $\mathrm{Var}\left[S(\mathbf{x})\right] = \sigma^2$ and $\mathrm{Corr}\left[S(\mathbf{x}_1), S(\mathbf{x}_2)\right] = \rho_k(||\mathbf{x}_1 - \mathbf{x}_2|| \,; \phi)$
- **nugget effect** Gaussian noise $\epsilon_i \sim N(0, \tau^2)$

Joint distibution

$$\mathbf{Y} \sim N\left(\mathbf{F}\boldsymbol{\beta}, \sigma^2\mathbf{R} + \tau^2\mathbf{I}\right)$$

where $R_{ij}(\phi, \kappa) = \rho(||\mathbf{x}_i - \mathbf{x}_j||)$.

Re-parametrise

- Signal to noise ratio: $\nu^2 := \frac{\tau^2}{\sigma^2}$
- $\mathbf{R}_*(\nu, \phi, \kappa) = \mathbf{R}(\phi, \kappa) + \nu^2 \mathbf{I}$
- $\mathbf{Y} \ \sim \ N\left(\mathbf{F}\boldsymbol{\beta}, \sigma^2 \mathbf{R}_*\right)$
- In this parametrisation σ will drop out of the likelihood.

log-likelihood

$$\ell(\boldsymbol{\beta}, \sigma^2, \nu, \phi, \kappa) = \\ -\frac{n}{2} \log 2\pi - \frac{1}{2} \log \left|\left|\sigma^2 \mathbf{R}_*\right|\right| - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{F}\beta)' \mathbf{R}_*^{-1} (\mathbf{y} - \mathbf{F}\beta)$$

Profile likelihood

- There is no closed-form analytical solution for the parameters which maximise the likelihood
- A numerical optimisation routine will have to be used
- If we know ν, ψ, and κ (the spatial correlation parameters) then there is an analytical solution for β and σ^2 as the model is linear.
- **The idea**: use the numerical optimiser on ν, ψ, and κ , using the analytic expressions for β and σ^2

$$\max_{\beta,\sigma^2,\nu,\phi,\kappa} \ell(\boldsymbol{\beta},\sigma^2,\nu,\phi,\kappa) = \max_{\nu,\phi,\kappa}\left(\max_{\beta,\sigma^2} \ell(\boldsymbol{\beta},\sigma^2,\nu,\phi,\kappa)\right)$$

By standard results of linear models (see 'Supplementary material'), the log-likelihood $\ell(\boldsymbol{\beta},\sigma^2|\nu^2,\phi,\kappa)$ is maximised at

$$\begin{aligned}
\hat{\boldsymbol{\beta}}(\nu,\phi,\kappa) &= \left(\mathbf{F}'\mathbf{R}_*^{-1}\mathbf{F}\right)^{-1}\mathbf{F}'\mathbf{R}_*^{-1}\mathbf{y} \\
\hat{\sigma^2}(\nu,\phi,\kappa) &= \frac{1}{n}\left(\mathbf{y}-\mathbf{F}\hat{\boldsymbol{\beta}}\right)'\mathbf{R}_*^{-1}\left(\mathbf{y}-\mathbf{F}\hat{\boldsymbol{\beta}}\right)
\end{aligned}$$

Profile Likelihood (cont)

Inserting the expressions for β and σ into the likelihood function gives the profile likelihood function

$$\begin{aligned} \ell^*(\nu, \phi, \kappa) &= -\frac{n}{2}\log 2\pi - \frac{1}{2}\log\left|\left|\hat{\sigma}^2\mathbf{R}_*\right|\right| - \frac{n}{2} \\ -2\ell^*(\nu, \phi, \kappa) + c &= n\log\hat{\sigma}^2 + ||\mathbf{R}_*|| \end{aligned}$$

where c is a constant.

- Use a numerical optimiser (such as `optim` in R) to find $\hat{\nu}, \hat{\phi}, \hat{\kappa}$
- Back substitution gives $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma^2}$.

- any reasonable version of the (linear) spatial Gaussian model has at least three parameters
  - A spatial variance parameter, for how close the process stays to the mean;
  - Observation error variance, to take care of uncorrelated noise; and
  - A range parameter so that changing between miles and km doesn't affect the model
- You need a lot of data (or contextual knowledge) to justify estimating more than three parameters
- the **Matérn** family adds a fourth, roughness, parameter.
  - Stein (1999)'s book shows, the roughness parameter has a lot of influence on the other parameter's estimates
  - Smooth surface $\Rightarrow$ low signal to noise ratio
  - It is recommended to try a small number of discrete κ e.g. $\{0.5, 1, 2\}$
  - The profile likelihood function for κ is usually fairly flat, and more precise estimation is usually not warranted.

Model Extensions (2) - Box-Cox

- The Gaussian model might be inappropriate for variables with asymmetric distributions.
- Log transforms often Normalise positive-valued data with a heavy right tail
- Squaring data works with a heavy left tail.
- The Box-Cox transformation has a parameter λ offering a continuous range of transformations.
- Box-Cox is commonly used in linear regression.
- Terminology: *Gaussian-transformed model*.

Box-Cox (continued)

The model is defined as follows:

- assume a variable $\mathbf{Y}^* \sim MVN(\mathbf{F}\boldsymbol{\beta}, \sigma^2 \mathbf{R}_*)$
- the data, denoted $\mathbf{y} = (y_1, ..., y_n)$, are generated by a transformation of the linear Gaussian model

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

The log-likelihood is:

$$\begin{aligned} \ell(\boldsymbol{\beta}, \sigma^2, \nu, \phi, \kappa, \lambda) = c & - & \frac{n}{2} \log \sigma^2 - \frac{1}{2} \log |R_*| \\ & + & (h_\lambda(\mathbf{y}) - \mathbf{F}\boldsymbol{\beta})' \{\sigma^2 \mathbf{R}_*\}^{-1} (h_\lambda(\mathbf{y}) - \mathbf{F}\boldsymbol{\beta})\} \\ & + & (\lambda - 1) \sum_{i=1}^{n} \log y_i \end{aligned}$$

Here $h_\lambda(\mathbf{y}) := (h_\lambda(y_1), \ldots, h_\lambda(y_n))'$.

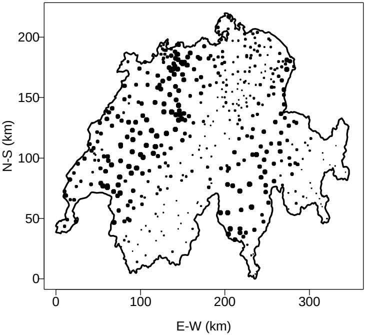
Notes:

- Requires all $y_i > 0$.
  - if some $y_i = 0$ simply through rounding then replace with ‘imputed’ low values.
  - if some $y_i = 0$ because there is a probability mass at 0 then the model is strictly inappropriate.
- Allowing any $\lambda \in \Re$ and simply maximising the log-likelihood can lead to difficulties in scientific interpretation.
  - Allow only a small set of interpretable values e.g. $\{-1, 0, 0.5, 1\}$.
  - Examine the profile log-likelihood for λ to choose the most appropriate value.
  - Transform the data then analyse as standard case.

- Optimisation is CPU intensive for large datasets
    - Most of the information for λ is in the marginal likelihood (i.e. ignoring spatial variation)
    - Obtain a point estimate by maximising

$$
\begin{aligned}
\ell^{*}(\boldsymbol{\beta}, \sigma^2, \lambda) &= c - \frac{n}{2} \log \sigma^2 \\
&\quad - \frac{1}{2\sigma^2}(h_\lambda(\mathbf{y}) - \mathbf{F}\boldsymbol{\beta})'(h_\lambda(\mathbf{y}) - \mathbf{F}\boldsymbol{\beta})\} + (\lambda - 1) \sum_{i=1}^{n} \log y_i
\end{aligned}
$$

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.

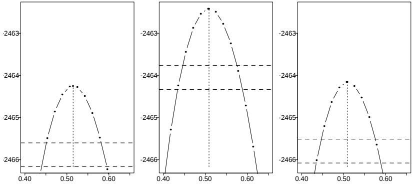
- 467 locations in Switzerland
- daily rainfall measurements on 8th of May 1986
- The data values are integers where the unit of measurement is 1/10 mm
- 5 locations where the value is equal to zero.

Swiss rainfall data (cont.)

MLE of Box-Cox parameter λ for different values of the Matérn roughness parameter κ.

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

Profile likelihoods for λ (–·–), with 90% and 95% confidence limits (- - -)



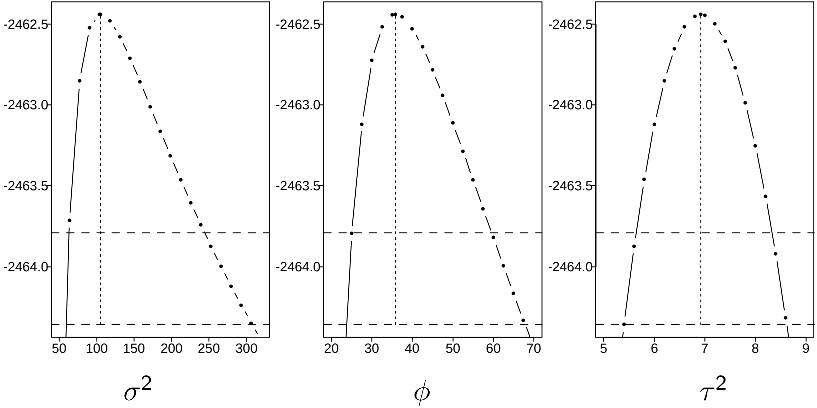
$\kappa = 0.5$, $\kappa = 1$, $\kappa = 2$.

- In all cases $\lambda = 0.5$ is within the interval but untransformed and log-transformed cases are not.

Parameter estimates 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

Profile likelihoods with $\kappa = 1$ and $\lambda = 0.5$



The log-transformation ($\lambda = 0$)

- Log Gaussian data tend to have sharp peaks and large shallow troughs.
- On the log scale, $Y^*(\mathbf{x}) = \log[Y(\mathbf{x})]$.

