European Journal of Soil Science

*European Journal of Soil Science*, June 2011, **62**, 454–466 doi: 10.1111/j.1365-2389.2011.01374.x

# From exploratory data analysis to geostatistical estimation: examples from the analysis of soil pollutants

C. de Fouquet

*MINES ParisTech, Géosciences-géostatistique, Ecole des mines de Paris, 35 rue Saint-Honoré, 77305 Fontainebleau Cedex, France*

## Summary

Data collected during sampling of the soils of former industrial waste lands are rarely scrutinized closely. However, exploratory analysis is an essential stage, allowing, *inter alia*, the detection of possible heterogeneities on the site, examination of the vertical variation in concentrations and detection of possible gaps in the survey. If available and identified, duplicated measurements can be used to characterize the magnitude of measurement errors. Subsequent spatial analyses with variograms aim to characterize and to quantify the spatial variability. For hydrocarbon pollution in soils, the spatial variability at small distances is very large and results in large uncertainties in the estimates. In this context, the usefulness and the limitation of linear estimation (kriging) are reviewed and illustrated. When mapping polluted zones defined by a quality threshold, such as a regulatory limit on acceptable concentrations, the kriged concentration map can be combined with the associated kriging standard deviation map to identify areas in which the uncertainties make it impossible to decide whether concentrations are greater or smaller than the quality threshold. Examples of exploratory and variogram analysis are given, followed by linear estimation of concentrations and comparison with a threshold, using data on hydrocarbon pollutants.

## De l’analyse exploratoire á l’estimation géostatistique pour des pollutions organiques des sols

### Résumé

Les données acquises lors de l’échantillonnage des sols de friches industrielles sont rarement examinées avec attention. Or l’analyse exploratoire est riche d’enseignements : mise en évidence de possibles hétérogénéités du site, examen de la variation verticale des concentrations, détection de possibles lacunes de la reconnaissance, utilité des mesures redondantes pour caractériser l’amplitude des erreurs de mesure, etc. L’étape suivante d’analyse variographique permet de caractériser la variabilité spatiale (anisotropies …) et de la quantifier. Pour des pollutions de sols par des hydrocarbures, la variabilité à petite distance apparaît très élevée, ce qui induit de fortes incertitudes sur l’estimation. Dans ce contexte, l’intérêt et les limites de l’estimation linéaire par krigeage sont rappelés et illustrés. Les zones polluées sont définies comme le lieu où les concentrations réelles (et dont le support est spécifié) dépassent une valeur limite ou ≪ seuil ≫. Combinée à la carte des concentrations estimées, la carte de l’écart-type de l’erreur d’estimation permet de délimiter les zones d’incertitude par rapport au dépassement de seuil, à un risque statistique près. Des exemples de pollution par des hydrocarbures illustrent les principales étapes de l’analyse exploratoire et variographique, ainsi que l’estimation linéaire et la comparaison à un seuil.

## Introduction

The soil, whether from agricultural land or former industrial sites, is a complex environment and its properties can vary strongly over short distances. Sampling is therefore needed to determine its properties in any particular parcel of land. Geo-referenced data thus obtained may be used for different purposes, including: (i) statistical characterization of variables such as physical properties or concentrations of substances, and sometimes information on their variation down the soil profile, or joint variation of two substances; (ii) local estimation, to draw maps of soil characteristics as accurately as possible; (iii) delimitation of zones where a variable exceeds, or is smaller than, a threshold value, in

Correspondence: C. de Fouquet. E-mail: chantal.de_fouquet@mines-paristech.fr

Received 11 December 2009; revised version accepted 5 April 2011

order to apply a fertilizer treatment or to implement remediation, for example and (iv) subdivision of land 'into parcels to be managed more or less differently according to these changes in the soil' (Webster & Oliver, 1990).

The data, which are expensive to collect, often only serve as an input to 'black box' models, to produce statistics or maps automatically. However, these uncritical treatments do not make full use of the data. With the help of simple statistical or geostatistical tools, exploratory and variogram analysis can additionally provide a synthetic description of the variable studied. Exploratory stages are difficult to formalize by a general flow chart and are often neglected, and in this paper I try to show their usefulness. The results of exploratory analysis are then used to guide modelling, including variogram fitting and estimation.

