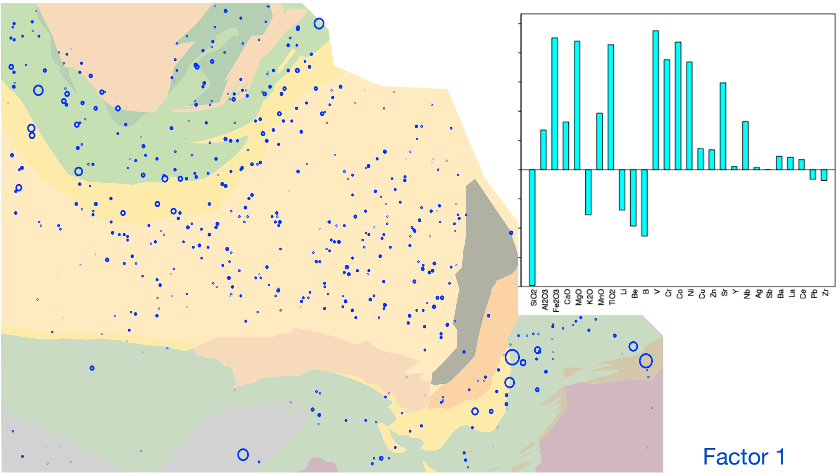
EPSC 552: Data analysis and Geostatistics:

Lecture XII: Spatial analysis of data

FA - processes in Massif Central dataset

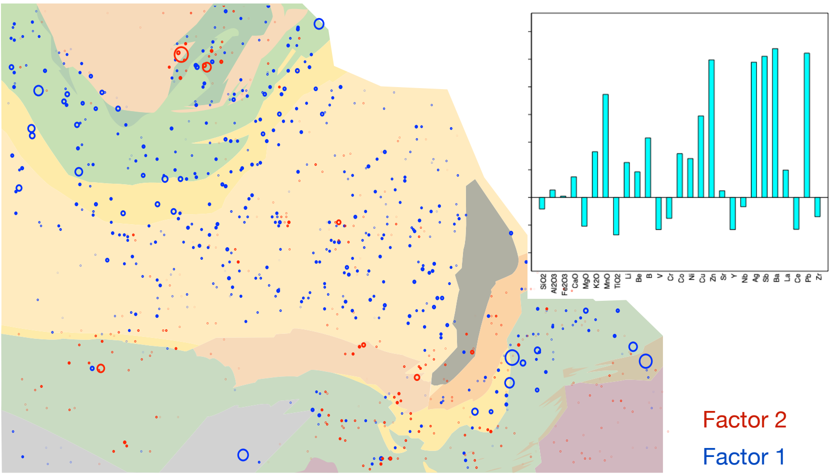
Loadings show the importance of that factor at each location



Factor 1

FA - processes in Massif Central dataset

Loadings show the importance of that factor at each location

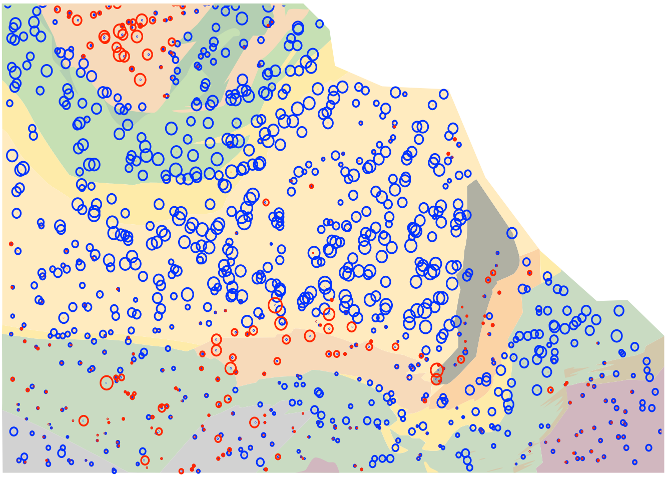


Factor 2

Factor 1

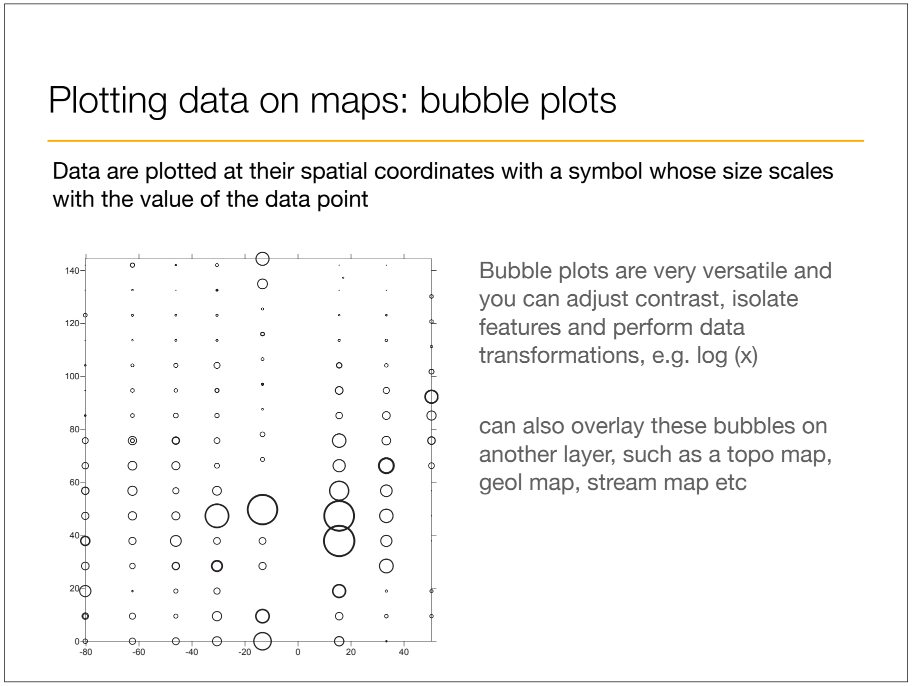
Clustering - groups in Massif Central dataset

