Recent Developments in Biostatistics: Space-Time Models

Tuesday: Geostatistics

Sonja Greven

Department of Statistics
Ludwig-Maximilians-Universität München

(with special thanks to Michael Höhle for some material from last year)

IBE LMU, 13 July 2010

Overview

Overview

Recap: Common Types of Spatial Data

- Spatial point patterns: Locations of events are themselves of interest → Friday
- Areal / lattice data: Outcome recorded in a number of geographical regions (e.g. counties, states) → Monday
- Geostatistical / point-referenced data: Outcome recorded at a number of locations → today

Goals of this Module

- Get a basic understanding of geostatistics, spatial processes, the semivariogram and the generalized linear model for geostatistical data
- Apply all methods to data examples using R.

Overview

Examples: Meuse River Heavy Metal Concentrations (1)

Topsoil heavy metal concentrations collected in a flood plain of the river Meuse, near the village of Stein (NL).

- Locations s: location where sample was taken, in Netherlands topographical map coordinates
- Outcomes Y: cadmium, copper, lead and zinc topsoil concentrations in mg/kg or "ppm" (continuous)
- Covariates $\mathbf{x}$: distance to the Meuse, elevation above river bed, flooding frequency, soil type, lime class, landuse class

Examples: Meuse River Heavy Metal Concentrations (2)

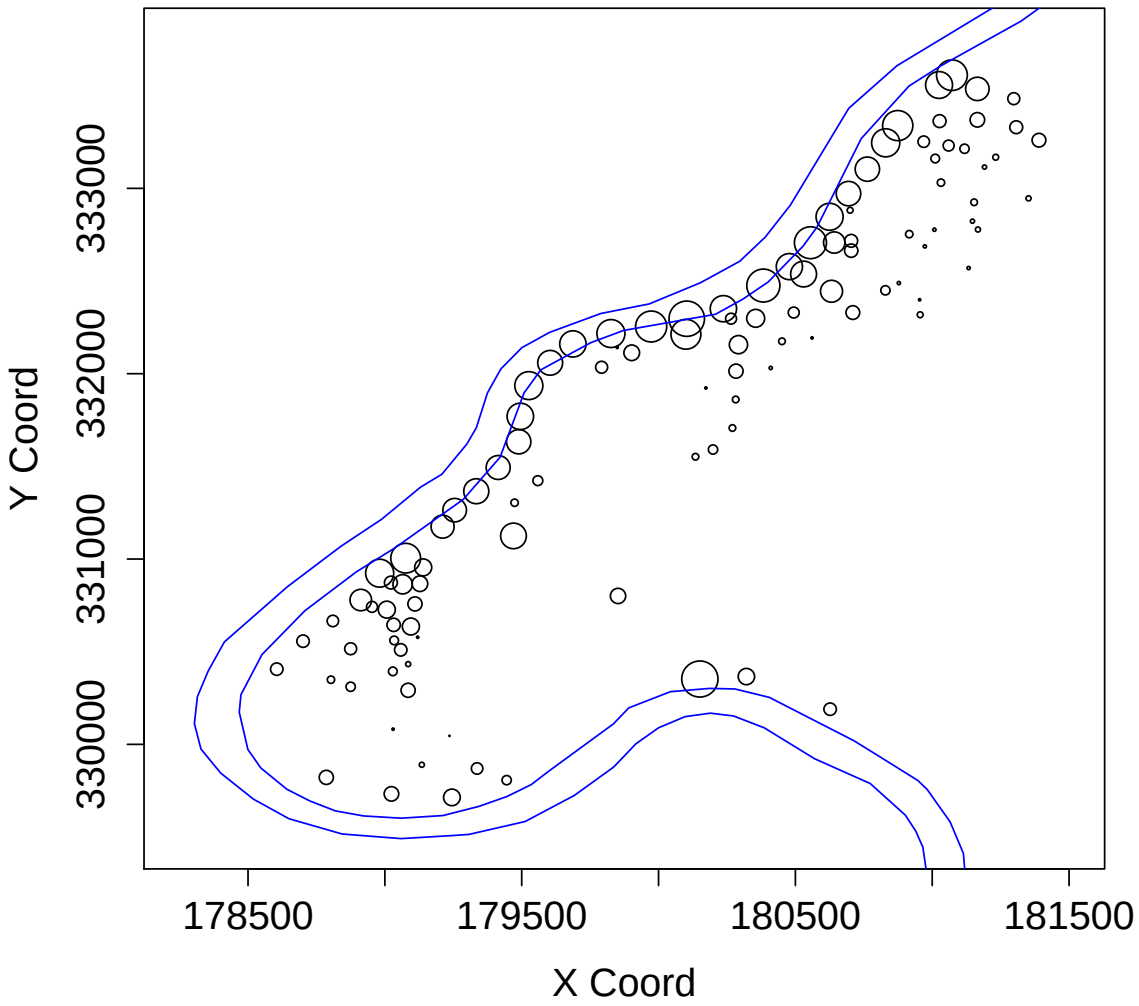


Figure: Circle radii are proportional to log(cadmium) level.

Examples: Meuse River Heavy Metal Concentrations (3)

Questions of interest:

1. Spatial patterns of the concentrations of the four heavy metals.
2. Relationship to factors that determine differences in sedimentation rate.

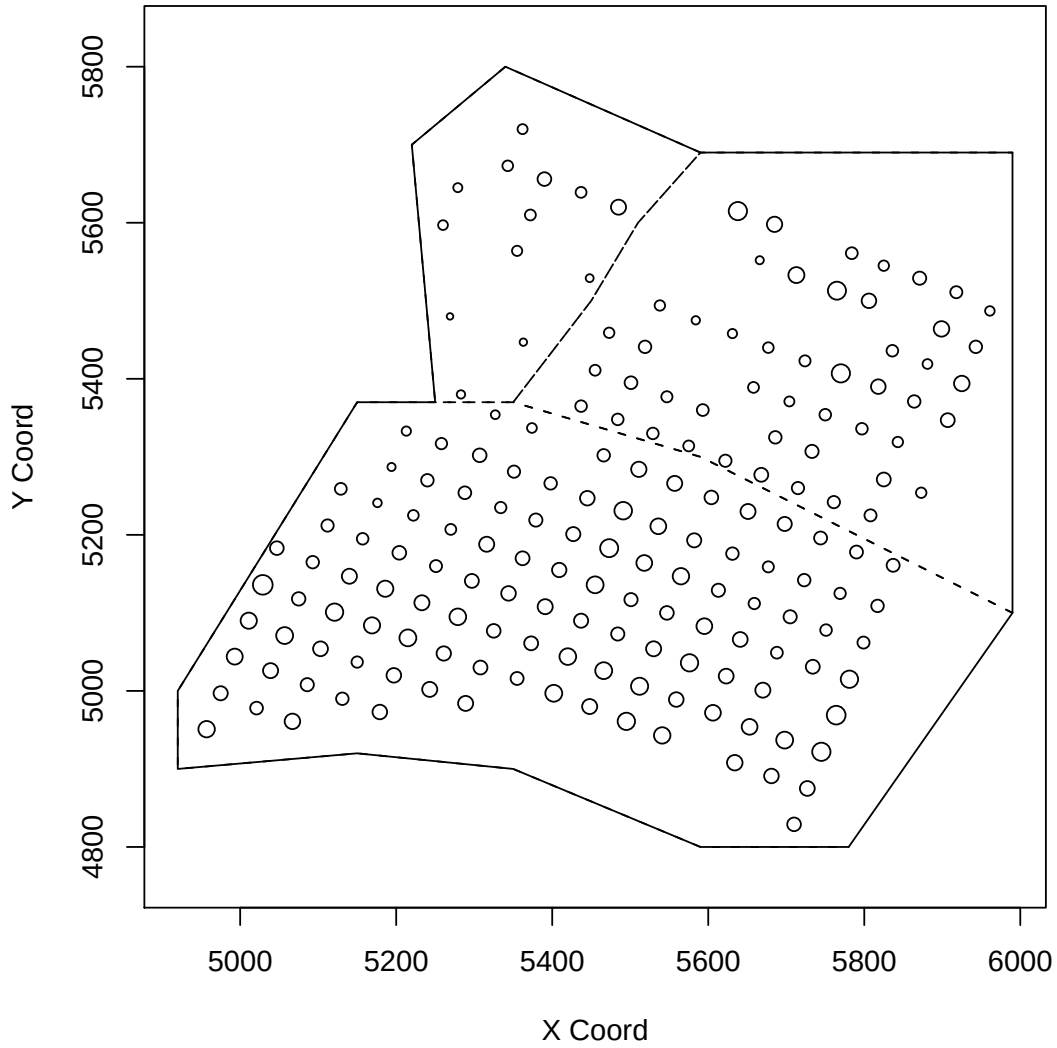
Examples: Soil Calcium Content (1)

Calcium concentrations (0-20 cm depth layer) measured in an area in Brazil. Sub-regions with different soil management histories:

1. typically flooded in rainy season, no longer used (upper left)
2. typically rice fields, fertilizers used, calcium recently added to neutralize effect of aluminum in soil (lower half)
3. experimental area, fertilizers used (upper right)

- Locations s: 178 soil core locations, incomplete regular lattice
- Outcome Y: calcium content in $mmol_c/dm^3$ (continuous)
- Covariates $\mathbf{x}$: sub-regions (3 levels), elevation in meters

Examples: Soil Calcium Content (2)



Examples: Soil Calcium Content (3)

Questions of interest:

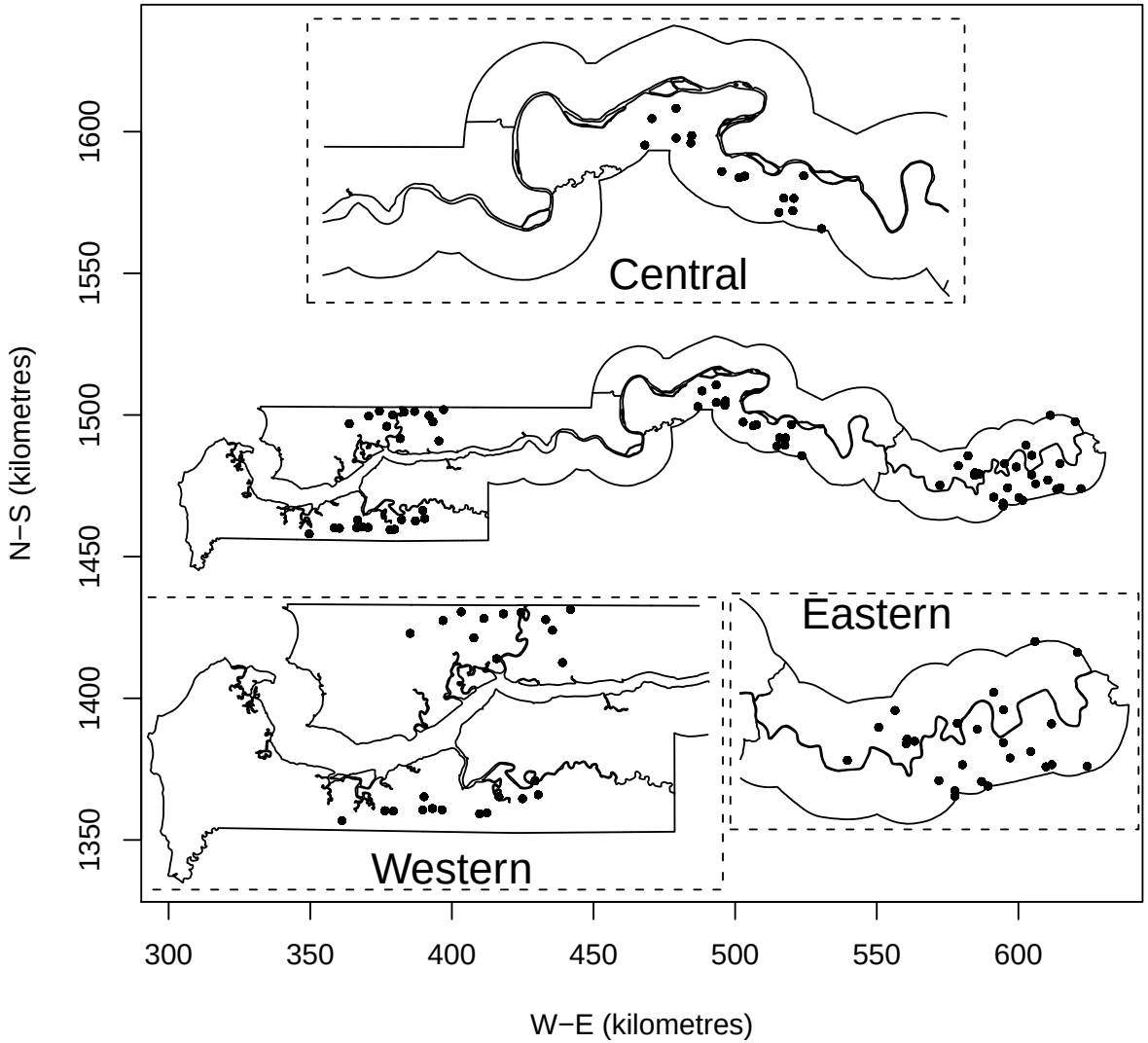
1. Construction of maps of the spatial variation in calcium.
   As measurements are taken from small soil cores and repeated sampling would yield different values, the map should not necessarily interpolate the data.
2. Relationship between calcium, study area and elevation.

