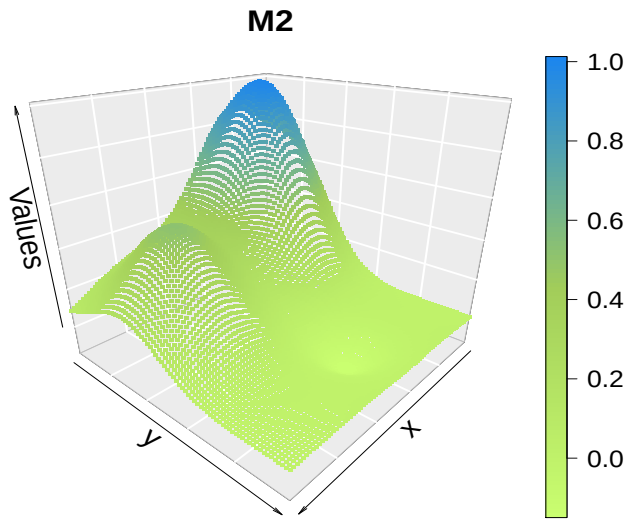
Statistical methods for spatial data analysis

Lecture 9: Interpolation of spatial data

Joaquin Cavieres

Geoinformatics, Bayreuth University



Outline

1. Introduction

Geostatistical data are data that could in principle be measured anywhere, but that typically come as measurements at a limited number of observation locations: think of gold grades in an ore body or particulate matter in air samples.

The pattern of observation locations is usually not of primary interest, as it often results from considerations ranging from economical and physical constraints to being ‘representative’ or random sampling varieties.

Problems what we can solve using Geostatistics are:

- The estimation of ore grades over mineable units, based on drill hole data
- Interpolation of environmental variables from sample or monitoring network data (e.g. air quality, soil pollution, ground water head, hydraulic conductivity)
- Interpolation of physical or chemical variables from sample data

But, what means interpolation?

But, what means interpolation?

Spatial interpolation is the prediction of a given phenomenon in unmeasured locations. In this case we need a spatial interpolation model to calculate predicted values of the variable of interest, given some observed data.