Fuzzy cluster assignment shows spatial grouping of samples



cluster 8

cluster 4

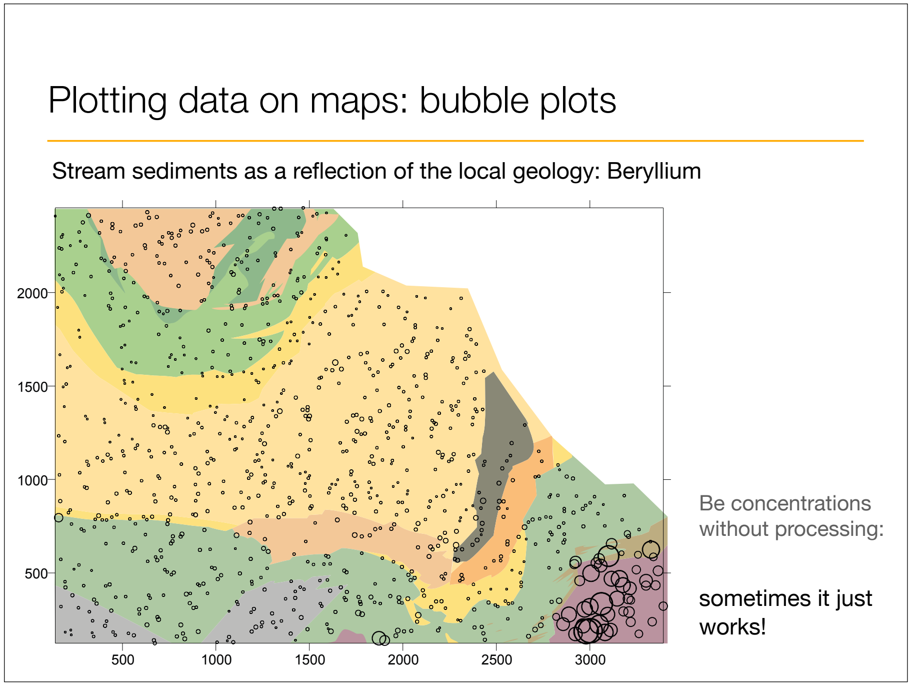


Plotting data on maps: bubble plots

Data are plotted at their spatial coordinates with a symbol whose size scales with the value of the data point

Bubble plots are very versatile and you can adjust contrast, isolate features and perform data transformations, e.g. log (x)

can also overlay these bubbles on another layer, such as a topo map, geol map, stream map etc

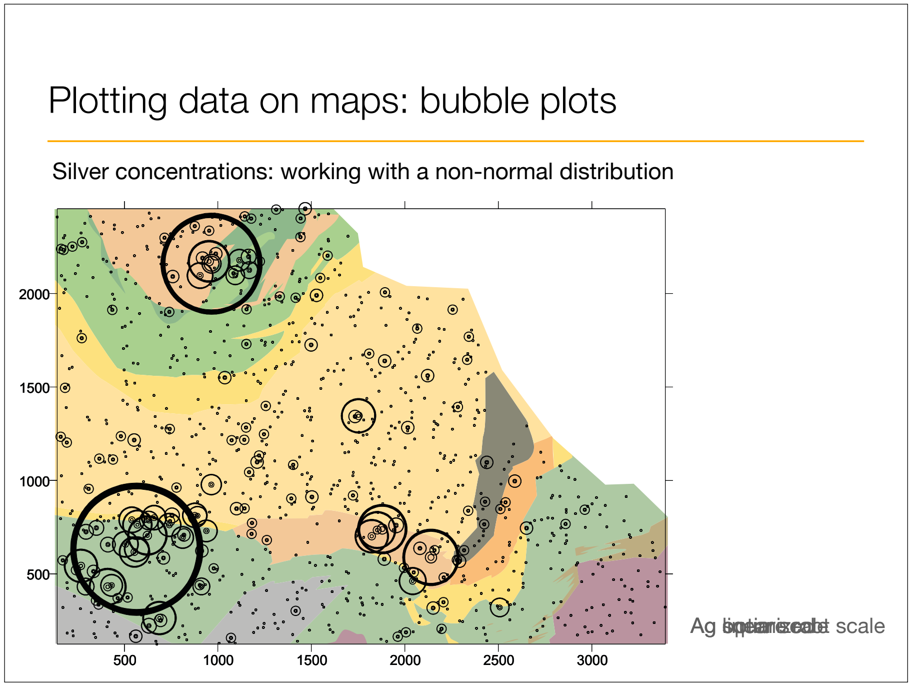


Plotting data on maps: bubble plots

Stream sediments as a reflection of the local geology: Beryllium

Be concentrations without processing:

