Good practices for getting reliable and useful information about soil contamination uncertainty from easily conducted statistical analyses

J.-B. Mathieu[1], M. H. Garcia[1] and V. Garcia[1]

[1] KIDOVA, 155 avenue Roger Salengro, 92370 Chaville, France

Abstract

In order to assess soil contaminations, it becomes common practice to collect more and more data, based on more systematic sampling strategies, and to combine on site, in situ and lab analyses. The on site or in situ contamination data are often used to better decide about soil samples that it would be worth analyzing in laboratory (avoiding to pay for lab analyses that would show no contamination at all).

From a number of available data that tends to increase, the soil contamination must be assessed in one way or the other to evaluate its environmental impact. Soil contaminant grades are especially required to assess environmental risks for human health and to delineate contaminated soils that should be remediated. Quantifying in-place mass of contaminants may also be needed to better know about the contamination source and decide about a remediation strategy.

To address such estimation problems, two types of approaches tend to be used and compete. On the one hand, the geostatistical approaches, which are recognized now for their usefulness and advantages, but also require specific expertise, dedicated software tools and additional work efforts that the engineering companies cannot easily provide. On the other hand, empirical approaches that are commonly used by environmental professionals and justified by the fact that an increasing number of more systematically collected data is available. The empirical approaches mainly consist in subdividing all or part of the potentially contaminated site into a set of "representative volumes", spatially distributed according to the data, and in assigning average contaminant grades to the volumes, the grades being estimated from the data located inside and possibly nearby each volume. The main drawback of these approaches is to be deterministic, i.e., they cannot provide any estimation error like a confidence interval, although the estimations are necessarily uncertain as related to the badly known spatial distribution of soil contaminants.

Between these two approaches, understandable and easy to use statistical methods exist that allow to provide estimations and estimation errors (uncertainty) without having to carry out geostatistical studies when they can be avoided, or before to decide that they would be helpful. These methods have the capability to quantify uncertainty on contaminated soil volumes, or on in-place masses of contaminants, over the whole site or any delimited zone. By making them available in software tools dedicated to contaminated sites and soils, this type of estimations can be obtained very quickly without needing advanced expertise in statistics.

This article presents the proposed statistical approach. After having defined the estimation problems to solve and introduced (or recalled) the basic (minimum) statistical notions that are required to study contaminated soils, it describes the few steps of the approach, in a simple way understandable to all. The application of the approach is then illustrated on a real case study about soils contaminated by persistent organic pollutants (organochlorines).

Key-words: soil contamination, in-place contaminant mass, exploratory data analysis, estimation, uncertainty, decision making

Objectives: Demonstrating that exploratory and statistical data analyses can easily be carried out to exploit at best available soil contamination data and derive from them relevant uncertainty results about in-place contaminant masses, or other soil contamination assessments of delimited zones, for decision making within diagnosis and remediation studies.

Innovative nature of the proposed topic: Taking advantage of easy to use and understand statistical tools that can advantageously be part of current practice of soil contamination studies to assess and remediate soils contaminated by persistent organic pollutants (POPs) or other contaminants.

Estimation of contaminated soil volume or in-place mass of contaminant: a spatial statistical issue

The volume and spatial distribution of contaminated soils must be estimated from sparse soil contamination data. Direct and indirect contamination data can be available. Direct data are contaminant grades, resulting from lab analyses on soil samples, and are usually considered as being the reference. Indirect data are on site or in situ contamination measurements that provide more or less precise "integrative" information, whether they measure several components together (e.g., light organic components), or cannot differentiate soil contamination from other water and air contaminations.

Whatever the number of data, their types and their locations, the estimation and delineation of contaminated soil volumes, or in-place contaminant masses, would require to use relevant spatial statistical tools and to quantify spatial uncertainty. If most environmental engineers are not familiar and confident with the calculation and use of statistics, they are even less prepared to address spatial statistics. Nevertheless, they are asked to provide soil contamination assessments, but also recommendations to decide whether additional exploration data are needed, or remediation strategies can already be devised.

Spatial statistics is often related to geostatistics, which provides a theoretical framework suitable to model spatial uncertainty (Chiles and Delfiner, 1999, Goovaert, 1997, Deutsch and Journel, 1997, Isaaks and Srivastava, 1989). The applications of geostatistics to contaminated soils are many (GeoSiPol, 2012, GeoSiPol, 2005). Geostatistics can be used to estimate local soil contaminant grades at soil sample scale, average contaminant grades of soil volumes like remediation blocks, volumes of contaminated soils where grades are higher than a critical threshold, or in-place masses of contaminants. This requires, however, to know enough about geostatistics, to have a good and user-friendly geostatistical software, but also to have time and budget to carry out a geostatistical study. In many situations, all these conditions are not met.

As a practical alternative to geostatistical approaches, the current tendency of environmental consulting or engineering companies, in accordance with recommendations from regulatory agencies (e.g., US EPA's Triad approach), is to multiply the number of data to be able to empirically delineate zones where soils are very likely to be contaminated and should require remediation solutions. The uncertainty about the actual level of soil contamination of delimited zones, whether it is expressed in terms of average contaminant grade or mass of contaminants, is seldom addressed. Basic statistics are simply calculated from the data, ignoring the fact that the data may be spatially correlated, hence partly redundant, and without quantifying any uncertainty about estimated statistics.

The aim of this article is to present and demonstrate simple and more advanced statistical tools that are useful to carry out exploratory data analysis (EDA) and to quantify the uncertainty about soil contamination assessments of delimited zones.

Introduction to basic statistics

Basics in statistics are required to understand and be able to quantify uncertainty. Related to contaminated site assessments, statistics are, or should be, used in the three following situations:

- to perform exploratory data analysis (EDA),
- to estimate local soil contamination (contaminant grades) over delimited zones or remediation blocks,
- to quantify uncertainty about contaminated soil volumes or in-place masses of contaminants.

A brief review of the statistical tools used in these situations is given below.

