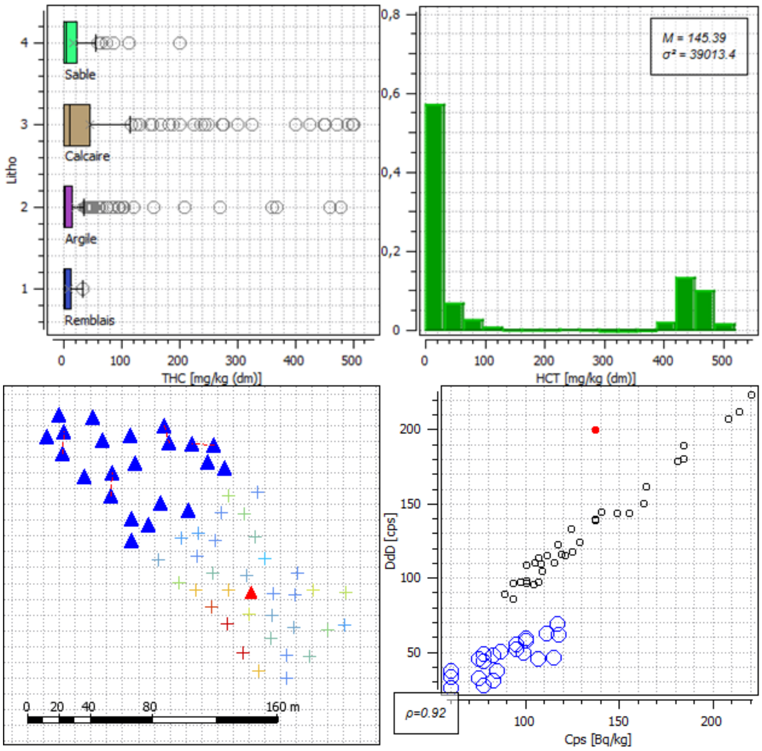
Geovariances

Where no one has gone before

White paper



Data quality control

Why it is essential during a contaminated soil study

Data quality control

Why it is essential during a contaminated soil study.

Be sure of the quality of your data prior to a detailed study using user-friendly visual tools.

Before starting an evaluation of the contamination leading to decision-making tools for site clean-up, it is essential to ensure data quality. Thus, based on solid and reliable foundations, uncertainties occurring during the different steps of the decision process will be better controlled and the final quality of the project improved.

The objective of this white paper is to show how to visualize and assess at best the validity of the data before going further into their analysis.

To do that, different tools can be used and are presented here: univariate and mutivariate statistics and visualization tools, which are then completed by spatial statistics and geostatistics. Applied by Geovariances for more than fifteen years, they have proved to be essential in such kind of projects.

Do you have confidence in the reliability of your data?

Data quality control: how? why?

First step in a contaminated soil study consists of checking the quality of the data in order to ensure the necessary reliability of the global project. This control therefore takes place before a potential modelling in order to know best the uncertainties, reduce the hazards during the decontamination works and so estimate the project costs, reducing the risk of unpleasant surprises.

The statistical and geostatistical tools presented here show how to go from a raw data table to a validated one, by detecting different kinds of errors or anomalies:

- measurement errors,
- recopy errors,
- outliers in statistical distributions,
- errors due to the location of the samples : coordinates problems, very high heterogeneity between close samples (erratic behavior)…

Moreover, these checks and corrections are always carried out in such a way that the initial data table is kept so that it remains a reference point.

By doing this, a second objective is also reached: **data are better understood** thanks to different displays, by nature more visual and informative than just tables.

Sampling

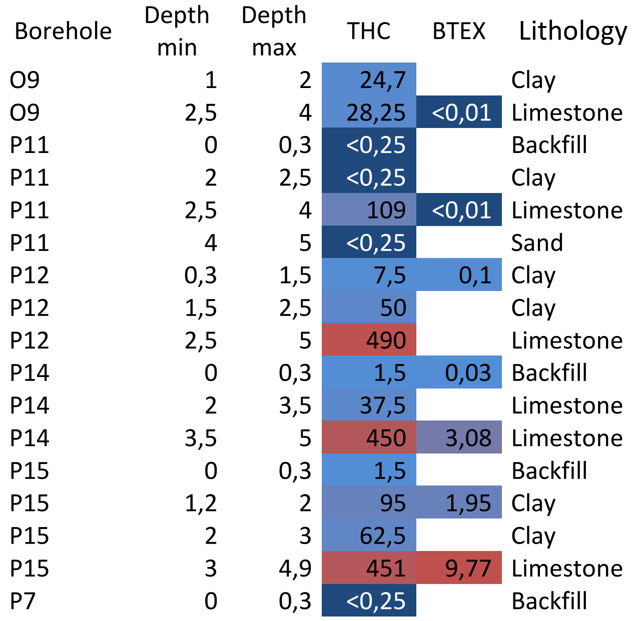
Data quality control is essential, but building a sampling strategy, well considered beforehand, is equally important.