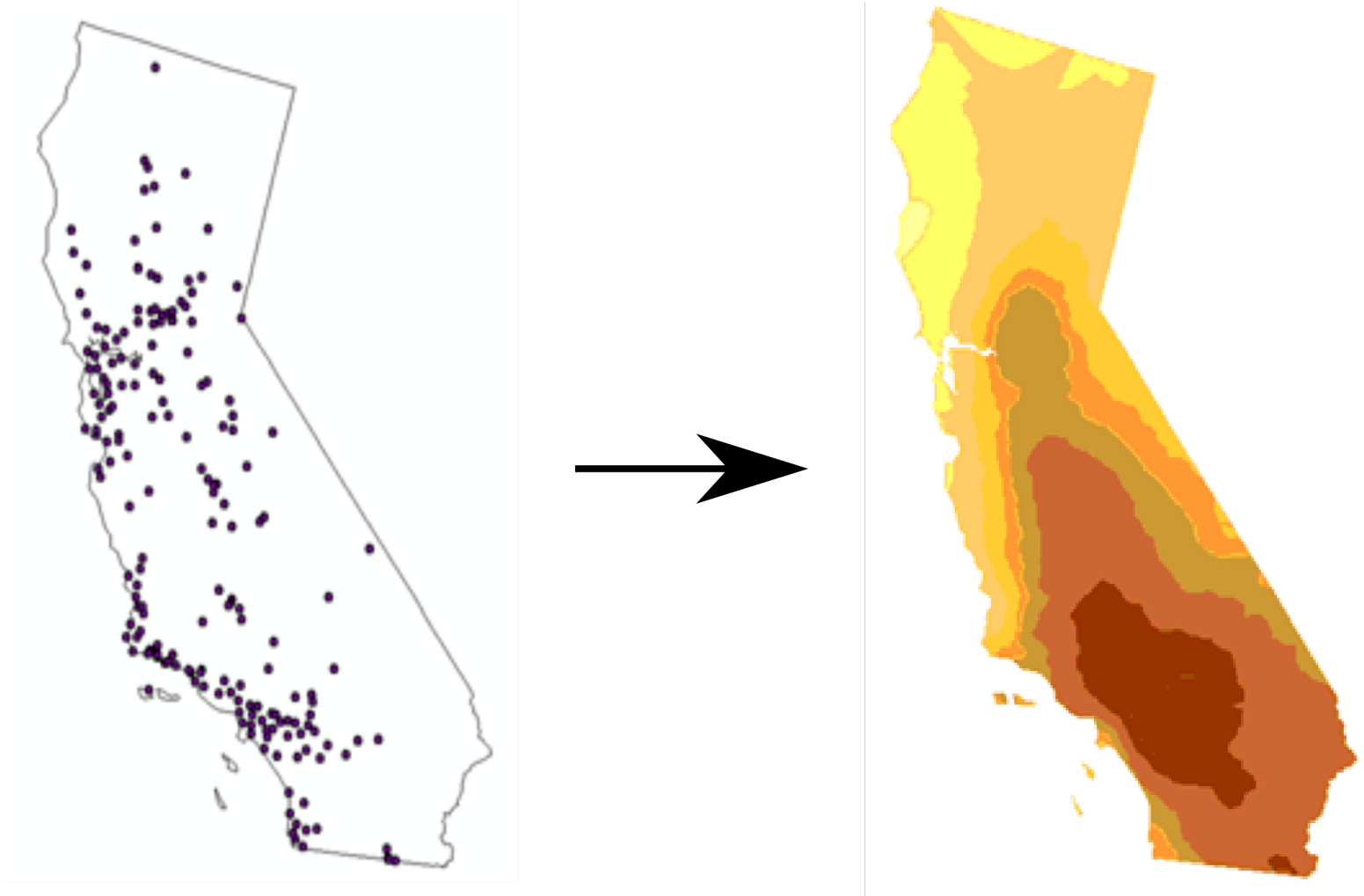


Figure 1: Spatial interpolation (Point locations of ozone monitoring stations, Interpolated prediction surface) (`https://geobgu.xyz/r-2020/spatial-interpolation-of-point-data.html`).

We can split spatial interpolation in two categories:

- Deterministic models: we uses arbitrary parameter values, for example: IDW, NNI.
- Statistical models: we uses parameters chosen objectively based on the data, for example: Kriging

And, what is Geostatistics?

And, what is Geostatistics? Well, first we have to consider two important concepts:

- Variability
- Stochastic methods

1.1. Variability

Natural phenomena generally occur in space, time or space/time. If we consider the space (spatial domain), for example, the topographic surface or a groundwater contamination one can observe high variability within small distances. However, this type of processes can not be described based on physical and chemical laws completely.

So, considering the underlying variability of the process, then we have to evaluate the following:

- Laboratory measurements are necessary to try quantify (in some way) the phenomena
- There is always a limited degree of explanation one can achieve since upscaling is non trivial

1.2. Stochastic methods

Usually we can use probabilistic and statistical methods for describing partly known (or sampled) natural parameters.

Measurement values of a certain parameter, obtained from different locations are often treated as different outcomes of the same random variable.

So, what is Geostatistics?

So, what is Geostatistics?

General definition

Geostatistics can be viewed as a collection of all statistical and probabilistic methods applied in geosciences.

What are Geostatistical data?

What are Geostatistical data?

Definition

Geostatistical data are measurements about a spatially continuous phenomenon that have been collected at particular sites.

This type of data may represent, for example, the temperature registered at several monitoring stations, the rain in some specific spatial area or the pollution of some pollutant.

The common steps to interpolate an spatial model for geostatistical data are:

- Analyse the data.
- Calculating the empirical semivariogram or covariance values.
- Fit (train) a model to the observed values.
- Predict values by the Kriging method and quantify the error (uncertainty) associated with them for every location in the output surface.

