Introduction to the Spatial Relationships Between Two Variables

Ronny Vallejos

Departamento de Matemática
Universidad Técnica Federico Santa María
Valparaíso, Chile
ronny.vallejos@usm.cl

Fundación Konrad Lorenz, Bogotá, Colombia
November, 2018

Outline

- Day 1
  1. Introduction and Motivation.
  2. Types of Problems to be Consider in This Shortcourse
  3. A Simple Solution
  4. Preliminaries and Notation (Spatial Statistics context)
- Day 2
  1. Introduction to R
  2. Uploading Images in R
  3. R Examples in a Spatial Context
  4. The Cross-Variogram ans its Properties
- Day 3
  4. The modified t test
  5. Correlation Between one Process and Several Others
  6. The Effective Sample Size

Outline

- Day 4
  1. The Codispersion Coefficient
  2. Properties and Asymptotic Results
  3. Hypothesis Testing
  4. Applications
- Day 5
  1. The Effective Sample Size Revisited
  2. Properties and Asymptotic Results
  3. Hypothesis Testing
  4. Applications

Material of the Course

The material contained in this short course and more can be found in:

- Vallejos, R., Osorio, F., Bevilacqua, M. (2019).
  Spatial Relationships Between Two Georeferenced Variables: With Applications in R.
  To appear in Springer. Number of pages $\approx$ 200.
- Book homepage `http://srb2gv.mat.utfsm.cl`
- Personal webpage `http://rvallejos.mat.utfsm.cl`
- e-mail: `ronny.vallejos@usm.cl`

1. Introduction and Motivation

- Consider two spatial processes $\{X(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ and $\{Y(\boldsymbol{s}) : \boldsymbol{s} \in D\}$, where $D \subset \mathbb{R}^2$.

- The available information is the observations $X(\boldsymbol{s}_1), \ldots, X(\boldsymbol{s}_n)$ and $Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)$.

- How to quantify the linear correlation between $X(\cdot)$ and $Y(\cdot)$ taking into account the georeferencing information?

- Is the Pearson's correlation coefficient enough in a spatial context?

Some Examples

Example 1: The Pinus Radiata Dataset

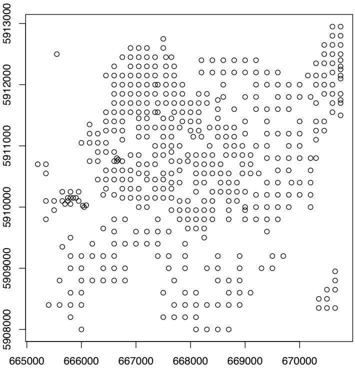


Figure: Locations where the samples were taken.

Example 1: The Pinus Radiata Dataset

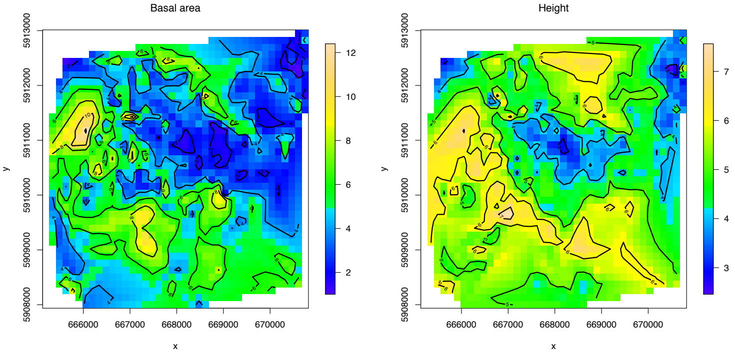


Figure: (Left) Bilinear interpolation of the three basal areas; (Right) Bilinear interpolation of the three heights.