Examples: Childhood Malaria in The Gambia (1)

Survey on 2035 children living in village communities in five regions in The Gambia

- Locations s: location of village
- Outcome Y: presence of malaria parasites in blood (binary)
- Covariates $\mathbf{x}$:
  - children: age, sex, mosquito net usage, mosquito net treatment with insecticide
  - villages: greenness of vegetation around village, belonging to primary health care structure of The Gambia Ministry of Health

Examples: Childhood Malaria in The Gambia (2)



Examples: Childhood Malaria in The Gambia (3)

Questions of interest:

1. Predictive model for malaria prevalence as a function of available explanatory variables.
2. Unexplained spatial variation between villages (might give clues on as-yet unmeasured environmental risk factors).

Examples: Residual contamination from nuclear weapons testing on Rongelap Island (1)

Measurements of residual contamination from the U.S. weapons testing program in the 1950s on Rongelap Island. Island uninhabited since the 1980s due to mounting health concerns, project to establish whether island safe for re-habitation.

- Locations s: spatial coordinates in meters. Nested grid design.
- Outcome Y: photon emission count (count)
- Covariates $\mathbf{x}$: time t in seconds over which the emission count was accumulated

Examples: Residual contamination from nuclear weapons testing on Rongelap Island (2)

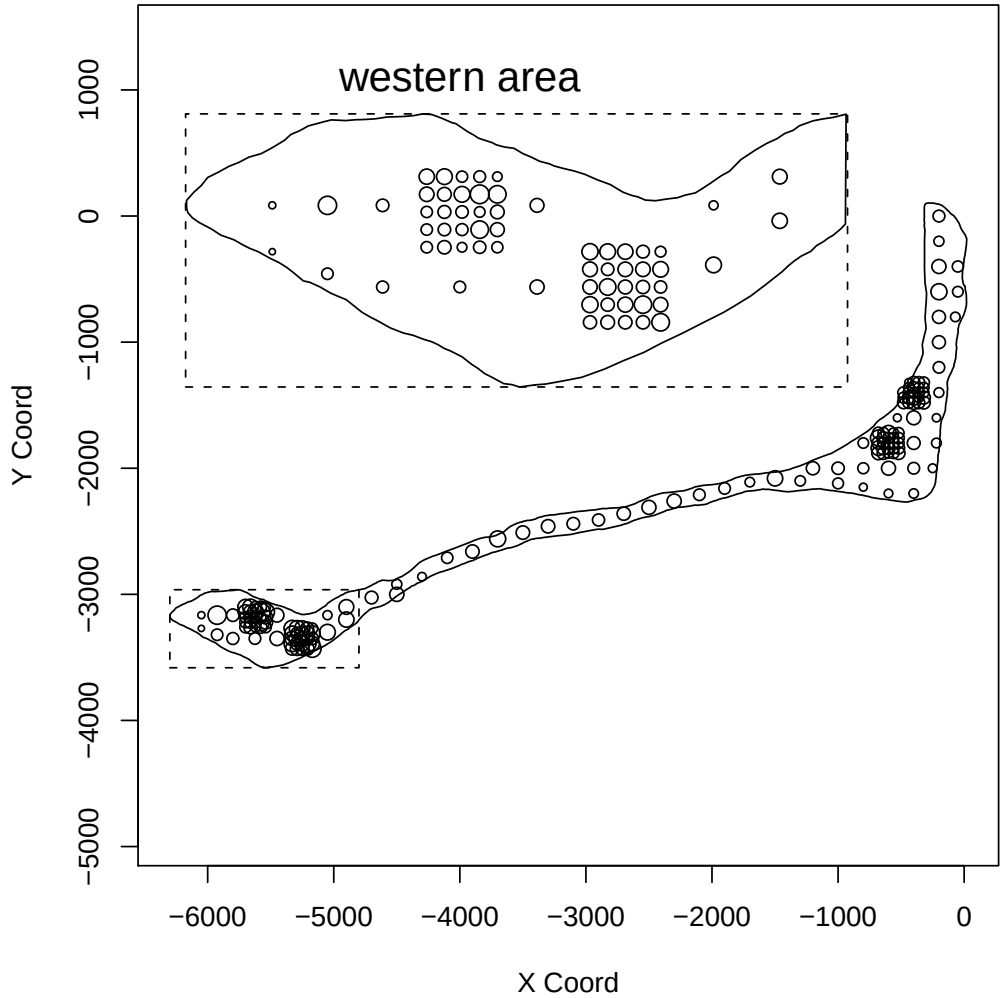


Figure: Circle radii are proportional to emission counts per unit time.

Examples: Residual contamination from nuclear weapons testing on Rongelap Island (3)

Questions of interest:

1. Estimated map of residual contamination.
2. Due to the health implications: Particular properties of the map such as location and value of the maximum, or areas where the contamination exceeds a prescribed threshold.

Data Sets

The first data set is taken from Applied Spatial Data Analysis with R by Bivand, Pebesma and Gómez-Rubio (2008) and available as part of the R package `sp`.

The other data sets are from Model-based Geostatistics by Diggle and Ribeiro (2007), which is available as ebook from the University library at `http://ebooks.ub.uni-muenchen.de/8793/`.
They are available as part of the R packages `geoR` and `geoRglm`.

R Code for Examples

```
library(sp); library(geoR); library(geoRglm)

### Meuse River Heavy Metal Concentrations ###
data(meuse); data(meuse.riv)
coordinates(meuse) <- c("x","y")
plot(coordinates(meuse),cex=log(meuse$cadmium),xlim=c(178250,181500),ylim=c(329500,333800),
     xlab="X Coord",ylab="Y Coord",cex.lab=1.4,cex.axis=1.4)
lines(meuse.riv,col=4)

### Soil Calcium Content ###
data(ca20)
points(ca20)
lines(ca20$reg1,lty=2)
lines(ca20$reg2,lty=2)
lines(ca20$reg3,lty=2)

### Childhood Malaria in The Gambia ###
data(gambia)
gambia.map()

### Residual contamination from nuclear weapons testing on Rongelap Island ###
data(rongelap)
points(rongelap)
rongwest <- subarea(rongelap,xlim=c(-6300,-4800))
rongwest.z <- zoom.coords(rongwest,xzoom=3.5,xoff=2000,yoff=3000)
points(rongwest.z,add=T)
rect.coords(rongwest$sub,lty=2,quiet=T)
rect.coords(rongwest.z$sub,lty=2,quiet=T)
text(-4000,1100,"western area",cex=1.5)
```

Overview

Data Structure (1)

- Locations $s_1, \ldots, s_n$
- Outcome variable recorded at those locations: $y_1, \ldots, y_n$
- Possibly vectors of predictors / covariates: $\mathbf{x_1}, \ldots, \mathbf{x_n}$

Here: data observed at locations in a two-dimensional domain, $s_i = (s_i^x, s_i^y) \in D \subset \mathbb{R}^2$. In contrast to areal data, the s_i are not restricted to a certain set of locations, but can vary continuously.

Data Structure (2)

Observations that are closer together tend to be more similar (e.g. due to underlying factors such as socio-economic variables, soil characteristics, shared health risks, ..., or due to processes such as diffusion or contagion) $\longrightarrow$ spatial correlation.

We assume that there is some common underlying (random) process $Y(s)$ that generates the y_i at the locations s_i.

$Y(s)$ can be continuous, discrete, ordinal, categorical or survival.

Assumptions

The locations $s_1, \ldots, s_n$ are either

a) fixed, i.e. chosen deterministically (e.g. on a grid), or

b) if sampled randomly, the sampling mechanism is (stochastically) independent of the outcome process Y.

An assumption that we are well advised to question in practice!

Example: Do we tend to have air pollution monitors (s_i) in cities/streets with high pollution levels (y_i)? What does that mean for learning about air pollution levels in the whole region/city?

Typical Questions of Interest

1. Relationship between an outcome $Y(s)$ and covariates $\mathbf{x}(s)$.
   $\rightarrow$ Take into account spatial correlation ("nuisance")
   (otherwise, parameter estimates may be less precise, and standard errors will be wrong, possibly leading to wrong conclusions)
   E.g. predictive model for childhood malaria prevalence

2. Prediction at locations s with no observations.
   $\rightarrow$ Spatial correlation allows learning from other locations!
   E.g. map of residual contamination on Rongelap Island

3. Description of unexplained spatial variability.
   $\rightarrow$ Spatial pattern itself of interest.
   E.g. unexplained spatial variation in malaria between villages

Model-based Geostatistics (1)

- Following Diggle and Ribeiro (2007), we will use a *model based* approach to geostatistics
- The generalized linear geostatistical model consists of two parts:
  1. A stationary Gaussian random field $Z(s)$ capturing spatial correlation
  2. Conditional on $Z(s)$, the $Y(s_1), \ldots, Y(s_n)$ are mutually independent random variables following a generalized linear model with

$$E(Y(s)|Z(s)) = \mu(s) = h^{-1}(Z(s) + \mathbf{x}(s)'\boldsymbol{\beta})$$

(e.g. linear model, Poisson log-linear model or logistic model)

Model-based Geostatistics (2)

Examples:

- For a linear Gaussian model h is the identity function and the error distribution is normal

$$Y_i|Z(s_i) \sim N(\mu_i, \tau^2).$$

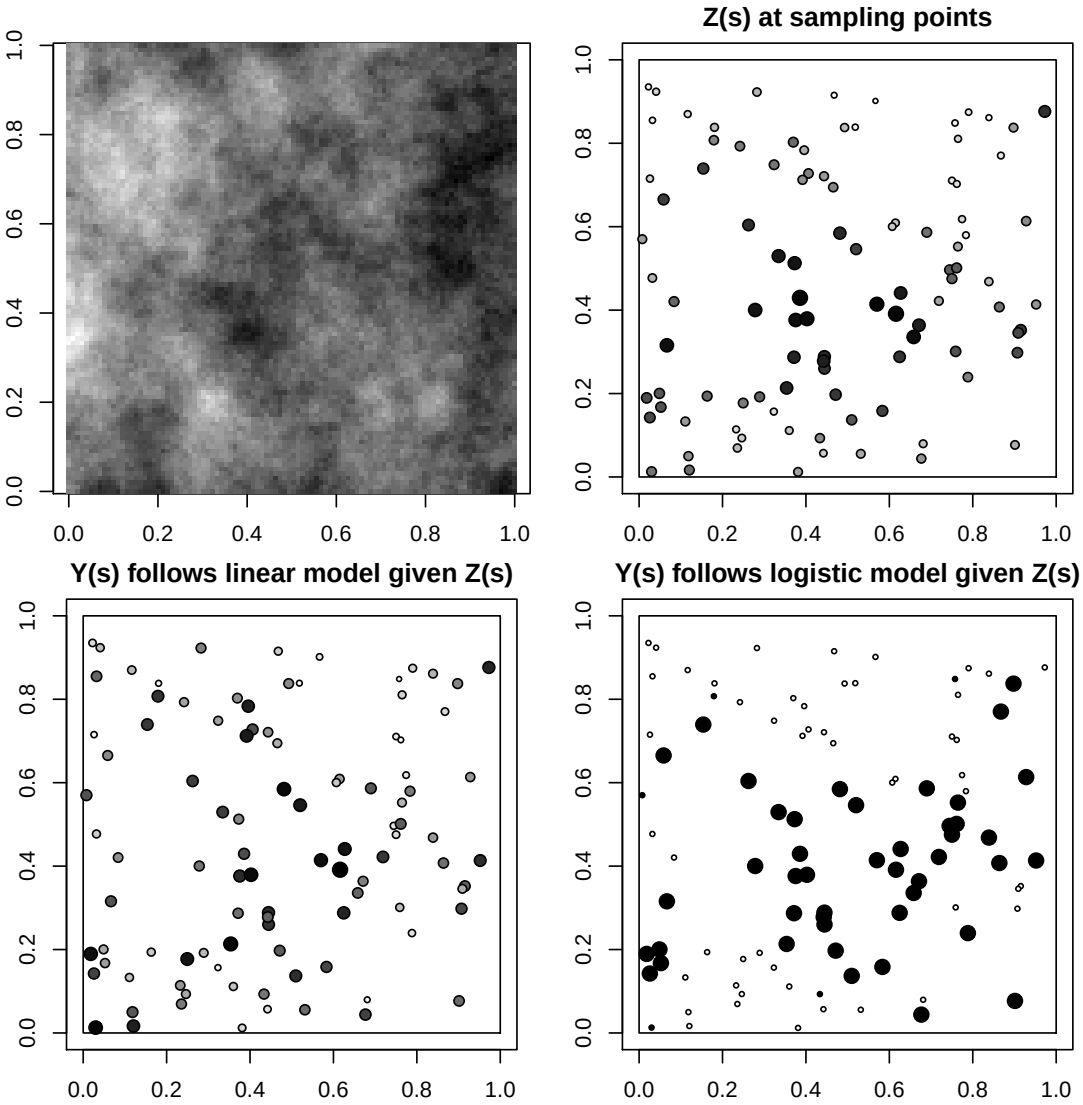
- For a Poisson log-linear model we have $h(\mu_i) = \log(\mu_i)$ and

$$Y_i|Z(s_i) \sim Pois(\mu_i).$$

- For a logit-linear model we have $h(\mu_i) = \log(\mu_i/(1-\mu_i))$ and

$$Y_i|Z(s_i) \sim Bin(\mu_i)$$

Model-based Geostatistics (3)



Model-based Geostatistics (4)

- *Estimation*: use data to deduce regression parameters β and parameters defining the covariance structure of $Z(s)$
- *Prediction*: refers to inference about the realization of the unobserved *signal process* $\mu(s)$, e.g.
  - prediction of $\mu(s)$ for a location $s \in D$ not in the data
  - prediction of summary statistics, such as the probability of $\mu(s)$ being above a threshold in an area
- *Hypothesis testing*: investigate whether a specific covariate has an effect on the distribution of $Y(s)$

Summing up

- Generalized linear geostatistical models (GLGMs)
  - assume a given distribution of $Y(s)|Z(s)$
  - formulate a model for $E(Y(s)|Z(s))$ based on covariates:

$$\mu(s) = h^{-1}(Z(s) + \mathbf{x}(s)'\boldsymbol{\beta})$$

- To use such models we need to
  - understand what a Gaussian random field is
  - figure out how to determine the unknown parameters of the model from data $\{s_i, y_i, \mathbf{x}_i\}$
- The difference between GLGM modeling and usual LM/GLM modeling is that we take the *spatial dependency* of the data into account

Overview

Overview

Random Field

$Y(s_1), \ldots, Y(s_n)$ are not independent, but spatially correlated. We assume that spatial correlation is due to an underlying random process or random field that is spatially correlated.