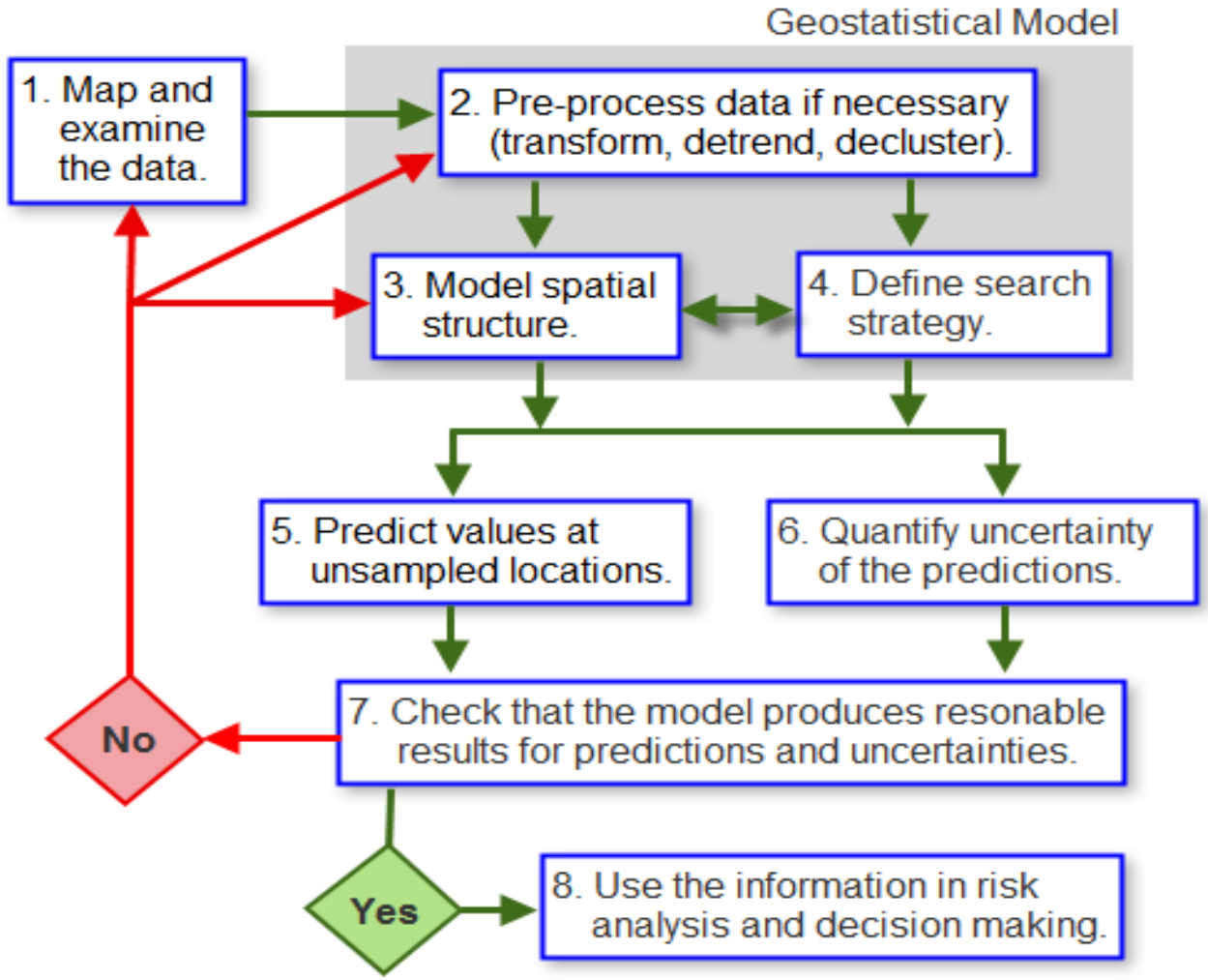


Figure 2: Source: `http://pro.arcgis.com/en/pro-app/help/analysis/geostatistical-analyst/the-geostatistical-workflow.htm`

For spatial interpolation, in this case we only will consider the statistical models for Geostatistical data.

2. Geostatistics

Let $Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)$ are observations of a spatial variable Y at the spatial locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$. Thus, we can assume that the spatial (geostatistical) data is a realization of a spatial stochastic process:

$$\{Y(\boldsymbol{s}) : \boldsymbol{s} \in D \subset \mathbb{R}^2\}, \tag{1}$$

where D is a fixed subset $\mathbb{R}^2$ and the spatial index $\boldsymbol{s}$ varies continuously throughout D.