The Pinus Radiata Dataset

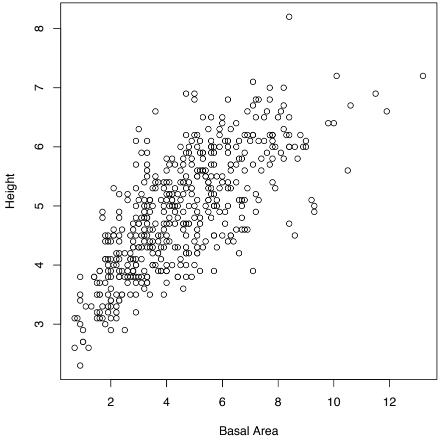


Figure: Height versus Basal Area (468 observations), $\rho = 0.7021$.

Example 2: The Murray Smelter Site Dataset

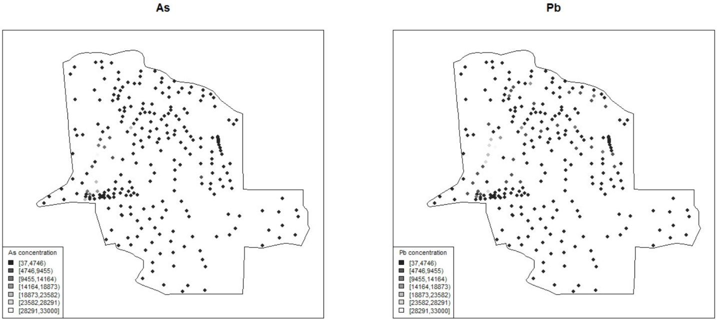


Figure: Locations of 253 geocoded aggregated surface soil samples collected in a 0.5 square mile area in Murray, Utah and their measured concentrations of As and Pb, $\rho = 0.5893$.

Example 3: Similarity Between Images

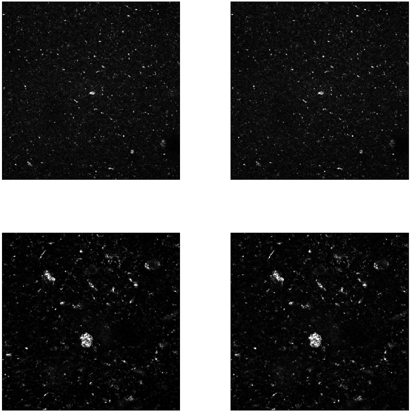


Figure: Dispersion of nanotubes. Images taken at NIST, USA. The images of the top were taken at the same distance. The images of the bottom were taken at the same distance but closer.

Example 4: Directional Contamination

Figure: (Left) Original image (Lenna); (Right) Image transformed into the direction $\boldsymbol{h} = (1, 1)$.

Philosophy of the Course

Throughout, I have attempted to follow one basic principle: never give an estimator without giving a confidence set.

Larry Wasserman
Department of Statistics
Carnigie Melon University

Easy Solution

The t Test

- Arrange the observations of each variable in a column. Under normality assume that cor$[X(\boldsymbol{s}), Y(\boldsymbol{s})] = \rho$ and consider the hypothesis testing problem:

$$\mathrm{H}_0 : \rho = 0 \text{ versus } \mathrm{H}_1 : \rho \neq 0.$$

- Use the statistic

$$t = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}},$$

where r is an estimator of ρ.

- Compute the p-value associated with the test.

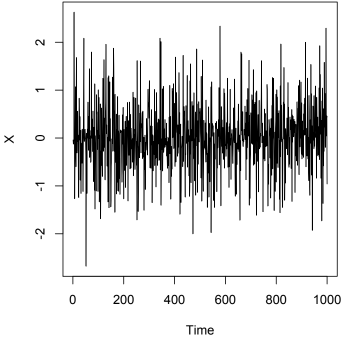
- For a fixed level α we reject H_0 if $p < \alpha$, otherwise H_0 is not rejected.