We typically assume a Gaussian random field.

Excursus: Multivariate Normal Distribution (1)

A vector $\mathbf{X} = (X_1, \ldots, X_p)$ is said to follow a multivariate normal distribution if

$$a_1 X_1 + \cdots + a_p X_p$$

follows a (univariate) normal distribution for any $a_1, \ldots, a_p \in \mathbb{R}$.

We write $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{A})$ with

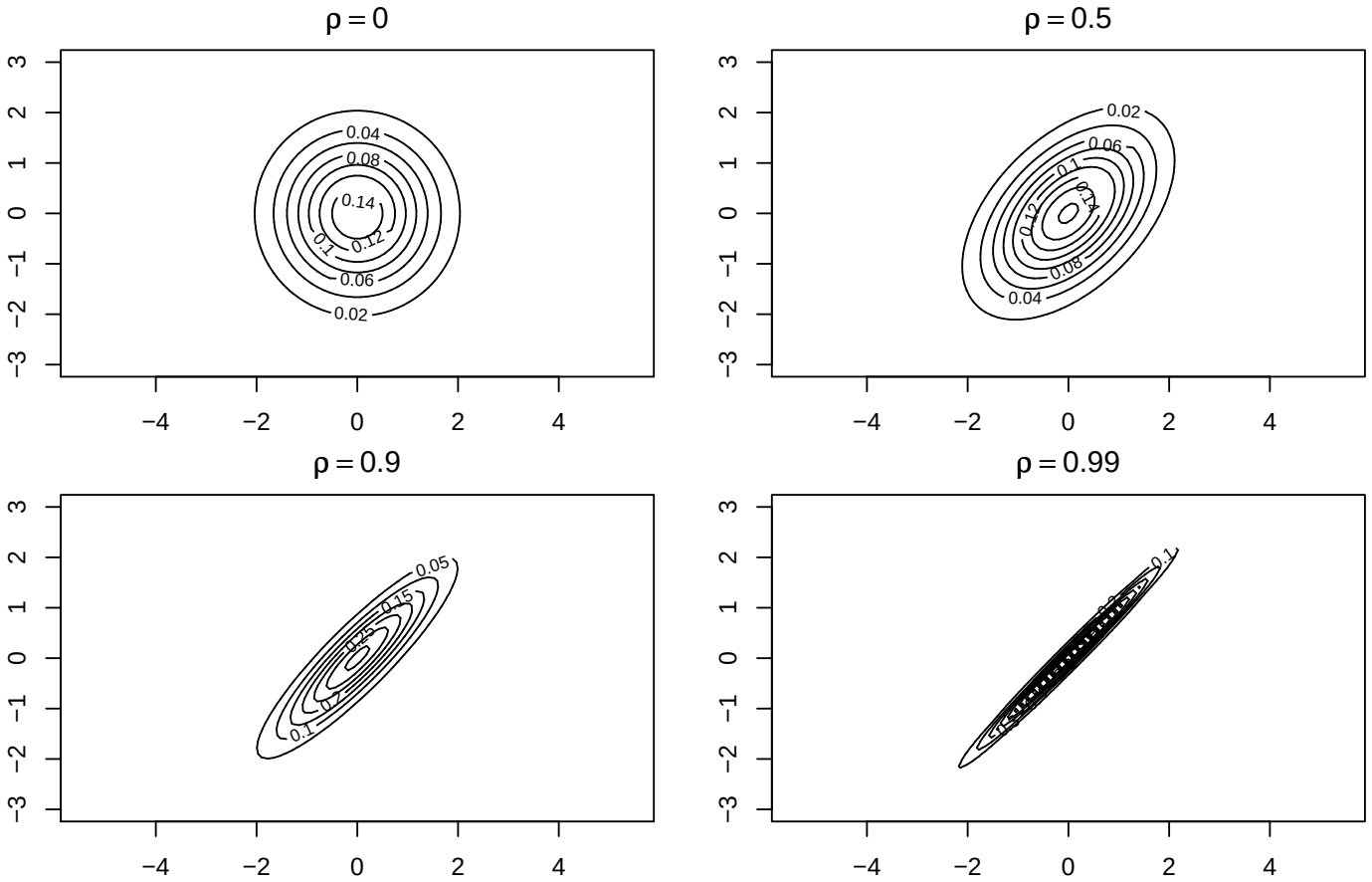
$$\begin{aligned} \boldsymbol{\mu} &= (E(X_1), \ldots, E(X_p)) \\ \mathbf{A} &= (Cov(X_i, X_j))_{ij}. \end{aligned}$$

The density of $\mathbf{X}$ then is

$$f(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} |\mathbf{A}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \mathbf{A}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right).$$

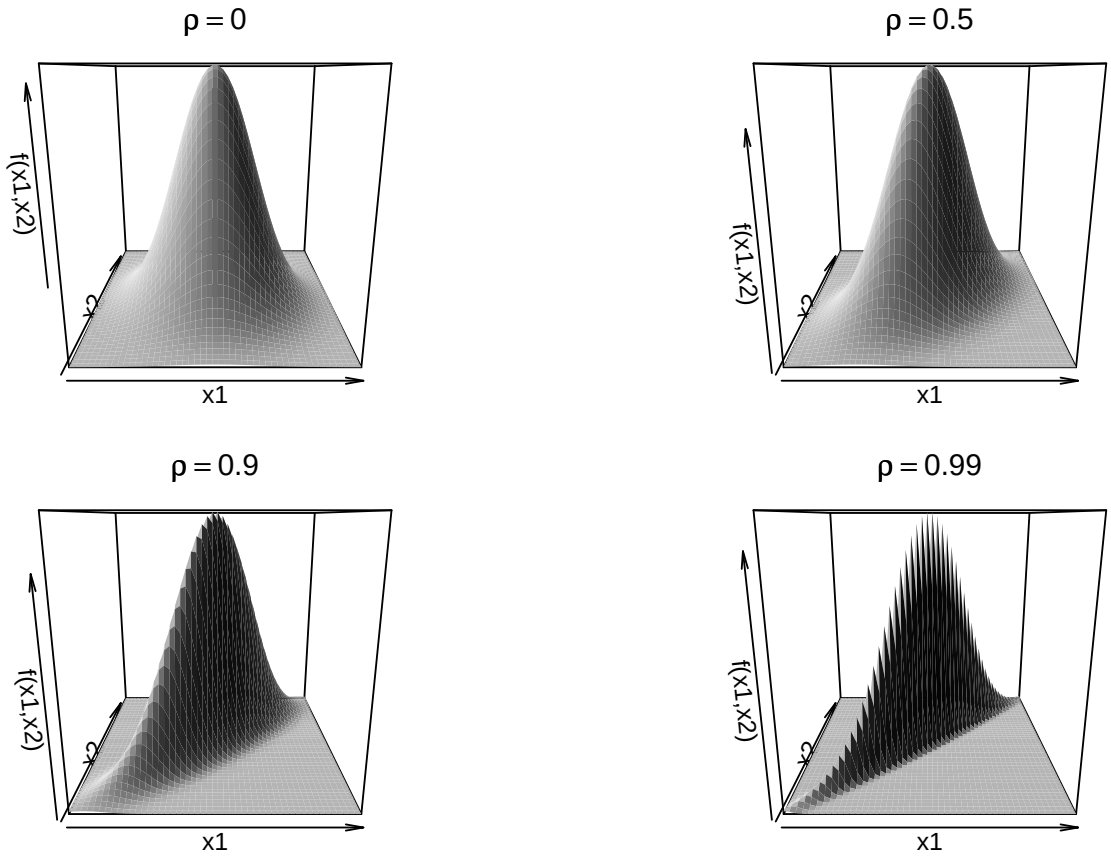
Excursus: Multivariate Normal Distribution (2)

- Example densities with equal variances for $p = 2$. ρ is the correlation $Corr(X_1, X_2) = Cov(X_1, X_2)/\sqrt{Var(X_1)Var(X_2)}$.



Excursus: Multivariate Normal Distribution (2)

- Example densities with equal variances for $p = 2$. ρ is the correlation $Corr(X_1, X_2) = Cov(X_1, X_2)/\sqrt{Var(X_1)Var(X_2)}$.



Gaussian Random Field (1)

A random field $\{Z(s) : s \in \mathbb{R}^2\}$ is a stochastic process in space, i.e. for each location $s \in \mathbb{R}^2$, $Z(s)$ is a random variable. ($Z(s)$ and $Z(t)$ may be correlated for $s \neq t$!)

For a Gaussian random field $\{Z(s) : s \in \mathbb{R}^2\}$, $(Z(s_1), \ldots, Z(s_n))'$ has a multivariate normal distribution for any $s_1, \ldots, s_n \in \mathbb{R}^2$.

The process then is characterized by its *mean function* $\mu(s) = \mathrm{E}(Z(s))$ and its *covariance function*

$$\gamma(s, s') = \mathrm{Cov}(Z(s), Z(s')).$$

Gaussian Random Field (2)

- A Gaussian random field is called (weakly) stationary if
  - the mean function $\mu(s) = \mu$ is constant
  - the covariance function $\gamma(s,t) = \gamma(s-t)$ depends only on $s-t$ (distance and direction).
- A stationary process is isotropic if $\gamma(s-s') = \gamma(||s-s'||)$, where $||\cdot||$ denotes the Euclidean ("straight-line") distance. (The direction of $s-t$ does not matter.)
- The *variance* of a stationary process is constant and equal to $\sigma^2 = \gamma(0)$. The correlation function is $\rho(u) = \gamma(u)/\sigma^2$.
- A process is called covariance stationary if $Z(s) - \mu(s)$ is stationary.

Assumptions

We typically assume (at least covariance) stationarity and isotropy.

Again, we are well advised to question these assumptions in practice. For example:

- Is the variance the same everywhere? (Or is there more unexplained variability in some areas than others?)
- Is the correlation the same in all directions?
  (E.g. does the wind typically blow in one direction? Then, air pollution levels down-wind might be more strongly correlated.)

Also: Non-Eucledian distances might sometimes be more meaningful (e.g. water contamination in river networks). For very large areas, the earth's curvature needs to be taken into account.

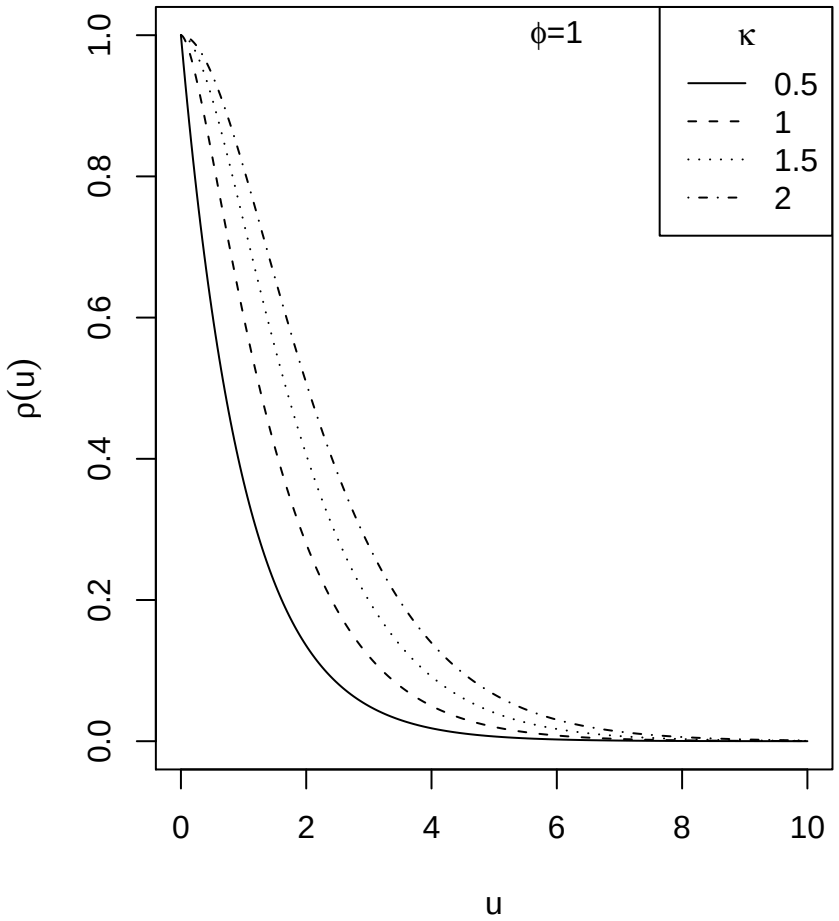
Isotropic Covariance Functions (1)