Statistical tools used in exploratory data analysis

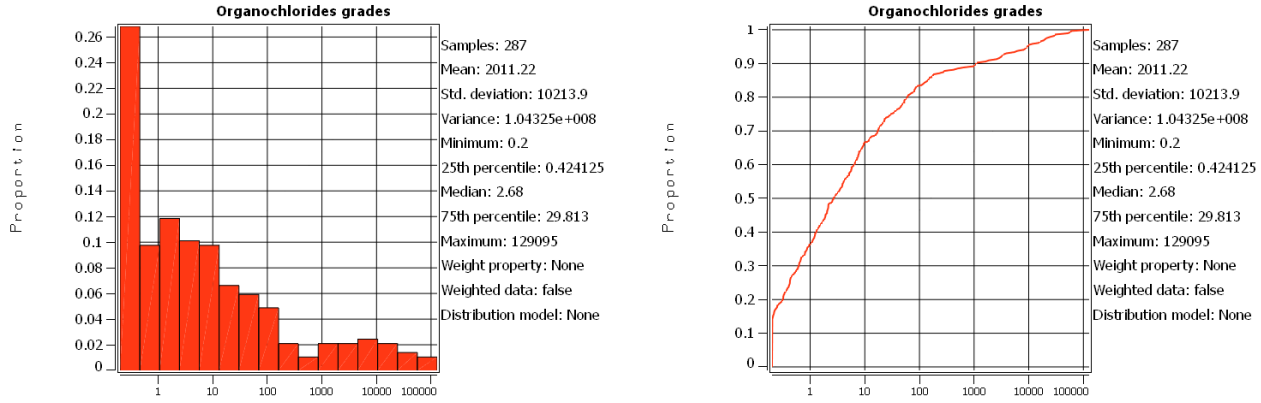
- Histogram and statistical summary (Figure 1.a): characterizing the variability of soil contaminant grade data and the complexity of the distribution (histogram), whose shape and skewness may be consequential for statistical estimation, checking the presence and meaningfulness of outlier data.
- Cumulative distribution function (cdf) (Figure 1.b): identifying the probability (or proportion) of soil contaminant grade data that are below or above any critical grade threshold. A cdf can be derived from a histogram by summing the bins.
- Scatterplot and correlation analysis (Figure 2): understanding the relationship and correlation, or lack of relationship, between contaminants or between direct and indirect soil contamination data, identifying, confirming or disconfirming outlier or suspicious data (see Harris *et al.*, 2014, for more advanced statistical methods).
- Data declustering (Figure 3): if contaminant grades are spatially correlated, as it is often the case, data closer to one another being more likely to be similar, the redundancy of data must be quantified and corrected to estimate reliable statistics, including histogram and cumulative distribution function. This is done by weighting the data, the more redundant (correlated) a data is with others, the smaller the weight.

Statistical tools used to estimate local soil contaminations

- Mean (average) contaminant grade: estimated over a delimited zone or a remediation block, from inside and ideally nearby direct and indirect contamination data. This is required to classify a volume of soils as being safe or contaminated (i.e., posing environmental or health risk for some particular land use), or to derive from it the total mass of contaminants within the volume.

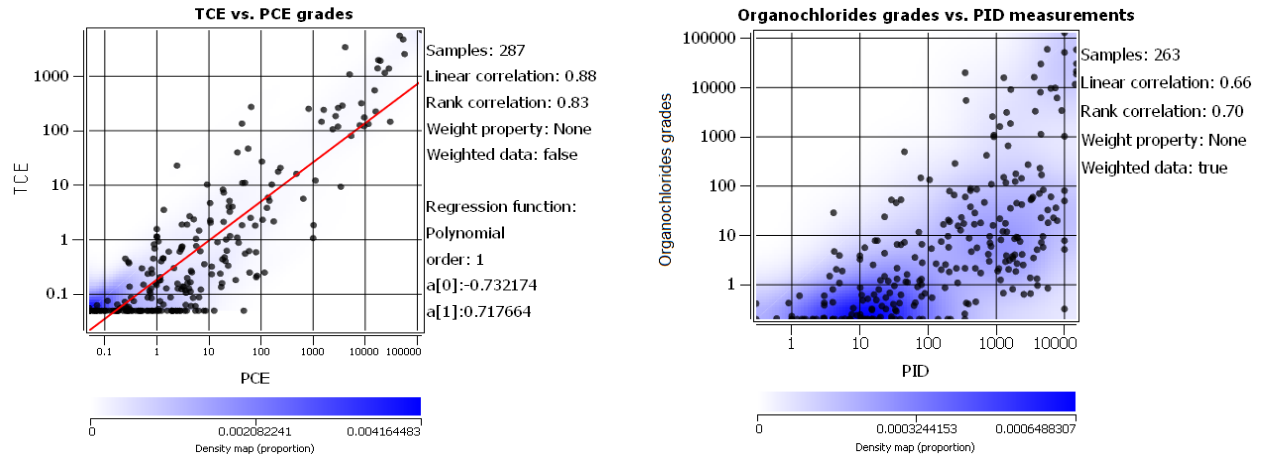
Statistical tools used to quantify uncertainty about contaminated soil volumes or in-place contaminant masses

- Standard-deviation: relative quantification of uncertainty, unless some good assumptions can be made about the uncertainty model (see for instance Chilès et Delfiner, 1999, §3.4.4).
- Confidence interval: related to a percentage of chances that the actual value (contaminated soil volume, in-place contaminant mass) be within an interval defined from minimum and maximum bounds (quantiles) derived from a cdf. Figure 4 shows a 95% confidence interval derived from a cdf.



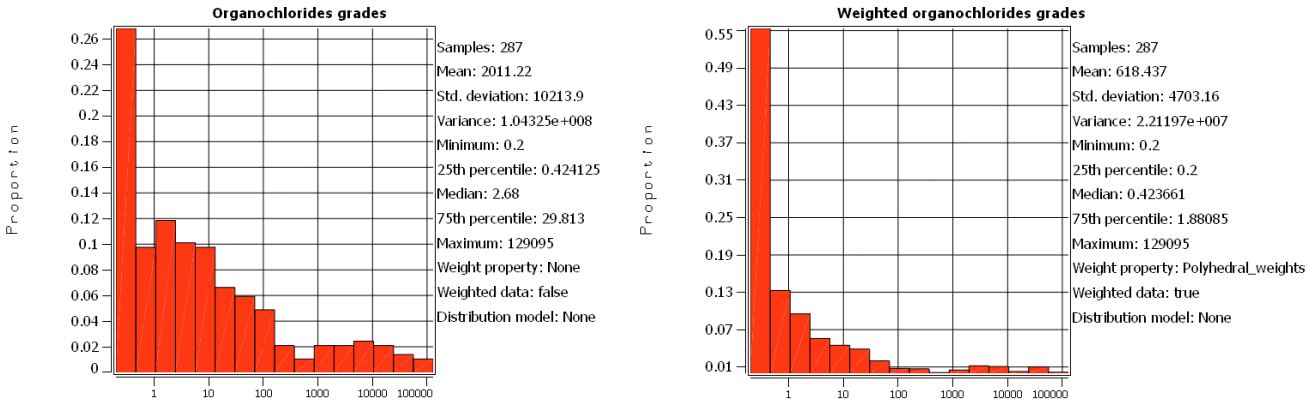
a) Histogram and statistical summary. b) Cdf and statistical summary.