Consistency between evaluation objectives, data analysis and sampling schemes is of major concern for the success of the project.

Visualization tools

Before doing any graphical displays of data, visualization of the table can already provide us with a better understanding of the data:

- sorting according to a parameter,
- selection of some values,
- conditional formatting with colors (Fig. 1) allowing identification of anomalies.



Borehole	Depth min	Depth max	THC	BTEX	Lithology
O9	1	2	24,7		Clay
O9	2,5	4	28,25	<0,01	Limestone
P11	0	0,3	<0,25		Backfill
P11	2	2,5	<0,25		Clay
P11	2,5	4	109	<0,01	Limestone
P11	4	5	<0,25		Sand
P12	0,3	1,5	7,5	0,1	Clay
P12	1,5	2,5	50		Clay
P12	2,5	5	490		Limestone
P14	0	0,3	1,5	0,03	Backfill
P14	2	3,5	37,5		Limestone
P14	3,5	5	450	3,08	Limestone
P15	0	0,3	1,5		Backfill
P15	1,2	2	95	1,95	Clay
P15	2	3	62,5		Clay
P15	3	4,9	451	9,77	Limestone
P7	0	0,3	<0,25		Backfill

Fig. 1 : Example of a formatted data table

How to visualize data.

If data have been **georeferenced**, a base map of data points using different sizes and colors (Fig. 2) very quickly highlights problems of value or locations of samples along with providing us with a better understanding of the contamination structure.

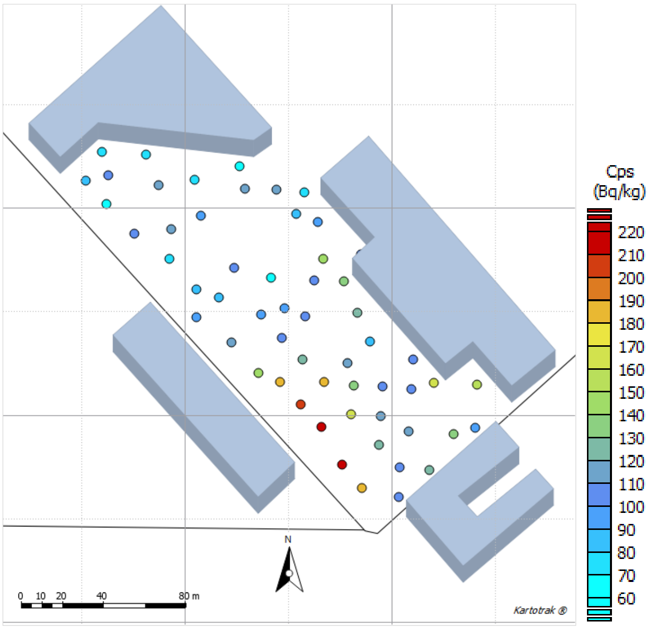


Fig. 2 : Base map with color scale depending on the radiological parameter of interest

In 3D, logs of boreholes may complete a base map or a 3D view (Fig. 3) by showing the link between contamination and depth in order to identify main impacted levels as well as the relation with other variables such as other contaminants or the lithology (Fig. 4).

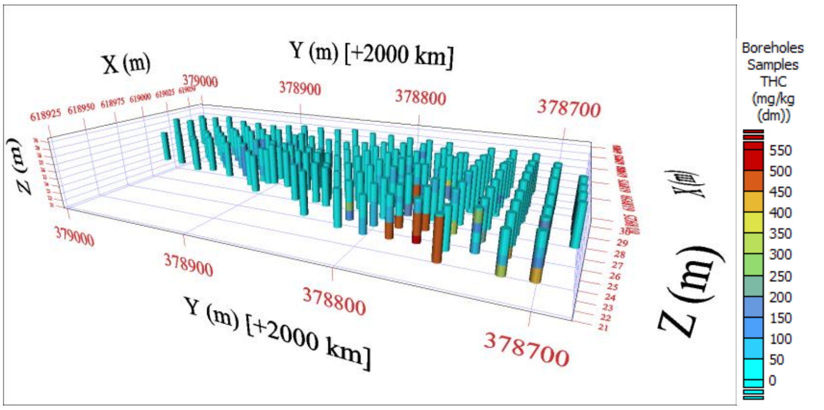


Fig. 3 : 3D visualization (hydrocarbons)