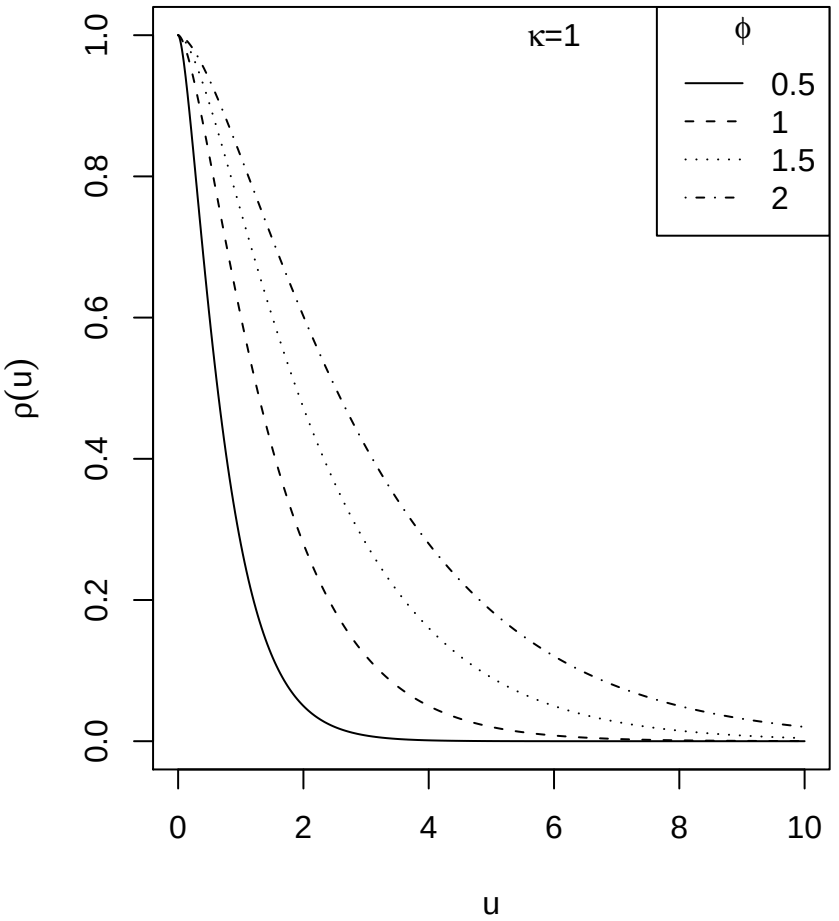
If we assume stationarity and isotropy, the covariance function $\gamma(u)$, or the correlation function $\rho(u) = \gamma(u)/\sigma^2$, describe how the correlation between two locations decreases with their distance u.

A flexible class of correlation functions that is often used in modeling is the Matérn family. This family has two parameters:

- $\kappa > 0$ controls the smoothness of the random field (shape parameter)
- $\phi > 0$ controls the range of spatial dependence (scale parameter).

Matérn Family Correlation Functions





Isotropic Covariance Functions (2)

- For $\kappa = 0.5$ the Matérn covariance function reduces to the exponential covariance function

$$\gamma(u) = \sigma^2 \exp(-u/\phi)$$

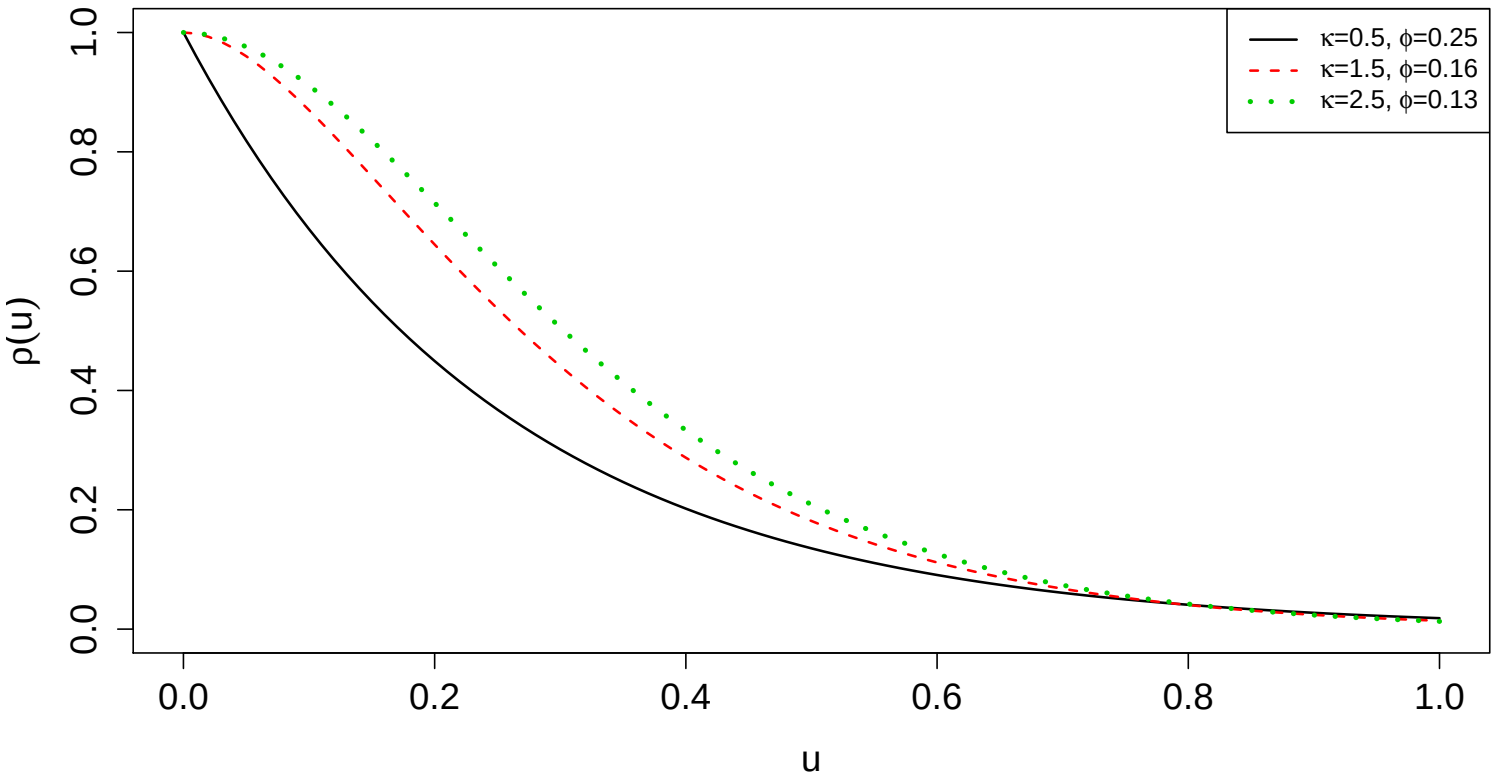
- For $\kappa \to \infty$ it reduces to the Gaussian covariance function

$$\gamma(u) = \sigma^2 \exp(-(u/\phi)^2)$$

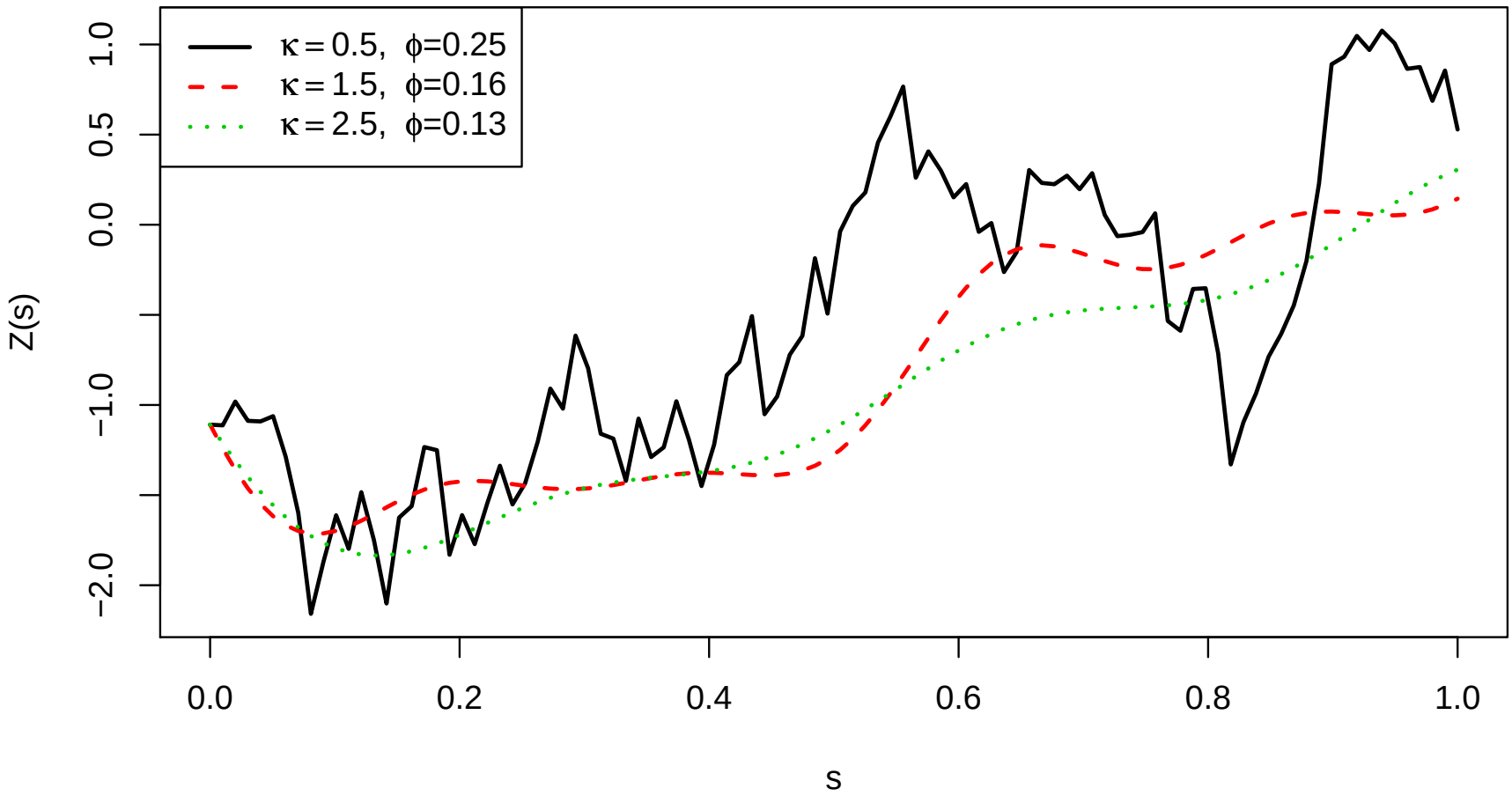
- The Matérn family and further covariance functions are implemented in the function `cov.spatial` in the R package `geoR`.

Simulations

The next slides show a one-dimensional Gaussian Random Process and Gaussian Random Fields with the following three Matérn covariance functions, chosen to have the same practical range of 0.75.



