Session 4.1: Introduction to Geostatistics

Geospatial Analytics using R and R-INLA

MRC
Centre for Environment & Health

Medical Research Council

Imperial College London

AIMS African Institute for Mathematical Sciences RWANDA

Learning Objectives

At the end of this session you should be able to:

- know the common models used for **geostatistical data** analysis, i.e. Gaussian fields (GF);
- understand the assumptions of stationarity and isotropy;
- understand and compute the variogram/semivariogram.

The topics covered in this lecture can be further investigated in the books:

- Section 6.4 of the book **Spatial and Spatio-Temporal Bayesian models with R-INLA**
- Chapter 12 of the book **Spatial Statistics for Data Science: Theory and Practice with R** https://www.paulamoraga.com/book-spatial/index.html
- Chapter 12, Sections 12.1-12.3 of the book **Spatial Data Science** https://r-spatial.org/book/part-1.html
- Chapter 2, Section 2.1 of the book **Hierarchical Modeling and Analysis for Spatial Data**

Outline

1. Introduction to spatial modeling for geostatistical data (based on GF)
2. Assumptions of stationarity and isotropy
3. The variogram and semivariogram

Introduction to spatial modeling for geostatistical data (based on GF)

Geostatistical data

Definition | Example

- The difference between models for **geostatistical** (or point referenced) data and the spatial models presented in the previous lectures is that here we treat space as continuous, not discretised (areas).
- We are concerned here with spatial data structures where the process of interest is a spatial field

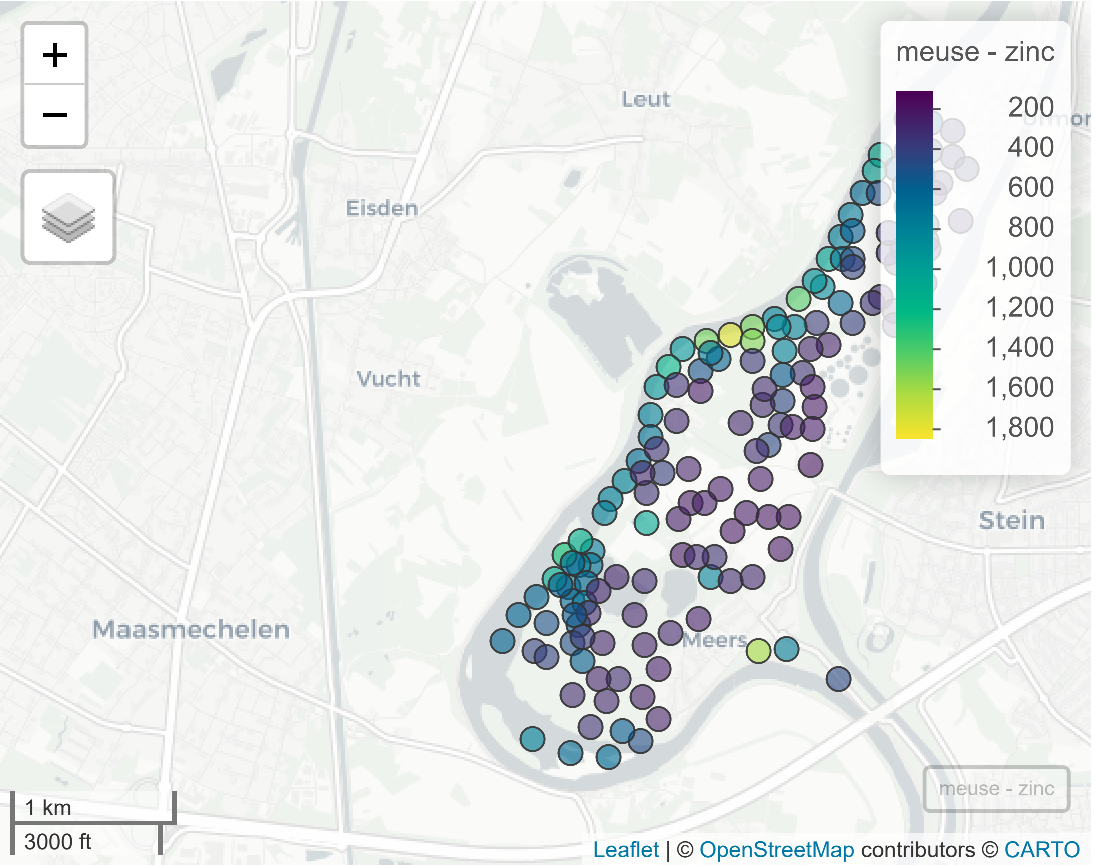
$$\{Z(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$$

  i.e. real values stochastic process characterized by a spatial index $\boldsymbol{s}$ which varies **continuously** in the fixed domain $\mathcal{D}$.