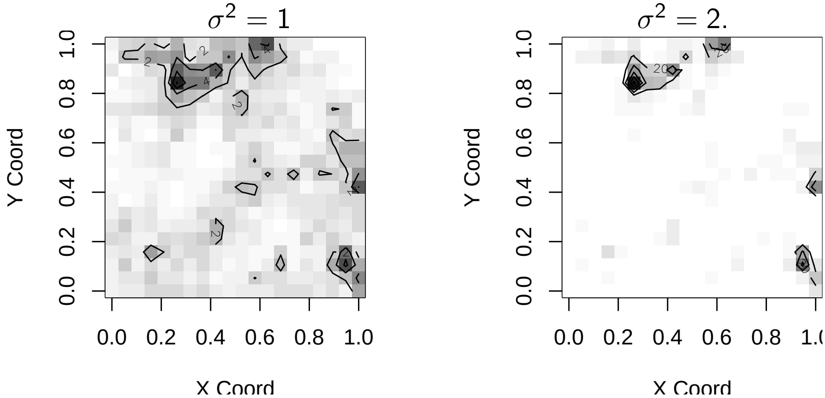
$$Y^*(\mathbf{x}) = \mu + S(\mathbf{x}) + \epsilon = \mu + \sigma^2 Z(\mathbf{x}) + \epsilon(x)$$

- On the natural scale,

$$Y(\mathbf{x}) = e^{Y^*(\mathbf{x})} = e^{\mu}(e^{Z(\mathbf{x})})^{\sigma^2} e^{\epsilon(x)}$$

  The larger σ^2 the sharper the peaks and softer the troughs.
- Remember $\mathrm{E}(Y(x)) = \exp[\mathrm{E}(Y^*(x))] + \mathrm{Var}\left[Y^*(x)\right]/2$.

Simulations of log-Gaussian processes

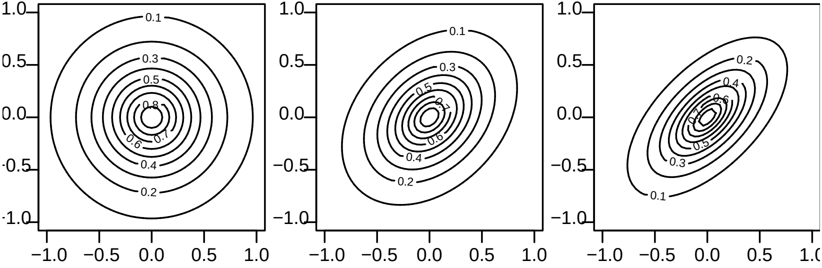


$\sigma^2 = 1$

$\sigma^2 = 2.$

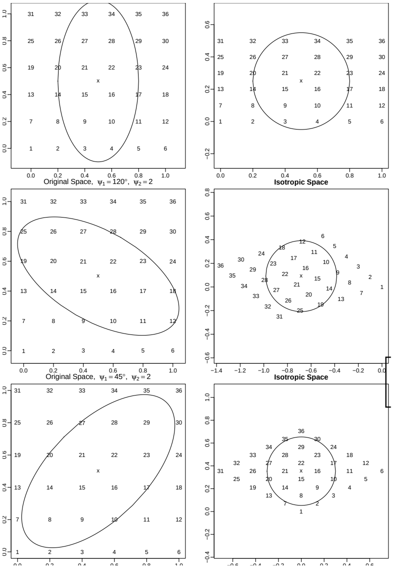
Model Extensions (3) - Anisotropy

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



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

Geometric Anisotropy



- two more parameters: the *anisotropy angle* ψ_A and the *anisotropy ratio* $\psi_R \geq 1$.
- transform the true co-ordinates (x_1, x_2) to new co-ordinates (x_1', x_2') by rotation and squashing.

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

- Analyse the data with respect to the new co-ordinate system.

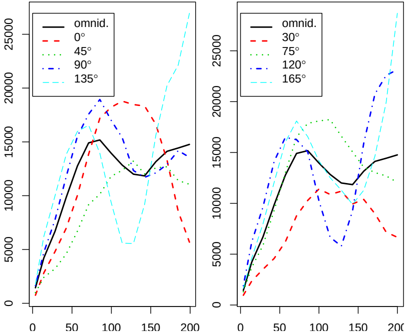
Parameter Estimation

Likelihood based methods

- Add two parameters (angle and squashing)
- Increases the dimension of the numerical minimisation problem
- In practice a lot of data might be needed

Variogram based exploration

- Compute variograms for different directions
- Angle bins, in particular for irregularly distributed data
- Directional variograms for the Swiss rainfall data. ⇒



Not covered here

Restricted maximum likelihood (REML)

- transform the data $\mathbf{Y} \rightarrow \mathbf{Y}^*$ so that $\mathbf{Y}^*$ does not depend on β
- estimate $(\sigma^2, \nu, \phi, \kappa)$ by maximum likelihood on the transformed data $\mathbf{Y}^*$
- leads to less biassed estimates in small samples
- MLE's are sensitive to mis-specification of $\mathbf{F}$ (the covariate mode for $\boldsymbol{\mu}$)

Also not covered

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.

Bayesian Estimation Of Parameters

As before:

the model: S is an SGP with $\mathbb{E}[S(\mathbf{x})] = 0$, $\text{Var}[S(\mathbf{x})] = \sigma^2$, and correlation function $\rho(u; \phi)$. Response is

$$Y(\mathbf{x}_i) = \mu + S(\mathbf{x}_i) + \epsilon_i$$

with i.i.d. Gaussian noise $\epsilon_i \sim N(0, \tau^2)$ ("nugget").

reparameterisation:

$$\nu^2 := \frac{\tau^2}{\sigma^2}$$

correlation matrix : has elements $R_{ij}(\phi) := \rho(||\mathbf{x}_i - \mathbf{x}_j|| ; \phi)$.

Define

$$\mathbf{R}_*(\phi, \nu) := \mathbf{R}(\phi) + \nu^2 \mathbf{I}$$

Bayesian stuff

A judicious choice of priors yields an convenient posterior

priors for ϕ and ν Choose a *discrete* prior for ϕ and ν

$$\pi_{\phi,\nu}(\phi,\nu)$$

prior for σ^2 and μ Choose continuous priors

$$\begin{aligned} \left(\frac{\sigma^2}{n_\sigma S_\sigma^2}\right)^{-1} |\phi,\nu & \sim \chi^2_{n_\sigma} \\ \mu|\sigma,\phi,\nu & \sim \mathsf{N}\left(m_\mu, \sigma^2 V_\mu\right) \end{aligned}$$

This is known as a Gaussian-scaled-Inverse-χ^2 distribution on (μ, σ^2).

posterior for (ϕ, ν) : a *discrete* posterior with

$$p(\phi, \nu|\mathbf{y}) \propto \pi_{\phi,\nu}(\phi, \nu) \ ||\mathbf{R}_*||^{-1/2} \, V_*^{1/2} \left(S_*^2\right)^{-\frac{1}{2}(n_\sigma+n)} \tag{1}$$

posterior for (μ, σ^2) : A Gaussian-scaled-Inverse-χ^2 posterior distribution for $\mu, \sigma^2|\phi, \nu$

$$\left(\frac{\sigma^2}{(n_\sigma+n)S_*^2}\right)^{-1} |\phi, \nu, \mathbf{y} \quad \sim \quad \chi^2_{n_\sigma+n} \tag{2}$$

$$\mu|\sigma^2, \phi, \nu, \mathbf{y} \quad \sim \quad \mathrm{N}\left(m_*, \sigma^2 V^*\right) \tag{3}$$

where

$$V_* \quad := \quad \left(V_\mu^{-1} + \mathbf{1}'\mathbf{R}_*^{-1}\mathbf{1}\right)^{-1}$$
$$m_* \quad := \quad V^*\left(m_\mu V_\mu^{-1} + \mathbf{1}'\mathbf{R}_*^{-1}\mathbf{y}\right)$$

and

$$S_*^2 := \frac{1}{n_\sigma + n}\left(m_\mu^2 V_\mu^{-1} + n_\sigma S_\sigma^2 + \mathbf{y}'\mathbf{R}_*^{-1}\mathbf{y} - m_*^2 V_*^{-1}\right)$$

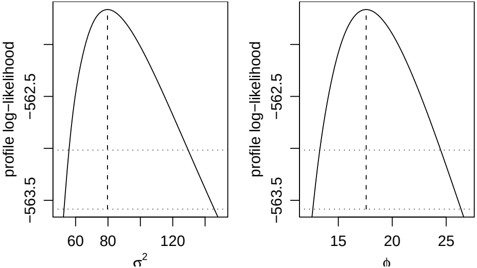
see handout for proof (the proof is ***not examinable****.)*

Monte Carlo Algorithm

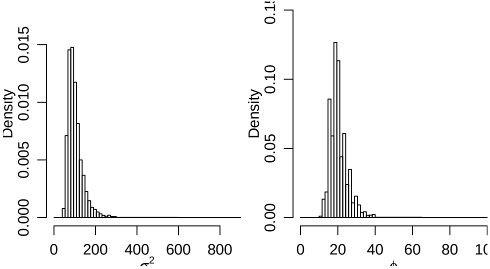
- Sum the right hand side of (**??**) over the finite number of possible values of (ϕ, ν) and divide by this to obtain the posterior $p(\phi, \nu|\mathbf{y})$ for each combination of (ϕ, ν).
- Simulate directly from the full posterior by the following Monte Carlo algorithm
  - simulate $\phi^{(i)}$ and $\nu^{(i)}$ from $p(\phi, \nu|\mathbf{y})$.
  - simulate from the posterior for $\sigma^2|\phi^{(i)}, \nu^{(i)}$ using (**??**).
  - simulate from the posterior for for $\mu|\sigma^{2\ (i)}, \phi^{(i)}, \nu^{(i)}$ using (**??**).
  - $i \leftarrow i + 1$; repeat.

A Case Study: Swiss rainfall, 100 data

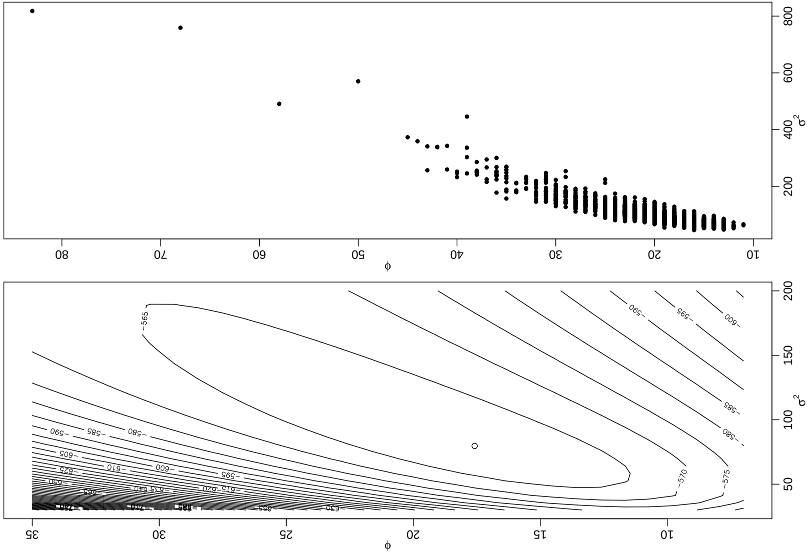
Profile Likelihoods



Posterior distributions



Joint posterior



Samples and contours

Extensions

Add covariates : with $\boldsymbol{\mu} = \mathbf{F}\boldsymbol{\beta}$, *but how does this affect the posterior?*

Vary all parameters : $\kappa, \lambda, \psi_R, \psi_A$ are fixed. In principal these could be included in the analysis, with discrete priors.

improper priors : certain simple improper conjugate priors for μ (*flat*) and σ^2 (*reciprocal*) are often chosen and still lead to proper posteriors (subject to reparameterisation to ν not τ)

- $\pi_\mu(\mu) \propto 1$ ('$V_\mu \to \infty$')
- $\pi_\sigma(\sigma^2) \propto 1/\sigma^2$ ('$n_\sigma \to 0$'). This is the commonly used Jeffrey's prior.
- but see note in section on GLSM's.