2. Preliminaries and Notation

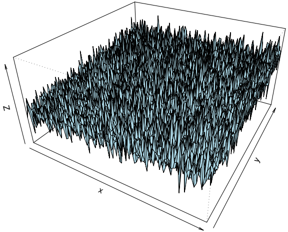
Definition

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space, and let D be an arbitrary index set. A spatial process is a function $X : (\Omega, \mathcal{F}, \mathcal{P}) \times D \longrightarrow \mathbb{R}$, such that for all $\boldsymbol{s} \in D$, $X(\boldsymbol{s})$ is a random variable. Commonly $D \subset \mathbb{R}^d$.

$X(s) = A\cos(\eta s + \phi)$, where A is a random variable independent of $\phi \sim \mathcal{U}(0, 2\pi)$, and η is a fixed constant. For the particular case when $A \sim \mathcal{N}(0, 1)$ and $\eta = 1$, 1000 observations were generated.



Realizations of a Spatial Process



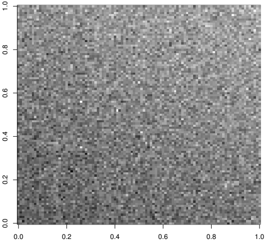


Figure: A realization from a spatial process defined on a finite grid of $D = \mathbb{Z}^2$. $Z(x, y) = \beta_1 x + \beta_2 y + \epsilon(x, y)$, where $\{\epsilon(x, y)\}$ is a collection of independent and identically distributed random variables with zero mean and variance σ^2.

Stationary Processes

- A condition that guaranties the existence of the mean, variance and covariance functions of a spatial process is the second order property:

$$\mathbb{E}[X^2(\boldsymbol{s})] < \infty, \text{ for all } \boldsymbol{s} \in D.$$

Theorem

Let X be a second order spatial process . Then

- $\mathbb{E}[X(s)] < \infty$, para todo $s \in D$.
- var$[X(\boldsymbol{s})] < \infty$, para todo $s \in D$.
- cov$[X(\boldsymbol{s}), X(\boldsymbol{t})] < \infty$, para todo $\boldsymbol{s}, \boldsymbol{t} \in D$.

- A second oreder spatial process is said weekly stationary if
  - i) $\mathbb{E}[X(\boldsymbol{s})] = \mu, \text{ for all } \boldsymbol{s} \in D$.
  - ii) $\mathrm{cov}[X(\boldsymbol{s}_i), X(\boldsymbol{s}_j)] = g(\boldsymbol{s}_i - \boldsymbol{s}_j) = C(\boldsymbol{h})$, for all $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$ and for some function g. An inmediate consequence is that var$[X(\boldsymbol{s})] = g(\boldsymbol{0})$, constant with respect to $\boldsymbol{s}$.

Covariance Functions

The covariance function of a stationary process is denoted as

$$C(\boldsymbol{h}) = \mathrm{cov}[X(\boldsymbol{s}), X(\boldsymbol{s}+\boldsymbol{h})], \;\; \boldsymbol{s}, \boldsymbol{h} \in D.$$

- The covariance function of a weakly stationary process must satisfy the positive-definiteness condition, i.e.,

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

for all $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$ and for all $a_i, a_j \in \mathbb{R}$.

- The corresponding covariance matrix of the vector $(X(\boldsymbol{s}_1), \ldots X(\boldsymbol{s}_n))^\top$ denoted as $\boldsymbol{\Sigma}$, satisfies that its ij-th element is

$$\boldsymbol{\Sigma}_{ij} = C(\boldsymbol{s}_i - \boldsymbol{s}_j).$$

Covariance Functions

Proposition

Let C be the covariance function of a stationary process. Then

- $C(\boldsymbol{h}) = C(-\boldsymbol{h}), \forall \boldsymbol{h} \in D$. (even function)
- $|C(\boldsymbol{h})| \leq C(\boldsymbol{0}), \forall \boldsymbol{h} \in D$. (bounded function)
- If C is a valid covariance function in $\mathbb{R}^d$, then C is valid in $\mathbb{R}^k$ for all integer $k < d$.
- If $C_j(\boldsymbol{h})$ are valid covarince functions, $j = 1, \ldots, k$, then $\sum_{j=1}^k b_j C(\boldsymbol{h})$ is a valid covarince function, if $b_j \geq 0$, for all j.
- If $C_j(\boldsymbol{h})$ are valid covarince functions, $j = 1, \ldots, k$, then $\prod_{j=1}^k$ it is also a valid covariance function.