Figure 1: Characterization of the variability of soil contamination grades from a) histogram and statistical summary, b) cumulative distribution function (cdf) and statistical summary.



a) PCE grade vs. TCE grade. b) PCE grade vs. PID measurement.

Figure 2: Use of scatterplot and correlation coefficients (measuring linear and rank correlations) to identify the presence of outlier or suspicious data and to characterize the relationships and correlation between a) contaminants, b) direct and indirect (photo-ionization detector or PID) contamination data.



a) Unweighted histogram & statistical summary. b) Weighted (corrected) histogram & statistical summary.

Figure 3: Histogram and statistical summary of a) unweighted data and b) weighted data, the weighted data begin corrected from preferential sampling and data redundancy.

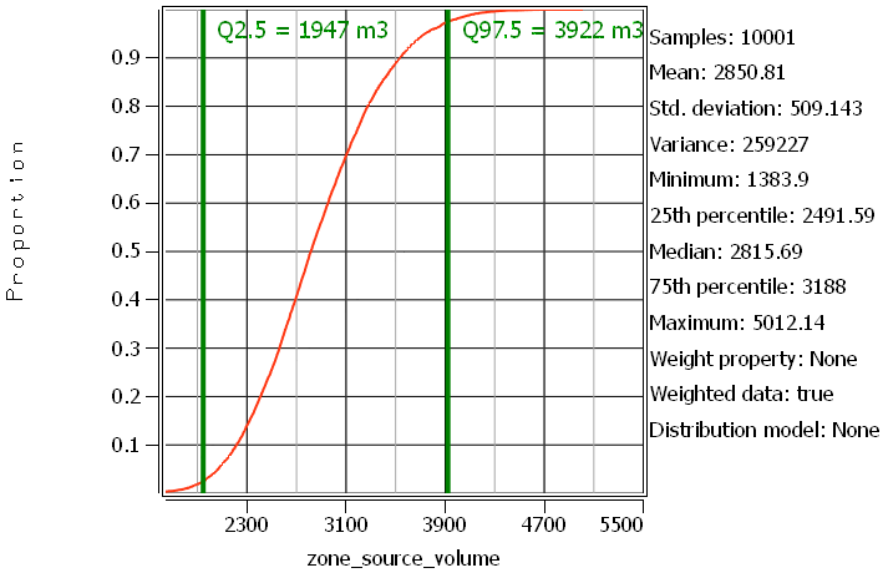


Figure 4: 95% confidence interval of [1 947; 3 022] derived from the cumulative distribution function (cdf) quantifying the estimation uncertainty about the volume of a soil contamination source. Q2.5 and Q97.5 denote the 2.5% and 97.5% quantiles, respectively, the true contamination source volume having 2.5% chances of being smaller than Q2.5 or higher than Q97.5.

From simple to advanced statistical methods

Bootstrap techniques (Efron 1993) have been a major progress to address statistical estimation problems in a simple way. They allow to numerically quantify the uncertainty about any or almost any statistics without having to make assumptions about the statistical models governing the studied phenomenon. Bootstrap techniques mainly consist in randomly resampling with replacement a dataset a large number of times. Each randomly drawn dataset is used to calculate the desired statistics. By repeating the statistical calculation on all randomly drawn datasets, as many statistics values are obtained as there are randomly drawn datasets. The set of statistics values can then be analyzed to quantify the uncertainty about the statistics estimated from the available dataset.

When applied to contaminated soil data, the particularity of this type of data is that they generally are spatially correlated: “two soil contaminant grades are more likely to be the same as the distance from each other decreases”. This is the basis of geostatistics which provides probabilistic tools to model spatial correlation and to take it into account to estimate local uncertainty (local cumulative distribution function) from spatially correlated data located in the neighborhood of the estimation point.

Spatial correlation is directly related to the notion of redundancy of information. Two spatially correlated data provide less information about a spatial phenomenon than two uncorrelated data because part of the information they provide is common (redundant). The closer data are to each other, the more redundant they are. Data redundancy can be seen as an issue for estimation purposes, part of the information being "paid twice" to acquire it from two different data. In reality, data redundancy is what makes it possible to estimate a spatial variable at a not sampled location from nearby spatially correlated measured data. Without spatial correlation or relationship between the data, no estimation of spatial phenomena would be possible.

Data redundancy, as associated with spatial correlation, is a concern for statistical analysis in the following situations: when preferential sampling is observed, as resulting from closer (clustered) data where soil contamination tends to be higher, when the boundaries of a delimited zone, regardless of the way the zone has been defined, must be taken into account to estimate the soil contamination of the zone, or when both situations occur (preferential sampling to characterize a delimited zone). In all cases, data redundancy must be corrected to estimate relevant soil contamination statistics and to quantify the uncertainty about them.

Data redundancy can simply be corrected by making use of a "declustering" method that consists in calculating weights attached to the data, the more redundant a data, the smaller the weight. Different declustering methods can be used, from the simplest one that does not need any user's input ("push button" method), to the most advanced one that requires some spatial statistical (variographic) analysis, but performs better. They all consist in somehow evaluating and translating the data layout, together with the delimited zone boundary, into weights that relatively measure the data redundancy.

The following methods can be used to calculate declustering weights (see Garcia *et al.*, 2014, Pyrcz & Deutsch, 2003). They are sorted from the easiest to the most advanced one.