Simulated Gaussian Random Processes



Simulated Gaussian Random Fields: Covariance

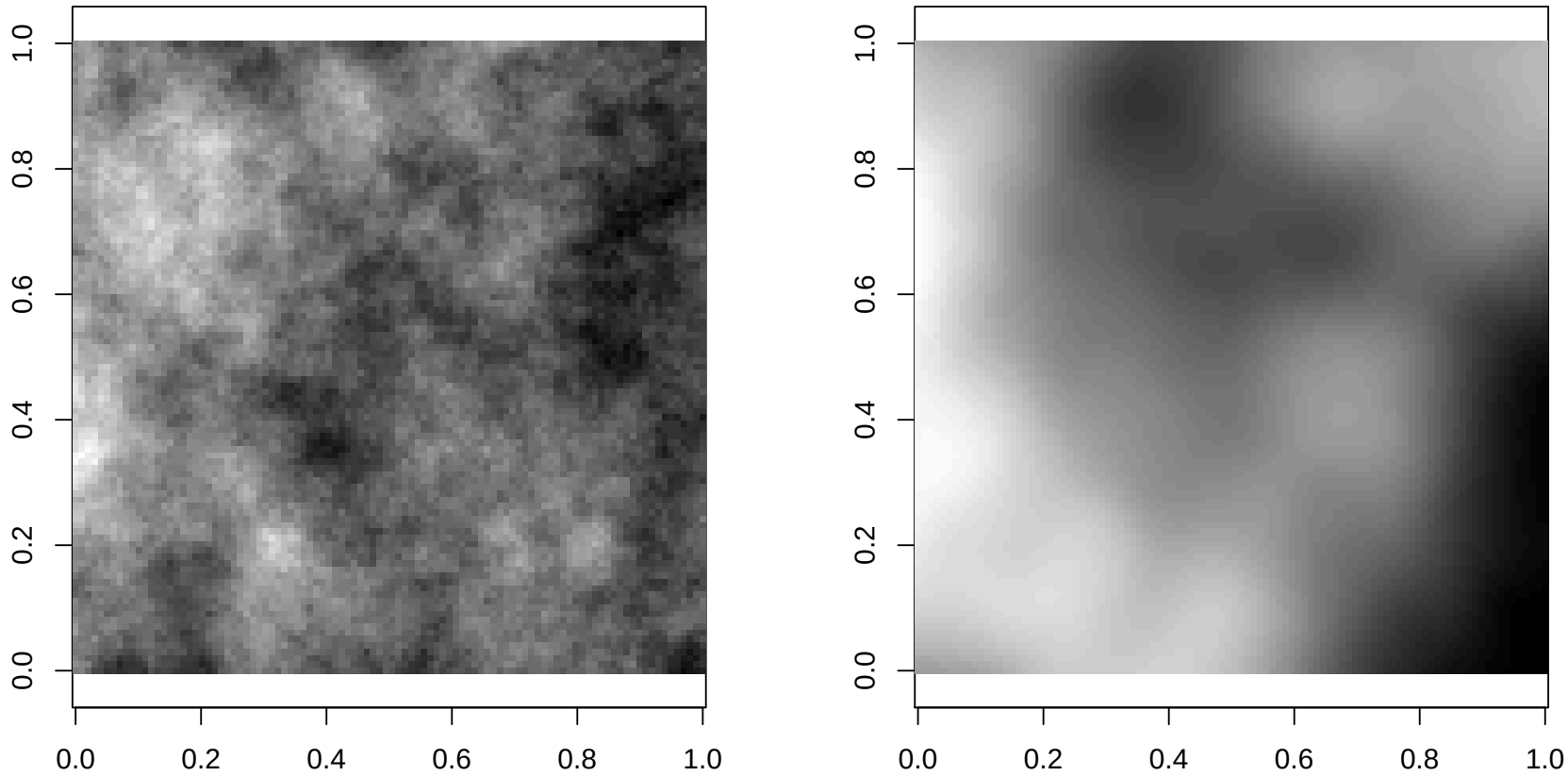


Figure: Simulated Gaussian random fields with Matérn covariance function. Left: $\kappa = 0.5, \phi = 0.25$. Right: $\kappa = 2.5, \phi = 0.13$.

Simulated Gaussian Random Fields: Anisotropy

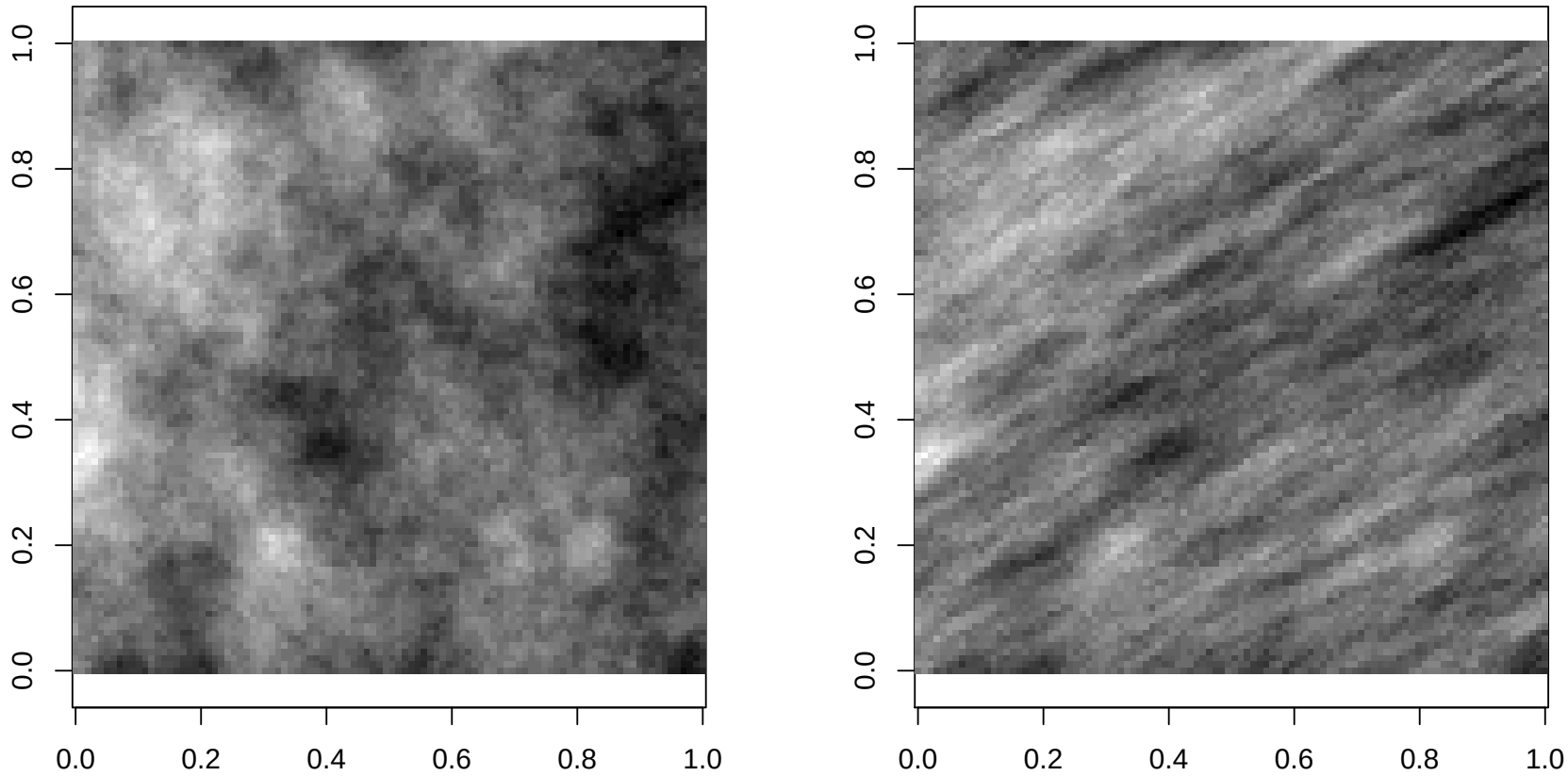


Figure: Isotropic (left) and anisotropic (right) simulated Gaussian random fields with Matérn covariance function with $\kappa = 0.5$ and $\phi = 0.25$.

R code for Simulation of Random Field

```
library(geoR)

### plot covariance functions ###
x <- seq(0,1,l=101)
plot(x,cov.spatial(x,cov.model="mat",kappa=0.5,cov.pars=c(1,0.25)),type="l",xlab="u",
    ylab=expression(rho(u)),ylim=c(0,1),lwd=1.6,cex.axis=1.4,cex.lab=1.4)
lines(x,cov.spatial(x,cov.model="mat",kappa=1.5,cov.pars=c(1,0.16)),lty=2,col=2,lwd=1.6)
lines(x,cov.spatial(x,cov.model="mat",kappa=2.5,cov.pars=c(1,0.13)),lty=3,col=3,lwd=3.2)
legend("topright",c(expression(paste(kappa,"=0.5, ",phi,"=0.25"),paste(kappa,"=1.5, ",phi,"=0.16"),
    paste(kappa,"=2.5, ",phi,"=0.13"))),lty=1:3,col=1:3,lwd=c(1.6,1.6,3.2))

### simulate random fields with different covariance functions ###
par(mfrow=c(1,2))
set.seed(159)
image(grf(100^2,grid="reg",cov.pars=c(1,0.25),cov.model="mat",kappa=0.5),
    col=gray(seq(1,0,l=51)),xlab="",ylab="")
set.seed(159)
image(grf(100^2,grid="reg",cov.pars=c(1,0.13),cov.model="mat",kappa=2.5),
    col=gray(seq(1,0,l=51)),xlab="",ylab="")

### simulate isotropic and anisotropic random fields ###
set.seed(159)
image(grf(100^2,grid="reg",cov.pars=c(1,0.25),cov.model="mat",kappa=0.5),
    col=gray(seq(1,0,l=51)),xlab="",ylab="")
set.seed(159)
image(grf(100^2,grid="reg",cov.pars=c(1,0.25),cov.model="mat",kappa=0.5,aniso.pars=c(pi/3,4)),
    col=gray(seq(1,0,l=51)),xlab="",ylab="")
```

Overview

Semivariogram (1)

How do we investigate the covariance of a random field?

For a stationary and isotropic random field, we have that for two locations s and t with distance u

$$V(u) = \frac{1}{2} Var(Z(s) - Z(t)) = \frac{1}{2} E((Z(s) - Z(t))^2) = \sigma^2(1 - \rho(u))$$

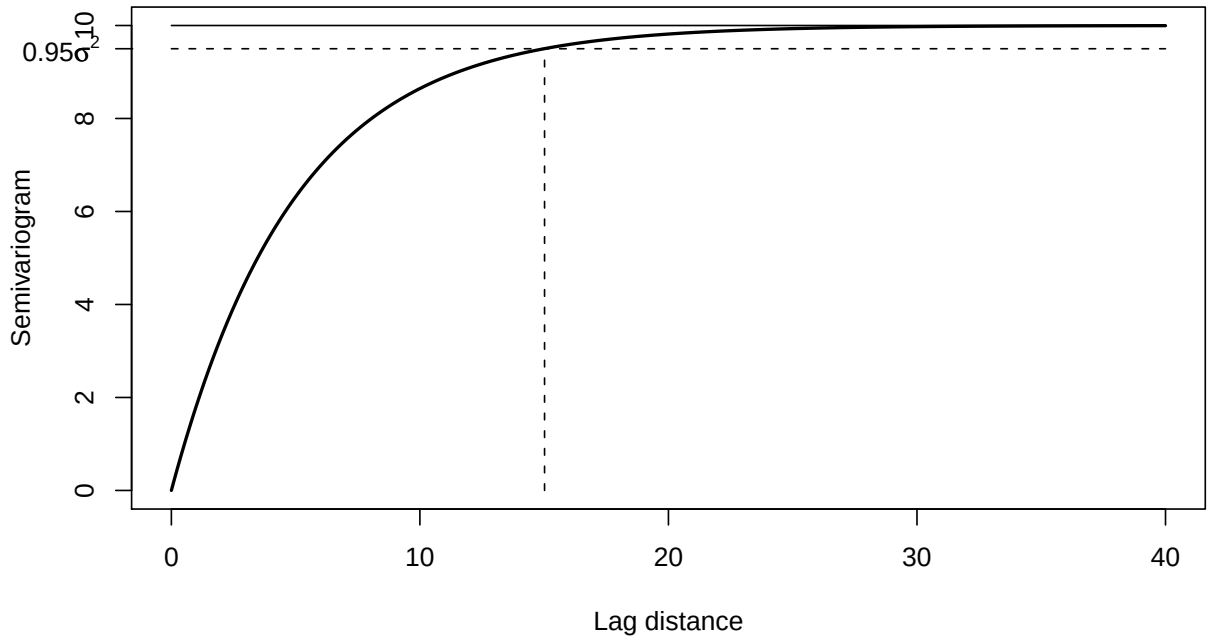
is a function of σ^2 and $\rho(u)$ - the semivariogram - which we can estimate by

$$\frac{1}{2}(z_i - z_j)^2$$

using all data pairs z_i and z_j with distance (approximately) u (more later).

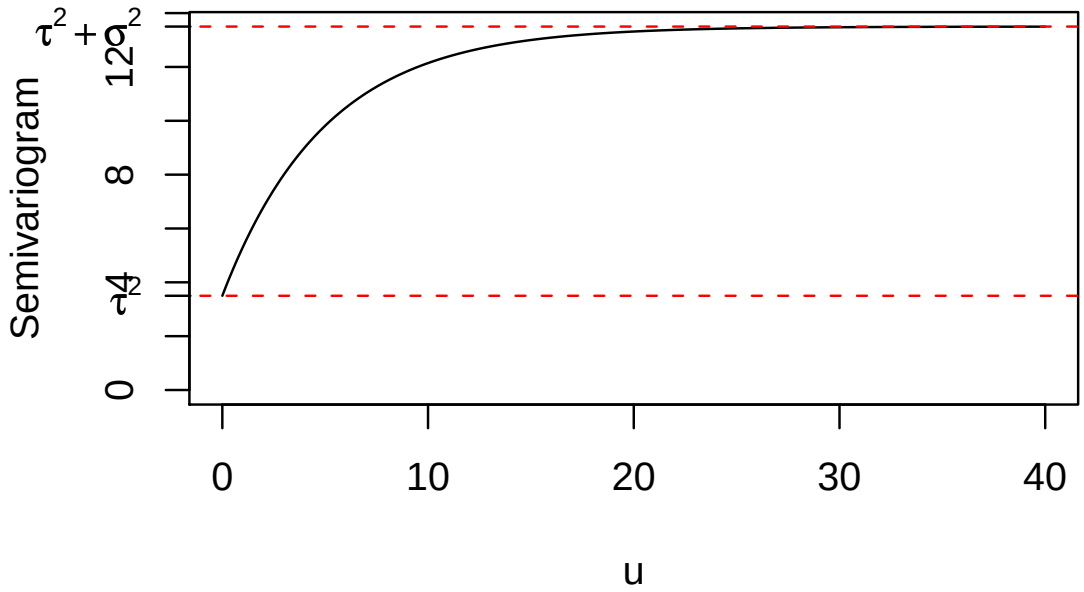
Semivariogram (2)

- $\lim_{u \to \infty} V(u) = \sigma^2$ is also called the sill (corresponding to the variance).
- The practical range is the distance u where $V(u) = 0.95 \cdot \sigma^2$, i.e. where the correlation has dropped to $\rho(u) = 0.05$.