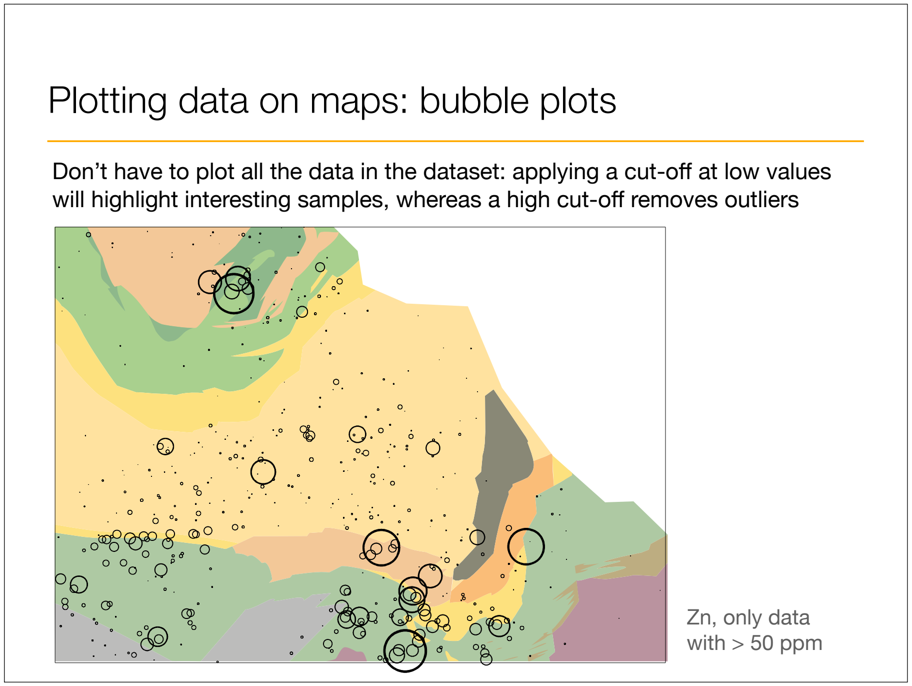
sometimes it just works!



Plotting data on maps: bubble plots

Silver concentrations: working with a non-normal distribution

Ag linear scale



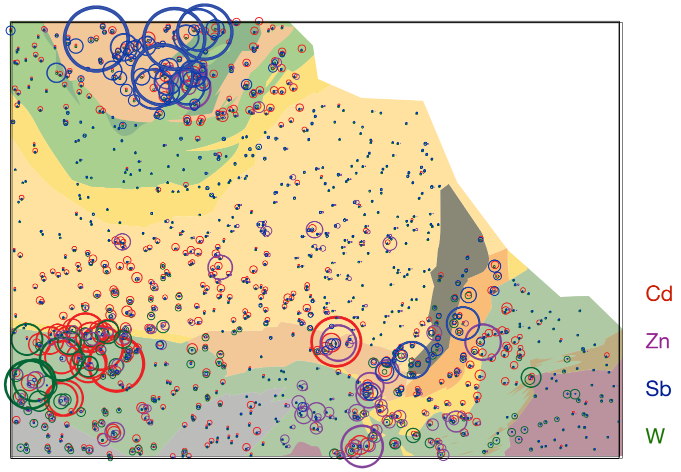
Plotting data on maps: bubble plots

Don’t have to plot all the data in the dataset: applying a cut-off at low values will highlight interesting samples, whereas a high cut-off removes outliers

Zn, only data with > 50 ppm

Plotting data on maps: bubble plots

Looking for element associations by combining bubble plots



Plotting data on maps

Combining elements by using multi-coloured bubble plots is useful, but fast becomes confusing: can lead you to miss interesting samples

Can also calculate such associations beforehand and plot them directly:

- Sb + Zn
- Sb / Zn

Or you can apply logical rules to the data before plotting:

- plot Sb if S > 200 ppm
- if SiO_2 > 60 wt% then plot K / Zr

Note that such properties are calculated much easier and faster in programs designed for such calculations: e.g. Excel or Quattro Pro

Plotting data on maps: QGIS and BC dataset

The BC survey makes a digital version of its geological map available onto which you can plot your geochemical data: need a GIS package (qgis.org)

Download the geol map as a shape file here: https://www2.gov.bc.ca/gov/content/industry/mineral-exploration-mining/british-columbia-geological-survey/publications/digital-geoscience-data

make a new file in QGIS, go to project > properties > CRS and set the coordinate system of the file to 3005

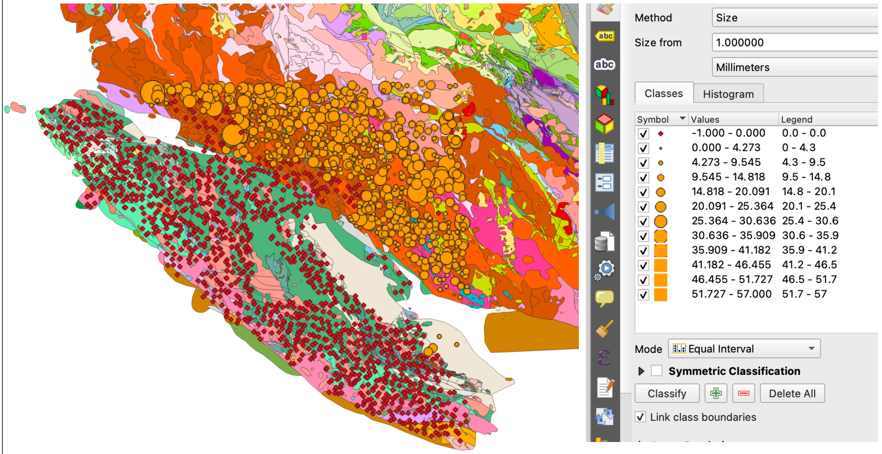
Drag the .shp file into the layers panel. To get the correct colours, go to layer properties > symbology > categorized > style > load style > open file: open the .qml file.