The context is soil pollution by hydrocarbons on former industrial sites. This differs from an agronomic context in that the studied variables are mainly pollutant concentrations, sometimes only one substance may be considered (a univariate case) and the sampling design is generally irregular or sparse. Diversified and successive activities located at various locations on the site, infrastructure such as roads or underground pipes and other installations generate heterogeneities in concentrations. Finally, during site decommissioning, levelling of the site and the use of infill material from various sources can further modify the spatial distribution of pollutants.

In this paper these problems, and solutions, are demonstrated in the context of three case studies. The principles and tools of exploratory analysis are reviewed. An example of kriging estimation is given and the consequences of uncertainties for the comparison with a quality threshold are discussed.

## Study sites

### *Former petrochemical complex*

At the request of the data supplier, only restricted information can be made available. The former petrochemical complex covers approximately 240 ha. The site survey consisted of about 60 sampling points, separated by some 150 m (Figure 1). In the zone of the former production units, sampling was locally intensified to an approximately 35-m spacing (about 130 sampling points). At most sampling points, two samples were taken, a 'surface' and a subsurface one (other details of sampling are not known).

The statistics for the concentrations of hydrocarbons at this site are summarized in Table 1. The empirical mean concentration is slightly larger in the zone of the production units (about 1600 mg kg$^{-1}$) than across the whole site (about 1350 mg kg$^{-1}$), but the vertical variation of concentrations is different. Across the whole site the mean subsurface concentration (1800 mg kg$^{-1}$) is twice as large as that of the surface (900 mg kg$^{-1}$), whereas the contrast between surface and subsurface concentrations is attenuated in the zone of the production units with respective means of about 1700 and 1500 mg kg$^{-1}$. In this zone mixing of material from the two depths might explain these differences (de Fouquet *et al*., 2004). The summary statistics show the complexity of this site.

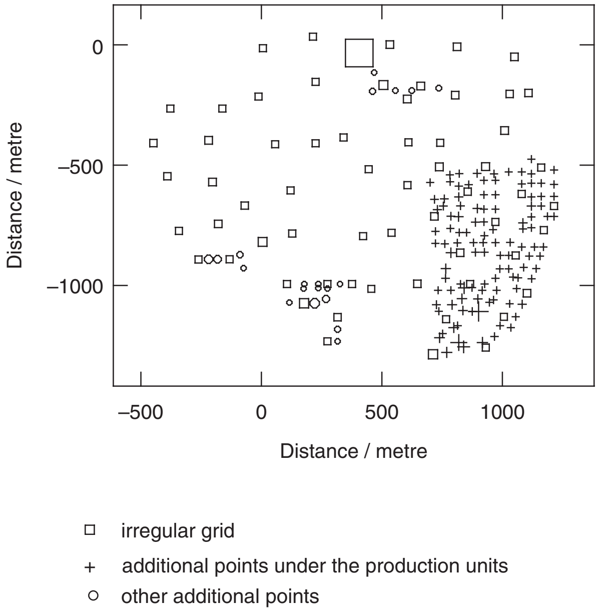


**Figure 1** Sampling design of the former petrochemical complex. Symbol size is proportional to subsurface sample concentration. The local origin of the coordinates is arbitrary.

### *Two sites from the LOQUAS project*

The LOQUAS project aimed to assess measurements of hydrocarbon concentrations in soils made in the field with the Pollut-Eval® system, based on pyrogram analysis (Blanchet *et al*., 2005; Benoit *et al*., 2008). The Pollut-Eval® analysis is performed on a small mass of soil (about 100 mg), which raises the questions of how to obtain a representative sample, and whether it is acceptable to form composite samples. In order to answer these questions it is necessary to characterize the spatial variability from scales of centimetres to tens of metres. Two sites were investigated using multiscale sampling on embedded grids.

*LOQUAS site 0.* A region of about 140 m$^2$ on a former test site for aircraft engines was sampled after the first 0.20 m of soil had been removed. The horizontal nested sampling scheme consists of a basic grid with 6-m spacing, locally intensified to 3.0, 1.50, 0.50 and 0.25-m intervals (Figure 2). The primary sample (Gy, 1975) consisted of a small soil core of depth 1.5 cm and square section with sides of approximately 10 cm. The sections were sampled according to a systematic 'reference' pattern, which is a set of five 'Pollut-Eval® points' located at the corners and centre of an 8-cm sided square. In addition, six of these sections were more intensively sampled by 25 'Pollut-Eval® points' spaced at 2-cm intervals.