Nugget effect

- For $u \approx 0$ and thus $\rho(u) \approx 1$, one would expect $V(u) \approx 0$.
- In geostatistical practice, one often sees a nugget effect $V(u) \approx \tau^2 > 0$ for small u, which can be interpreted as the variance of (spatially uncorrelated) measurement error.
- Then, $V(u) = \tau^2 + \sigma^2(1 - \rho(u))$.



Overview

Overview

Data

- Assume $Y(s_1), \ldots, Y(s_n)$ are generated by a stationary and isotropic random field Z with mean μ and variance σ^2 as

$$Y(s_i) = Z(s_i) + \epsilon_i, \quad i = 1, \ldots, n$$

where the ϵ_i are independent and identically distributed with zero mean and variance τ^2.

- The semivariogram of the observation process is

$$\begin{aligned} V_Y(u_{ij}) &= \frac{1}{2} \mathrm{E}\left[(Y(s_i) - Y(s_j))^2\right], \quad u_{ij} = ||s_i - s_j|| \\ &= \tau^2 + \sigma^2 \left\{1 - \rho(u_{ij})\right\}. \end{aligned}$$

Here, $\rho(\cdot)$ denotes the autocorrelation function of the random field Z.

Estimating the Semivariogram

- To learn about the semivariogram one could plot $v_{ij} = \frac{1}{2}(y_i - y_j)^2$ against $u = ||s_i - s_j||$.
- A more easily interpretable plot is obtained by averaging the v_{ij} within distance bins $(u_{k-1}, u_k]$, $k = 1, \ldots, K$, with $u_0 = 0$. The empirical semivariogram value for the kth bin then is

$$\widehat{V}_k = \frac{1}{2|N_k|} \sum_{i,j \in N_k} (y_i - y_j)^2,$$

  where N_k includes all pairs i and j with distance $||s_i - s_j||$ in $(u_{k-1}, u_k]$.
- In R, the `variog` function from package `geoR` does the work.

Estimating the Semivariogram: Soil Calcium Data (1)

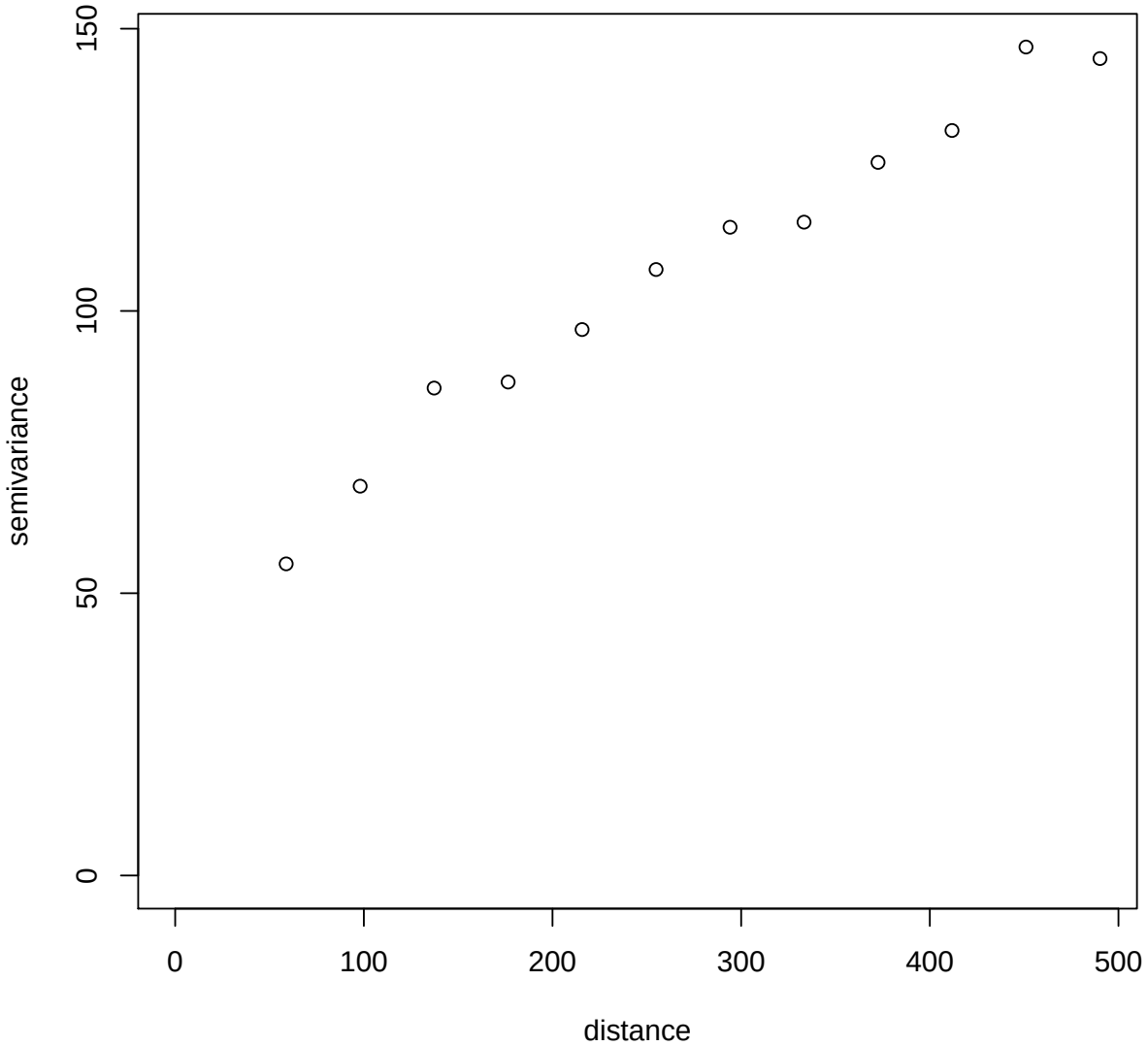


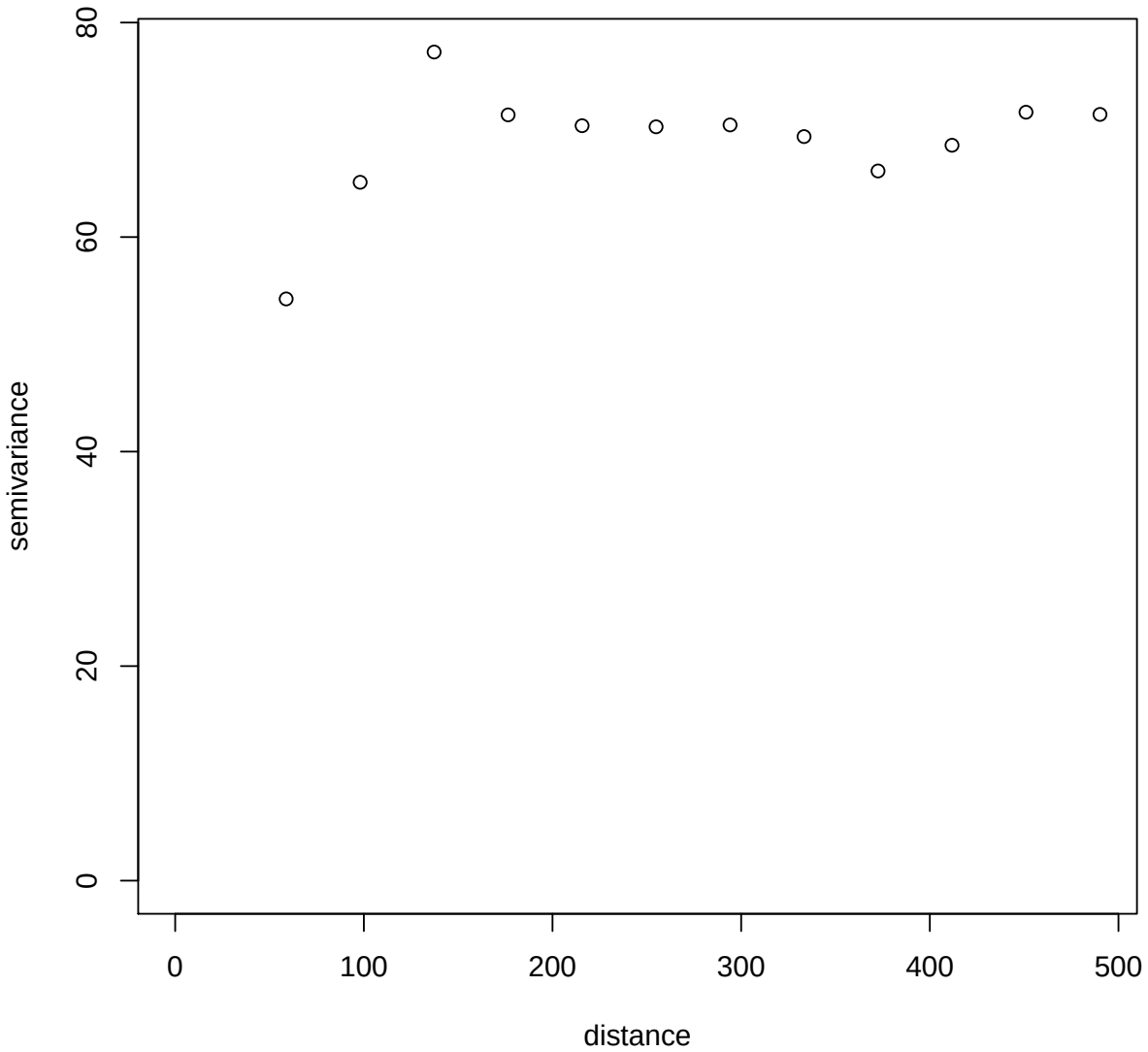
Figure: Empirical semivariogram for calcium in the soil calcium data

Pitfalls using the Semivariogram

- If directionality is suspected (i.e. the process is not isotropic), a map of binned values v_{ij} taking into account direction can be used instead, although more data is needed to obtain reasonable estimates.
- If the mean of the process is not constant (spatial trends, covariate effects), the empirical variogram can give misleading results. One can estimate the semivariogram from residuals, regressing first on spatial terms and covariates, but a better approach is to estimate everything jointly using maximum likelihood.

Estimating the Semivariogram: Soil Calcium Data (2)

Figure: Empirical semivariogram for calcium in the soil calcium data, adjusting for area, altitude and quadratic spatial trend



Overview

Maximum Likelihood Estimation

- In classical geostatistics, the semivariogram estimators are not only used for exploratory analysis, but also for parameter estimation.
- A statistically more sound approach is the use of likelihood-based methods.
- Given a probability model for the observations, a *likelihood function* $L(\boldsymbol{\theta}; \mathbf{y})$ is formulated, where $\boldsymbol{\theta}$ are the parameters of the model.
- The aim is to find those values of $\boldsymbol{\theta}$ which maximize the (log-)likelihood of the actual observed data.

Gaussian models (1)

Recall the *Gaussian model* from the introduction:

- Let $Z(s)$ be a zero mean stationary and isotropic Gaussian random field and let $\mathbf{Z} = (Z(s_1), \ldots, Z(s_n))'$.
- Conditionally on $\mathbf{Z}$, the observations $\mathbf{Y}$ are independent and follow a normal distribution

$$Y(s_i)|Z(s_i) \sim N(\mu(s_i), \tau^2),$$

where

$$\mu(s_i) = Z(s_i) + \mathbf{x}(s_i)'\boldsymbol{\beta}.$$

- Let $\mathbf{X}$ be an $n \times p$ matrix having rows $\mathbf{x}(s_i)$, $i = 1, \ldots, n$. For $\boldsymbol{\mu} = (\mu(s_1), \ldots, \mu(s_n))$ we can write

$$\boldsymbol{\mu} = \mathbf{Z} + \mathbf{X}\boldsymbol{\beta}.$$

Gaussian models (2)

- The marginal distribution of $\mathbf{Y} = (Y(s_1), \ldots, Y(s_n))$ (taking into account the distribution of the $Z(s_i)$) then is

$$\mathbf{Y} \sim N_n(\mathbf{X}\beta,\ \sigma^2\mathbf{R}(\phi) + \tau^2\mathbf{I}),$$

  where $\mathbf{I}$ is the $n \times n$ identity matrix and the $n \times n$ matrix $\mathbf{R}$ depends on a vector of parameters ϕ of the covariance function.
- The covariance of $\mathbf{Y}$ thus has two components: one from the random field $Z(s)$ capturing the spatial correlation, and one from additional uncorrelated random noise.