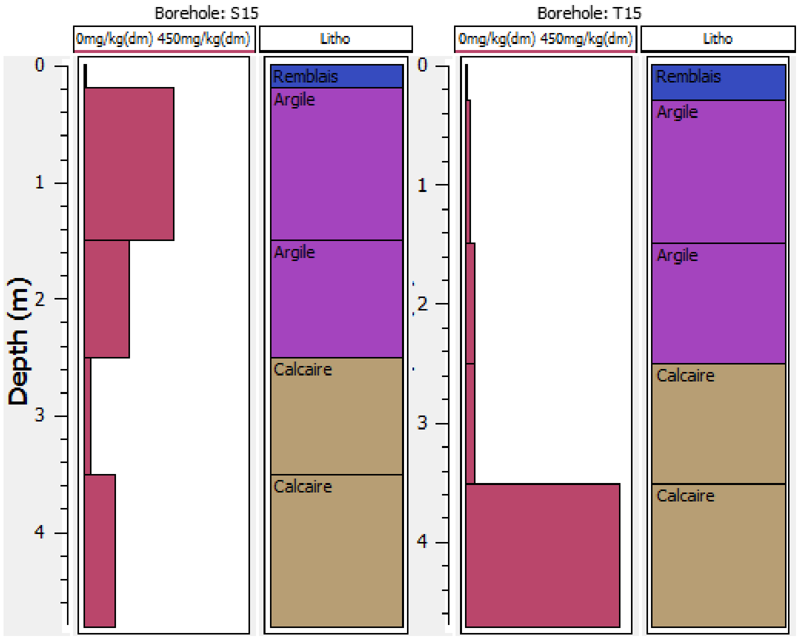


Fig. 4 : Borehole logs of the variable of interest and the lithology

Statistical tools

Several statistical tools exist for studying the distribution of samples. Among them, **histograms and boxplots** are designed to examine one variable at a time.

How to detect outliers.

The histogram (Fig. 5) shows the statistical distribution of data. In contaminated soil studies, this kind of distribution is frequently observed: a large proportion of low values and few high values. Outliers or unexpected distributions may also be detected.

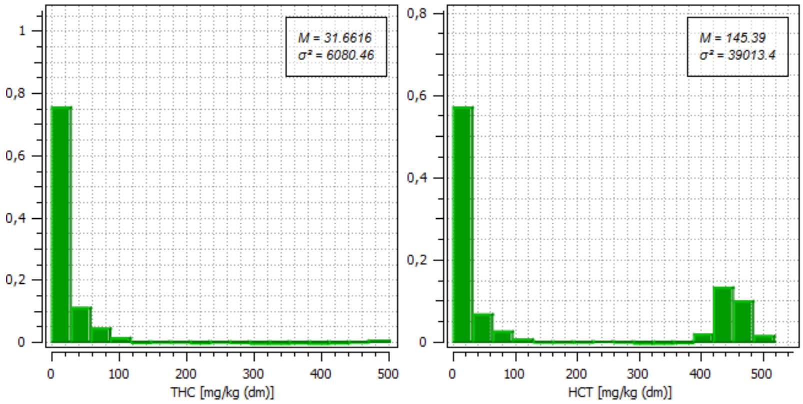


Fig. 5 : Examples of histograms (hydrocarbons)

The histogram on the right is said to be bimodal. It generally corresponds to two sub-populations. Such a histogram should raise questions: was this distribution expected? is it the same laboratory for all analyses? do all the samples come from the same matrix? with the same sampling support? are there several dates?

Clustering effects or preferential sampling can also be detected using histograms, especially if they are dynamically linked to a base map (see principle on Fig. 7).

Boxplots can also complete the histograms: being more synthetic they enable us to analyze the distribution of quantiles and especially to spot the outliers. Drawn by lithology or domain such as in Fig. 6, they provide us with interesting elements on the distribution of the contamination, for example in the different soil horizons.

