Exercise 1 - Basic Statistics

Office hours: Friday 09:00-12:00

A. Berne alexis.berne@epfl.ch
G. Ghiggi gionata.ghiggi@epfl.ch
M. Guidicelli matteo.guidicelli@epfl.ch

The objectives of this first exercise are twofold: (1) illustrate the basic statistical concepts (exploratory data analysis, distributions, moments, etc...) described during the lecture and (2) become familiar with R, an open-source and multiplatform statistical software, and in particular with the gstat package (`http:/www.gstat.org`) that will be used throughout this course.

For this first exercise, we will use the "meuse" dataset included in gstat. The dataset consists of measurements of heavy metal concentrations (in ppm), along with a number of soil and landscape variables at different locations, collected in a flood plain of the river Meuse, near the village of Stein (NL). The data were collected by Ruud van Rijn and Mathieu Rikken and compiled by Edzer Pebesma.

A R code (ex1.R) showing how to read and analyze the data is already provided. Information about the syntax and arguments of R functions is available via the *help* within R. A short but useful tutorial is also available on moodle.

1. Plot and show the map of cadmium, copper, lead, zinc and organic matter concentrations for the studied domain. Hint: the code for the cadmium is already provided. What can you say about the spatial distribution of these variables?

2. Plot and show the empirical probability density function (pdf) estimated from the values of cadmium, copper, lead, zinc and organic matter. Hint: the code for the cadmium is already provided. Is normality a reasonable hypothesis? Explain your answer.

3. Compute and give the values of the mean, the standard deviation, the median and the 90% percentile of the cadmium, copper, lead, zinc and organic matter concentrations. Which variable exhibits the minimum relative variability? And the maximum? Hint: Use the functions mean(), sd(), median() and quantile().

4. Compute and give the minimum/maximum euclidean distance between each pair of data in the studied domain. How many pairs are separated by 200 m or less? Hint: use the functions dist(), min(), max() and which().

5. Demonstrate the following equality: $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X,Y]$.