**Table 1** Statistical summary of pollutant concentrations on the common data points for both depth intervals of the former petrochemical complex

| Area | Depth | Number of data | Min. | Max. | Median | Mean | Standard deviation | Coefficient of variation |
|---|---|---|---|---|---|---|---|---|
| Approximately 150-m grid | Surface | 63 | 1 | 15 935 | 120 | 900 | 2590 | 2.9 |
| | Subsurface | | 20 | 53 415 | 270 | 1810 | 6695 | 3.7 |
| Area of former production units (PU) | Surface | 127 | 1 | 32 005 | 250 | 1480 | 4090 | 2.8 |
| | Subsurface | | 1 | 31 410 | 210 | 1700 | 3965 | 2.3 |

Units are mg kg$^{-1}$ (Min. = minimum; Max. = maximum).

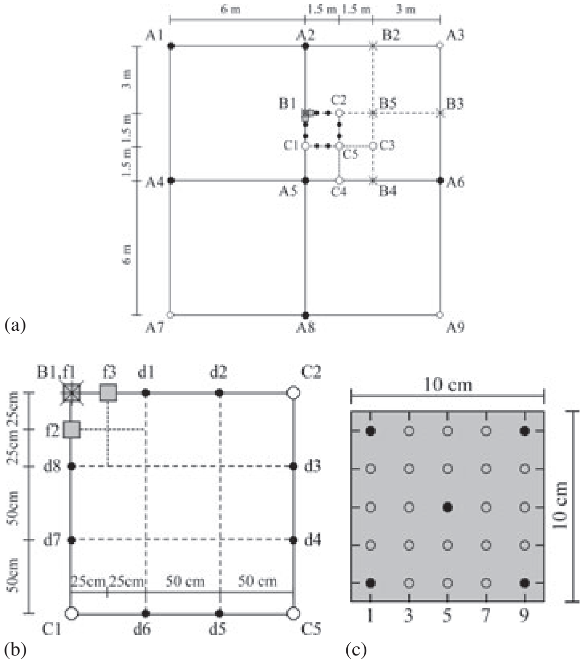


**Figure 2** Site LOQUAS 0. (a) Multi-scale grids, (b) zoom on the central zone and (c) detail of the five-point pattern (full circles) and of the six quasi-exhaustive 25-point patterns (A3, A7, A9, B1 = f1, f2, f3) on the 10 cm × 10 cm soil sections. The local origin of the coordinates is arbitrary.

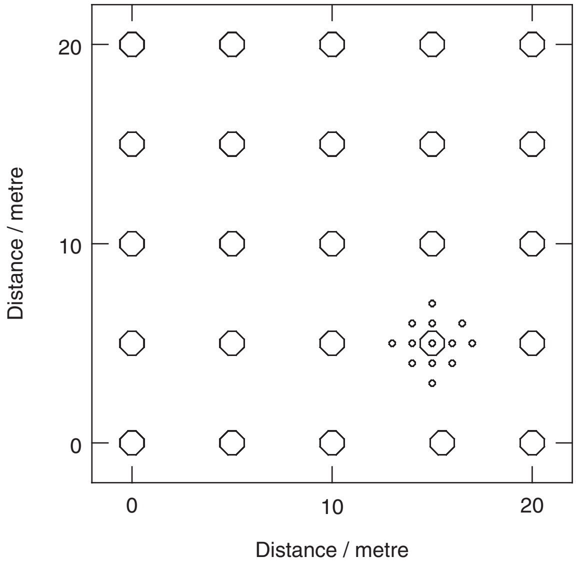


**Figure 3** Site LOQUAS 2. Five-metre horizontal grid and local 1-m grid. Each circle represents the location of one vertical borehole of 4-m length, cut into four 1-m cores. The local origin of the coordinates is arbitrary.

Statistics were calculated with only nine data. From the basic 6-m grid, the mean of the nine sections (empirical mean concentration on the five-point patterns) is about 1300 mg kg$^{-1}$, with a coefficient of variation of 70% (Table 2). Mean and variance increased with scale from 6-m to the 3-m to 1.5-m grids, but the coefficient of variation did not change (Table 2). This might indicate the presence of a proportional effect, which is discussed later.