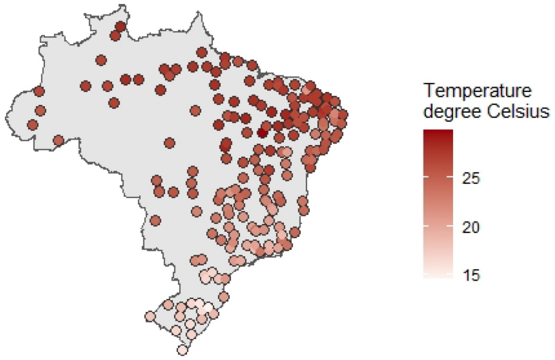
- Data are measured (possibly with error) at n spatial locations $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$ and are denoted by $\boldsymbol{Z} = (Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n)) = (Z_1, \ldots, Z_n)$.

Definition | Example

- Examples:
  - in environmental science: rainfall, air pollution concentrations, radioactive emission in soil, etc.
  - in epidemiology: prevalence of a disease measured at different villages distributed over a region of interest.
  - in ecology: density of mosquitoes responsible for disease transmission measured using traps placed at different locations.



- To **reconstruct the spatial field** from a finite set of **noisy** observations taken at a finite number of spatial locations.
- To use the spatial dependence to **predict values** of the spatial field (together with associated uncertainty) at locations where there are no observations.
- Example below: Each point represents a weather station in Brazil, and we have an associated $Z(s)$, which is air temperature.
- Using geostatistical methods we can reconstruct a latent spatial field from the finite set of observations taken at a finite number of spatial locations.



Observed

Predicted

Gaussian fields

- The common methodological framework to geostatistical models is that of **Gaussian fields (or processes)**, which are based on the multivariate Normal distribution.

- A spatial process $Z(\boldsymbol{s})$ is a **Gaussian field** (GF) if for any $n \geq 1$ and for each set of locations $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n)$, the vector $(Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n))$ follows a multivariate Normal distribution with mean $\boldsymbol{\mu} = (\mu(\boldsymbol{s}_1), \ldots, \mu(\boldsymbol{s}_n))$ and spatially structured covariance $\boldsymbol{\Sigma}$.

- The generic element of $\boldsymbol{\Sigma}$ is defined by a **spatial covariance function** (sometimes called a kernel) $\mathcal{C}(\cdot, \cdot)$ such that $\Sigma_{ij} = \mathrm{Cov}\left(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)\right) = \mathcal{C}(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j))$.

Stationarity and Isotropy

Stationarity

- An important concept is geostatistical data analysis is given by stationary, that refers to the stability (i.e. equilibrium) of the statistical properties of spatial process.
- In simple terms, stationarity means that the random field (i.e. spatial process) looks similar in different parts of the domain.
- We are going to explore different degrees of stationarity:
  – Strict (or strong) stationarity
  – Weak (or second-order) stationarity
  – Intrinsic stationarity

Strict (or strong) stationarity

- The spatial process (or random process) is called **strict (or strong) stationary** if it is invariant under translation of the coordinates (i.e. invariant to shifts in space).

- In particular, for any set of locations $\mathbf{s}_i$, $i = 1, \ldots, N$ and any displacement, $\mathbf{h} \in \mathbb{R}^2$, the distribution of $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_N)\}$ is the same as that of $\{Z(\mathbf{s}_1 + \mathbf{h}), \ldots, Z(\mathbf{s}_N + \mathbf{h})\}$.

- A strictly stationary random field repeats itself throughout the domain.

Note that here $\mathbf{h}$ refers to the spatial separation of the locations (i.e. the lag-vector: $\mathbf{s}_i - \mathbf{s}_j$).

Weak (or second-order) stationarity and Isotropy

Weak stationarity or second order stationarity

- It imposes conditions only on the mean and covariance, which are translation invariant.
- The spatial process is called **weak (or second-order) stationary** if
  - $\boldsymbol{\mu}$ is constant, i.e., $\mu(\boldsymbol{s}_i) = \mu$ for each i
  - the spatial covariance function depends on length and orientation of the vector $\boldsymbol{h}$ linking two points $\boldsymbol{s}_i$ and $\boldsymbol{s}_j = \boldsymbol{s}_i + \boldsymbol{h}$, i.e. $\mathrm{Cov}\left(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)\right) = \mathcal{C}(\boldsymbol{s}_i - \boldsymbol{s}_j) = \mathcal{C}(\boldsymbol{h})$

