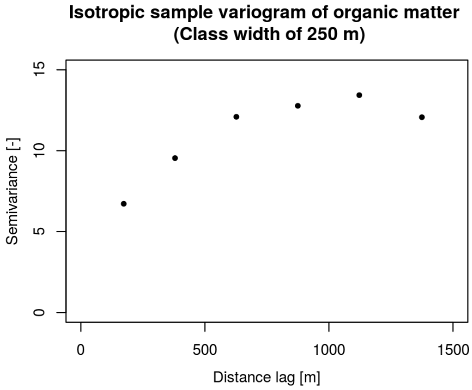
Exercise 2 - Structural analysis - Solutions

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

1. *Compute and plot the isotropic sample variogram of the percentage of organic matter using regularly spaced distance classes between 250 m and 1500 m. Roughly identify the value of the nugget, the sill and the range.*

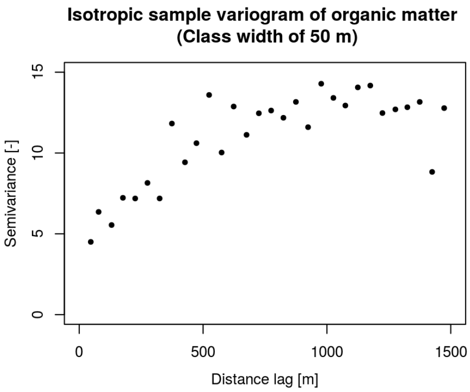


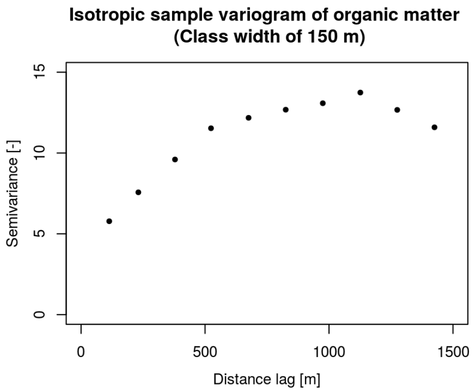
The estimated nugget is about 2, the sill about 13 and the range about 1000 m.

2. *How many pairs of points are there in each distance class?*

distance [m]	0-250	250-500	500-750	750-1000	1000-1250	1250-1500
number of pairs [-]	482	1070	1276	1313	1145	1021

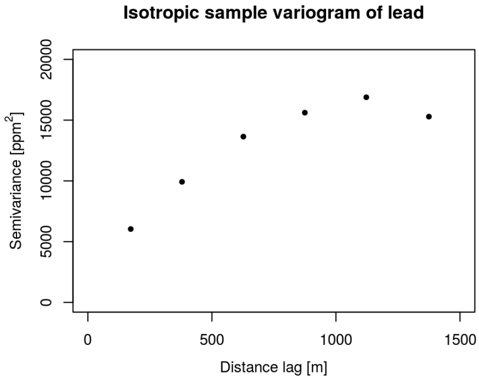
3. *Compute the same variogram using a class width of 150 m and 50 m. Explain the differences and specify which class width you think is most appropriate.*





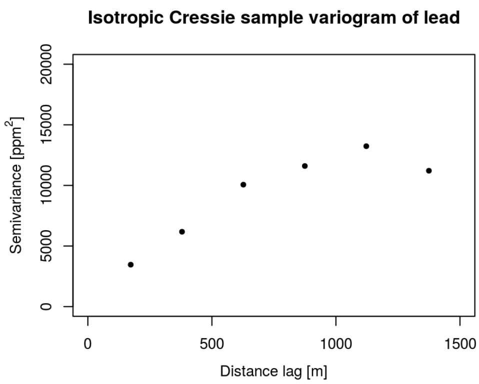
Smaller class widths allow to better estimate the spatial structure (especially at small scales) but the corresponding sample variograms are more noisy due to the larger sampling uncertainties (i.e., the semivariance estimates are less reliable because there are fewer point in each distance class). In our case, both 250 m and 150 m width are adequate. They provide reasonable spatial resolutions and allow to reliably estimate the spatial structure of the field. The large number of data points in each distance class (at least 482 for the 250 m width and 164 for the 150 m width) further increases our confidence in the semivariance estimates. The 50 m width on the other hand does not yield a satisfactory sample variogram. The semivariance estimates are clearly affected by large sampling uncertainties (e.g., as shown by the small semivariance value between 1400 and 1450 m). Moreover, the number of data points in the first 2 classes (2 and 50) is not sufficient to obtain reliable semivariance estimates.

4. *Compute the isotropic sample variogram of the lead concentrations using regularly spaced distance classes of 250 m and a maximum distance of 1500 m. Roughly identify the value of the nugget, the sill and the range.*



The estimated sill is 15500 ppm^2 and the nugget is about 2500 ppm^2 (16.1% of the sill). The range is about 950 m.

5. *Compute the same variogram using Cressie's estimator. Analyze and explain the differences. Which one is more representative? and why?*



The values of the semivariances obtained using Cressie's estimator are smaller than the standard semivariance estimates. The estimated sill is about 12000 ppm^2 and the estimated nugget about 700 ppm^2 (5.8% of the sill). The range is practically identical (about 1000 m). The variogram obtained using Cressie's estimator is more representative because it is less influenced by the large concentrations values of lead (i.e., Cressie's estimator gives less weight to large values)

6. *Estimate the variance of the lead concentrations and compare it to the sill of the variograms obtained in 4 and 5. What can you say?*

   The estimated variance of the lead concentrations is 12392.15 ppm^2, which is very close to the sill of the sample variogram obtained using Cressie's estimator. This shows that Cressie's estimator is less biased than the standard variogram estimator (which has an estimated sill of about 15500 ppm^2).

7. *Demonstrate the following equality:* $\mathrm{Cov}[X,Y] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]$

$$\begin{aligned}
\mathrm{Cov}[X,Y] &= \mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])] \\
&= \mathrm{E}[XY - \mathrm{E}[X]Y - X\mathrm{E}[Y] - \mathrm{E}[X]\mathrm{E}[Y]] \\
&= \mathrm{E}[XY] - 2\mathrm{E}[X]\mathrm{E}[Y] + \mathrm{E}[X]\mathrm{E}[Y] \\
&= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]
\end{aligned}$$