2.1. Gaussian random fields

A Gaussian random field (GRF) is a collection of random variables where the observations occur in a continuous domain, and the collection of the random variables has a multivariate normal distribution.

$$\{Y(\boldsymbol{s}) : \boldsymbol{s} \in \mathscr{D} \subset \mathbb{R}^2\} \tag{2}$$

Additionally, a stochastic (spatial) process is strictly stationary if for a set of locations $\boldsymbol{s}_i$, $i = 1, \ldots, n$, it has a constant mean and invariant to translations, that is, the distribution of $\{Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)\}$ is the same that of $\{Y(\boldsymbol{s}_1 + \boldsymbol{h}), \ldots, Y(\boldsymbol{s}_n + \boldsymbol{h})\}$ for any $\boldsymbol{h} \subset \mathbb{R}^2$.

A less restrictive condition is given by the second-order stationarity. Under this condition, the spatial process has constant mean

$$E[u(\boldsymbol{s})] = \mu, \forall \boldsymbol{s} \in \mathscr{D}, \tag{3}$$

and the covariance

$$\text{Cov}(u(\boldsymbol{s}), u(\boldsymbol{s}+\boldsymbol{h})) = C(\boldsymbol{h}), \forall \boldsymbol{s} \in \mathscr{D}, \forall \boldsymbol{h} \in \mathbb{R}^2, \tag{4}$$

only depending on the differences between locations. The covariance matrix of a GRF specifies its dependence structure and it is constructed from a covariance function.

In addition, if the covariances are functions only of the distances between locations and not of the directions, the process is called isotropic. Other way, it is anisotropic. A process is said to be intrinsically stationary if in addition to the constant mean assumption it satisfies

$$Var[Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_j)] = 2\gamma(\boldsymbol{s}_i - \boldsymbol{s}_j), \forall \boldsymbol{s}_i, \boldsymbol{s}_j. \tag{5}$$

where $2\gamma(\cdot)$ is the variogram and $\gamma(\cdot)$ is the semivariogram ([3]).

Considering the assumption of intrinsic stationarity, the constant-mean assumption implies:

$$2\gamma(\boldsymbol{h}) = Var(Y(\boldsymbol{s}+\boldsymbol{h}) - Y(\boldsymbol{s})) = E[(Y(\boldsymbol{s}+\boldsymbol{h}) - Y(\boldsymbol{s}))^2], \qquad (6)$$

and the semivariogram (empirical):

$$2\hat{\gamma}(\boldsymbol{h}) = \frac{1}{|N(\boldsymbol{h})|} \sum_{N(\boldsymbol{h})} (Y(\boldsymbol{s}_i) - Y(\boldsymbol{s}_j))^2, \qquad (7)$$

where $|N(\boldsymbol{h})|$ is the number of distinct pairs in $N(\boldsymbol{h}) = \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : \boldsymbol{s}_i - \boldsymbol{s}_j = \boldsymbol{h},\ i,j = 1,\ldots,n\}$. If the process is isotropic, the semivariogram is a function of the distance $h = ||\boldsymbol{h}||$

Once calculated the variogram, it is standard practice to smooth the empirical semivariogram by fitting a parametric model to it, and the parameters used to this are: the *Nugget*, the *Sill* and the *Range*.

- Nugget: The random error process indicated by the height of the jump of the semivariogram at the discontinuity at the origin.
- Sill: The limit of the variogram tending to infinity lag distances.
- Range: The distance in which the difference of the variogram from the sill becomes negligible.