Gaussian models (3)

- The interesting part of the loglikelihood function is

$$l(\beta, \tau^2, \sigma^2, \phi) = -\frac{1}{2}\log(|\mathbf{S}|) + (\mathbf{y} - \mathbf{X}\beta)'\mathbf{S}^{-1}(\mathbf{y} - \mathbf{X}\beta),$$

  where $\mathbf{S} = \sigma^2\mathbf{R}(\phi) + \tau^2\mathbf{I}$ and $|\cdot|$ denotes the determinant.
- Let $\nu^2 = \tau^2/\sigma^2$ and $\mathbf{V} = \mathbf{S}/\sigma^2 = \mathbf{R}(\phi) + \nu^2\mathbf{I}$. Given $\mathbf{V}$, the above likelihood is maximized in β and σ^2 using weighted least squares

$$\beta(\mathbf{V}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y},$$
$$\sigma^2(\mathbf{V}) = \frac{1}{n}(\mathbf{y} - \mathbf{X}\beta(\mathbf{V}))'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta(\mathbf{V}))$$

Gaussian models (4)

- Plugging this result back into the loglikelihood yields the *profile loglikelihood* for ϕ and ν^2:

$$l_p(\nu^2, \phi) = -\frac{n}{2}\log(\sigma^2(\mathbf{V})) - \frac{1}{2}\log(|\mathbf{V}|)$$

- By numerically optimizing $l_p(\nu^2, \phi)$ one obtains $\widehat{\phi}$ and $\widehat{\nu}^2$.
- By back substitution one obtains $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$.
- An alternative for estimating (ν^2, σ^2, ϕ) is *restricted maximum likelihood* estimation.

Gaussian models (5)

- When using a Matérn covariance function, the order parameter κ is often poorly identified. Therefore, one often estimates κ by choosing the value from a discrete set, e.g. $\kappa = \{0.5, 1.5, 2.5\}$, which obtains the highest likelihood.
- An advantage of the likelihood framework is that, based on asymptotic normality, the variability of the estimators can be quantified, e.g. using $\text{Cov}(\widehat{\beta}_i)$.
- A $(1-\alpha)\cdot 100\%$ Wald confidence interval for a regression parameter β_i, $i = 1, \ldots, p$, is then

$$\widehat{\beta}_i \pm 1.96\sqrt{\text{Var}(\widehat{\beta}_i)}.$$

Transformed Gaussian models

- The applicability of the Gaussian model can be extended by assuming that the model holds after a marginal transformation of the response variable, i.e. $Y^*(s_i) = g(Y(s_i))$.
- Examples are the log, $\sqrt{}$, logit and probit-transformation of the data. A general class is the Box-Cox transformation:

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

- Transformations used to be a common strategy to model non-Gaussian data. However, a more modern approach is GLM based modeling (see next section).

geoR

The R package `geoR` allows the fitting of Gaussian models.

- Function `variog` can compute a (binned) empirical semivariogram.
- Function `likfit` can fit the Gaussian geostatistical model and simultaneously estimate regression parameters $\boldsymbol{\beta}$, variance and correlation parameters σ^2 and ϕ.
- Function `boxcox.fit` estimates λ for the Box-Cox transformation.
- Spatial prediction (next subsection) of $\mu(s)$ can be done using the function `krige.conv`.

More information on `geoR`: `http://www.leg.ufpr.br/geoR/`.

Linear Geostatistical Model: Soil Calcium Data (1)

- Estimating λ for a Box-Cox transformation yields $\lambda = 1.1 \approx 1$. We proceed with the untransformed data.
- Comparing the maximized likelihoods for a linear geostatistical model with a Matérn covariance function and $\kappa = 0.5$, 1.5 or 2.5 yields best results for $\kappa = 0.5$ (exponential covariance).
- Fitting linear geostatistical models with exponential covariance function and different covariates yields the following results:

Linear Geostatistical Model: Soil Calcium Data (2)

Mean model	Parameters	2 logL
constant	4	-1265.36
soil type	6	-1257.49
soil type, altitude	7	-1257.32
soil type, linear spatial trend	8	-1255.34
soil type, altitude, linear spatial trend	9	-1255.21

A log-likelihood-ratio test for nested models yields a significant effect of soil type, but no significant additional improvement when including altitude or a linear spatial trend $\longrightarrow$ second model.

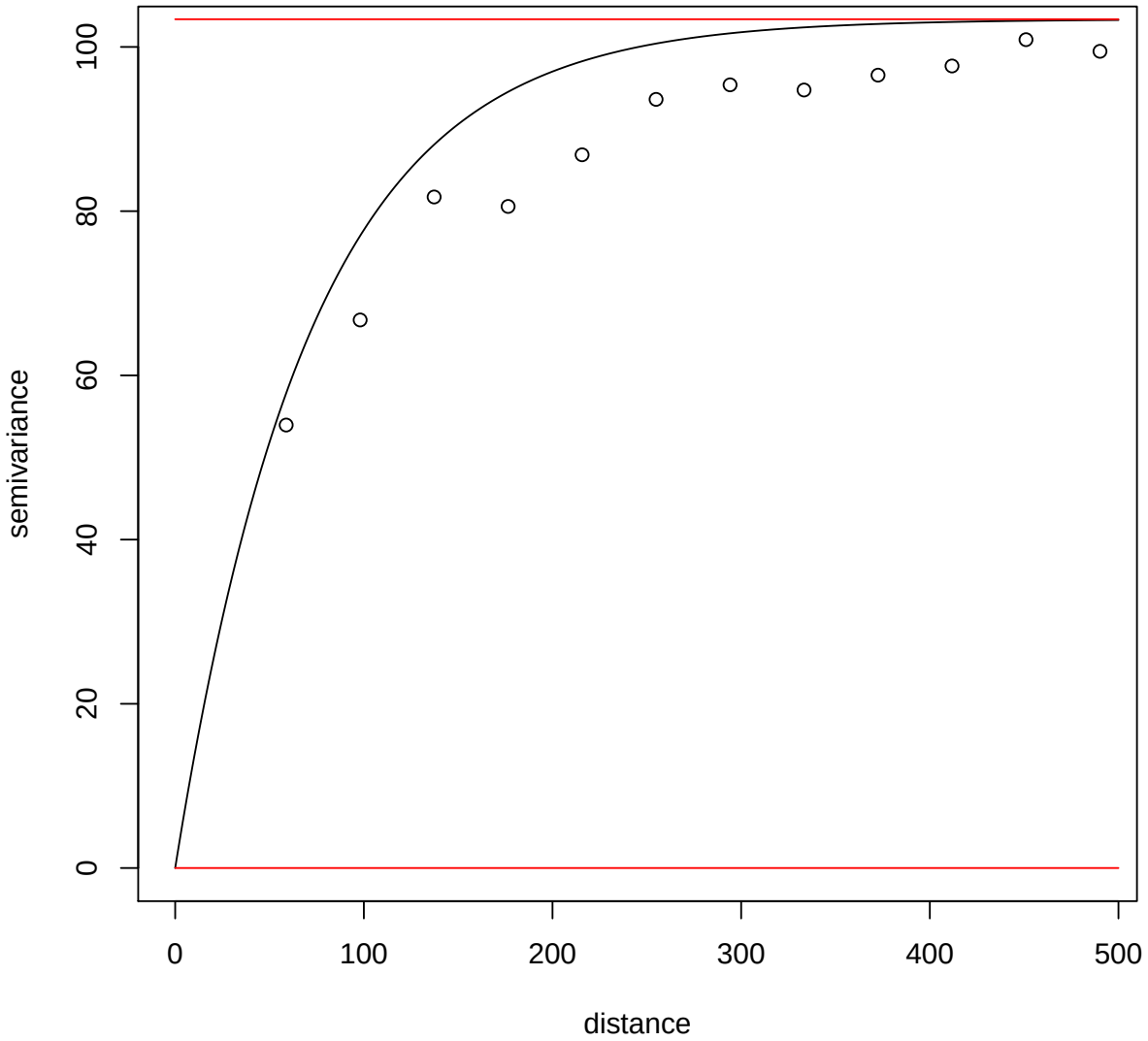
Linear Geostatistical Model: Soil Calcium Data (3)

For the second model, the parameter estimates are:

Parameter	Estimate
area 1	38.42
area 2	46.96
area 3	54.00
τ^2	0
σ^2	103.4
ϕ	71.78

Linear Geostatistical Model: Soil Calcium Data (4)

Figure: Empirical (adjusted) semivariogram and fitted theoretical semivariogram for soil calcium data



R code for Analyzing the Soil Calcium Data

```
### Box-Cox transformation? ###
boxcox.fit(ca20$data)  #no transformation (lambda=1) seems appropriate
hist(ca20$data)

### semivariogram ###
plot(variog(ca20,max.dist=510))  # no covariates
t.all <- trend.spatial(ca20,trend=~area + altitude,add="2nd")
plot(variog(ca20,max.dist=510,trend=~t.all))  # adjusting for area, altitude and quadratic spatial trend

### linear geostatistical models: deciding on kappa ###
likfit(ca20, ini = c(10,200), nug=50,kappa=0.5)  # largest likelihood
likfit(ca20, ini = c(10,200), nug=50,kappa=1.5)
likfit(ca20, ini = c(10,200), nug=50,kappa=2.5)

### linear geostatistical models ###
m1 <- likfit(ca20, ini = c(10,200), nug=50)
m2 <- likfit(ca20, trend=~area, ini = c(60,100), nug=40)  # chosen model
m3 <- likfit(ca20, trend=~area + altitude, ini = c(60,100), nug=40)
m4 <- likfit(ca20, trend=~area + coords, ini = c(60,100), nug=40)
m5 <- likfit(ca20, trend=~area + altitude + coords, ini = c(60,100), nug=40)

### semivariogram and fitted covariance model ###
plot(variog(ca20,max.dist=510,trend=~area))
x <- seq(0,500,1)
lines(x,m2$tausq + m2$sigmasq - cov.spatial(x,cov.model="mat",kappa=m2$kappa,
   cov.pars=c(m2$sigmasq,m2$phi)))
lines(x,rep(m2$tausq,length(x)),col=2)
lines(x,rep(m2$tausq+m2$sigmasq,length(x)),col=2)
```

Overview

Spatial Prediction for the Gaussian model (1)

Recall the Gaussian geostatistical model: Conditionally on $\mathbf{Z}$, the observations $\mathbf{Y}$ are independent and follow a normal distribution

$$Y(s_i)|Z(s_i) \sim N(\mu(s_i), \tau^2),$$

where

$$\mu(s_i) = Z(s_i) + \mathbf{x}(s_i)'\boldsymbol{\beta}.$$

Interest is now in predicting the value of $\mu(s_0)$ at a new location s_0 (or a set of new locations, such as in a predictive map).

Spatial Prediction for the Gaussian model (2)

$\mu(s_0)$ and $\mathbf{Y}$ jointly follow a multivariate normal distribution. The best predictor of $\mu(s_0)$ (in terms of *mean square prediction error* and for known parameters) then is the conditional expectation

$$\widehat{\mu}(s_0) = E(\mu(s_0)|\mathbf{Y}) = \mathbf{x_0'}\beta + \mathbf{r'}(\mathbf{R} + \nu^2\mathbf{I})^{-1}(\mathbf{Y} - \mathbf{X}\beta).$$

Here, $\mathbf{x_0}$ are the covariates at s_0, and $\mathbf{r}$ is a vector with elements $r_i = \rho(||s_0 - s_i||), i = 1, \ldots, n$. In practice, we plug in the estimated values for all parameters.

The covariate-based estimate $\mathbf{x_0'}\hat{\beta}$ is adjusted up- or downward according to the deviations $\mathbf{Y} - \mathbf{X}\hat{\beta}$ – weighted by the correlations between the locations $s_1, \ldots, s_n$ and s_0.

Spatial Prediction for the Gaussian model (3)

- One can also derive the prediction variance for $\widehat{\mu}(s_0)$.
- Likewise, one can derive best predictors for other linear and non-linear targets, such as
  - the cumulative values of $\mu(s)$ over a sub-area A
  - the maximum of $\mu(s)$
  - the set of locations for which $\mu(s)$ exceeds some threshold of interest.

  For non-linear targets one has to resort to Monte Carlo simulation for a grid of prediction points to estimate the target.