ML v Bayes

Bayes

- Allows for for parameter uncertainty to carry over to predictions
- Less damage caused by inclusion of poorly identified parameters
- More exact parameter confidence intervals (ML is asymptotic)
- Can incorporate prior information
- Bayes is necessary for non-Gaussian responses (more on that later).

ML

- Not affected by priors
- Computationally simpler

PART V: SPATIAL PREDICTION FOR GAUSSIAN MODELS

1. **Stochastic Process Prediction**
2. **Prediction under the Gaussian Model**
3. **What does Kriging Actually do to the Data**
4. **Prediction of linear Functionals**
5. **Plug-in Prediction**
6. **Model Validation and Comparison**

Stochastic Process Prediction

General results for prediction

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

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

the mean square error (MSE) of $\hat{T}$ is

$$MSE(\hat{T}) = \mathbb{E}\left[(T - \hat{T}(\mathbf{Y}))^2\right]$$

(expectation over both T and $\mathbf{Y}$)

the MMSE predictor minimises the MSE

Theorem

The minimum mean square error predictor of T is

$$\hat{T} = \mathbb{E}\left[T|\mathbf{Y}\right]$$

at which point

$$\mathbb{E}\left[(T-\hat{T})^2\right] = \mathbb{E}_{\mathbf{Y}}\left[\mathrm{Var}\left[T|\mathbf{Y}\right]\right]$$

(the prediction variance is an estimate of the MSE) □
See handout for proof.
Also, directly from the second tower property

$$\mathrm{Var}\left[T\right] = \mathbb{E}_{\mathbf{Y}}\left[\mathrm{Var}\left[T|\mathbf{Y}\right]\right] + \mathrm{Var}_{\mathbf{Y}}\left[\mathbb{E}_{\mathbf{Y}}\left[T|\mathbf{Y}\right]\right]$$

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

Comments

- We call $\hat{T}$ the *least squares predictor* for T, and $\text{Var}\,[T|\mathbf{Y}]$ its *prediction variance*
- $\text{Var}\,[T] - \text{Var}\,[T|\mathbf{Y}]$ measures the contribution of the data (exploiting dependence between T and $\mathbf{Y}$)
- point prediction and prediction variance are summaries
- complete answer is the distribution $[T|\mathbf{Y}]$
- not transformation invariant: $\hat{T}$ the best predictor for T does NOT necessarily imply that $g(\hat{T})$ is the best predictor for $g(T)$.

Prediction Under The Gaussian Model

- **assume all the parameters $\boldsymbol{\beta}, \sigma^2, \tau^2, \phi, \kappa$ are known**
- assume that the target for prediction is $T = S(\mathbf{x}')$
- $\hat{T} = \mathbb{E}\left[T|\mathbf{Y}\right]$, $\mathrm{Var}\left[T|\mathbf{Y}\right]$ and $\left[T|\mathbf{Y}\right]$ can be easily derived from a standard result.

Under the Gaussian model $Y(\mathbf{x}_i) = \mu_i + S(\mathbf{x}_i) + \epsilon_i$

$$\begin{bmatrix} T \\ \mathbf{Y} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \boldsymbol{\mu} \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & \mathbf{r}' \\ \mathbf{r} & \mathbf{R} + \nu^2 \mathbf{I} \end{bmatrix} \right)$$

$\boldsymbol{\mu} = \mathbf{F}\boldsymbol{\beta}$ and $\mathbf{r}$ is a vector with elements
$r_i = \rho_\kappa(||\mathbf{x}' - \mathbf{x}_i||; \phi) : i = 1, \ldots, n$ Again define $\mathbf{R}_* = \mathbf{R} + \nu^2 \mathbf{I}$

Conditional Distribution

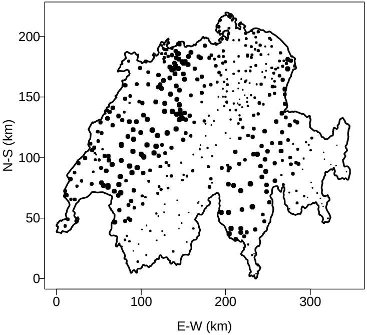
Using *background results on partitioning the MVN* with $\mathbf{Z}_1 = T$ and $\mathbf{Z}_2 = \mathbf{Y}$, we find that the minimum mean square error predictor for $T = S(\mathbf{x})$ is

$$\begin{aligned} \hat{T} &= \sigma^2 \mathbf{r}' \, (\sigma^2 \mathbf{R}_*)^{-1} (\mathbf{y} - \boldsymbol{\mu}) \\ &= \mathbf{r}' \, (\mathbf{R}_*)^{-1} (\mathbf{y} - \boldsymbol{\mu}) \qquad (4) \end{aligned}$$

with prediction variance

$$\begin{aligned} \mathrm{Var}\left[T | \mathbf{Y}\right] &= \sigma^2 - \sigma^2 \mathbf{r}' \, (\sigma^2 \mathbf{R}_*)^{-1} \sigma^2 \mathbf{r} \\ &= \sigma^2 \left(1 - \mathbf{r}' \, (\mathbf{R}_*)^{-1} \mathbf{r}\right) \qquad (5) \end{aligned}$$

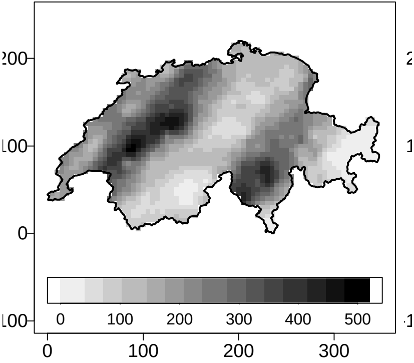
Exampe:Swiss rainfall data



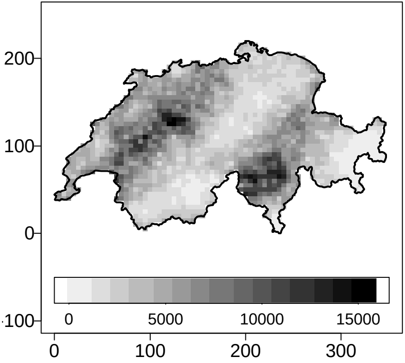
- Locations shown as points with size proportional to the value of the observed rainfall.
- 467 locations in Switzerland
- daily rainfall measurements on 8th of May 1986
- $\hat{\kappa} = 1$, $\hat{\mu} = 20.13$, $\hat{\sigma}^2 = 105.06$, $\hat{\phi} = 35.79$, $\hat{\tau}^2 = 6.92$

Swiss rainfall data (cont.)

Predictions $\mathrm{E}(S(x)|Y_1 \dots Y_n)$



Variances $\mathrm{Var}\,[S(x)|Y_1 \dots Y_n]$



Notes

1. Applies whether $\mathbf{x}'$ is a new point or a data point.
2. Because the conditional variance does not depend on $\mathbf{Y}$, the prediction mean square error is equal to the prediction variance.
3. Equality of prediction mean square error and prediction variance (for any $\mathbf{y}$) is a special property of the multivariate Gaussian distribution, not a general result.
4. In conventional geostatistical terminology, construction of the surface $\hat{T} = \mu(\mathbf{x}) + \hat{S}(\mathbf{x})$ using (**??**) is called *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).
5. Easy to extend to finding the expectation and joint covariance matrix $\boldsymbol{\Sigma}$ of the signal at a set of points: $\mathbf{S}_G := [S(\mathbf{x}'_1), \ldots, S(\mathbf{x}'_g)]'$ given the data (this is a complete specification of the distribution since $\mathbf{S}_G \sim MVN$).

What Does Kriging Actually Do?

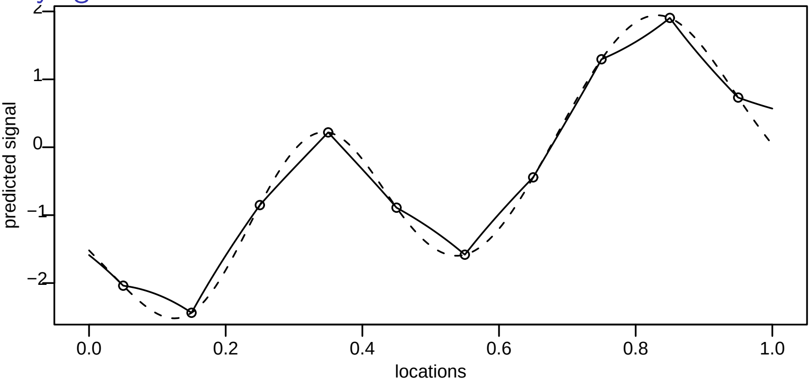
The minimum mean square error predictor for the mean + signal $\mu(\mathbf{x}') + S(\mathbf{x}')$ is given by

$$\hat{T}(\mathbf{x}') \;\; = \;\; \mu(\mathbf{x}') + \sum_{i=1}^{n} w_i(\mathbf{x}')(Y_i - \mu(\mathbf{x}_i))$$

- the predictor $\hat{T}(\mathbf{x}')$ is a compromise between the unconditional mean $\mu(\mathbf{x}')$ and the deviations of the observed data $\mathbf{Y}(\mathbf{x}_i)$ from their means $\mu(\mathbf{x}_i)$
- the nature of the compromise depends on the target location $\mathbf{x}'$;, the data-locations $\mathbf{x}_i$ and the values of the model parameters.
- the $w_i(\mathbf{x}')$ are called the *prediction weights*.