Weak (or second-order) stationarity and Isotropy

Weak stationarity or second order stationarity

- It imposes conditions only on the mean and covariance, which are translation invariant.
- The spatial process is called **weak (or second-order) stationary** if
  - $\boldsymbol{\mu}$ is constant, i.e., $\mu(\boldsymbol{s}_i) = \mu$ for each i
  - the spatial covariance function depends on length and orientation of the vector $\boldsymbol{h}$ linking two points $\boldsymbol{s}_i$ and $\boldsymbol{s}_j = \boldsymbol{s}_i + \boldsymbol{h}$, i.e. $\text{Cov}\left(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)\right) = \mathcal{C}(\boldsymbol{s}_i - \boldsymbol{s}_j) = \mathcal{C}(\boldsymbol{h})$

Isotropy

- A second-order stationary spatial process is **isotropic** if the covariance does not depend on the direction but just on the Euclidean distance $||\boldsymbol{s}_i - \boldsymbol{s}_j||$ (where $|| \cdot ||$ denotes the Euclidean norm, i.e. Euclidean distance)
- In other words, the covariance depends only on the distance between two points irrespective of the geographical direction (north-south or east-west) of one from the other
- When covariance functions exhibit different behavior in different directions, the random fields are called anisotropic

Intrinsic stationarity

- Intrinsic stationarity is a relaxed form of stationarity, which is based on the difference in the process between locations.
- For a choice of two locations, we assume that:
  – the difference in means will be zero (i.e. constant mean assumption)
  – the variance of increments depends only on $\boldsymbol{h}$:

$$Var(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})) = 2\gamma(\mathbf{h})$$

- The function $2\gamma(\mathbf{h})$ is called Variogram and $\gamma(\mathbf{h})$ is called Semivariogram

- The gamma symbol, γ, is the standard symbol for variability in a variogram. Commonly, the terms variogram and semivariogram are used interchangeably, although the semivariogram is half of the variogram (we adopt this common terminology here).

Variogram

- The (semi)variogram is a useful devices in exploratory data analysis for geostatistical data, as it summarizes the strength of association as function of the distance (and in case of anisotropy, direction).

- The variogram analysis is often associated with kriging, a widely used interpolation method for geostatistical data.

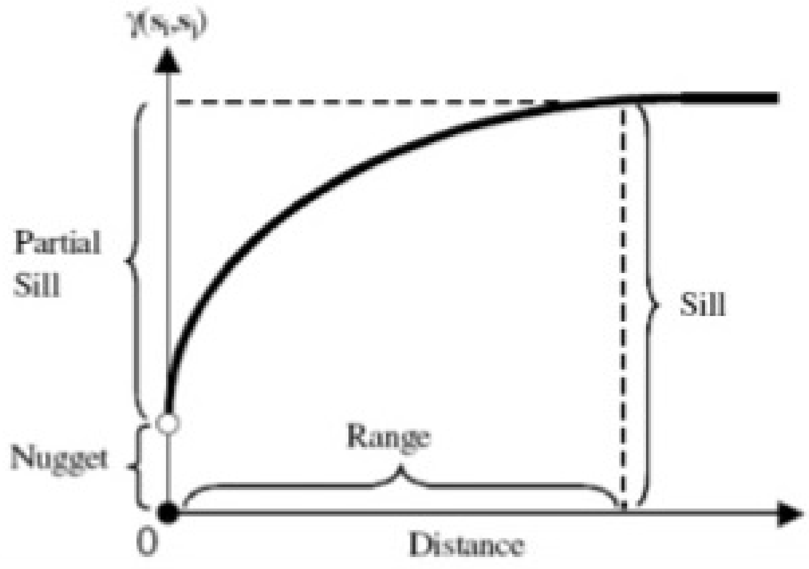
- The variogram is generally estimated by the experimental (or empirical) variogram, $\hat{\gamma}(\mathbf{h})$, which measures the similarity of values as a function of the distance between their locations:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i}^{N(\mathbf{h})} (Z(\mathbf{s}_i + \mathbf{h}) - Z(\mathbf{s}_i))^2$$

where $\mathbf{h}$ is the distance class and $N(\mathbf{h})$ is the set of pairs of points; therefore the sum is taken over all sample points separated by a lag vector of magnitude $\mathbf{h}$.

Characteristics of the variogram

- Sill: the value at which the variogram levels off (corresponds to the overall variability of the data)
- Range: the distance at which the variogram reaches the sill (corresponds to the distance at which observations become independent)
- Nugget: The nugget represents the combination of sampling error and short-range variability that causes two samples apparently taken from the same location to have different values.