- **Polyhedral declustering:** geometrical declustering only based on the data locations and the geometry of the delimited zone. The weight assigned to each data is then proportional to the volume of the closest neighborhood, i.e. the neighborhood where all points are closest to the data than to any other data. This weighting method can be considered as very similar to the usual practice that consists in applying the grade values measured on soil samples to the "volumes of influence" (or representative volumes) of the samples (see Figure 5). The higher is the volume of influence, the higher is the weight assigned to the data. This method does not need any user's input.
- **Cell declustering:** geometrical declustering based on a grid definition to calculate weights that are inversely proportional to the number of data falling into each grid cell. The higher is the number of data belonging to the same cell, the smaller is the weight assigned to all these data. In order to avoid grid effects due to a particular choice of grid (cell size, grid origin, grid orientation), cell size-dependent average weights are calculated by repeating the calculation of weights for different grid cell sizes, grid origins and possibly grid orientations. A scatterplot of cell declustered (weighted) contaminant grade mean versus grid cell size may then be helpful to check and interpret biased data (Figure 6). This method requires that the user defines the various grid definitions to test. The user must also decide about the grid cell size that is best to calculate relevant weights. Looking at the shape of the cloud of points may then be helpful to make good choices (see the example in Figure 6).
- **Bivariate declustering:** deriving weights from a correlated (explicative) variable that is exhaustively known (known everywhere within the studied zone), or at least much more sampled and known than the one to estimate. This method is very efficient provided that the correlation with the explicative variable is good enough. The explicative variable can then be seen as providing information about spatial trends of the estimated variable, and the declustering weights as correcting the trends. This method requires that a bivariate distribution model, relating the estimated variable to the explicative one, be inferred from the bivariate data.
- **Kriging-based declustering:** geostatistical declustering taking into account the spatial correlation and being able to include data outside the delimited zone. This method requires that variographic analysis (analysis and modeling of spatial correlation) be carried out. Local neighborhood parameters (neighborhood size and orientation, maximum number of conditioning data) may also need to be defined (kriging parameters).

Once declustering weights are calculated, an appropriate bootstrap technique must then be used to quantify the uncertainty about statistics estimated from weighted data.

The illustration of the approach on a case study is presented in the next section. The polyhedral and cell declustering methods are compared with each other and with a reference geostatistical model. Comparisons with the more advanced bivariate and kriging-based declustering methods will be presented in Garcia *et al.* (2014).

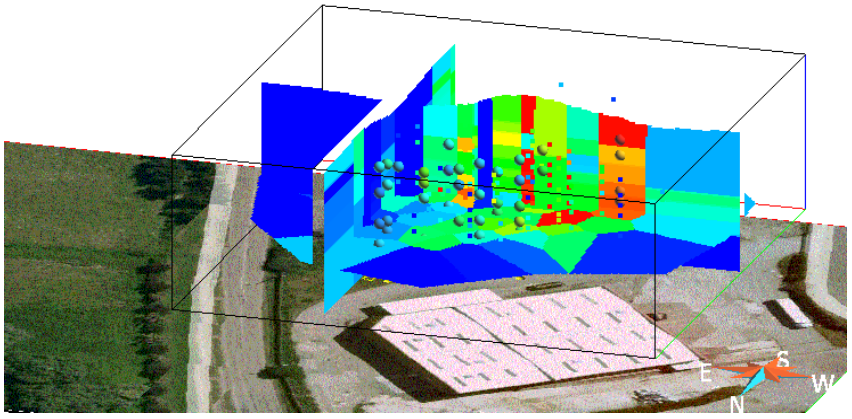


Figure 5: Three-dimensional view of volumes of influence (representative volumes) attached to sample contaminant grade data. The color scale is related to the contaminant grades.

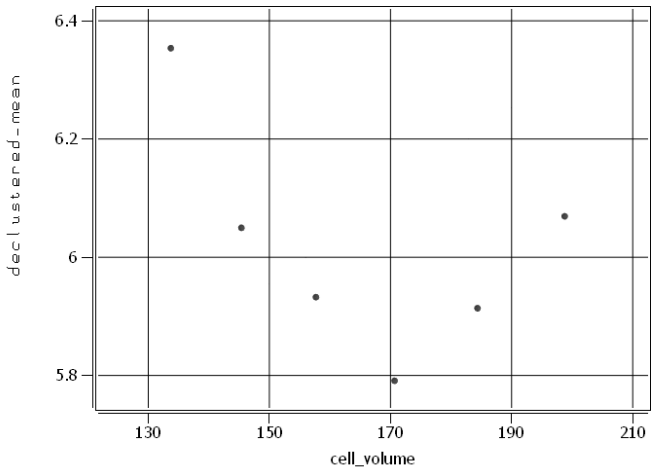


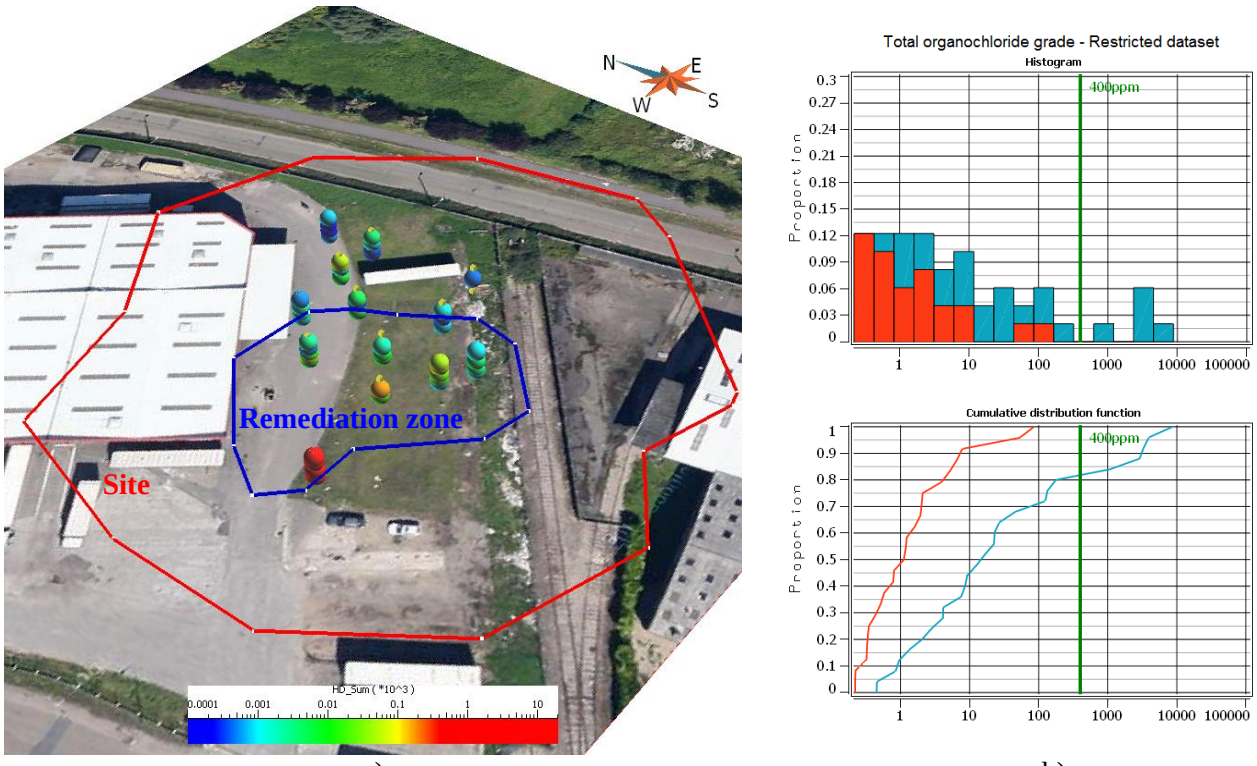
Figure 6: Scatterplot of cell declustered (weighted) contaminant grade mean vs. cell size. The U-shape of the cloud of points indicates that contaminated soils or zones have been sampled preferentially. An optimum grid cell size would then be the one minimizing the mean grade.