Strongly Stationary Processes

Definition

A spatial process X is said strictly stationary if $\forall k \in \mathbb{N}$, $\forall(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_k) \in D^k$, and $\forall \boldsymbol{h} \in D$, the distribution of the vector $(X(\boldsymbol{s}_1 + \boldsymbol{h}), \ldots, X(\boldsymbol{s}_k + \boldsymbol{h}))^\top$ is independent of $\boldsymbol{h}$.

Theorem

Let X be a second order process.
X *strictly stationary* $\Rightarrow$ X *weakly stationary*

Definition

A spatial process X is Gaussian if, $\forall k \in \mathbb{N}$, the vector $(X(\boldsymbol{s}_1), \ldots, X(\boldsymbol{s}_k))^\top$ has a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We will use the notation $(X(\boldsymbol{s}_1), \ldots, X(\boldsymbol{s}_k))^\top \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Theorem

X *strictly stationary* $\iff$ X *weakly stationary and Gaussian*

Intrinsically Stationary Processes

Definition

A spatial process $\{X(\boldsymbol{s}) : \boldsymbol{s} \in D\}$ is called intrinsically stationary if for all $s \in D$, $\mathbb{E}[X(\boldsymbol{s})] = \mu$, and

$$\text{var}[X(\boldsymbol{s}+\boldsymbol{h}) - X(\boldsymbol{s})] = 2\gamma(\boldsymbol{h})$$

is only a function of $\boldsymbol{h}$. In such a case, the function $2\gamma(\boldsymbol{h})$ is called a variogram of the process and $\gamma(\boldsymbol{h})$ is called a semivariogram.

- If $\mathbb{E}[X(\boldsymbol{s}+\boldsymbol{h}) - X(\boldsymbol{s})] = 0$, then $\text{var}[X(\boldsymbol{s}+\boldsymbol{h}) - X(\boldsymbol{s})] = 2\gamma(\boldsymbol{h}) = \mathbb{E}[X(\boldsymbol{s}+\boldsymbol{h})) - X(\boldsymbol{s})]^2$.
- In practice, small variance is expected for points that are close in space, while large variance is expected for points with a large separation in space according to the first law of geography (Tobler, 1970).

Properties of the Semi-Variogram

The semi-variogram γ of an intrinsically stationary process X satisfies:

- $\gamma(\boldsymbol{h}) = \gamma(-\boldsymbol{h})$ (even function) and $\gamma(\mathbf{0}) = 0$;
- If A is a linear function defined on $\mathbb{R}^d$, then the function $h \longrightarrow \gamma(A\boldsymbol{h})$ it is also a semi-variogram.
- If γ is continous in $\mathbf{0}$, then γ is continous for all $\boldsymbol{h}$ γ is locally bounded.
- If γ is bounded ina neighbor of $\mathbf{0}$, then there exist $a, b \in \mathbb{R}$ such that for all $x \in D$, $\gamma(\boldsymbol{x}) \leq a||\boldsymbol{x}||^2 + b$.
- Conditionally Negative Definite. For all $n \in \mathbb{N}$, for all $\boldsymbol{a} \in \mathbb{R}^n$ such that $\sum_{i=1}^{n} a_i = 0$, and for all $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n) \in D^n$, $\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) \leq 0$.
- A continuous function γ sobre $\mathbb{R}^d$ such that $\gamma(\mathbf{0}) = 0$ is a semi-variogram iff for all $a > 0$, $\boldsymbol{h} \longrightarrow e^{-a\gamma(\boldsymbol{h})}$ is a covariance function.