To get your data in, export the excel file as a .csv. Then in QGIS > add layer > add delimited text layer > open your .csv file. Make sure longitude and latitude are selected as x and y fields and set the CRS to 4326 (WGS 1984)

To do fun stuff: click on symbology > graduated > method:size > value:Co > mode:equal interval > classify > apply. You now have a bubble plot for Co

Plotting data on maps: QGIS and BC dataset

The BC survey makes a digital version of its geological map available onto which you can plot your geochemical data: need a GIS package (qgis.org)



Method: Size

Size from: 1.000000

Millimeters

Classes | Histogram

Symbol	Values	Legend
☑ ●	-1.000 - 0.000	0.0 - 0.0
☑ ●	0.000 - 4.273	0 - 4.3
☑ ●	4.273 - 9.545	4.3 - 9.5
☑ ●	9.545 - 14.818	9.5 - 14.8
☑ ●	14.818 - 20.091	14.8 - 20.1
☑ ●	20.091 - 25.364	20.1 - 25.4
☑ ●	25.364 - 30.636	25.4 - 30.6
☑ ●	30.636 - 35.909	30.6 - 35.9
☑ ■	35.909 - 41.182	35.9 - 41.2
☑ ■	41.182 - 46.455	41.2 - 46.5
☑ ■	46.455 - 51.727	46.5 - 51.7
☑ ■	51.727 - 57.000	51.7 - 57

Mode: Equal Interval

□ Symmetric Classification

☑ Link class boundaries

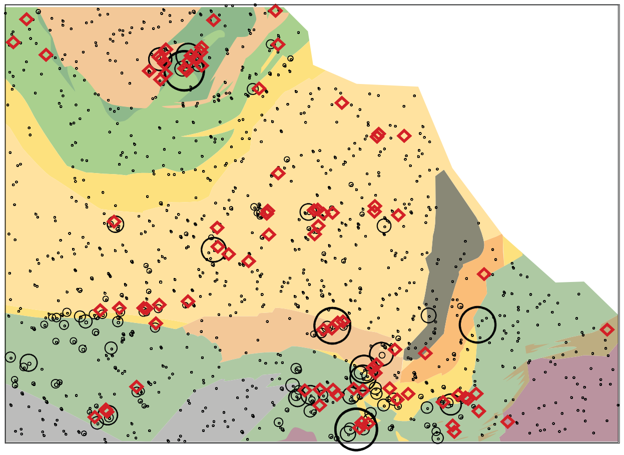
Plotting data on maps

Not limited to plotting data, but can also plot derived properties such as the mean, median, standard deviation, etc

and not just values, but also other observations: geol code / vegetation / mode in multi-modal distribution

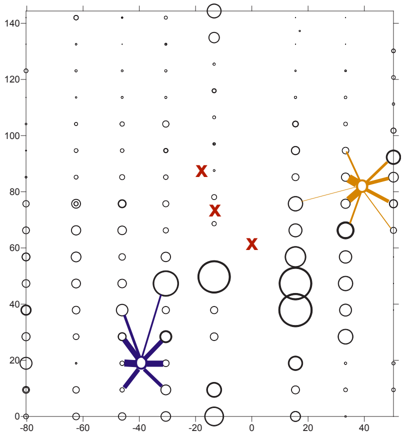
Plotting data on maps: bubble plots

Plotting processed data - standard deviation: the variability at a sample site



Spatial data visualization

To be able to calculate contours and surfaces: interpolation



need to know the concentration at any point in the sampling space to be able to draw smooth contours:

interpolate between values

interpolation on As content grid;

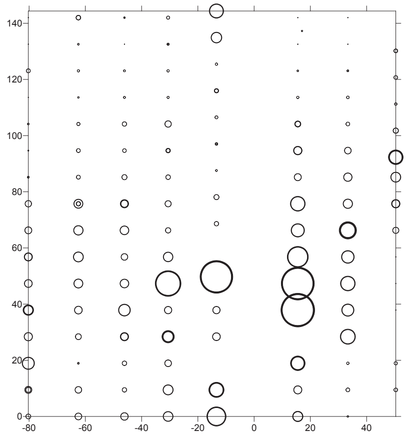
x nearest neighbour

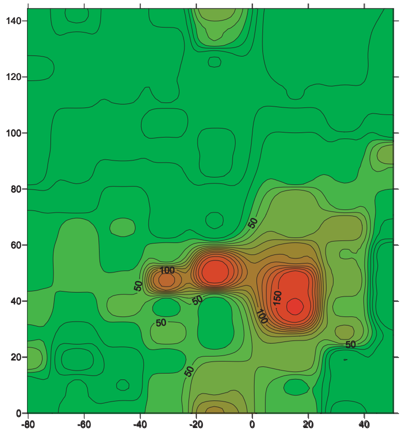
o radius technique: 1/r

o radius technique: $1/r^2$

Spatial data visualization

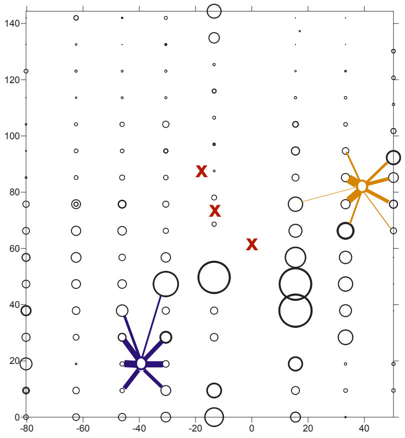
Results of different interpolation techniques:





Spatial data visualization

To be able to calculate contours and surfaces: interpolation



interpolation on As content grid;

x nearest neighbour

o radius technique: 1/r

o radius technique: $1/r^2$

main issue: what samples should be included in the interpolation:

what should the maximum radius be?