Application to an actual case study

The case study is a former storage, packaging and expedition unit for the chemical industry. The site was investigated to meet the needs of the ATTENA research project (Kaskassian *et al.*), supported by ADEME and led by the BRGM in partnership with INERIS, BURGEAP, TOTAL PETROCHEMICALS, ARCELOR MITTAL and RHODIA. The ATTENA dataset has also been used by the GeoSiPol working group as a demonstration case study about the application of geostatistics to contaminated sites and soils (GeoSiPol 2012). The objective of this case study was to quantify and delineate the soil contamination source where a mobile organic non aqueous phase liquid (NAPL) is present (contamination source). The NAPL is primarily composed of tetrachloroethene (PCE), trichloroethene (TCE) and 1,2cis-dichloroethene (DCE). It is considered as mobile (above residual saturation) if the overall soil organochlorine grade is greater than or equal to 400 ppm.

In this article, the experimental ATTENA dataset is used to define the following exploration and remediation assessment scenario.

- The potentially contaminated site is studied to delineate contamination zones that should be remediated to remove the soil contamination source. The area of the site is 6 000 m².
- Two successive exploration surveys are conducted:
  - a preliminary one based on 13 boreholes from which 49 soil samples are collected for lab analyses and 89 for on site PID measurements,
  - a supplementary one providing 238 additional lab analyses and 813 PID measurements on soil samples from 33 new boreholes and the ones from the preliminary survey (collected soil samples not analyzed at the beginning).
- Based on the exploration survey data and historical information about the industrial activity on the site, the soil contamination study concludes that a single zone needs to be remediated to remove most of the contamination source.
- Irrespective of the method used to delineate the remediation zone (geostatistical approach as in GeoSiPol, 2012, engineering approach as in Mathieu *et al.*, 2014), the environmental engineer in charge of the study is asked to provide an assessment of in-place mass of organic contaminants within and outside the delimited zone and to quantity the uncertainty. These results are required to decide about the remediation strategy and costs.
- Two situations are compared, whether the estimations are made at the end of the preliminary survey (based on the **restricted dataset**), or after the supplementary one (based on the **full dataset**).



a) b)

Figure 7: Information about the **restricted dataset** with a) a 3D view of the soil samples analyzed in laboratory, and b) statistical plots (histogram and cumulative distribution function) of overall organochlorine grade data. The color and size of the soil sample spheres in the 3D view are related to the overall organochlorine grade measured on the samples (color defined by the color scale, size increasing with the grade). The statistical plots differentiate the data that are located inside the remediation zone (blue histogram bins and cdf line) and outside (red histogram bins and cdf line).

Figure 7 and Figure 8 show the 3D location of overall organochlorine grade data (from lab-analyzed soil samples) and statistical plots (histograms and cdf) for the two restricted and full datasets, respectively. The corresponding statistical summaries are given in Table 1. All statistics are for unweighted data, i.e., without correcting any data redundancy and preferential sampling. The 3D views also display the boundaries of the site (red line) and the remediation zone (blue line) for which in-place mass of contaminants must be

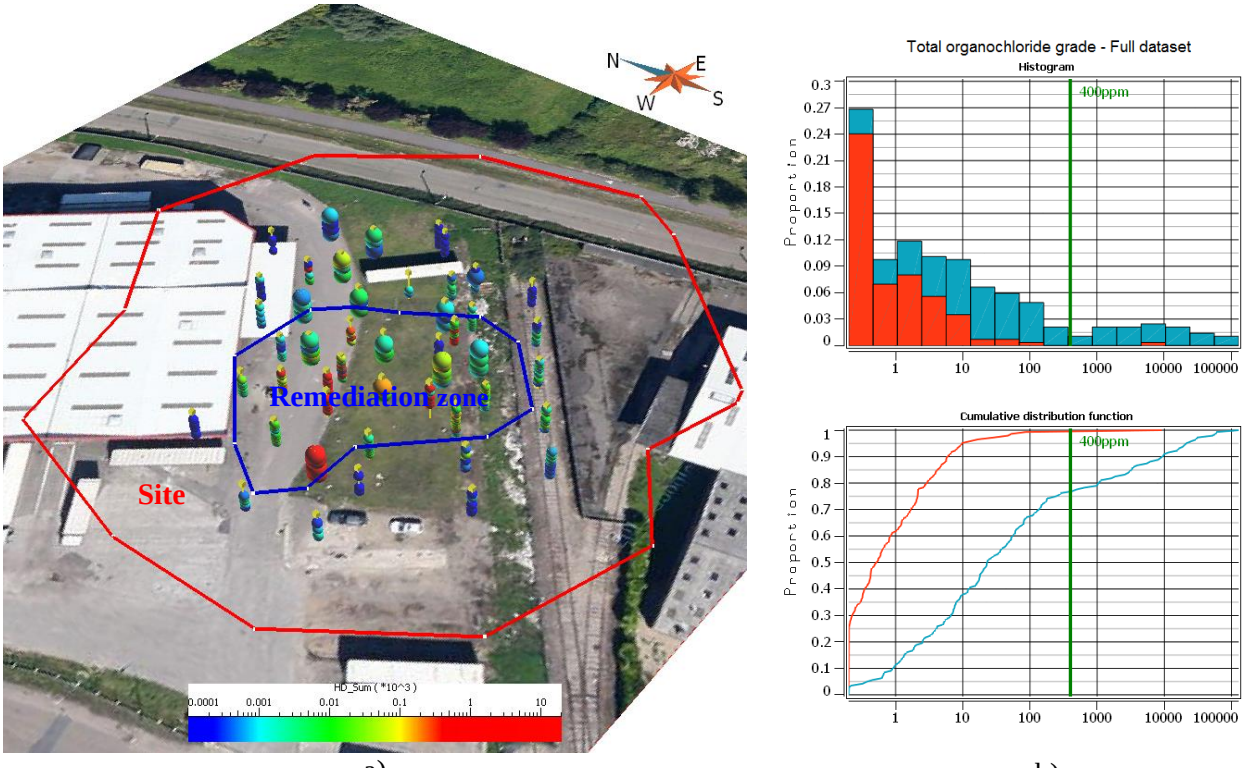
estimated. The statistical plots differentiate the data located inside the remediation zone (blue histogram bins and cdf line) and outside (red histogram bins and cdf line).