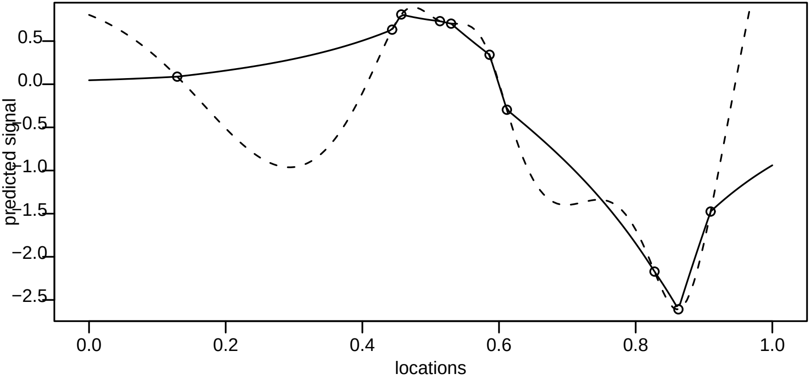
Effects on predictions

Varying the correlation function



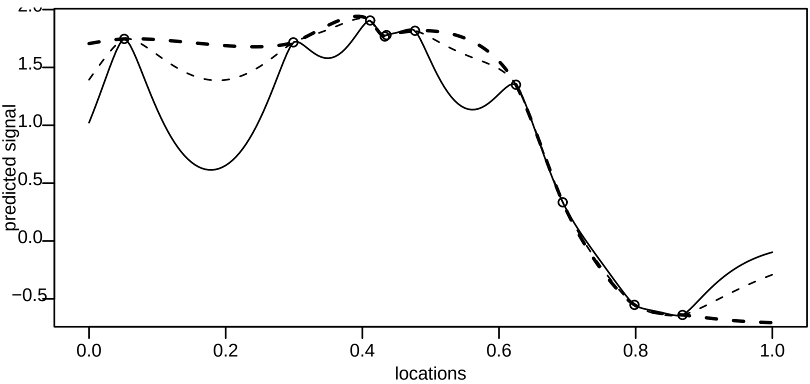
Predictions from 10 equally spaced data-points using exponential (—) or Matérn of order 2 (- - -) correlation functions.

Unequally spaced data



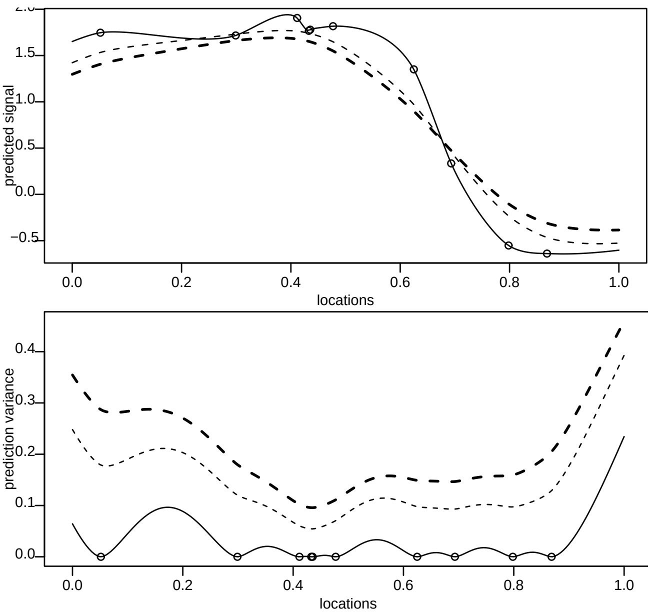
Predictions from 10 randomly spaced data-points using exponential (—) or Matérn of order 2 (- - -) correlation functions.

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 (—), 0.1 (- - -) and 0.5 (▬ ▬ ▬).

Varying the noise-to-signal ratio



$\Leftarrow \mathrm{E}(S|Y)$

$\tau^2 =$
0 (—),
0.25 (- - -),
and
0.5 (━ ━ ━).

$\Leftarrow \mathrm{Var}\,[S|Y]$

Prediction of Linear Functionals

Let T be any *linear* functional of $S^* := \mu + S$,

$$T = \int_A c(\mathbf{x}) S^*(\mathbf{x}) d\mathbf{x}$$

for some prescribed weighting function $c(\mathbf{x})$.
e.g. the average of $S^*(\cdot)$ over a region,

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

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

Conditional Distribution

Under the Gaussian model:

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

$$\mathsf{E}(T|\mathbf{Y}) = \int_A c(\mathbf{x})\left(\mu(\mathbf{x}) + \mathbb{E}\left[S(\mathbf{x})|\mathbf{Y}\right]\right) d\mathbf{x}$$

$$\mathsf{Var}\left[T|\mathbf{Y}\right] = \int_A \int_A c(\mathbf{x})c(\mathbf{x}')\mathsf{Cov}\left[S(\mathbf{x}), S(\mathbf{x}')\right] \; d\mathbf{x}d\mathbf{x}'$$

Note in particular that

$$\hat{T} = \int_A c(\mathbf{x})(\mu(\mathbf{x}) + \hat{S}(\mathbf{x}))d\mathbf{x}$$

Explanation in words

- Given a predicted surface $\hat{S}(\mathbf{x})$, 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 $S(x)$
- Replace the unknown S with the known $\hat{S}$ in the formula for T
- it is *NOT* legitimate to do this for prediction of non-linear properties
- for example, the maximum of $\hat{S}(\mathbf{x})$ is a very bad predictor for the maximum of $S(\mathbf{x})$

Prediction of non-linear Functionals

- Let T be *any* functional of $S^* := \mu + S$, e.g.
  - the proportion of the area over which $S^* > 5$
  - the maximum value of S^* over the region.
  - When using the Box-Cox transform, predicting $\mathrm{E}(Y)$ rather than $\mathrm{E}[h_\lambda(Y)]$
- Substituting $\hat{S}$ in the linear case "worked" because $\mathrm{E}(aY + b) = a\,\mathrm{E}(Y) + B)$.
- This doesn't work for non-linear transforms, for instance $Y \sim N(0, 1)$ results in $\mathrm{E}(Y^2) = 1$.
- The solution is to simulate from the conditional distribution and transform the simulated values.

The algorithm

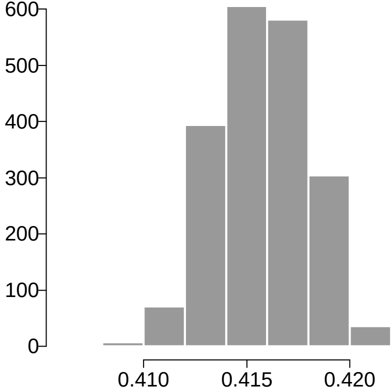
- Define a prediction grid $G = \{\mathbf{x}'_1, \ldots, \mathbf{x}'_g\}$ to cover the area of interest
- Dimulate a realisation of $\mathbf{S}_G := [S(\mathbf{x}'_1), \ldots, S(\mathbf{x}_g)]$ from the conditional distribution $[\mathbf{S}_G|\mathbf{Y}]$
- add on any mean effects $\boldsymbol{\mu}_G = \mathbf{F}_G\ \beta$ where $\mathbf{F}_G$ is a matrix of covariates at the points in G.
- calculate $t^{(1)}$ from this simulation
- repeat to obtain $t^{(2)}, \ldots, t^{(m)}$ - a sample from the distribution of T.

How fine should we make the prediction grid?

- As fine as your computer will allow and you have the patience for!
- fine enough to pick up all the features
- not so fine as to make the computation time and memory requirements prohibitive
- pragmatic strategy: stop when the finer of two grids makes no signficant difference to the quantity of interest (e.g. to posterior mean or median)

Swiss rainfall data

Prediction of $\mathcal{F}_{200}(S)$, the percentage of the area where $Y(x) \geq 200 = 0.4157$



Samples from the predictive distribution of $\mathcal{F}_{200}(S)$. NB Possible difficulties with negative values and back-transformation (simply set to zero in geoR code - crude but alternative is computationally intensive).

Plug-In Prediction

- Usually the model parameters are in fact **unknown**.
- The **plug-in prediction** consists of replacing the true parameters by their estimates.

Comments

- we will use ML estimates
- could also use REML estimates
- The conventional approach to kriging 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 $T = S(\mathbf{x})$
  - but often under-estimates prediction variance
  - and can produce poor results when predicting other targets T

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 too expensive to be left out

“Leaving-one-out”

- Also called Jackknifing
- Write $\mathbf{Y}_{-i} = \{Y_j; j \neq i\}$
- One by one, for each datum:

  1. remove the datum from the data-set
  2. (re-estimate model parameters)
  3. predict at the datum location

  $$\mathrm{E}(Y_i|\mathbf{Y}_{-i}), \ \mathrm{Var}\left[Y_i|\mathbf{Y}_{-i}\right]$$

  4. compute standardised residuals

  $$\mathrm{E}(Y_i|\mathbf{Y}_{-i})/\{\mathrm{Var}\left[Y_i|\mathbf{Y}_{-i}\right]\}^{1/2}$$

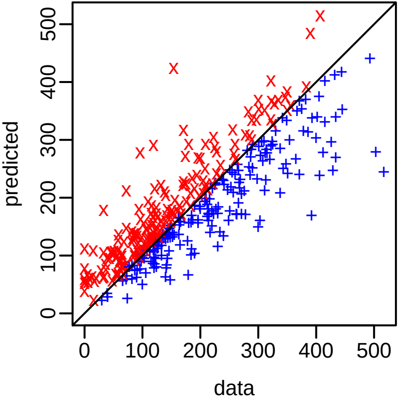
- Compare original data with predicted values.

What to use cross-validation for

- Comparing two models or estimation procedures
  - Compare total sums of squares of prediction errors
- As a diagnostic, particularly when the dataset is small.
  - Check for Normality
  - Check for a constant variance
- The `R` function `xvalid` does this

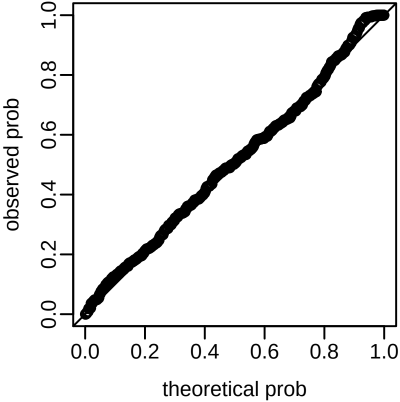
100 fitting data + 367 validation data

Y_i v $\mathrm{E}(Y_i|\mathbf{Y}_{-i})$



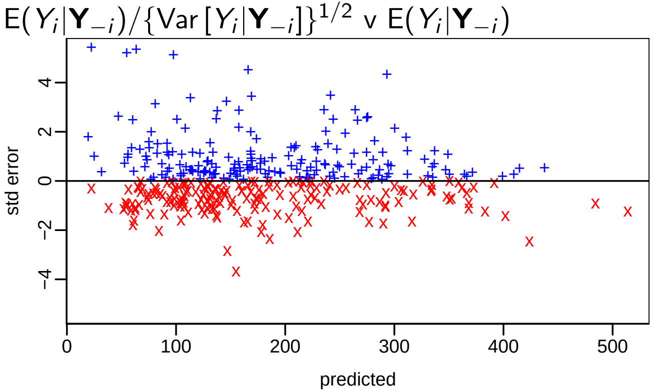
Underestimating big values

QQ plot of standardised residuals



Roughly Gaussian apart from
heavy end to the right tail

Standerdised residuals



$E(Y_i|\mathbf{Y}_{-i})/\{\mathrm{Var}[Y_i|\mathbf{Y}_{-i}]\}^{1/2}$ v $E(Y_i|\mathbf{Y}_{-i})$

Bayesian prediction

We wish to make inferences about functional T based on the posterior distribution

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

This is a weighted sum of the distribution of T given the data and a particular set of parameter values, taken over all possible parameter values and using the parameters' posterior distribution as the weight.