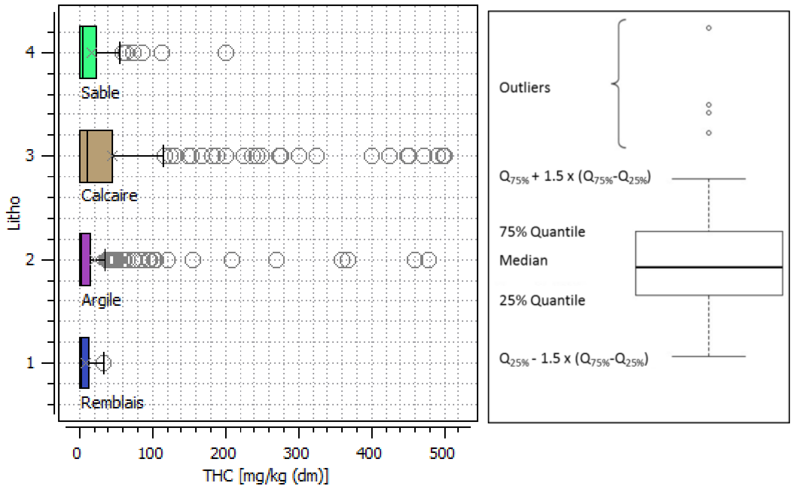


Fig. 6: Boxplots by lithology (hydrocarbons) and defintion of the boxplot

Some outliers may not be detected using these univariate tools, for example when some samples present abnormal values for one or another variable. **Scatter plots** are useful to study the links between variables and to spot these abnormal values or clouds split in several parts, symptomatic of subpopulations.

Fig. 7 corresponds to a base map and a scatter plot between two variables. We can observe a linear relation with a good correlation. However, the red sample is outside of the cloud and seems to present an abnormal value for at least one variable (which would be invisible on a histogram). A different behavior can also be observed in the bottom part of the cloud. By selecting dynamically the points on both plots, it appears that they are all located in the same part of the area studied. These two points should raise questions followed by some checks: is there any problem with the analysis of the red sample? is it really a very particular point? is there any difference in laboratory, soil matrix, date between the two subpopulations? etc.

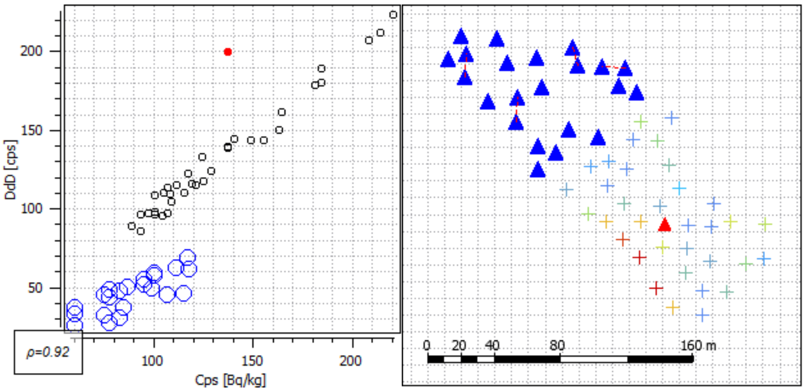


Fig. 7 : Scatter plot (radiological parameters) with two subpopulations and one outlier (with correspondinng selection on the base map)

Are there inconsistencies in data location?

Other tools may also be used during this phase: principal component analysis (PCA), classification, etc. Being more advanced, they are usually kept for advanced statistical analyses after data validation.

Other tools: spatial statistics and geostatistics

To complete statistical tools, **spatial statistics and geostatistics** are used to address another aspect of data quality control: assuming some spatial consistency, are there some samples with abnormal values given their coordinates?

To study this, tools like the **variogram cloud** (showing the square of the difference between two points depending on the distance between them) and the **variogram** (built using the variographic cloud, it shows the spatial continuity of the phenomenon) are very appropriate. Clues like a very high nugget effect (discontinuity at the very beginning of the variogram showing large differences between close points) or very high points on the variogram cloud corresponding to pairs on the base map with always one common sample (Fig. 8) often reveal a particular value regarding its neighbors. This measured value should then be checked as well as its location: X or Y coordinates errors, inversion or translation of coordinates, depth error in case of 3D dataset…

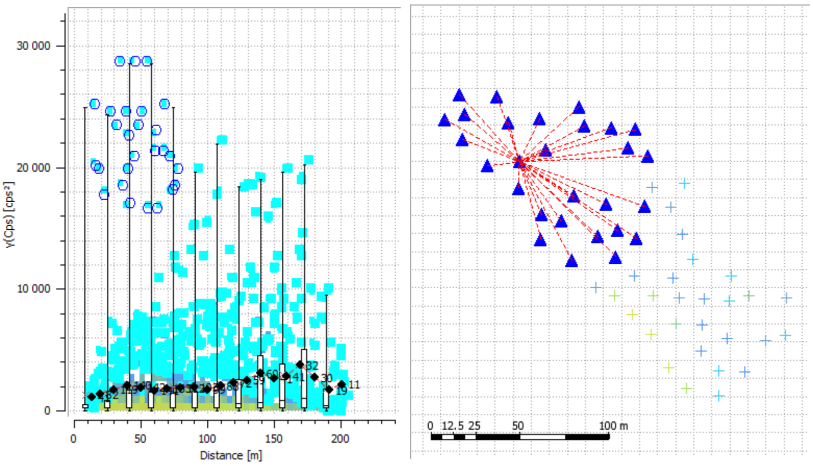


Fig. 8: Variographic cloud with selection of some pairs involving one single sample on the base map