Covariance and Semi-Variograms

- For weakly stationary processes, there is a relationship between $\gamma(\boldsymbol{h})$ and $C(\boldsymbol{h})$:

$$\begin{aligned} 2\gamma(\boldsymbol{h}) &= \mathrm{var}[X(\boldsymbol{s}+\boldsymbol{h}) - X(\boldsymbol{s})] \\ &= \mathrm{var}[X(\boldsymbol{s}+\boldsymbol{h})] + \mathrm{var}[X(\boldsymbol{s})] - 2\mathrm{cov}[X(\boldsymbol{s}+\boldsymbol{h}), X(\boldsymbol{s})] \\ &= 2(C(\boldsymbol{0}) - C(\boldsymbol{h})). \end{aligned}$$

Thus,

$$\gamma(\boldsymbol{h}) = C(\boldsymbol{0}) - C(\boldsymbol{h}).$$

Example

- Consider the AR(1) process $X(t) = \phi X(t-1) + \epsilon(t)$, where $|\phi| < 1$, and $\epsilon(t) \sim WN(0, \sigma^2)$.
- $C(h) = \sigma^2 \phi^h / (1-\phi^2)$, $C(0) = \sigma^2/(1-\phi^2)$.
- $\gamma(h) = \sigma^2(1-\phi^h)/(1-\phi^2)$.

Covariance and Semi-Variogram Plots

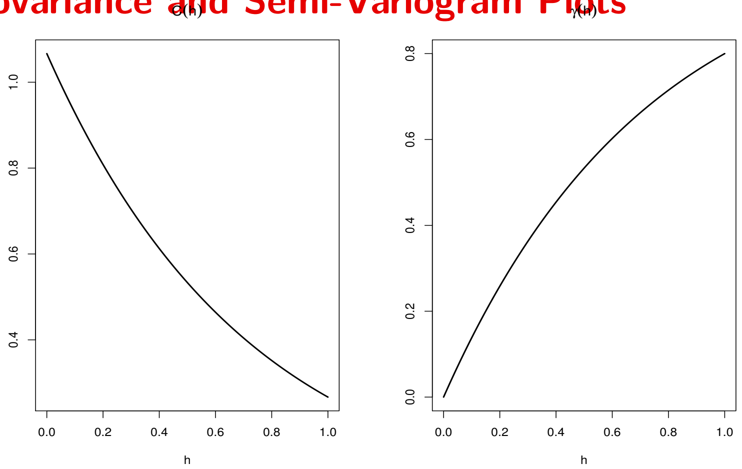


Figure: $C(h)$ and $\gamma(h)$ for an AR(1) model.

Isotropy

- If a variogram (semivariogram) depends on $||h||$, the process $\{X(\boldsymbol{s})\}$ is called isotropic; otherwise, the process $\{X(\boldsymbol{s})\}$ is called anisotropic.

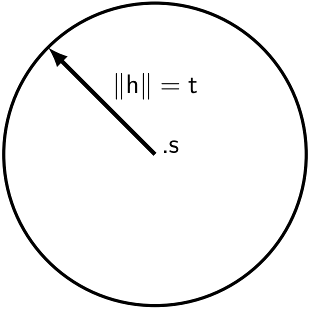


Figure: Circular (isotropic) correlation structure.