Before describing an efficient algorithm to sample from the posterior, note:

- Conditional on knowledge of $\mathbf{S} := [S(\mathbf{x}_1), \ldots, S(\mathbf{x}_n)]$ the signal at all data points...
- ... the distribution of the signal at a grid of points $\mathbf{S}_G := [S(\mathbf{x}'_1), \ldots, S(\mathbf{x}'_g)]$ depends on σ^2 and ϕ
- **but** does not depend on $\boldsymbol{\beta}$, $\nu = \tau^2/\sigma^2$, or the data $\mathbf{y}$.

This is because

$$\left[\begin{array}{c} \mathbf{S}_G \\ \mathbf{S} \end{array} \right] \sim N\left(\mathbf{0}, \sigma^2 \mathbf{R}_{dg}\right)$$

where the element of $\mathbf{R}_{dg}$ corresponding to any two locations $\mathbf{x}^*_1$ and $\mathbf{x}^*_2$ is simply $\rho(||\mathbf{x}^*_1 - \mathbf{x}^*_2|| / \phi)$

Predictive sampling algorithm

Define a predictive grid $G := \{\mathbf{x}'_1, \ldots, \mathbf{x}'_g\}$.

At the i^{th} iteration

- simulate $\boldsymbol{\beta}^{(i)}, \sigma^{2\ (i)}, \phi^{(i)}, \nu^{(i)}$ from their posterior distribution (as described under Bayesian parameter estimation in Chapter IV).
- simulate the signal at all data points $\mathbf{S}^{(i)} = [S(\mathbf{x}_1), \ldots, S(\mathbf{x}_n)]$ using $\boldsymbol{\beta}^{(i)}, \sigma^{2\ (i)}, \phi^{(i)}, \nu^{(i)}$ and $\mathbf{y}$ (as described under Prediction, earlier in this chapter).
- simulate the signal at all grid points $\mathbf{S}_G^{(i)} = [S(\mathbf{x}'_1), \ldots, S(\mathbf{x}'_g)]$ using $\mathbf{S}^{(i)}, \sigma^{2\ (i)}, \phi^{(i)}$.
- calulate the mean at all grid points $\boldsymbol{\mu}_G = [\mu(\mathbf{x}'_1), \ldots, \mu(\mathbf{x}'_g)]$ using $\boldsymbol{\beta}^{(i)}$ and $\mathbf{F}_G$, the covariates matrix at predictive grid points.
- calculate $t^{(i)}$ from the extended grid of values $(\boldsymbol{\mu}^{(i)} + \mathbf{S}^{(i)}, \boldsymbol{\mu}_G^{(i)} + \mathbf{S}_G^{(i)})$ - this is a sample from the posterior distribution for T.
- repeat ...

Notes:

- The sampled values from one iteration to the next are *independent* - this is not MCMC!
- Computation of moments of $\mathbf{S}_G$ (mean, variance,...) can be performed more simply as a mixture of multivariate t distributions since some of the integrals can be computed analytically given $\phi^{(i)}, \nu^{(i)}$:
  Integrate out $\boldsymbol{\beta}, \sigma$ then use knowledge of analytical distribution -

$$\begin{aligned} \mathbf{F}_G\boldsymbol{\beta} + S_G|\mathbf{y}, \boldsymbol{\beta}^{(j)}, \sigma^{2(j)}, \phi^{(j)}, \nu^{(j)} &\sim \text{m.v. Gaussian} \\ \mathbf{F}_G\boldsymbol{\beta} + S_G|\mathbf{y}, \phi^{(j)}, \nu^{(j)} &\sim \text{m.v. } t \end{aligned}$$

- Simulation of the signal at data points could also have been used in the algorithm for estimating non-linear functionals.

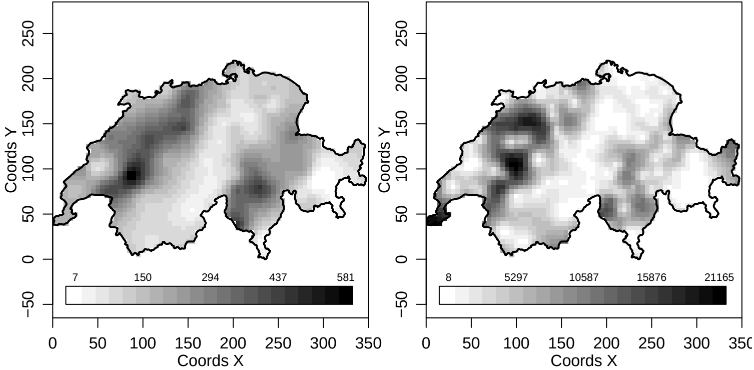
Comparing plug-in and Bayesian

- 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.

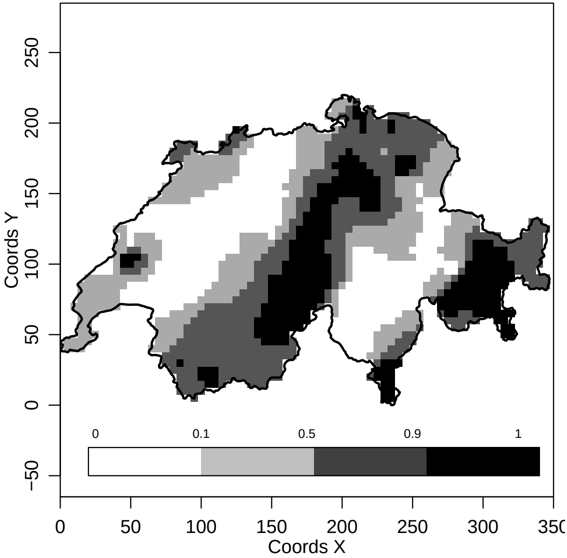
Bayesian prediction is usually more cautious than plug-in prediction, or in other words:

- allowance for parameter uncertainty usually results in wider prediction intervals

Swiss rainfall: prediction results

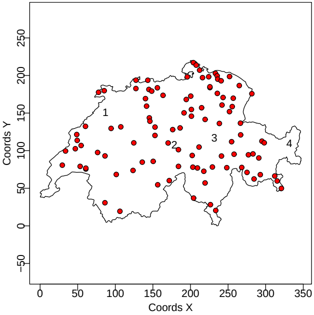


Predicted signal surfaces and associated measures of precision for the rainfall data: (a) posterior mean; (b) posterior variance

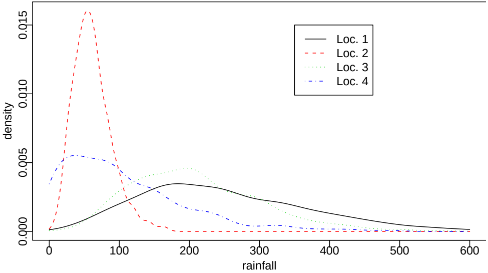


Posterior probability contours for levels 0.10, 0.50 and 0.90 for the random set $T = \{x : S(x) < 150\}$

Swiss rainfall: prediction results (cont.)



Recording stations and selected prediction locations (1 to 4)

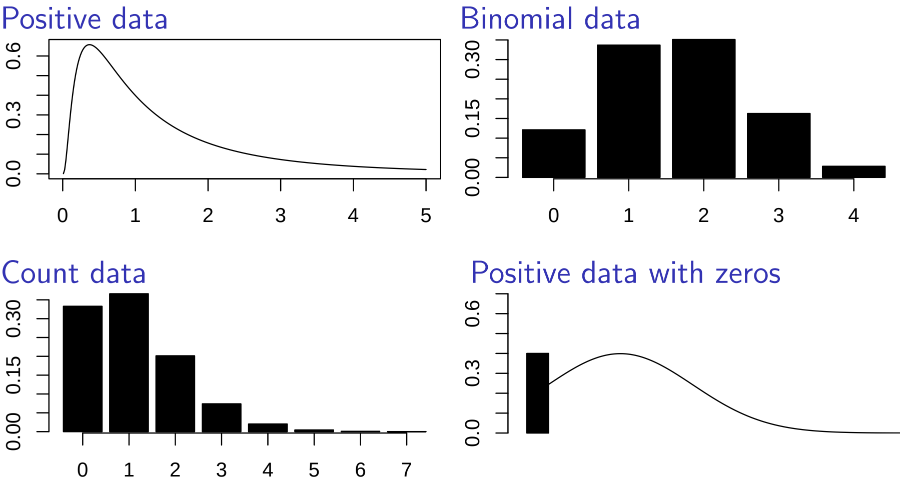


Bayesian predictive distributions for average rainfall at selected locations.

PART VI: GENERALIZED LINEAR SPATIAL MODELS

1. **Non Gaussian data**
2. **Generalized linear geostatistical models**
3. **Application of MCMC to Generalized Linear Prediction**
4. **Case-study: Rongelap Island**
5. **Case-study: Gambia Malaria**

Non-Gaussian data



Positive data

Binomial data

Count data

Positive data with zeros

Towards a model specification

- Consider the linear model

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$$

- Re-write it as

$$Y_i \sim N(\mu_i, \sigma^2)$$
$$\mu_i = \sum_{j=1}^{p} F_{ij}\beta_j = \mathbf{f}_i'\boldsymbol{\beta}$$

- Generalise the model as

$$Y_i \sim Q(\mu_i, \ldots)$$
$$h(\mu_i) = \sum_{j=1}^{p} F_{ij}\beta_j = \mathbf{f}_i'\boldsymbol{\beta}$$

  - Q is a distribution in the exponential family
  - $h(\cdot)$ is a (pre-specified) link function

- Generalized Linear Models (GLM)

Generalized Linear Geostatistical Models

Classical generalized linear model has

- $Y_i : i = 1, ..., n$ mutually independent, with $\mu_i = \mathrm{E}[Y_i]$
- $h(\mu_i) = \mathbf{f}_i'\boldsymbol{\beta}$ 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
- $h(\mu_i) = \mathbf{f}_i'\boldsymbol{\beta} + U_i$ 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 latent spatial stochastic process $\mathbf{S} := [S(\mathbf{x}_1), \ldots, S(\mathbf{x}_n)]$
- $h(\mu_i) = \mathbf{f}(\mathbf{x}_i)'\boldsymbol{\beta} + S(\mathbf{x}_i)$ for known link function $h(\cdot)$

Examples

$\mathbf{x}_1, \ldots, \mathbf{x}_n$ locations with observations

Poisson-log

- $[Y(\mathbf{x}_i) \mid S(\mathbf{x}_i)]$ is Poisson with density

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

- link: $E[Y(\mathbf{x}_i) \mid S(\mathbf{x}_i)] = \mu_i$, $\mathbf{f}(\mathbf{x}_i)'\boldsymbol{\beta} + S(\mathbf{x}_i) = \log \mu_i$.