Finally, experimental variogram don't always show the same spatial continuity for each direction. Vertical variability is usually greater than the horizontal one. When a horizontal anisotropy appears it is often linked to a physical phenomenon like a flow. If it cannot be explained, it will be necessary to check if this anisotropy is not artificially induced by distinct subpopulations.

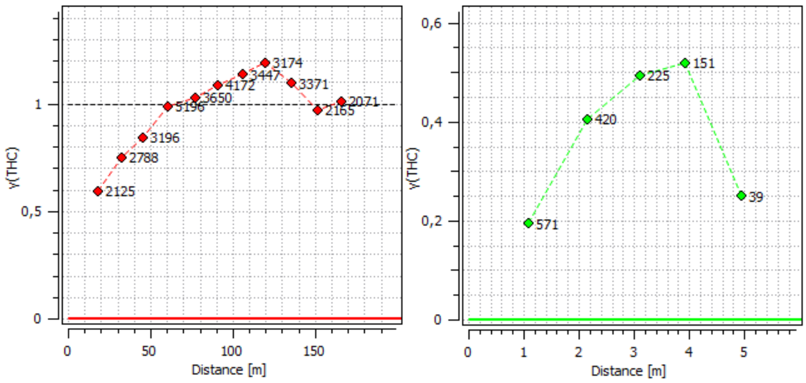


Fig. 9 : Horizontal variogram in red and vertical on in green, showing an anisotropy

References

- Y. Desnoyers (Geovariances), D. Dubot (CEA) : Data Analysis for Radiological Characterisation: Geostatistical and Statistical Complementarity – Workshop on "Radiological Characterisation for Decommissioning", Studsvik, Sweden (April 2012)
- C. Faucheux, Y. Desnoyers (Geovariances) et P. De Moura (CEA) : Characterization of a deep radiological contamination: integration of geostatistical processing and historical data – AquaConsoil, Barcelona, Spain (April 2013)

Who is Geovariances?

Geovariances is a specialist geostatistical software, consulting and training company. We have over 45 staff, including environmental consultants and statisticians.

Geovariances develops and sells two software solutions for contamination characterization:

- **Kartotrak** is an integrated software solution dedicated to the characterization of sites contaminated with chemical or radioactive substances.
- **Isatis** is the accomplishment of +25 years of dedicated experience in geostatistics. It is the global software solution for all geostatistical questions.

Unique expertise

Geovariances is a world leader in developing and applying new and practical geostatistical solutions to the environmental industry. We have strong experience in site characterization and have gained trust from the leading environmental and consulting companies.

Conclusion

Data quality control is the first essential step of a contaminated site study. To carry it out, numerous easy-to-use visual tools exist.

Histograms and boxplots are used to detect outliers for one given variable. Plotting scatter plots can help detect problems, taking into account the links and correlations between variables. Finally, spatial statistics and geostatistics complete this control by checking the spatial consistency of the samples.

Once done, this quality control ensures a more reliable site study and its possible modelling, with a better understanding of the uncertainties and makes it more dependable to predict a budget and plan the decontamination works.

Our expertise

Kartotrak is the first all-in-one software solution for contaminated site characterization and is born out of a +10 year partnership between the French Alternative Energies and Atomic Energy Commission CEA and Geovariances. Easy to use, the software offers an integrated workflow which guides the user through each step of his project, from data loading and quality control, contamination mapping to contaminated soil volume estimation and uncertainty quantification.

Geovariances offers a unique expertise based on more than fifteen years of experience in applying geostatistics to site characterization and remediation projects issues. Most projects have been conducted for main site owners, public institutes, consultancies and remediation companies.

For more information

Let us help you understand the geostatistics added-value for your remediation projects or your contamination mapping.

Contact our consultants: consult-env@geovariances.com.

Geovariances
49 bis, av. Franklin Roosevelt
77215, Avon Cedex
France
T +33 1 60 74 90 90
F +33 1 64 22 87 28

Geovariances Pty Ltd
Suite 3, Desborough House
1161 Hay Street
WEST PERTH, WA 6005
Australia
T +61 8 9321 3877

www.geovariances.com