Miscellaneous

- The term *ordinary Kriging* is used for Kriging with constant mean.
- The term *Trans-Kriging* is used for Kriging of transformed Gaussian models.
- Kriging with non-constant mean $\mu(s) = \mathbf{x}(s)'\boldsymbol{\beta}$ is called *universal Kriging*.
- Note the plug-in of $\boldsymbol{\beta}$, variance and correlation parameters means that their estimation uncertainty is ignored $\rightarrow$ Bayesian Kriging

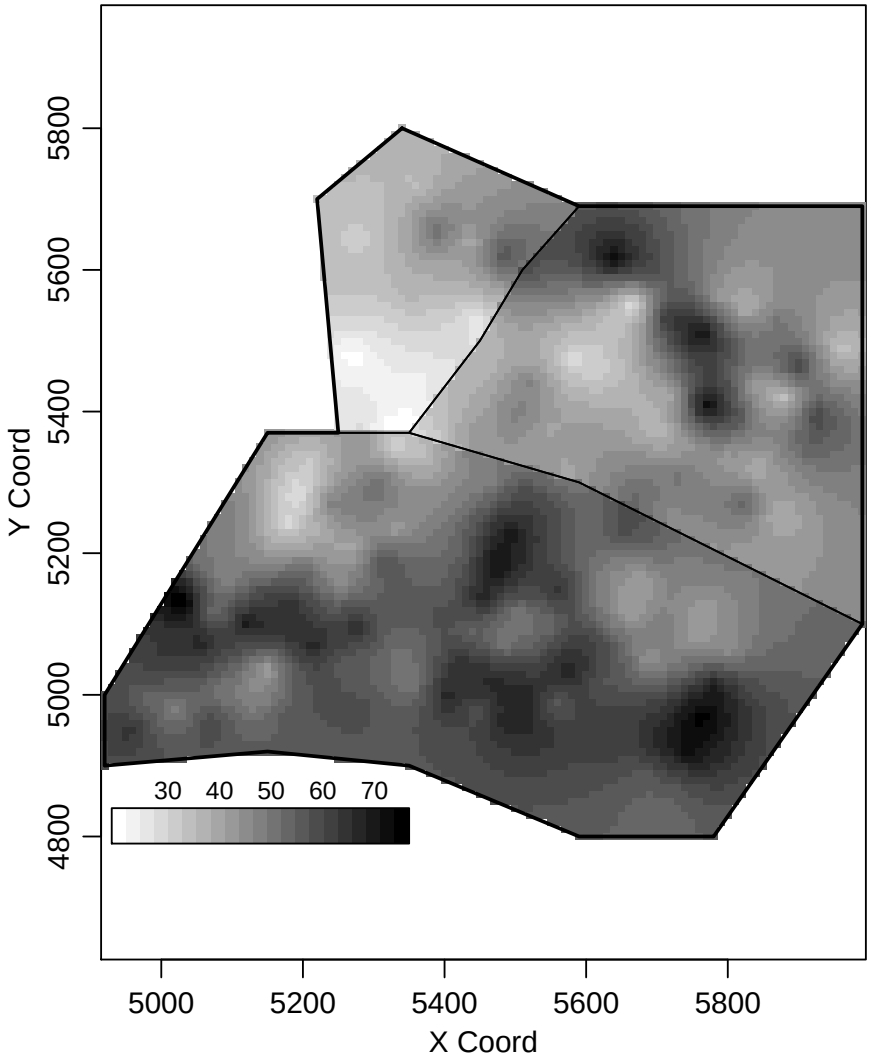
Spatial Prediction: Soil Calcium Data (1)

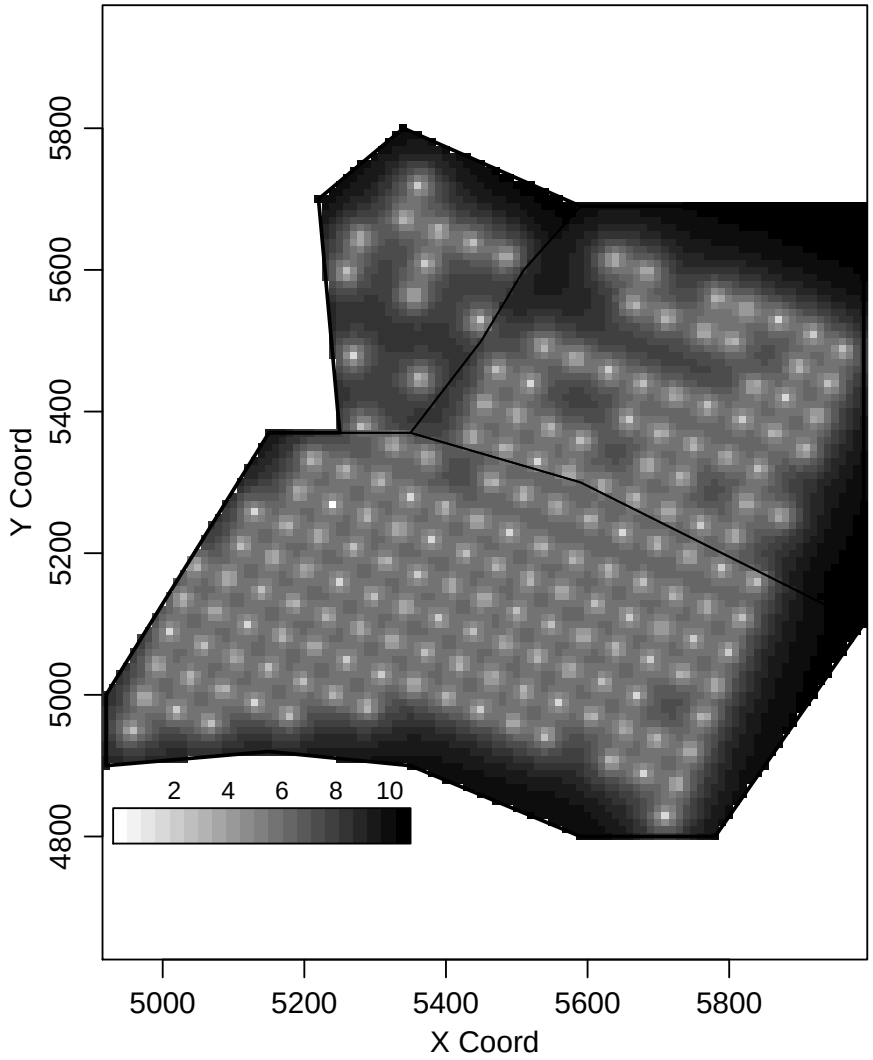
- We obtain a predictive map for calcium, computed at points across the whole study area with a spacing of 10 m.
- We use the model treating sub-areas as a factor with 3 levels.
- The maps on the next slide show the predictive map (left) and corresponding prediction standard errors (right).

Note that

- there are discontinuities of the map at the sub-area boundaries, as we allow for area-specific means.
- prediction standard errors are lowest at observation points.
- predictions and measured values at these points may differ.

Spatial Prediction: Soil Calcium Data (2)





R code for Spatial Prediction for Soil Calcium Data

```
### define 10 m grid, and select points within study area ###
gr <- pred_grid(ca20$borders,by=10)
gr0 <- polygrid(gr,borders=ca20$border,bound=T)

### build covariate vector for prediction locations indicating subarea ###
ind.reg <- numeric(nrow(gr0))
ind.reg[.geoR_inout(gr0,ca20$reg1)] <- 1
ind.reg[.geoR_inout(gr0,ca20$reg2)] <- 2
ind.reg[.geoR_inout(gr0,ca20$reg3)] <- 3
ind.reg <- as.factor(ind.reg)

### predict at prediction locations, accounting for subareas ###
KC <- krige.control(trend.d = ~area, trend.l = ~ind.reg, obj.model = m2)
ca20pred <- krige.conv(ca20, loc = gr, krige = KC)

### plot prediction map and associated prediction standard errors ###
par(mar = c(2.8, 3.1, 0.5, 0.5), mgp = c(1.8, 0.7, 0), mfrow = c(1, 2))
image(ca20pred, loc = gr, col = gray(seq(1, 0, l = 21)), x.leg = c(4930, 5350), y.leg = c(4790, 4840))
polygon(ca20$reg1)
polygon(ca20$reg2)
polygon(ca20$reg3)
image(ca20pred, loc = gr, val = sqrt(ca20pred$krige.var), col = gray(seq(1, 0, l = 21)),
   x.leg = c(4930, 5350), y.leg = c(4790, 4840))
polygon(ca20$reg1)
polygon(ca20$reg2)
polygon(ca20$reg3)
```

Overview

Overview

Recap: Generalized Linear Model

For independent outcomes $Y_1, \ldots, Y_n$ and corresponding vectors of explanatory variables $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the generalized linear model assumes

- each Y_i has a density f from the exponential family.
- $\mathrm{E}(Y_i) = h^{-1}(\mathbf{x}_i'\beta)$ for a known link function $h(\cdot)$.

Typical examples:

- linear regression: f normal density, g identity-link
- logistic regression: f binomial density, g logit-link
- Poisson regression: f Poisson density, g log-link

Generalized Linear Geostatistical Models (1)

The generalized linear geostatistical model consists of two parts:

1. A stationary Gaussian random field $Z(s)$ (spatial correlation)
2. Conditional on $Z(s)$, the $Y(s_1), \ldots, Y(s_n)$ are mutually independent, following a generalized linear model with

$$E(Y(s)|Z(s)) = \mu(s) = h^{-1}(Z(s) + \mathbf{x}(s)'\boldsymbol{\beta})$$

Examples:

- Geostatistical logistic model:

$$P(Y(s) = 1) = logit^{-1}(\mu(s)) = logit^{-1}(Z(s) + \mathbf{x}(s)'\boldsymbol{\beta})$$

- Geostatistical log-linear Poisson model:

$$E(Y(s)) = exp(\mu(s)) = exp(Z(s) + \mathbf{x}(s)'\boldsymbol{\beta})$$

Generalized Linear Geostatistical Models (2)

- Conditionally on $\mathbf{Z}$, the $Y(s_i)$ are independent, and the conditional likelihood is thus the product

$$L(\theta|\mathbf{Z}) = \prod_{i=1}^{n} f_i(y_i|\mathbf{Z}, \theta).$$

- In contrast to the Gaussian geostatistical model, there is generally no closed-formed expression for the marginal likelihood, which involves integrating over $\mathbf{Z}$ – an n-dimensional integral!
- Two possible approaches are Monte Carlo maximum likelihood and Bayesian inference using MCMC. The latter has the additional advantage of avoiding ignoring uncertainty in the plug-in prediction discussed before.

Further R packages

- `geoRglm`: Extension of `geoR` to generalized linear geostatistical models.
  - `likfit.glsm` performs Monte Carlo maximum likelihood based on a Monte Carlo sample from the conditional distribution.
  - `binom.krige` and `pois.krige` perform conditional simulation (by MCMC) and spatial prediction in the logistic / Poisson geostatistical model for fixed covariance parameters.
  - `binom.krige.bayes` and `pois.krige.bayes` perform posterior simulation (by MCMC) and spatial prediction in the logistic / Poisson geostatistical model.
- `gstat`: Similar functionality to `geoR`. See `www.gstat.org`.

Further Reading

- Bivand, R. S., Pebesma, E. J., and Gómez-Rubio, V. (2008). Applied Spatial Data Analysis with R. Springer.
  Book's website: `http://www.asdar-book.org/`.
  R-packages: `sp`.
- Diggle, P. J. and Ribeiro, P. J. (2007). Model-based Geostatistics. Springer.
  As ebook: `http://ebooks.ub.uni-muenchen.de/8793/`.
  Book's website: `http://www.leg.ufpr.br/mbgbook/`.
  R-packages: `geoR` and `geoRglm`.

Overview

Introduction

In many settings, phenomena of interest and corresponding data are both spatial as well as temporal in nature. Examples:

- air pollution concentration monitoring in different cities over time
- monitoring of disease cases at several locations over time
- ...

Modeling of such data has to account for both spatial and temporal correlation, and can use both types of information to obtain parameter estimates or predictive maps.

Geostatistical Space-Time Models

We are now dealing with a spatio-temporal process $Y(s,t)$.

In principle, a modeling approach can follow a similar strategy as in the spatial case, with a model consisting of two parts:

1. A Gaussian spatio-temporal random process $Z(s,t)$ (accounting for spatial and temporal correlation)
2. Conditional on $Z(s,t)$, the $Y(s_1,t_1),\ldots,Y(s_n,t_n)$ are mutually independent, following a generalized linear model with

$$E(Y(s,t)|Z(s,t)) = \mu(s,t) = h^{-1}(Z(s,t) + \mathbf{x}(s,t)'\boldsymbol{\beta})$$

Covariances in Space and Time

The covariance between $Z(s_1, t_1)$ and $Z(s_1, t_2)$ depends on both space (s_1 and s_2) and time (t_1 and t_2). Typically, we need additional simplifying assumptions in practice. The covariance function of a random field Z is called

- separable if it is the product of spatial and temporal covariance functions

$$Cov\{Z(s_1, t_1), Z(s_2, t_2)\} = \gamma_S(s_1, s_2)\gamma_T(t_1, t_2).$$

- stationary if it depends on s_1 and s_2 through $s_1 - s_2$ and on t_1 and t_2 through $t_1 - t_2$, i.e.

$$Cov\{Z(s_1, t_1), Z(s_2, t_2)\} = \gamma(s_1 - s_2, t_1 - t_2).$$

- isotropic in space, if additionally

$$Cov\{Z(s_1, t_1), Z(s_2, t_2)\} = \gamma(||s_1 - s_2||, |t_1 - t_2|).$$

Inference for Geostatistical Space-Time Models

For an isotropic stationary separable random field Z, we thus have

$$Cov\{Z(s_1, t_1), Z(s_2, t_2)\} = \gamma_S(||s_1 - s_2||)\gamma_T(|t_1 - t_2|).$$

These assumptions are convenient, but often questionable in practice, and tests for the assumptions exist.

An empirical spatio-temporal semivariogram can be computed analogous to the spatial semivariogram, but is now a two-dimensional function of distances in both space and time.

Inference for geostatistical space-time models is less developed than for spatial models, due to the larger model-complexity, often large sizes of the data sets, and less existing software. For more information, see e.g. Schabenberger and Gotway (2005) or Finkenstädt et al (2007).

Other Spatio-Temporal Data

Spatio-temporal extensions also exist for the other two types of spatial data discussed.

For point processes, see P. Diggle in Finkenstädt et al (2007). Examples:

- The U.K. 2001 epidemic of foot-and-mouth disease: Occurrence of cases in farms across England over time.
- Spatio-temporal distribution of gastroenteric disease in the county of Hampshire over the years 2001-2002.

For areal data, see Banerjee, Carlin and Gelfand (2004). Examples:

- Lung cancer mortality in Ohio counties in 1968-1988.
- Usage of cancer screening in Bavarian counties 2006-2008.

Further Reading

- Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004). Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall / CRC.
- Finkenstädt, B., Held, L. and Isham, V. (Eds.) (2007). Statistical Methods for Spatio-Temporal Systems. Chapman & Hall / CRC.
- Schabenberger, O. and Gotway, C. A. (2005) Statistical Methods for Spatial Data Analysis. Chapman & Hall / CRC.