Note that the partial sill is the sill minus the nugget

Some isotropic variogram functions [1]

- After obtaining the empirical variogram, we fit a model to it (i.e. theoretical variogram) to get a smooth fit. Popular variogram functions are:

Linear

$$\gamma(h) = \begin{cases} t^2 + \sigma^2 h & \text{if } h > 0 \\ 0 & \text{if } h = 0 \end{cases}$$

Spherical

$$\gamma(h) = \begin{cases} t^2 + \sigma^2 & \text{if } h \geq 1/\phi \\ t^2 + \sigma^2[\frac{3}{2}\phi h - \frac{1}{2}(\phi h)^3] & \text{if } 0 < h \geq 1/\phi \\ 0 & \text{if } h = 0 \end{cases}$$

Exponential

$$\gamma(h) = \begin{cases} t^2 + \sigma^2(1 - \exp(-\phi h)) & \text{if } h > 0 \\ 0 & \text{if } h = 0 \end{cases}$$

For details, see Banerjee, Carlin, and Gelfand (2014), Sections 2.1.2 - 2.1.4

Some isotropic variogram functions [2]

- t^2 is the nugget
- $t^2 + \sigma^2$ is the sill, therefore σ^2 is the partial sill
- the value $1/\phi$ is the range, while ϕ is called the decay parameter

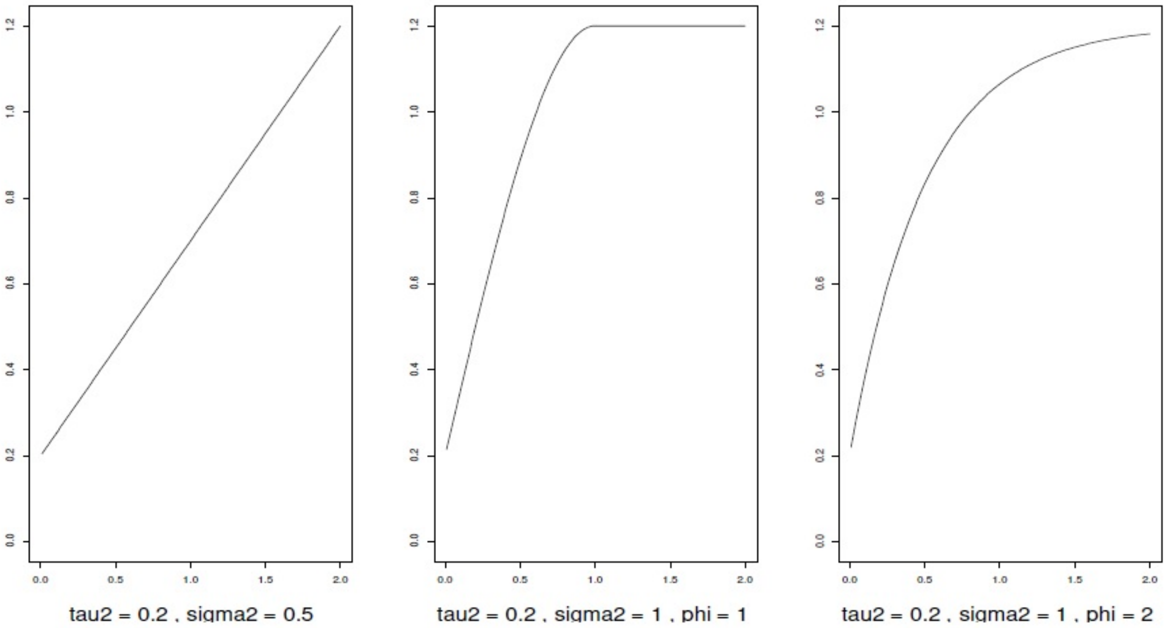
- For the exponential model, strictly speaking, the range is infinite. In this case the notation of effective range, h_0, is often used.
- It is the distance at which the correlation is negligible, dropping to 0.05. Setting:

$$
\begin{aligned}
\exp(-h_0\phi) &= 0.05 \\
\implies h_0 &= -\log(0.05)/\phi \\
\implies h_0 &\approx 3/\phi
\end{aligned}
$$

since $\log(0.05) \approx -3$.