Binomial-logit

- $[Y(\mathbf{x}_i) \mid S(\mathbf{x}_i)]$ is binomial with density

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

- link: $\mathrm{E}[Y(\mathbf{x}_i) \mid S(\mathbf{x}_i)] = \mu_i$, $\mathbf{f}(\mathbf{x}_i)\boldsymbol{\beta} + S(\mathbf{x}_i) = \log(\mu_i/(r-\mu_i))$

Likelihood function

$$L(\boldsymbol{\beta}, \sigma^2, \phi) =$$

$$\int_{\mathbb{R}^n} \prod_{i}^{n} f(y_i; h^{-1}(\mathbf{f}_i' \boldsymbol{\beta} + s_i)) f(\mathbf{s} \mid \sigma^2, \phi) ds_1 \ldots ds_n$$

High-dimensional integral !!!

Inference For The Generalized Linear Geostatistical Model

- likelihood evaluation involves high-dimensional numerical integration
- approximate methods: Breslow and Clayton (1993), Geyer and Thompson (1992), Geyer (1994) are of uncertain accuracy but useful for exploratory analysis
- MCMC is feasible, although not routine.
- `geoRglm` and WinBUGS have greatly improved the accessibility of MCMC for spatial models.

Application of MCMC

Start with

- data $y_i = y(\mathbf{x}_i) \quad (i = 1, \ldots, n)$
- matrix of covariates at data points $\mathbf{F}$
- (optional) a grid $G := \{\mathbf{x}'_1, \ldots, \mathbf{x}'_g\}$ of points at which we wish to sample the signal, and covariate mx $\mathbf{F}_G$
- covariance model (e.g. Matern)
- initially assume *no random effects* (i.e. $\nu^2 = 0$)
- initially assume *fixed* κ

We must

- specify priors for regression parameters $\boldsymbol{\beta}$ and covariance parameters $\boldsymbol{\theta} := [\sigma^2, \phi]$
- choose initial parameter values $\boldsymbol{\beta}^{(0)}, \sigma^{(0)}, \phi^{(0)}$
- choose inital values for the signal at data points $\mathbf{S} := [S(\mathbf{x}_1), \ldots, S(\mathbf{x}_n)]'$

The goal

- posterior distribution of $\boldsymbol{\beta}, \sigma, \phi$
- functionals of the mean + signal at data and/or prediction points (e.g. mean/max over an area or proportion of region over a certain threshold)

Implementation

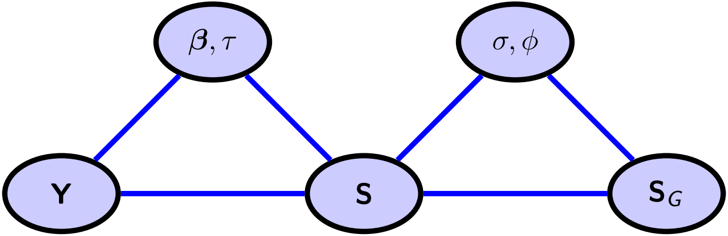
priors - exactly as for analysis of Gaussian model

- discrete prior for ϕ
- Gaussian-scaled-Inverse-χ^2 for $(\boldsymbol{\beta}, \sigma^2)$

initial values

- choose $\phi^{(0)}$ and $\sigma^{(0)}$ from their priors - either sensibly from their support or by direct sampling (if priors are proper)
- set $\boldsymbol{\beta}^{(0)}$ by fitting a standard GLM to the data
- set $s^{(0)}(\mathbf{x}_i) = h(y_i) - \mathbf{f}(\mathbf{x}_i)\boldsymbol{\beta}^{(0)}$

Conditional independence structure



Generic MCMC scheme

- update $\beta^{(i-1)}, \tau^{(i-1)}$ to $\beta^{(i)}, \tau^{(i)}$ conditional on $\mathbf{y}, \mathbf{s}^{(i-1)}$
- update $\sigma^{(i-1)}, \phi^{(i-1)}$ to $\sigma^{(i)}, \phi^{(i)}$ conditional on $\mathbf{s}^{(i-1)}$
- update $\mathbf{s}^{(i-1)}$ to $\mathbf{s}^{(i)}$ conditional on $\mathbf{y}, \beta^{(i)}, \sigma^{(i)}, \phi^{(i)}$
- (optional) sample $\mathbf{s}_G^{(i)}$ directly from its conditional distribution given $\mathbf{s}^{(i)}, \sigma^{(i)}, \phi^{(i)}$

For a simple MCMC scheme based on independence sampling for σ^2, ϕ, and $\mathbf{s}$, see Diggle, Tawn and Moyeed (1998).

Prediction

- Use output from chain to construct posterior probability statements about $[T|\mathbf{Y}]$, where $T = \mathcal{F}(\mathbf{S}_G, \mathbf{S}, \beta)$.
- Two approaches are possible for estimating **expectations** (rather than obtaining full posterior distributions).
- For simplicity just consider expectations of some function of the prediction grid.

Full Monte Carlo

After m iterations approximate $\mathbb{E}\left[T(\mathbf{F}_G\boldsymbol{\beta} + \mathbf{S}_G)|\mathbf{y}\right]$ by

$$\frac{1}{m}\sum_{j=1}^{m} T(\mathbf{F}_G\boldsymbol{\beta}^{(j)} + \mathbf{s}_G^{(j)})$$

Using analytic distributions

- Integrate out $\boldsymbol{\beta}, \sigma$ then use knowledge of analytical distribution:

$$\begin{aligned} \mathbf{F}_G\boldsymbol{\beta} + S_G|\mathbf{s}^{(j)}, \boldsymbol{\beta}^{(j)}, \sigma^{2(j)}, \phi^{(j)} &\sim \text{ multivariate Gaussian} \\ \mathbf{F}_G\boldsymbol{\beta} + S_G|\mathbf{s}^{(j)}, \phi^{(j)} &\sim \text{ multivariate } t \end{aligned}$$

- If it is possible to do so, calculate $\mathbb{E}\left[T(\mathbf{S}_G)|\mathbf{s}^{(j)}, \phi^{(j)}\right]$, $j = 1, \ldots, m$ directly, and estimate $\mathbb{E}\left[T(\mathbf{S}_G)|\mathbf{y}\right]$ by

$$\frac{1}{m}\sum_{j=1}^{m}\mathbb{E}\left[T(\mathbf{S}_G)|\mathbf{s}^{(j)}, \phi^{(j)}\right]$$

This is preferable to Monte Carlo as it eliminates the portion of the Monte Carlo error due to sampling $\mathbf{S}_G, \boldsymbol{\beta}, \sigma$.

A more efficient MCMC scheme

The scheme implemented in `geoRglm` is documented in Diggle, Ribeiro and Christensen (2003).
Note that conditional on $\phi^{(i)}$ the posterior for

$$\boldsymbol{\beta}, \sigma^2 \mid \mathbf{S} + \mathbf{F}\boldsymbol{\beta}$$

is Gaussian-scaled-inverse-χ^2 (as with the Gaussian case).

- update $\phi^{(i-1)}$ to $\phi^{(i)}$ conditional on $\mathbf{s}^{(i-1)}$ using a random walk Metropolis step
- update $\sigma^{(i-1)}, \boldsymbol{\beta}^{(i-1)}$ to $\sigma^{(i)}, \boldsymbol{\beta}^{(i)}$ conditional on $\mathbf{F}\boldsymbol{\beta}^{(i-1)} + \mathbf{s}^{(i-1)}$ by sampling exactly from the posterior
- update $\mathbf{s}^{(i-1)}$ to $\mathbf{s}^{(i)}$ conditional on $\mathbf{y}, \boldsymbol{\beta}^{(i)}, \sigma^{(i)}, \phi^{(i)}$ using a truncated Metropolis adjusted Langevin algorithm (MALA)

Also contains a cunning reparameterisation $\mathbf{S} \rightarrow \mathbf{Z}$ where $\mathbf{S} = \sigma \mathbf{R}^{1/2}\mathbf{Z}$ makes the updates to $\mathbf{S}$ more efficient.

Notes

- The optimal acceptance rate for many high dimensional MALA algorithms is $\approx 60\%$ - tune the **S** scaling parameter to achieve this.
- The optimal acceptance rate for many high dimensional RWM algorithms is $\approx 23\%$ but this algorithm is one-dimensional so tune the ϕ scaling parameter to $\approx 30\% - 40\%$.

Extensions

- discrete prior for κ e.g. $\kappa = \{0.5, 1, 1.5, 2.5\}$ with probabilities $\{0.25, 0.25, 0.25, 0.25\}$
- random effects (return of the nugget) e.g. n villages and m_i people measured in the i^{th} village. The mean for the j^{th} person in the i^{th} village is given by

$$h(\mu_{ij}) = \mathbf{f}_{ij}\boldsymbol{\beta} + S(\mathbf{x}_i) + Z_i \quad \text{where} \quad Z_i \;\sim\; N(0, \tau^2)$$

  - extra village-level *non-spatial* effect (e.g. missing covariates).
  - require a discrete prior on $\nu = \tau/\sigma$.
- more general priors for $\boldsymbol{\beta}$ and σ can be accomodated but require random walk Metropolis steps for these parameters (NB RWM on $log(\sigma)$).

Improper prior and improper posterior

- In a generalised linear mixed model, the improper prior $\pi(\sigma^2) \propto 1/\sigma^2$ leads to an improper posterior for σ^2 (Natarajan & Kass, 2000 - JASA).
- Generalised linear geostatistical models are generalised linear mixed models with a specific covariance structure. Therefore *avoid the Jeffrey's prior for* σ^2.
- The Gaussian model with a nugget effect is an example of a generalised linear mixed model. However in this case (and only in this case) the reparameterisation $\nu^2 = \tau^2/\sigma^2$ gets round the mathematical detail and leads to a proper prior for σ^2.
- The whole idea of an improper prior is (arguably) dubious. It is safer to use diffuse but proper priors.