Typical Semi-Variogram Plot

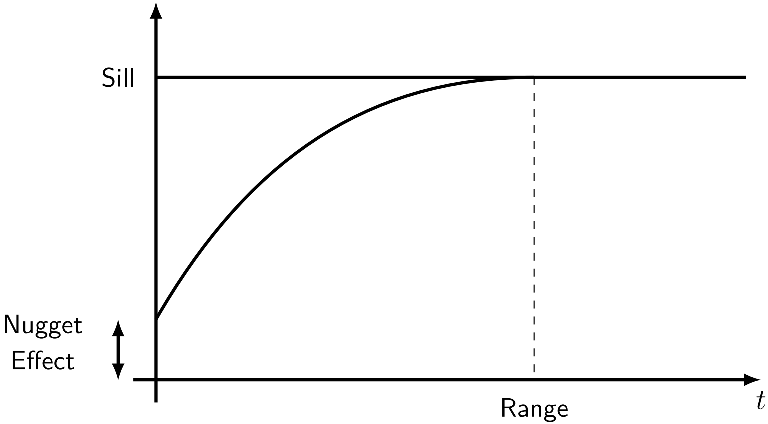


Figure: Behavior of a typical semivariogram model. $||\boldsymbol{h}|| = t$

Parametric Covariance Models

- There are several parametric covariance and semi-variogram models that are valid.
- The Matérn covariance family is described by

$$C(t) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\frac{t}{r}\right)^{\nu} \kappa_{\nu}\left(\frac{t}{r}\right),$$

  where κ_ν is a modified Bessel function of the second kind of order ν, Γ is the gamma function, $r > 0$ is the range, and $\nu > 0$ is the smoothness parameter.
- If $\nu = 1/2$, then $C(t) = \sigma^2 exp(-t/r)$.
- The covariance matrix of a spatial process with Matérn covariance function is positive definite.

Parametric Covariance Models

Model	Covariance Function $C(t)$
Lineal	$C(t)$ does not exist.
Spheric	$C(t) = \begin{cases} 0, & \text{if } t \geq 1/\phi, \\ \sigma^2 \left(1 - \frac{3}{2}\phi t + \frac{1}{2}(\phi t)^3\right), & \text{if } 0 < t \leq 1/\phi, \\ \tau^2 + \sigma^2, & \text{otherwise.} \end{cases}$
Exponential	$C(t) = \begin{cases} \sigma^2 \exp(-\phi t), & \text{if } t > 0, \\ \tau^2 + \sigma^2, & \text{otherwise.} \end{cases}$
Gaussian	$C(t) = \begin{cases} \sigma^2 \exp(-(\phi t)^2), & \text{if } t > 0, \\ \tau^2 + \sigma^2, & \text{otherwise.} \end{cases}$
Wave	$C(t) = \begin{cases} \sigma^2 \frac{\sin(\phi t)}{\phi t}, & \text{if } t > 0, \\ \tau^2 + \sigma^2, & \text{otherwise.} \end{cases}$
Matérn	$C(t) = \begin{cases} \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(2\sqrt{\nu}t\phi)^\nu K_\nu(2\sqrt{\nu}t\phi), & \text{if } t > 0, \\ \tau^2 + \sigma^2, & \text{otherwise.} \end{cases}$

Estimation of the Variogram

- For n sampling sites $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$, a natural and unbiased estimator of the semivariogram of a spatial process $\{X(\boldsymbol{s}) : \boldsymbol{s} \in D \subset \mathbb{R}^2\}$ is

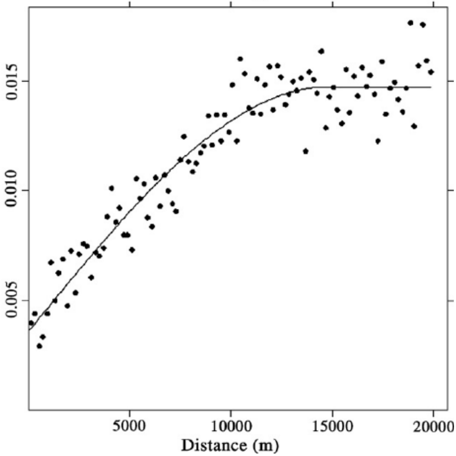
$$\widehat{\gamma}(\boldsymbol{h}) = \frac{1}{2|N(\boldsymbol{h})|} \sum_{N(\boldsymbol{h})} (X(\boldsymbol{s}_i) - X(\boldsymbol{s}_j))^2,$$

  where $N(\boldsymbol{h}) = \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : \boldsymbol{s}_i - \boldsymbol{s}_j = \boldsymbol{h}, 1 \leq i, j \leq n\}$, and $|\cdot|$ denotes the cardinality of a set.