Some isotropic variogram functions [3]

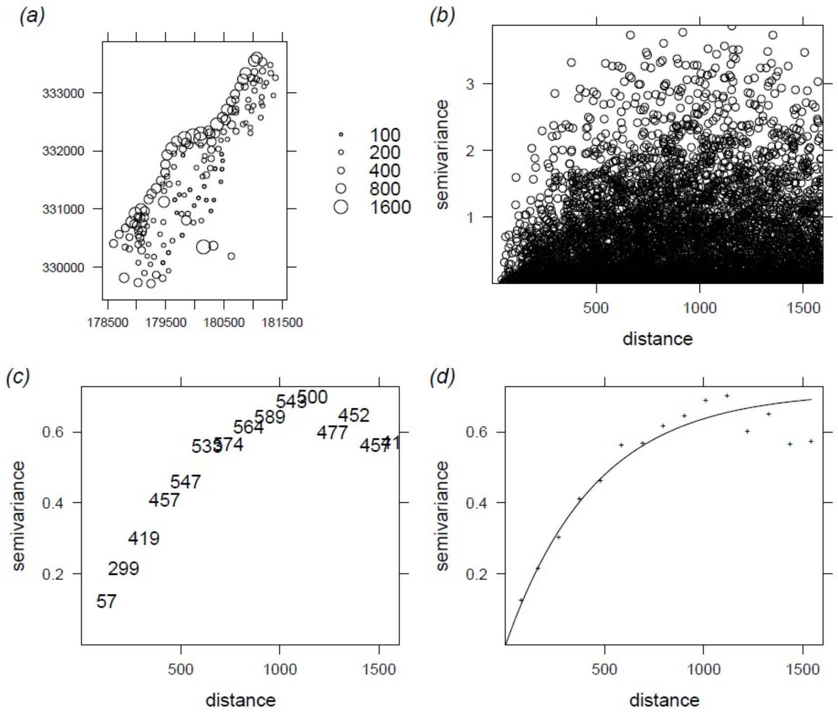
Theoretical variograms for three models, respectively linear, spherical, and exponential



Source: Banerjee, Carlin, and Gelfand (2014)

Steps of the variogram analysis

We use here the Meuse data set, which includes four heavy metals measured in the top soil in a flood plain along the river Meuse (it runs from France to the Netherlands)



Source: Hengl (2009), https://edepot.wur.nl/485517

- Steps: (a) sampling locations and measured values of the target variable, (b) (semi)variogram cloud showing semivariances for all pairs (log-transformed variable), (c) semivariances aggregated to lags of about 100 m, and (d) the final (semi)variogram model fitted using `gstat`.

Basic code for the computation of the variogram in R [1]

To demonstrate the computation of the variogram in R, we use the Meuse data set. Among the metals, we work with zinc

```
> library(sp)
> library(gstat)
>
> # We use Meuse dataset, which includes concentrations of zinc
> # measured at 155 sampling sites within the Meuse River plain
> data(meuse)
>
> # Transform the dataframe into a SpatialPointDataFrame
> coordinates(meuse) = ~x+y # the function coordinates
>                           # promotes the data.frame meuse
>                           # into a SpatialPointsDataFrame
>
> # Bubble plot
> bubble(meuse, "zinc", col=c("#00ff0088", "#00ff0088"),
+        main = "zinc concentrations (ppm)")
>
> hist(meuse$zinc) # we see a strong right skew in the data, so we log-transform them
>
> # Lagged scatter plot
> hscat(log(zinc)~1, meuse,(0:9)*100) # the correlation is quite strong when the lag
>                                     # is between 100 meters, then decrease with distance
```

Basic code for the computation of the variogram in R [2]

```
> # Construct the variogram
> meuse.vgm = variogram(log(zinc)~1, meuse) # we assume a constant trend for
>                                            # the variable log(zinc)
>
> # Plot the experimental variogram
> plot(meuse.vgm)
> plot(meuse.vgm, plot.numbers = TRUE, pch = "+") # The numbers of points in the
>                               # lag group used to compute the corresponding value of gamma(h)
>
> # Fit a variogram model
> model.1 = fit.variogram(meuse.vgm, vgm("Sph"))
> plot(meuse.vgm, model=model.1)
>
> # Look at the result of the fit
> model.1
>
> # We can also specify a set of models. In this case the best fitting is returned
> model.2 = fit.variogram(meuse.vgm, vgm(c("Exp", "Sph")))
> model.2 # here the spherical model with nugget=0.051, partial sill =0.591 and range=897 is chosen
>
> # Specify theoretical variogram with its characteristics
> model.final = fit.variogram(meuse.vgm, vgm(psill=0.59,"Sph",range=897,nugget=0.05))
> plot(meuse.vgm, model=model.final)
```

References

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2014). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Hengl, T. (2009). "A practical guide to geostatistical mapping".