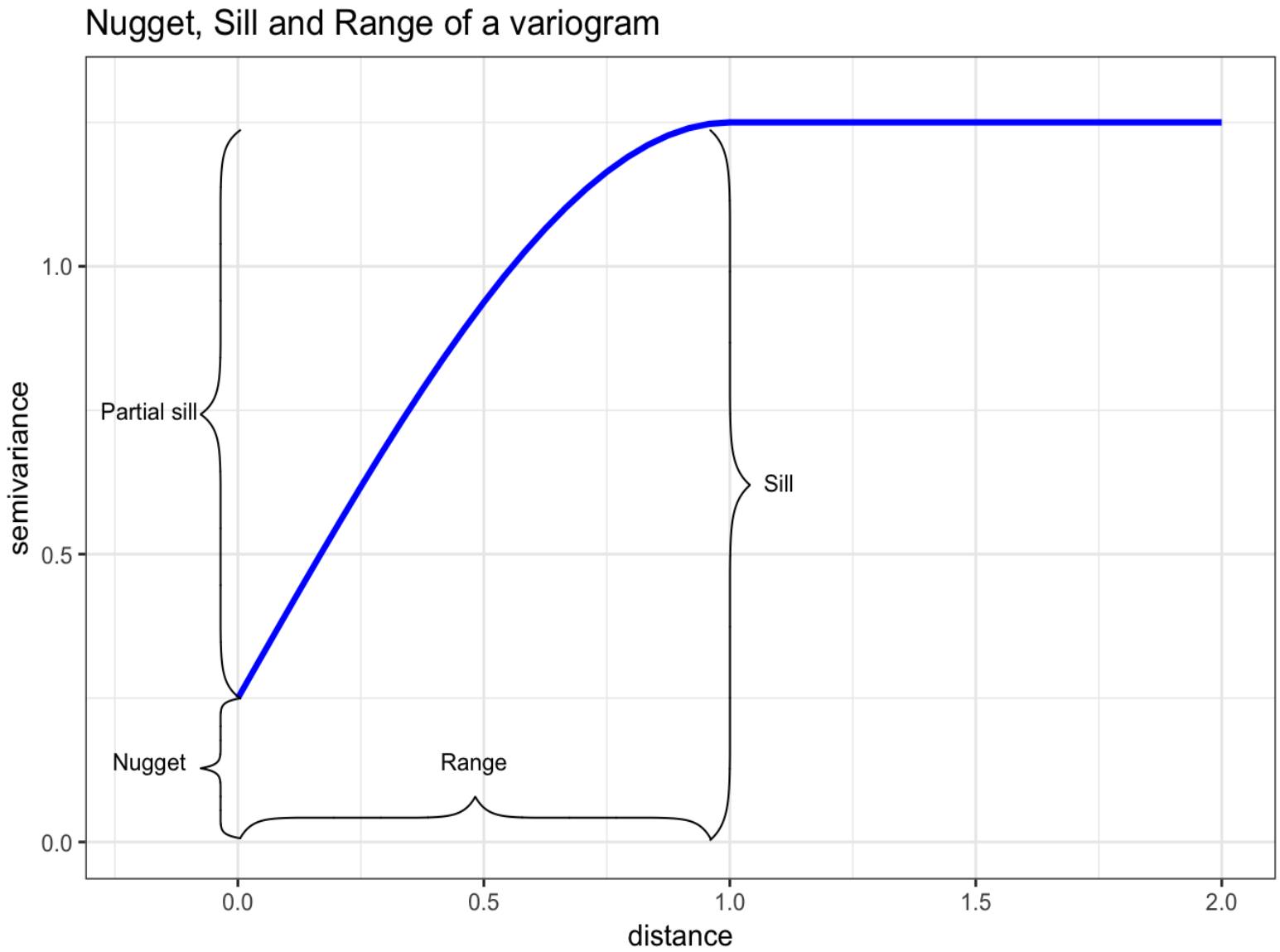


Figure 3: Nugget, Sill and Range of a variogram (Source: Freie Universitat Berlin)

There are different parametric models to fit a variogram, for example:

- Spherical: $\gamma(h,\theta) = \begin{cases} \theta_1(\frac{3h}{2\theta_2} - \frac{h^3}{2\theta_2^3}) & \text{for } 0 \leq h \leq \theta_2 \\ \theta_1 & \text{for } h > \theta_2 \end{cases}$
- Exponential: $\gamma(h,\theta) = \theta_1\{1 - \exp(-h/\theta_2)\}$
- Gaussian: $\gamma(h,\theta) = \theta_1\{1 - \exp(-h^2/\theta_2^2)\}$
- Matérn: $\gamma(h,\theta) = \theta_1\left(1 - \frac{(h/\theta_2)^\nu \kappa_\nu(h/\theta_2))}{2^{\nu-1}\Gamma(\nu)}\right)$,

  where $K_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν.