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

- approximately 2500 miles south-west of Hawaii
- contaminated by nuclear weapons testing in 1954
- residents evacuated after the test, many died
- 1957 Rongelap declared safe, residents returned.
- Leukemia and thyroid-tumors followed. Greenpeace evacuates residents in 1985
- now safe for re-settlement?
- Radiation measurements taken, spatial maps made
- After some removal of soil, radiation levels have fallen
- Reconstruction is underway with resettlement expected soon.

Statistics in Rongelap

The Problem

- field-survey of ^{137}Cs measurements
- estimate spatial variation in ^{137}Cs radioactivity
- compare with agreed safe limits

The model

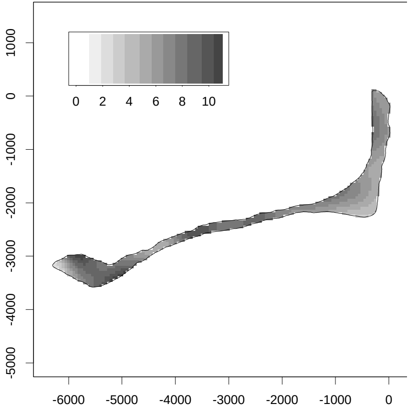
- Basic measurements are net counts Y_i over time-intervals t_i at locations $\mathbf{x}_i$ $(i = 1, ..., n)$
- Suggests following model:
  - $S(\mathbf{x}) : \mathbf{x} \in R^2$ stationary Gaussian process (local radioactivity)
  - $Y_i | \{S(\cdot)\} \sim \text{Poisson}(\mu_i)$
  - $\mu_i = t_i \lambda(\mathbf{x}_i) = t_i \exp\{\beta_1 + S(\mathbf{x}_i)\}$.

Aims

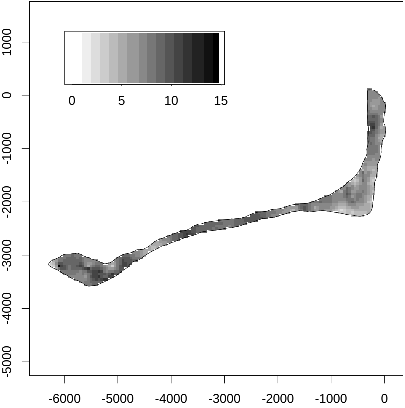
- predict $\lambda(\mathbf{x})$ over whole island
- predict max $\lambda(\mathbf{x})$
- predict argmax $\lambda(\mathbf{x}))$

Predicted radioactivity surface

log-Gaussian kriging

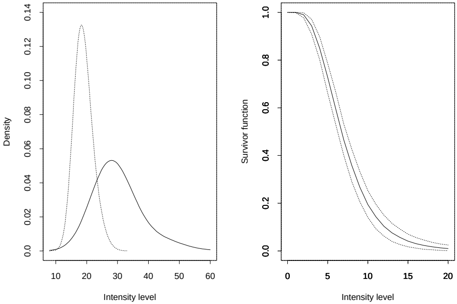


Poisson log-linear model with latent Gaussian process



- 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



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

Posterior distributions from the Poisson model of the maximum radioactivity based on:

- The fully Bayesian analysis incorporating the effects of parameter uncertainty in addition to uncertainty in the latent process (solid line)
- Fixing the model parameters at their estimated values, ie allowing for uncertainty only in the latent process

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
- $\mathbf{f}_{ij}$ = child-specific covariates
- $\mathbf{w}_i = \mathbf{w}(\mathbf{x}_i)$ village-specific covariate
- $\mathrm{logit} P(Y_{ij} = 1 | S(\cdot)) = \mathbf{f}'_{ij}\boldsymbol{\beta}_1 + \mathbf{w}_i{}'\boldsymbol{\beta}_2 + S(\mathbf{x}_i)$

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:

- $U_i \sim \mathrm{N}(0, \nu^2)$
- $\mathrm{logit} P(Y_{ij} = 1 | S(\cdot), \mathbf{U}) = \mathbf{f}'_{ij}\boldsymbol{\beta}_1 + \mathbf{w}_i{}'\boldsymbol{\beta}_2 + U_i + S(\mathbf{x}_i)$

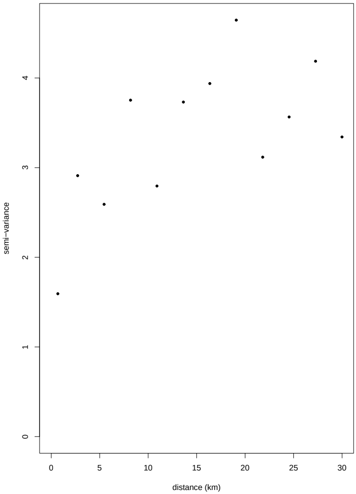
Exploratory analysis

- fit standard logistic linear model, ignoring $S(\mathbf{x})$ and perhaps U
- compute for each village:

$$\begin{aligned}N_i &= \textstyle\sum_{j=1}^{n_i} Y_{ij}\\ \mu_i &= \textstyle\sum_{j=1}^{n_i} \hat{P}_{ij}\\ \sigma_i^2 &= \textstyle\sum_{j=1}^{n_i} \hat{P}_{ij}(1-\hat{P}_{ij})\end{aligned}$$

- 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

Fixed effects

α = intercept

β_1 = age

β_2 = bed-net use

β_3 = treated bed-net

β_4 = green-ness index

β_5 = presence of public health centre in village

Random effects

$\nu^2 = \mathrm{Var}[U_i]$, non-spatial

$\sigma^2 = \mathrm{Var}[S(x)]$, spatial

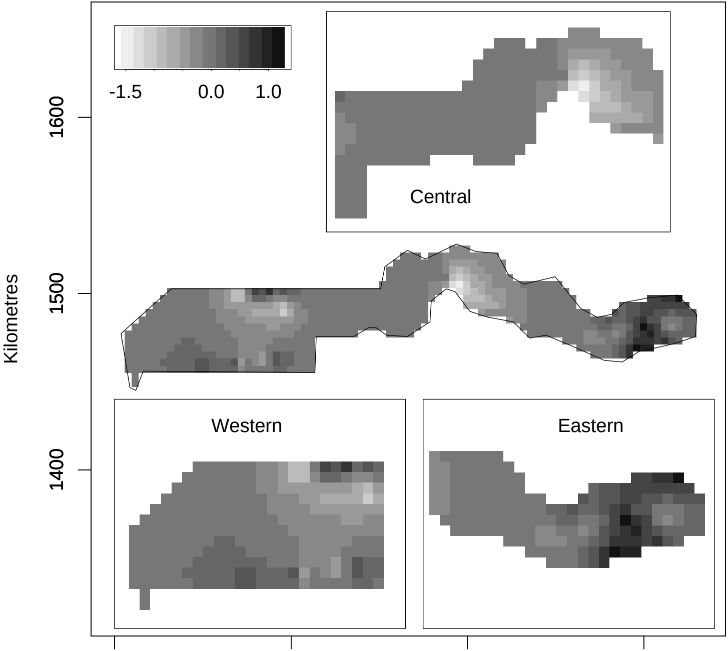
ϕ = spatial range

κ = Matérn shape

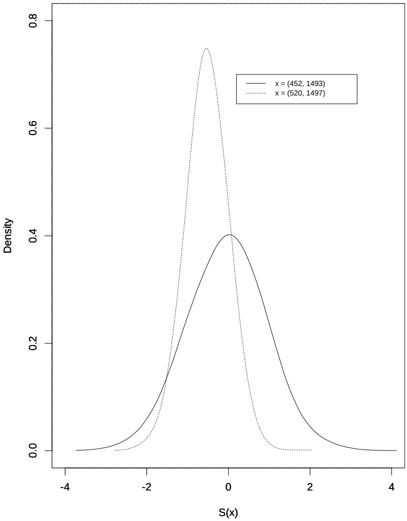
	2.5%	97.5%	Mean	Median
α	-4.23	1.11	-1.66	-1.70
β_1	0.00044	0.00092	0.00068	0.00068
β_2	-0.68	-0.084	-0.380	-0.39
β_3	-0.78	0.055	-0.36	-0.36
β_4	-0.040	0.072	0.019	0.020
β_5	-0.79	0.18	-0.32	-0.32
ν^2	$2 \cdot 10^{-6}$	0.52	0.12	0.019
σ^2	0.24	1.66	0.79	0.74
ϕ	1.24	53.35	11.65	7.03
κ	0.15	1.96	0.94	0.83

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

Posterior mean of $S(x)$



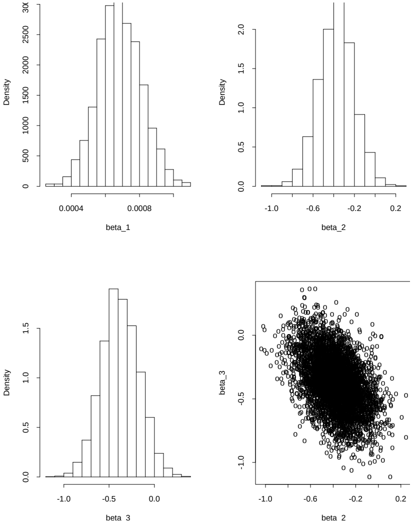
Posterior density estimates for $S(\mathbf{x})$ at two selected locations.



— Remote location (452, 1493)

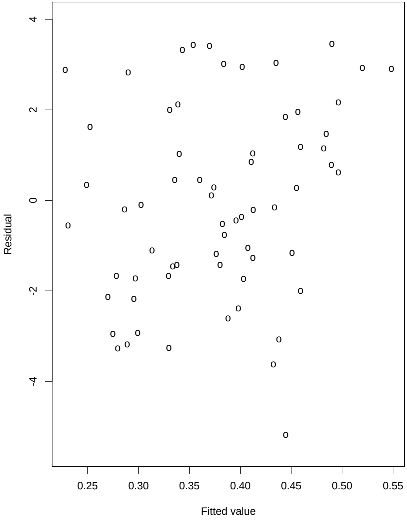
- - - dashed curve – location (520, 1497), close to observed sites in central region.