Interpolation radius

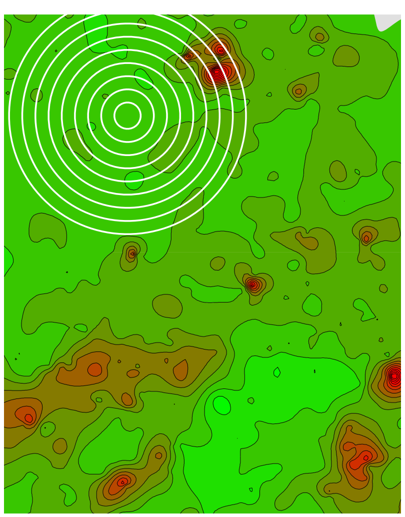
Spatial data have a very useful property:

adjacent samples should be most similar, whereas samples that are far apart can be distinctly different, or:

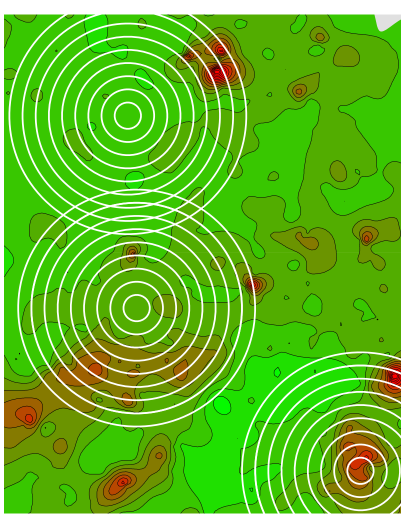
- the variance for a small interpolation radius is small, as the variance between adjacent samples is small
- the variance increases as the interpolation radius increases (i.e. as samples further away from the point of interest are included)
- at some radius the variance will no longer increase as we have reached the overall variance, which is called the “regional variance”

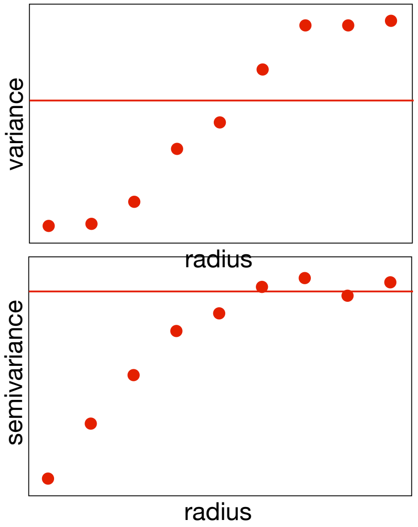
including values beyond the regional variance radius is pointless as such samples do not contain any information on the value at the point of interest

Interpolation radius

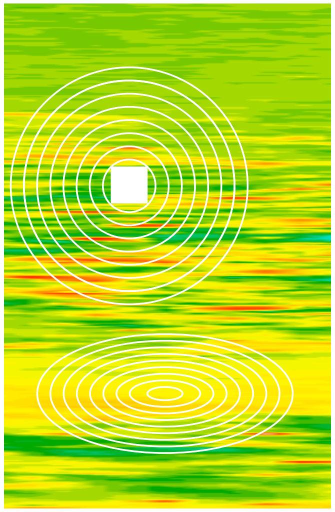


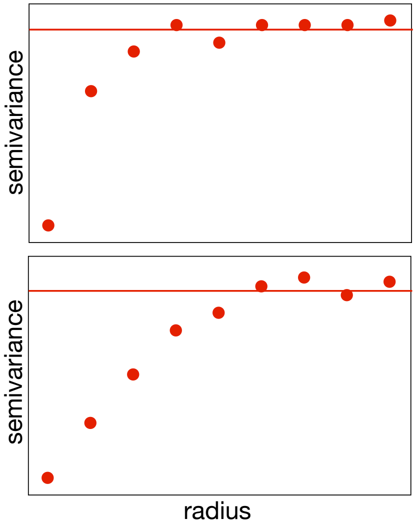
Interpolation radius





Interpolation radius





Semivariance and semivariograms

This concept is semivariance and is shown in a semivariogram

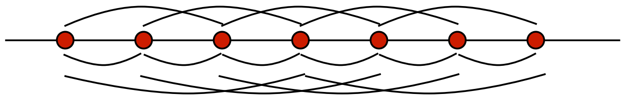
semivariance: the variance between samples a specified interval or distance apart

as the interval increases, the semivariance will approach the total variance of the data set, so it is a spatially controlled partial variance of the data