It can already be seen from the statistical summaries of Table 1 that the two datasets do not give the same picture of the soil contamination. Indeed, the full dataset reveals much higher grade values than the restricted one: the maximum grade is about 15 times higher and the mean 5 times. Such a difference can result from preferential sampling (contaminated soils more frequently sampled), but partly also from the fact that the distribution of grade is strongly positively skewed with a small proportion of very high values (coefficient of variation > 1). In both cases, it is already an indication that precautions must be taken to estimate statistics.

Table 1: Statistical summaries of overall organochlorine grade data (ppm) for the two datasets. CV = coefficient of variation (std dev / mean) that measures the distribution skewness of a positive variable.

Dataset	Nb data	Mean	Std dev.	CV	Min	Q25	Median	Q75	Max
Restricted dataset	49	425	1 460	3.4	0.2	0.8	3.4	23	8 800
Full dataset	287	2 010	10 200	5.1	0.2	0.4	2.7	30	129 000

The threshold lines of 400 ppm, depicted on the cdf plots of Figure 7 and Figure 8, allow to evaluate the proportions of data that are above the threshold: 18% in the remediation zone and 0% outside for the restricted dataset, 23% in the remediation zone and very few (less than 1%) outside for the full dataset. The data being possibly biased due to preferential sampling and data redundancy, these proportions should, however, be taken cautiously. The histogram of Figure 8.b shows in which (red) bin the high grade data, seen by the full dataset outside the remediation, is located (just below 10 000 ppm).



a) b)

Figure 8: Information about the **full dataset** with a) a 3D view of the soil samples analyzed in laboratory, and b) statistical plots (histogram and cumulative distribution function) of overall organochlorine grade data. The color and size of the soil sample spheres in the 3D view are related to the overall organochlorine grade measured on the samples (color defined by the color scale, size increasing with the grade). The statistical plots differentiate the data that are located inside the remediation zone (blue histogram bins and cdf line) and outside (red histogram bins and cdf line).

Based on the above scenario, the benefits of using simple or more advanced statistical tools at the following steps of the study are discussed.

- Estimation step: estimating the in-place mass of organochlorines within the remediation zone.

- Uncertainty quantification step: quantifying the uncertainty about the estimated in-place mass of organochlorines.

To do so, comparison of results are made for statistics estimated from unweighted data (no correction of biased data set due to preferential sampling and data redundancy) and from weighted data for weights calculated using different polyhedral and cell declustering methods. The quantification of (spatial) uncertainty about in-place mass of organochlorines is obtained by applying a bootstrap technique to the different sets of unweighted and weighted data. Each dataset is resampled 10 000 times. The uncertainty results are then compared with geostatistical results that are supposed to be a reference.

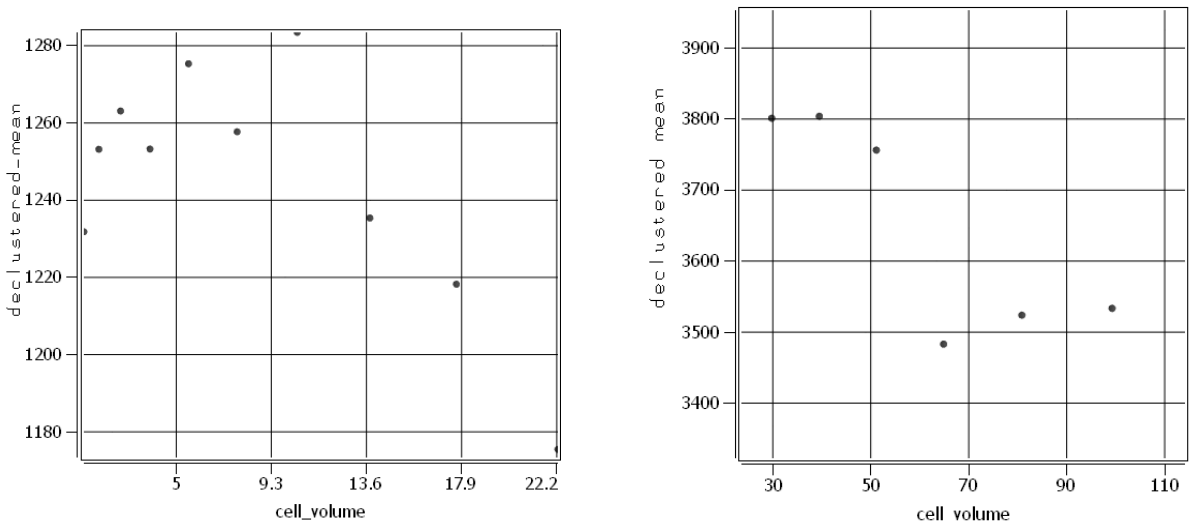
Estimation step

The aim here is to estimate the in-place mass of organochlorines within the remediation zone. Depending on the available data, two methods can be used.