Fitting a variogram (see the example in R)

3. Kriging

Spatial prediction refers to the prediction of unknown quantities $Y(\boldsymbol{s}_0)$, based on sample data $Y(\boldsymbol{s}_i)$ and assumptions regarding the form of the trend of Y and its variance and spatial correlation.

If we think about the behaviour of Y, in this quantity the large scale and small scale contribute to its variation. So, continuing with this idea, we can model this stochastic spatial process (spatial random field) $\{Y(\boldsymbol{s}), \boldsymbol{s} \in D \subset \mathbb{R}^d\}$ as:

$$Y(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \epsilon(\boldsymbol{s}), \quad \boldsymbol{s} \in D, \tag{8}$$

where $\mu(\boldsymbol{s}) = E(Y(\boldsymbol{s}))$, the mean function, and $\epsilon(\boldsymbol{s})$ a zero-mean random error process.

- $\mu(\boldsymbol{s})$ is the first-order structure.
- $\epsilon(\boldsymbol{s})$ is the the second-order structure.

Here $\mu(\boldsymbol{s})$ accounts for large-scale spatial variation (*global trend*), and $\epsilon(\boldsymbol{s})$ for the small-scale spatial variation (*spatial dependence*). As $\epsilon(\boldsymbol{s})$ typically has no spatial structure we explicitly separate it from the spatially dependent component. Thus, the geostatistical model can be written as:

$$Y(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \eta(\boldsymbol{s}) + \epsilon(\boldsymbol{s}) \tag{9}$$

where

- $\eta(\boldsymbol{s})$ is the spatially dependent component and the spatially uncorrelated mean zero errors.
- $\epsilon(\boldsymbol{s})$ is the measurement error.

Note that the processes $\eta(\boldsymbol{s})$ and $\epsilon(\boldsymbol{s})$ are independent, and it is often referred to as a *nugget effect*.

3.1. Estimation of $\mu(s)$

The mean function $\mu(\boldsymbol{s})$ is specified by a parametric model, $\mu(\boldsymbol{s};\beta)$, then:

$$\mu(\boldsymbol{s};\beta) = \boldsymbol{X}(\boldsymbol{s})^T\beta \tag{10}$$

where $\boldsymbol{X}(\boldsymbol{s})$ is a vector of covariates (explanatory variables) observed at $\boldsymbol{s}$, and β is a parameter vector.

The standard method for fitting the mean function $\mu(\boldsymbol{s})$ is the ordinary least squares (OLS). The mathmatical expression is the following:

$$\hat{\boldsymbol{\beta}}_{OLS} = \text{argmin} \sum_{i=1}^{n} [Y(\boldsymbol{s}_i) - \boldsymbol{X}(\boldsymbol{s}_i)^T \beta]^2 \tag{11}$$

Fitted values and fitted residuals at data locations are given by the equation $\hat{Y} = \boldsymbol{X}^T \hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\epsilon} = Y - \hat{Y}$.

3.2. Ordinary Kriging (OK)

We will consider the model:

$$Y(\boldsymbol{s}) = \mu + \epsilon(\boldsymbol{s}) \tag{12}$$

where μ is the constant stationary function (global mean) and $\epsilon(\boldsymbol{s})$ is the spatially correlated stochastic part of variation

The predictions are done using the following equation:

$$\hat{Y}_{OK}(\boldsymbol{s}_0) = \sum_{i=1}^{n} w_i(\boldsymbol{s}_0) \cdot Y(\boldsymbol{s}_i) = \boldsymbol{\lambda}_0^T \cdot \boldsymbol{Y} \tag{13}$$

where $\boldsymbol{\lambda}_0^T$ is the vector of Kriging parameters (weights, (w_i)), $\boldsymbol{Y}$ is the vector of n observations at primary locations.

See the example in R

Thanks!...

- Moraga, P., 2023. Spatial Statistics for Data Science: Theory and Practice with R
- Pebesma, E., & Bivand, R. (2023). Spatial Data Science: With Applications in R. CRC Press.
- Cressie, N. (2015). Statistics for spatial data. John Wiley & Sons.
- Bivand, R. S., Pebesma, E. J., Gomez-Rubio, V., & Pebesma, E. J. (2008). Applied spatial data analysis with R (Vol. 747248717, pp. 237-268). New York: Springer.