$$\gamma_h = \frac{\sum (z_i - z_{i+h})^2}{2(n - h)}$$

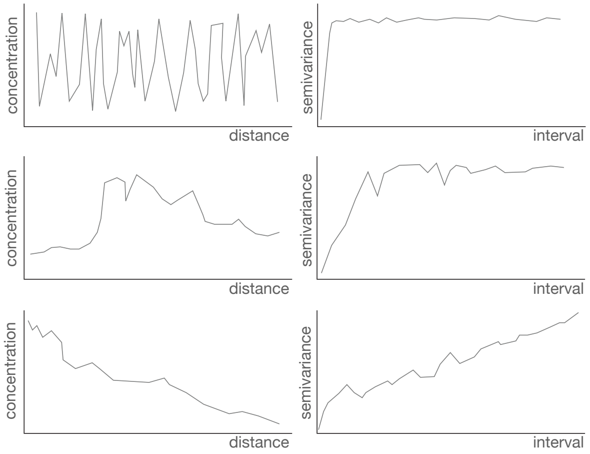
with: γ = semivariance for interval h
n = total number of samples
z_i = value at position i

as h increases, the relatedness of the samples decreases and the variance will therefore increase:



Semivariance and semivariograms

plotting the semivariance against h: semivriogram



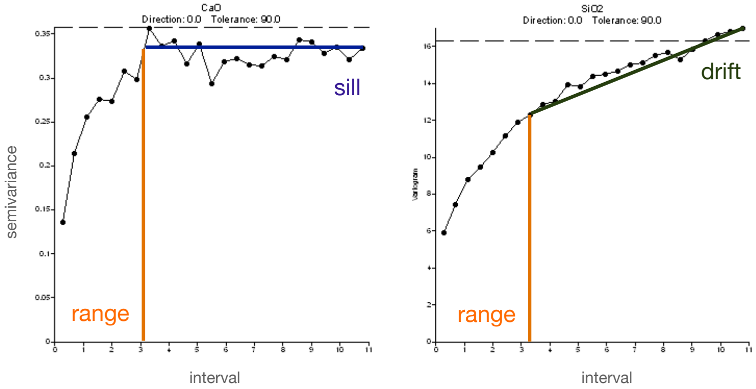
no relation with distance: random

gradual changes in concentration

continuous variation with distance: trend

Semivariance and semivariograms

properties of a semivariogram :



the range is the interval within which there is similarity between the samples

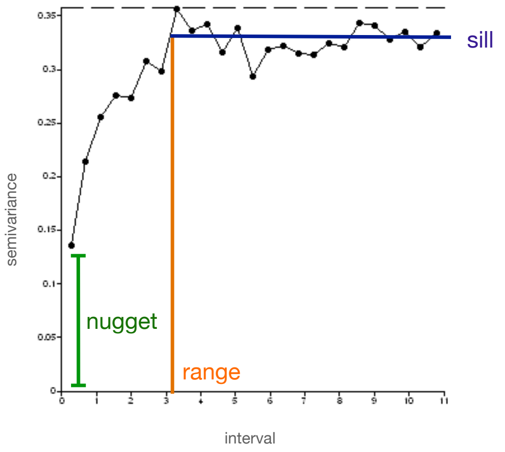
Semivariance and semivariograms

Semivariograms provide our maximum radius criterion: only samples that fall within the range are included in interpolation

before we continue, a few notes:

- semivariograms have to be determined for each variable as each has its own range: interpolation has to be performed separately as well
- semivariograms are generally different for different spatial directions (N, SW, etc). Such anisotropy can point to an underlying geological phenomenon such as layering or a fault control on conc. This can be corrected for either manually by stretching the coordinate system perpendicular to the main axis, or automatically by kriging software
- most semivariagrams have an apparent cut-off at zero distance that has a semivariance ≠ 0. This is called the nugget effect and is caused by sample heterogeneity (= field duplicate variance)

Nugget effect in semi-variograms



There is always some uncertainty at a given sample site, which you could quantify by taking field duplicates.

This sample site variance is the "nugget" in a semivariogram (in essence the variance at zero distance)

Every element will have such a nugget, but the effect is strongest for elements that are heterogeneously distributed, such as gold present as nuggets in a sediment because we use mean + var

Using semivariogram information: kriging

The interpolation technique that employs the range information as obtained from semivariograms is called kriging

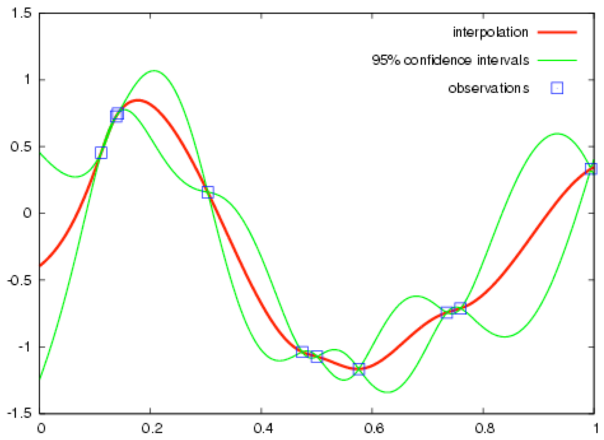
in kriging, only samples that are within the range are used to determine the value at a given intermediate position and the weighing for each sample is derived from its associated semivariance

$$A\ (x_i,y_i) = wt_1 \cdot A\ (x_1,y_1) + wt_2 \cdot A\ (x_2,y_2) + wt_3 \cdot A\ (x_3,y_3) + \ldots$$

as an added bonus this also gives us the variance associated with each interpolated value (the uncertainty), so we can immediately see where our interpolations are reliable and where they are not

because weights are based on the semivariance, obvious trends in the data should be removed as this leads to a continuous rise in the semivariance: can be done by first subtracting a trend surface