- Accurate method: by directly estimating the average mass of organochlorine per unit volume to account for spatially varying soil density. This requires, however, that both contaminant grade and soil density data are available for all lab-analyzed samples to calculate data of organochlorine mass per unit volume as the product of organochlorine grade times soil density. The average mass of organochlorine per unit volume must then be multiplied by the volume of the remediation zone.
- Approximate method: by estimating the average organochlorine grade over the remediation zone and by multiplying it by a relevant (average) soil density value and the volume of the zone.

The latter and most common method is applied here using a mean soil density of 1.55 t/m^3.

Estimating the average organochlorine grade over the remediation zone is an estimation problem that requires to correct (to weight) the organochlorine grade data from possible bias due to preferential sampling and data redundancy. The two polyhedral and cell declustering methods have been applied to compute weights and compare statistics (results from the more advanced bivariate and kriging-based declustering methods will be discussed in Garcia *et al.*, 2014). Results from these declustering methods are compared with geostatistical estimations from the GeoSiPol demonstration study (GeoSiPol, 2012)



a) Restricted dataset. b) Full dataset.

Figure 9: Scatterplots of cell declustered (weighted) mean grade vs. cell size for the data located within the remediation zone from a) the restricted dataset and b) the full dataset. The hump-shape of the restricted dataset can be interpreted as resulting from a preferential sampling of not or poorly contaminated soils, while the U-shape of the full dataset is to be related to preferentially sampled contaminated soils.

The efforts required to implement the different declustering methods are recalled below (see also previous section).

- Common to both methods: delineation of the remediation zone to be evaluated.

- Specific to the polyhedral declustering method: simplest method that does not need any user's input.
- Specific to the cell declustering method: method of little complexity that only requires to test several grid cell sizes until one allows to identify a possible preferential sampling situation. Looking at scatterplots of declustered (weighted) contaminant grade mean versus grid cell size (Figure 9), it appears that the restricted dataset preferentially samples poorly contaminated soils within the remediation zone (hump-shaped cloud of points), while the full dataset shows a preferential sampling of contaminated soils (U-shaped cloud of points).

These efforts are to be compared with those required to carry out the geostatistical approach: uni and multivariate statistical analyses, variographic analysis (analysis and modeling of spatial correlation), geostatistical (co)simulation of the three correlated organochlorine contaminants, and some post-processing to extract and average simulated contaminant grade values. It must be noted, however, that the geostatistical approach directly provides a quantification of the uncertainty, which is not the case of the declustering methods that requires to proceed with bootstrap analysis (see next section).

The calculation of declustering weights cannot be done manually but requires to use an efficient and ideally user-friendly software. Soil Remediation *Data Analysis*, developed by KIDOVA, was used here.

Table 2 presents the in-place mass of organochlorines estimated for the remediation zone, from the restricted and full datasets, by considering unweighted (raw) data, weighted data declustered to correct them from preferential sampling and data redundancy, and geostatistical simulations (demonstration study from GeoSiPol, 2012).

Table 2: In-place masses of organochlorines (in t), estimated for the remediation zone, from the restricted and full datasets, by considering unweighted (not declustered) data, weighted (declustered) data from polyhedral and cell declustering methods, and geostatistical simulations.

Dataset	Unweighted data	Polyhedral declustering	Cell declustering	Geostatistical simulation
Restricted dataset	5.6	8.0	8.7	25.3
Full dataset	27.2	24.7	23.3	20.0

Having in mind that better estimates should be obtained with the full dataset and that the geostatistical model should be the reference (better and more complete integration of the data), the expected in-place mass of organochlorines should be of about 20 t. The following comments can then be made about the other estimation results.

- As expected, the unweighted data sets are biased and give the worst predictions: the in-place mass of organochlorines is greatly underestimated with the restricted dataset (5.6 t), which preferentially samples poorly contaminated soils, and greatly overerestimated with the full dataset (27.2 t), which preferentially samples contaminated soils.
- As also expected, the two declustering methods correct, at least partly, the statistical biases and improve the estimations by predicting higher masses from the restricted dataset, and smaller masses from the full dataset. Slightly better results are obtained with the cell declustering method.
- Applying the declustering methods to the restricted dataset, the discrepancy between the reference in-place contaminant mass of 20 t and the estimated ones are still very important (up to 2.5 times smaller).
- The full dataset allows to highly improve the estimations based on declustered (weighted) statistics (overestimation of 16.5% with the cell declustering method).
- As already mentioned before, the major difference between the two datasets is the much higher grade values measured on the full dataset. Especially, the maximum grade is already about 15 times greater (see Table 1). The calculation of in-place mass of organochlorines is then necessarily sensitive to the definition of extreme values (distribution tail), which is a complex issue in general, and even more for small datasets.
- Only the geostatistical method appears to provide relevant estimations from the restricted dataset (estimated mass 25% higher than the reference one). The performances of the geostatistical model strongly relies, however, on the reliability of statistical models that must be inferred from the restricted dataset to be representative of the whole site. As discussed in GeoSiPol (2012) and above, the restricted

dataset provides an incomplete image of the soil contamination. Additional data or assumptions are required to properly simulate the correlated contaminant grades.

Having in-place masses of organochlorines estimated, the next step is to quantify the uncertainty about them.

Uncertainty quantification step

A bootstrap technique able to account for unweighted or weighted data is used to quantify the uncertainty about estimated in-place mass of organochlorines within the remediation zone (see Garcia *et al.*, 2014). Each unweighted and weighted data set is resampled 10 000 times, thus leading to 10 000 estimated values of in-place mass of organochlorines from which statistics can be calculated to derive a 95% confidence interval (95% chances that the true in-place mass value is inside the interval).

The bootstrapped confidence intervals are compared with each other and with the reference confidence interval calculated from the 100 geostatistical (stochastic) simulations of soil contaminant grades (demonstration study from GeoSiPol, 2012).

The uncertainty results quantified by bootstrap or geostatistical simulation are presented in Table 3 for the restricted dataset and Table 4 for the full dataset.

Table 3: Uncertainty about in-place mass of organochlorines within the remediation zone (in t), estimated by bootstrap on unweighted and weighted data and by geostatistical simulation (GeoSiPol, 2012) from the **restricted dataset**.