- Robust estimators have been studied by Cressie (1980), Genton (1998) and García-Soidán (2004) among others.

- the computation of the semivariogram is available in the R packages `GeoR`, `RandomFields`, and `sgeostat`.

Estimation of the Variogram



Estimation

- Let us assume a parametric model for the semi-variogram of the form $\{\gamma(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p\}$. Commonly $\boldsymbol{\theta} = (\sigma^2, \tau, \phi)$.

- OLS estimation. The least squares estimator of $\boldsymbol{\theta}$ is:

$$\widehat{\boldsymbol{\theta}}_{OLS} = \text{argmin}_{\boldsymbol{\theta} \in \Theta} \sum_{k=1}^{K} (\widehat{\gamma}(\boldsymbol{h}_k) - \gamma(\boldsymbol{h}_k))^2$$

- GLS Estimator. The generalized least squares estimator of $\boldsymbol{\theta}$ is:

$$\boldsymbol{\theta}_{GLS} = \text{argmin}_{\boldsymbol{\theta} \in \Theta} (\widehat{\gamma}_K - \gamma_K(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}_{\widehat{\gamma}_K}^{-1}(\boldsymbol{\theta})(\widehat{\gamma}_K - \gamma_K(\boldsymbol{\theta})),$$

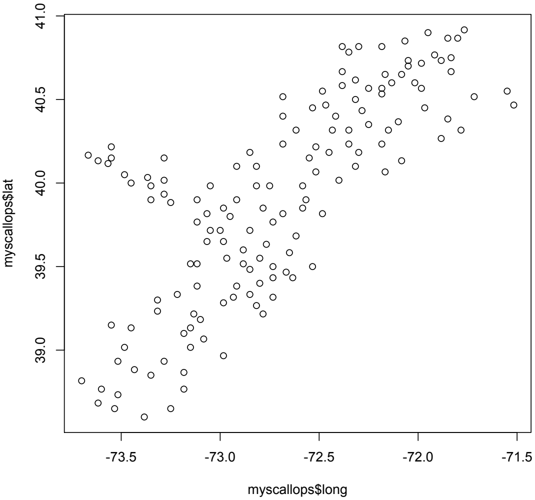
where $\widehat{\gamma}_K = (\widehat{\gamma}(\boldsymbol{h}_1), \ldots, \widehat{\gamma}(\boldsymbol{h})_K)^\top$,
$\gamma_K(\boldsymbol{\theta}) = (\gamma(\boldsymbol{h}_1, \boldsymbol{\theta}), \ldots, \gamma(\boldsymbol{h}_K, \boldsymbol{\theta}))^\top$ and $\Sigma_{\widehat{\gamma}_K}^{-1}(\boldsymbol{\theta})$ is the inverse of the covariance matrix of $\widehat{\gamma}_K(\boldsymbol{\theta})$.

An Example: Scallops Dataset

- Scallops (ostiones) is a dataset collected in the atlantic ocean in the north part of the United States.
- 148 observations (latitud, longitud, catches) were collected in total.
- The sampling procedure was carried out between 1990 and 1993.
- The dataset has been provided by the National Marine Fisheries Service of the USA.

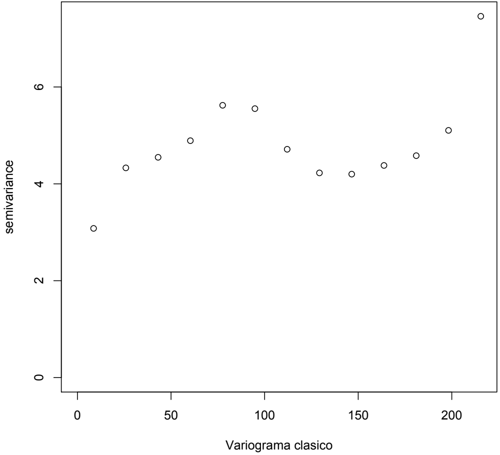
Banerjee et al. (2004).Hierarchical Modelling and Analysis for Spatial Data. Chapman & Hall/CRC, Boca Ratón.