Estimate of uncertainty for each interpolated value

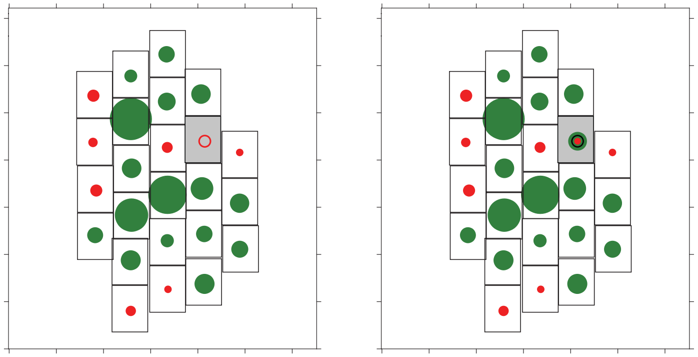


source: wikipedia.org

Uncertainty in block kriging of grades

Kriging is commonly applied to estimate the grade of blocks in open pit mining using a sample grid or the grade of adjacent blocks (or both).

In such cases it is invaluable to know the uncertainty on the grade estimate



Flavours of kriging

There are many flavours of kriging and discussing them all would be a course in its own right. A few terms that you come across commonly:

Simple/Ordinary kriging: no trend in the data, so there is a constant mean in the dataset and the variance is calculated as the difference from this mean. This mean is either known (Simple) or calculated from the data (Ordinary)

Universal kriging: there is a spatial trend in the data, so the mean varies with the spatial coordinates. Instead of using universal kriging, you can also remove the trend in pre-processing of the data

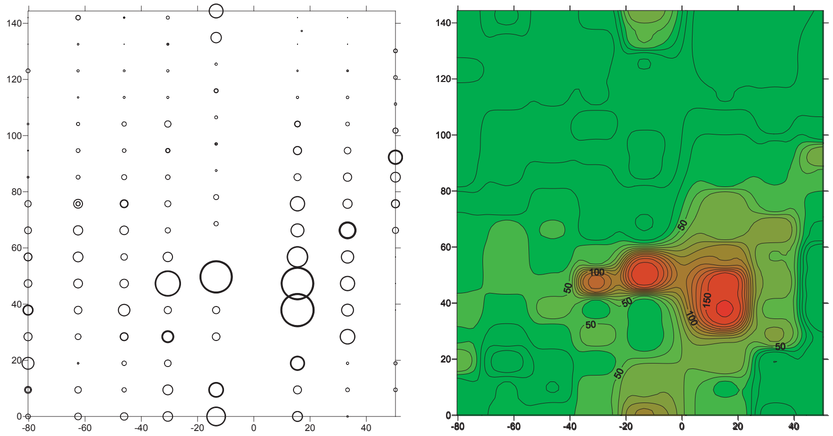
Indicator kriging: rather than estimating a numerical value at a given point, you estimate if it is higher or lower than a set value, and the prob. of this

Co-kriging: a second variable is included in the kriging which is correlated with the first variable. This should improve estimates of the first and main variable

Good kriging resource: Clark & Harper (2000) Practical Geostatistics ISBN 0970331703, or you can download the 1979 original at http://www.kriging.com/pg1979_download.html

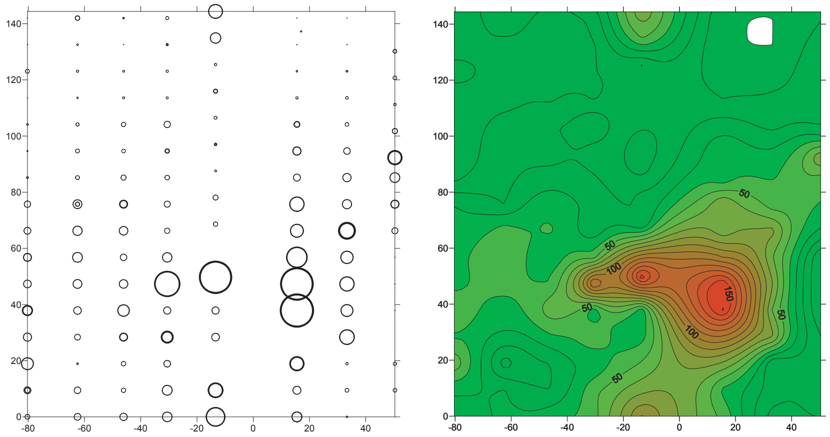
Back to our example

Results of different interpolation techniques:



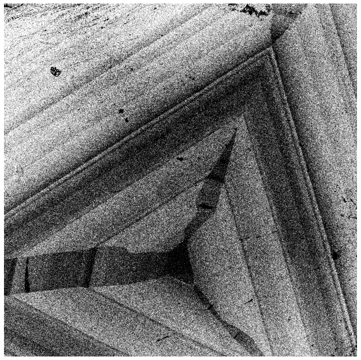
And now using kriging as the interpolation method

Results of kriging on this data set:

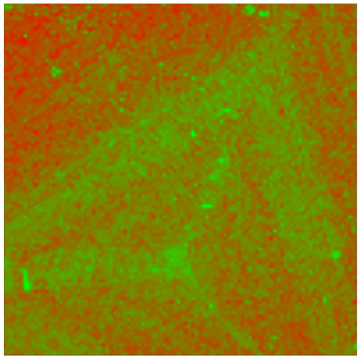


Some data are not suited to interpolation/kriging

There is a strong tendency to directly start with the most complex or fancy technique, such as kriging. However, kriging is not always appropriate !