*LOQUAS site 2.* A small region of about 400 m$^2$ at a former commercial fuel station was sampled with vertical drillings located at 25 nodes on a horizontal 5-m square grid, with some gaps in this sampling regime (Figure 3). Vertical 4-m drill holes were cut into four 1-m sections, which were homogenized on site. Three samples were taken from each homogenized core. The 97 data are the empirical means of these three Pollut-Eval® measurements.

**Table 2** Statistical summary of pollutant concentrations on the multiscale grids from the LOQUAS site 0

| Grid size (m) | Number of data | Min. | Max. | Median | Mean | Standard deviation | Coefficient of variation |
|---|---|---|---|---|---|---|---|
| 6 | 9 | 530 | 3700 | 1040 | 1300 | 910 | 0.7 |
| 3 | 9 | 840 | 3700 | 1040 | 1660 | 1080 | 0.7 |
| 1.5 | 9 | 830 | 3700 | 970 | 1700 | 1160 | 0.7 |
| 0.5 | 12 | 400 | 13 670 | 1290 | 2650 | 3510 | 1.3 |

Means are on a five-point pattern. Notations are as used in Table 1. Units are mg kg$^{-1}$ (Min. = minimum; Max. = maximum).

**Table 3** Statistical summary of concentrations for two grids of the site LOQUAS 2. Four-metre vertical drillholes, cut into 1-m cores

| Grid (m) | Number of elementary measurements | Number of data | Min. | Max. | Median | Mean | Standard deviation | Coefficient of variation |
|---|---|---|---|---|---|---|---|---|
| 5 m | Three on homogenized core | 97 | 130 | 1590 | 520 | 580 | 320 | 0.54 |
| 1 m | Four in the middle of non-homogenized core | 49 | 120 | 1710 | 390 | 470 | 360 | 0.76 |

The grid refers to the horizontal design. Data are the means of, respectively, three or four measurements, depending on the grid. Units are mg kg$^{-1}$.

The overall site mean is 580 mg kg$^{-1}$, with a coefficient of variation of 55% (Table 3).

In some areas, the horizontal sampling was intensified to 1-m intervals, in order to compare vertical and horizontal variability. The additional data are the means of Pollut-Eval® measurements of four samples taken in the horizontal section in the middle of the vertical 1-m core, before homogenization. The additional data are thus located on a three-dimensional 1-m grid. Their mean is smaller and their standard deviation larger than those of the 5-m grid.

## Exploratory analysis

The aim of exploratory analysis is to understand how the variable studied is distributed on the site. This stage is an essential first step to guide the subsequent modelling or, even if there is no subsequent modelling, to obtain a summary description of the pollution.

### *Preliminary stages*

The site history indicates which pollutants should be investigated. The plans of the former facilities indicate possible areas of large concentrations. Site history provides useful but often incomplete supporting information: accidental spills and temporary storage of polluted materials have generally been forgotten. A geophysical survey can locate buried objects such as previous infrastructure components and other large materials such as barrels. However, the results are not directly related to the concentrations of pollutants and their spatial variability depends on depth, the acquisition technique and the method used for the return of previously removed soil. Geophysical survey results should not be used as an approximate map of concentrations. As with the site history, these suggest the possible presence of pollution, but are less reliable as indicators of where pollution is absent.

On polluted sites the initial sampling scheme is usually irregular and may have been intensified in areas where larger concentrations were expected. Statistics calculated from such preferential sampling cannot be extrapolated *a priori* to the whole site. Statistics calculated from a regular sampling design are more reliable and a subsequent systematic sampling may identify other areas with large concentrations on the site. Exploratory analysis is divided into the following three stages.

*Visualization of the data.* After coding available information (lithology, drill-hole length and classes of concentration), mapping the data helps to answer important questions. (i) Sampling characteristics: is sampling regular, or are irregularities or intensifications of the sampling deliberate in some regions, which might introduce bias? (ii) Vertical extension of the survey: what is the maximum depth and why (were there drilling problems, changes of the lithology, etc.) and are some depths less fully sampled? (iii) Does sampling differ according to some criteria, for example lithology (superficial infill material or underlying soil)? (iv) When successive sampling campaigns were carried out, how were they organized?