Method	Nb data	Estimated (mean)	95% confidence interval
Bootstrap on unweighted data	25	5.6	[1.3; 11.5]
Bootstrap on polyhedral declustered data	25	8.0	[1.6; 16.9]
Bootstrap on cell declustered data	25	8.7	[1.9; 17.3]
Geostatistical simulation	49	25.3	[5.3; 63.9]

Table 4: Uncertainty about in-place mass of organochlorines within the remediation zone (in t), estimated by bootstrap on unweighted and weighted data and by geostatistical simulation (GeoSiPol, 2012) from the **full dataset**.

Method	Nb data	Estimated (mean)	95% confidence interval
Bootstrap on unweighted data	143	27.2	[12.9; 45.4]
Bootstrap on polyhedral declustered data	143	24.7	[12.6 ; 40.5]
Bootstrap on cell declustered data	143	23.3	[11.3; 36.1]
Geostatistical simulation	287	20.0	[11.6; 32.9]

Keeping in mind that the geostatistical model based on the full dataset is the reference solution, the following comments can be drawn from the 95% confidence intervals of Table 3 and Table 4.

- As expected, the uncertainty is higher for the restricted dataset (max to min ratio ≥ 9), than for the full dataset (max to min ratio < 3.5). This is also true with the geostatistical models. **In that sense, all approaches, either based on bootstrap resampling or geostatistical modeling, tend to show that the restricted dataset should not be sufficient by itself to precisely estimate the soil contamination**.
- With the restricted dataset, none of the confidence intervals estimated by bootstrap reaches the expected in-place mass of 20 t predicted by the most complete (reference) geostatistical model. Underestimation is observed, as already noticed at the estimation step. The declustered data sets, corrected from preferential sampling and data redundancy by weighting the data, allow, however, to greatly extend, hence to improve, the estimated confidence intervals. In this example, best results are obtained with the cell declustering method.
- Using the restricted dataset, only a geostatistical approach, which relies on spatial correlation models and can integrate additional information from data outside the remediation zone, would be fully efficient. The alternative would be to use more advanced kriging-based or bivariate declustering methods, together with bootstrap technique, to integrate more information (purpose of Garcia *et al*, 2014).
- With the full dataset, all unweighted and weighted data provide better and consistent confidence intervals. As already observed at the estimation step, the unweighted data, not corrected from the preferential sampling of contaminated soils, keep being those that give the largest overestimation of in-

place mass of organochlorines, followed by the polyhedral declustered data and the cell declustered ones. **Compared with the geostatistical simulation results, the uncertainty derived from the cell declustered data appears as being a relevant or good enough approximation of the uncertainty for decision making.**

- It must be noted that the full dataset has not been designed to supplement the restricted dataset, but results from experimental objectives of the ATTENA project. In other words, the additional 238 data would have been better located elsewhere if the primarily objective was to decrease the spatial uncertainty. This is the reason why a rather high uncertainty prevails from the restricted to the full dataset.

Conclusion

It has been showed that statistical methods can easily be understood and implemented to analyze soil contamination data and to derive from them a relevant quantification of the uncertainty about in-place mass of contaminants, or other soil contamination assessments, attached to delimited remediation zones. This requires, however, to properly correct the data from preferential sampling and data redundancy, yet taking into account the geometry of the remediation zone. To do so, different declustering methods are available, some being simpler than others, the more advanced the method is, the better generally it performs.

As expected, the uncertainty about soil contamination assessments is directly related to the size of the dataset and the spatial distribution of the data. The proposed approach can give imprecise and possibly biased results with too small or not ideally defined datasets. Nevertheless, its application is enough to identify highly uncertain situations that would require additional information to provide reliable and precise enough estimations.

Compared with the more thorough but also more demanding geostatistical approach, the proposed statistical methods appear as a good alternative at the early stage of a soil contamination study to generate preliminary uncertainty quantification results, but also to help building more reliable geostatistical models as part of the statistical analysis work.

References

Chilès, J.-P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*, John Wiley & Sons: New York.

Deutsch, C.V. and Journel, A.G. (1997). *GSLIB: Geostatistical Software Library and Users Guide*, Oxford University Press, New York, second edition.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall/CRC. ISBN 0-412-04231-2

Garcia, M.H., Mathieu, J.-B. and Garcia, V. (2014). Methodological and practical aspects of geostatistical bootstrap for quantifying global and local soil contamination uncertainty, accepted at *geoENV 2014*, Paris, July 2014.

GeoSiPol (2012). *Etudes de démonstration de l'intérêt de la géostatistique dans le domaine des sites et sols pollués*, GeoSiPol[1].

GeoSiPol (2005). *Géostatistique appliquée aux sites et sols pollués – Manuel méthodologique et exemples d'application*, GeoSiPol[1].

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*, Oxford university press, New York.

Harris, P., Brundson, C., Charlton, M., Juggins, S. and Annemarie, C. (2014). Multivariate spatial outlier detection using robust geographically weighted methods, *Math Geosci* (2014) 46:1-13.

Isaaks, E.H. and Srivastava, R. M. (1989). *Introduction to applied geostatistics*, Oxford University Press.

Kaskassian, S., Gleize, T., Chastanet, J. and Côme, J.-M. (2013). *Projet ATTENA – Phase 2, Tâche 3.1.2 : mise en œuvre des guides méthodologiques MACAOH par BURGEAP sur le Site 1bis (solvants chlorés)*, rapport final, 3 vol., 296 p, www.attena.org.

Mathieu, J.-B., Kaskassian, S. and Garcia, M.H. (2014). Apport de la géostatistique au diagnostic des sites et sols pollués : prolongement d'un cas d'étude de démonstration GeoSiPol, submitted at *3ème Rencontres Nationales de la Recherche sur les Sites et Sols Pollués*, ADEME, Paris, novembre 2014.

Pyrcz, M. J. and Deutsch, C. V. (2003). Declustering and debiasing, in *Searston S(ed) Newsletter 19*, October 2003, Geostatistical Association of Australasia, Melbourne.

Acknowlegement

We thank BURGEAP for providing us with the ATTENA research project dataset. We are also grateful to GeoSiPol for allowing us to present results from the demonstration geostatistical studies.

[1] Association and working group co-founded by KIDOVA, Geovariances and Mines Paristech, see www.geosipol.org.