Empirical posterior distributions for regression parameters



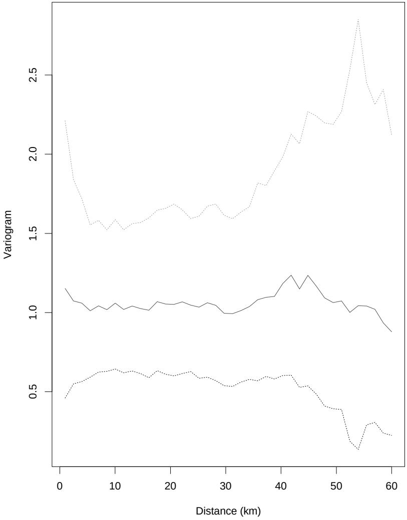
- β_1 = effect of age
- β_2 = effect of untreated bed-nets
- β_3 = additional effect of treated bed-nets

Goodness-of-fit for Gambia malaria model



Village-level residuals against fitted values.
(Diggle et al., 2002)

- $r_{ij} = (Y_{ij} - \hat{p}_{ij})/\sqrt{\{\hat{p}_{ij}(1-\hat{p}_{ij})\}}$
- $r_i = \sum r_{ij}/\sqrt{n_i}$
- intended to check adequacy of model for p_{ij}
- more sensitive to individual 'unlikely data' than $(N_i - \mu_i)/\sigma_i$ which was used in exploratory spatial analysis (so perhaps less preferable).

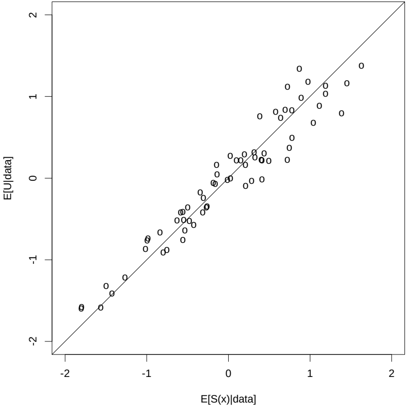


Standardised residual empirical variogram plot (village-level data and pointwise 95% posterior intervals constructed from simulated realisations of fitted model).

Simulate realisations from the fitted model and calculate

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

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

PART VII: FURTHER TOPICS, HISTORY AND PHILOSOPHY

1. **Sampling design**
2. **Multivariate methods**
3. **Space-time models**
4. **Marked point processes**
5. **Philosophy and History**
6. **Closing remarks**

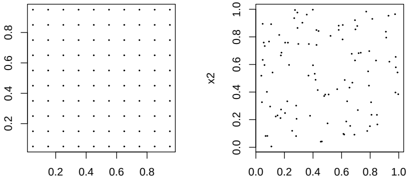
Sampling design

How do we choose the sample points $\mathbf{x}_1, \ldots, \mathbf{x}_n$?

Grid types

Should be stochastically independent of the signal $S(\mathbf{x})$. Possibilities include

- *uniform* e.g. a square grid - with the centre positioned at random
- *random* - e.g. a Poisson process



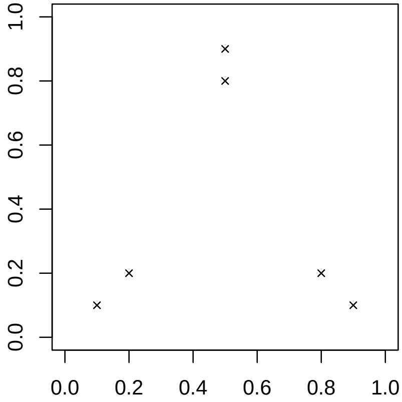
Design grids with 100 points - regular square lattice (left) and generated by a homogenous Poisson process (right).

Prediction considerations

- For two points $\mathbf{x}_1$ and $\mathbf{x}_2$ close together $S(\mathbf{x}_1)$ and $S(\mathbf{x}_2)$ will be very similar and so the second point adds little information about $S(\mathbf{x})$ in that neighbourhood
- therefore if prediction of a homogenous spatial average is required choose a homogenous regular grid
- if certain subregions are more important then sample more heavily in those subregions

Parameter considerations

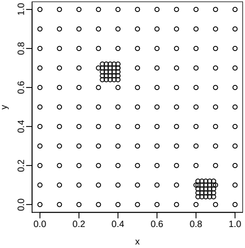
Consider the extreme grid below.



- the difference in $S(\mathbf{x})$ between points close together is informative about about τ^2, ϕ and any anisotropy parameters.
- for close pairs, $S(\mathbf{x}_2)$ provides little extra information (over $S(\mathbf{x}_1)$) about any overall mean μ or variance σ^2

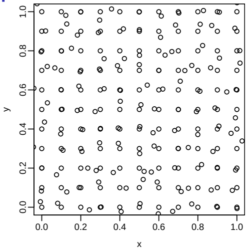
Compromise lattices

Lattice plus infill



- Nested sub-lattices
- As with Rongelap

Lattice plus close pairs



- Extra points randomly located within a disk of radius δ around each lattice point

- Infill risks committing too many points to the infill squares, the rest of the grid becomes too sparse
- Parameters badly estimated if range is bigger than the infills but smaller than the grid.
- Diggle and Lophaven (2006)'s design criteria has close pairs being much better than a simple grid or a grid plus infills.

Multivariate methods

Motivation

- two or more related repsonses are measured at each location (e.g. Cancer and Heart Disease cases)
- covariate is missing at some data points
- Y_2 is of no direct interest but is correlated with the response of interest Y_1, and is much cheaper to measure

Approach

- 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 - must ensure positive definiteness of linear combinations of both/all responses

Bivariate model

$$\left[\begin{array}{c} Y_1(\mathbf{x}) \\ Y_2(\mathbf{x}) \end{array}\right] = \left[\begin{array}{c} \mathbf{f}(\mathbf{x})\boldsymbol{\beta}_1 \\ \mathbf{f}(\mathbf{x})\boldsymbol{\beta}_2 \end{array}\right] + \left[\begin{array}{cc} a_{11} & a_{12} \\ a_{21} & a_{22} \end{array}\right] \left[\begin{array}{c} S_1(\mathbf{x}) \\ S_2(\mathbf{x}) \end{array}\right] + \left[\begin{array}{c} U_1 \\ U_2 \end{array}\right]$$

Here $S_1(\mathbf{x})$ and $S_2(\mathbf{x})$ are independent SGP's with $\mu = 0$, $\sigma^2 = 1$ and positive definite correlation functions $\rho(u; \phi_1, \kappa_1)$ and $\rho(u, \phi_2, \kappa_2)$. Also

$$\left[\begin{array}{c} U_1 \\ U_2 \end{array}\right] \sim N(\mathbf{0}, \boldsymbol{\Sigma})$$

for some covariance matrix $\boldsymbol{\Sigma}$.

NB Number of parameters increases rapidly with dimension of response - detailed implementation should use structure of the specific problem.

Shared component model

- Suppose Y_{1i}, Y_{2i}, Y_{ni} are death rates from different causes at location x_i.
- Death rates are affected by a common surface S_0^* (perhaps relating to environmental pollution), and separate surfaces $S_1^*(x)$ to $S_n^*(x)$.

$$S_j(s) = S_0^*(s) + S_j^*(s);\ j = 1 \dots n$$

- Dimension of the problem doesn't grow so quickly with n
- Even for $n = 2$, perhaps this is a more intuitive interpretation than the bivariate model?

Space-time models

- Emerging space-time data-sets are **big**, and present severe computational challenges.
- Specific models are best defined in context.
- Calibration of radar reflectance against ground-truth rainfall intensity (Brown, Diggle, Lord and Young, 2001).
  1. $Y_{it} : i = 1, ..., n$ – ground-truth log-rainfall intensity at small number of sites x_i
  2. $U(x, t) : x \in A$ – log-radar reflectance measured effectively continuously over a study region A
  3. Empirical model,

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

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

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

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

1. Data give population density $\lambda_0(x)$ (approximately), plus locations of daily incident cases
2. 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.

Philosophy

What is the signal $S(\mathbf{x})$?

What justification is there for imagining a real world phenomenon as a single realisation of a spatially correlated stochastic process?

- $S(\mathbf{x})$ is not some underlying truth - it is a convenient model
- surrogate for covariates that are spatially correlated but that we have not thought to measure (e.g. elevation in Swiss rainfall data)
- representing some real process that spreads spatially (e.g. a snapshot of the occurences of some disease e.g. root fungus in a potato field)

What is the nugget effect $\epsilon \sim N(0, \tau^2)$?

Two possible interpretations

- measurement error
- spatial variation on a scale smaller than can be captured by our observation grid

For the Gaussian model only repeated measurements at the same point can hope to resolve the difference between the two.
e.g. for each soil sample, divide it into three and measure calcium content on each subsample.
For GLGM's there is already variability in the response - if there is a nugget it is even harder to determine whether it represents measurement error or small scale variability.
For the Poisson or binomial GLGM we can in principal discern the *need* for a nugget without repeated measurements but would need a lot of data.

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

Traditional geostatistics:

- avoids explicit references to the parametric specification of the model
- inference via variograms
- complex variogram structures are often used
- concentrates on linear estimators
- the *kriging menu*

Model-Based Geostatistics:

Term coined by Diggle, Tawn and Moyeed (1997)
Model based geostatistics means that we adopt a model-based approach to this class of problems; an explicit stochastic model is declared and from this associated methods of parameter estimation, interpolation and smoothing are derived by the application of general statistical principles.

Use of variograms

For parameters $\boldsymbol{\theta}$ (e.g. $\mu, \sigma^2, \phi, \tau^2$) estimate $\tilde{\theta}$ to minimise a particular criterion

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.

Potentially misleading (and even biased) because of inherent correlations amongst successive $\bar{V}_k$.

Example: Swiss rainfall data

Traditional Linear Geostatistics

Suppose the **target** for prediction is $T = \mu(\mathbf{x}') + S(\mathbf{x}')$
The **predictor** which minimises MSE is

$$\hat{T}(\mathbf{x}') = \mu(\mathbf{x}') + \mathbb{E}\left[S(\mathbf{x}')|\mathbf{Y}\right]$$

- **Traditional (linear) geostatistics:**
  - Assume that $\hat{T}$ is linear in $\mathbf{Y}$, so that

$$\hat{T}(\mathbf{x}') = b_0(\mathbf{x}') + \sum_{i=1}^{n} b_i(\mathbf{x}') Y_i$$

  - Choose b_i to minimise $MSE(\hat{T})$ within the class of linear predictors
- **Model-based geostatistics:**
  - specify a probability model for $[\mathbf{Y}, T]$
  - choose $\hat{T}$ to minimise $MSE(\hat{T})$ amongst all functions $\hat{T}(\mathbf{Y})$

But under the Gaussian geostatistical model:

$$\hat{T}(\mathbf{x}') = \mu(\mathbf{x}') + \mathbf{r}' \, (\mathbf{R}_*)^{-1}(\mathbf{y} - \boldsymbol{\mu})$$

n

Closing remarks

- "Essentially, all models are wrong, but some are useful." George E.P. Box & Norman R. Draper, Empirical Model-Building and Response Surfaces (Wiley 1987) p. 424:
- 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.