An Example: Scallops



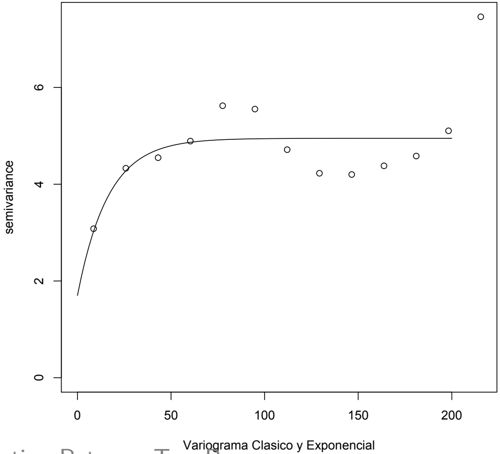
An Example: Scallops

Empirical Semi-Variogram



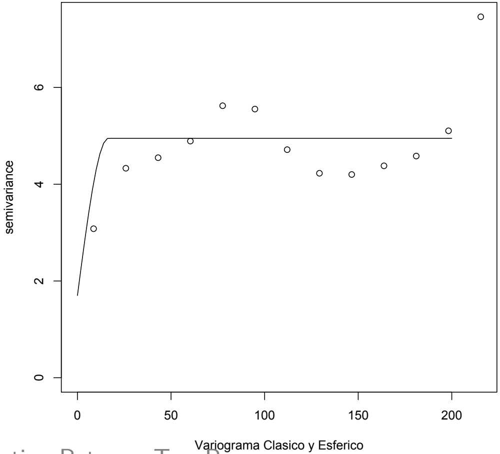
An Example: Scallops

Exponencial Semi-Variogram



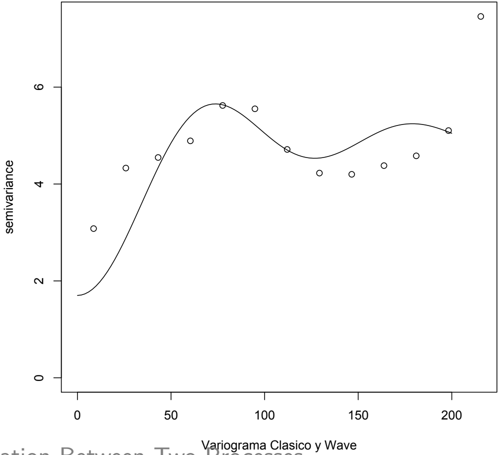
An Example: Scallops

Shperical Semi-Variogram



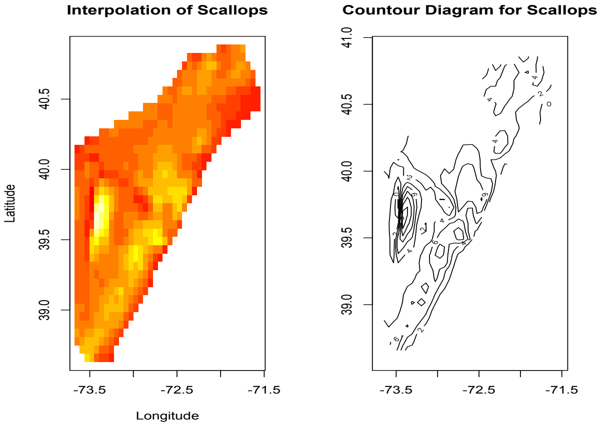
Un Ejemplo: Scallops

Wave Model: $\gamma(\boldsymbol{h}) = c \cdot \sin(a||\boldsymbol{h}||)/(a||\boldsymbol{h}||)$, for $||\boldsymbol{h}|| > 0$.



Semi-Variogram Applications

Kriging interpolation



Day 2