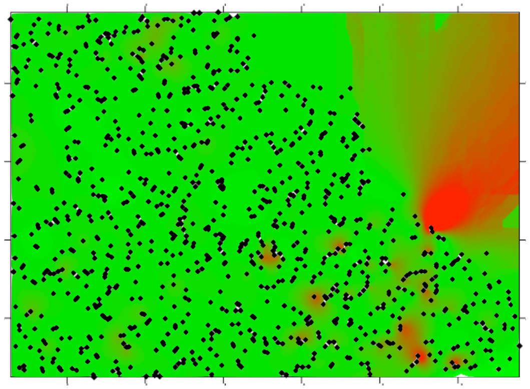
raw concentrations plotted



optimized kriging map

Kriging and sample coverage

Kriging works best when you have a high sample density and a more or less uniform distribution of data over the sample are. If not ➛ get artefacts



Areas without samples need to be blanketed out, not just removed afterwards

Geostatistics and Data Analysis

Course summary

Eigenvector techniques - highlights

Techniques to locate the principle directions in your dataset

- useful to reduce the dimensionality of a dataset to its principle directions - greatly facilitates the interpretation
- the principle directions generally represent the underlying processes that control the data distribution - process identification

some practicalities:

two main techniques: principle component analysis and factor analysis - very similar, but PCA is a true data transformation (no loss of info) whereas FA retains only a subset of the variance

eigenvector techniques are basically a clustering of the variables based on their correlation/covariance similarity - high cor/cov: same trend, low cor/cov: different trend

have to carefully decide the number of significant factors - use the scree plot. If a more appropriate interpretation can be made using more or less factors than the number suggested by the scree plot - no problem

Spatial analysis - highlights

Spatial analysis of data is a great technique to:

- interact with your data, spot trends, correlations, outliers, clustering, and thereby suggests ways to analyze and interpret your data
- link your data to all kinds of other spatial information, such as position of roads, towns, rivers, ice cover, topography, geology, soil type, vegetation, etc
- disseminate your results to others: easy to understand

some practicalities:

advanced methods need a dedicated sampling design, otherwise stick to the more basic techniques such as bubble plots

when a dense uniform sampling grid is available, best results for Earth Science datasets are generally obtained by using kriging and semivariance

trend surfaces are a further powerful technique to interpret spatial data and de-trending should be performed before kriging

Clustering techniques - highlights

Clustering of data is used to:

- split up multi-modal datasets so they can be analyzed with other statistical techniques, such as t-tests and ANOVA
- look for homogeneous groups in the data, which can tell you something about the main separating processes acting upon the data
- classify samples: assign samples to pre-determined groups

some practicalities:

many varieties of separation techniques: DFA, hierarchical clustering, fixed or sought cluster means, partitioning clustering using hard and fuzzy rules, etc

fuzzy clustering is the most powerful for geochemical datasets as it gives the partial membership to each cluster, thereby being able to cope with intermediate samples

as in eigenvector techniques, the main difficulty is in deciding the number of clusters. A variety of parameters can help you make that decision, but feel free to deviate (e.g. outliers commonly get their own cluster)

Regression analysis - highlights

Regression analysis is a technique:

- that allows you to fit a quantitative model to data that can subsequently be used in mathematical models. Also allows for inter- and extrapolation
- that allows you to determine whether a variable explains a significant part of the variance in the dataset: in other words, whether it belongs in the model
- test what the best model is to describe your data (linear, quadratic, logarithmic, exponential, multiple linear, etc)

some practicalities:

the best regression fit has maximum variance along the regression line and minimal on either side. The ratio of explained over total variance is R^2.

important assumptions in regression analysis that have to be met: always check normality of residuals, multi-collinearity, significance of coefficients, etc

Testing - highlights

Statistical testing:

- test the validity of a hypothesis at a specified confidence interval α
- rejection of the null-hypothesis is the stronger results: choose your hypotheses carefully
- all techniques work in exactly the same way: each test has a probability distribution: exceed the critical probability (α) and the hypothesis is rejected, otherwise: no reason to reject the null hypothesis
- crucial to keep the errors in mind when testing: type I - known, specified as the confidence interval in testing results; type II - unknown
- many statistical tests, optimized for specific hypotheses, data distributions, etc (e.g. t-test, Z-test, F-test, ANOVA, Kolmogorov-Smirnov, χ^2-test)
- most commonly used: t-test/ANOVA - determine whether a number of groups/clusters are significantly different from each other
  χ^2-test - determine whether two data distributions or curves are significantly different

Basic techniques - highlights

data description:

central value: mean, median, mode

measures of spread: range, stdev, IQR, percentile, accuracy vs. precision

normal versus robust techniques

type of distribution: normal, lognormal, multi-modal, outliers

data visualization: histograms, boxplots, scatter diagrams, violin plots, etc

correlation:

correlation between variables expressed by a Pearson or Spearman correlation coefficient. To quickly assess correlations for a complex data matrix: cor matrix

error propagation:

technique to propagate the uncertainty on the measured values to the property you are deriving. Easiest way to do this: split up the equation to its most basic operators: add - subtract - multiply - divide

The end....

If you take nothing else away from this course,
remember these:

garbage in = garbage out

most scientists use statistics as the drunkard uses a lamppost; for support rather than illumination