Some other information is also essential for interpreting the data, including whether the samples were simple or composite, whether the materials were homogenized or not before sampling, whether the sampling technique was similar for all the data and whether they have the same support (the size and shape of the volume of soil to which the data correspond). For example, on the LOQUAS 2 site, the preparation of samples (taken before the cores were homogenized or after) and their supports differed between the 5-m and 1-m grids. We also need to know whether the samples were located in the saturated or unsaturated zone.

*Statistical summary and histograms.* The statistical summaries (minimum and maximum values, mean, standard deviation and coefficient of variation, and percentiles) can be calculated overall and for some specific classes such as lithology or depth interval. When the concentrations range over several orders of magnitude, it is convenient to examine log-transformed values.

The statistical summary provides information about the site characteristics such as (i) the range and variability of the concentrations and the criteria according to which these vary and (ii) the presence of one or more populations, in particular when the histogram is multimodal. In this case, we need to know whether the different classes or modes are separated or mixed on the site.

The empirical variance represents a dispersion variance; that is, the variance of one data point uniformly taken from the dataset. This variance depends on the support of the data and on their relative location (Chilès & Delfiner, 1999). Except for the mean, all characteristics of the distribution linked with dispersion also depend on the support of the samples and on the extent of the sampled domain. For this reason, data obtained from different measurement techniques or with different supports should not be mixed for statistical calculations.

*Scatter diagrams.* Scatter diagrams show how one variable varies with another one, for example, how concentrations vary with depth. Plotting the empirical regression (the empirical mean of the ordinate by class of the abscissa) makes it possible to examine whether the link between the variables is linear or not. When the data are represented by the same symbols or colours on the different figures, simple plots such as location map, histograms and scatter diagram complemented by the empirical regression are very informative.

When calculated from regular sampling, the scatter diagram plotted between local mean and variance allows us to detect and identify any relationship between these two quantities (Chilès & Delfiner, 1999). The proportional effect is presented and discussed for the site LOQUAS 0 in a subsequent section.

### *Consequences of sampling characteristics*

What the data can or cannot show depends on how the sample locations were selected and on the depth sampled. Because of the spatial correlation, the data should be regularly distributed over the domain so that statistics reflect reality.

Empirical quantities (mean and variance) can be seen as a discrete approximation of associated 'regional quantities' (Chilès & Delfiner, 1999). For the variable $z$, the empirical mean $\frac{1}{n}\sum_{i=1}^{n} z_i$ of the $n$ data located in the domain $V$ is a discrete approximation of the spatial integral $\frac{1}{V}\int_V z(\mathbf{x})\mathrm{d}\mathbf{x}$. This approximation is more precise when the $n$ data cover $V$ regularly, and it can be biased when data are irregularly or preferentially located. In the same way, the empirical variance is an approximation of the dispersion variance of a point (randomly and uniformly taken) in the domain $V$. Declustering techniques should then be used to extract data subsets to obtain an approximately regular grid, or weighting the data to avoid losing some of them (Chilès & Delfiner, 1999). Nevertheless, the bias induced by preferential sampling cannot always be corrected, for example if the large or small values are not measured.

Initial sampling of polluted sites is generally restricted to shallow depths. A survey of superficial material can be justified when the risk is restricted to these depths, for example for dust generation. However, an accurate characterization of the vertical distribution of concentrations is needed to evaluate the volumes of soil polluted or to study the risk of transfer to groundwater. For many sites, polluted volumes remain unknown at the end of the investigation (de Fouquet, 2006).

When various operators have worked on a site, the resulting sampling design becomes irregular, but the less-sampled areas are not always apparent. They are highlighted by the kriging standard deviation map, even if an approximately fitted variogram model is used.

### *Application 1: variation of concentrations with depth*

On some sites, the sampling technique varies according to depth. As the support is variable, care should be taken when comparing the statistics for concentration at various depths.

*Example 1: former petrochemical complex.* The comparison of the histograms for concentrations measured at various depths on samples with a common support is enlightening. On the former petrochemical complex, the statistical summaries for the global 150-m grid show that concentrations at depth were, on average, twice as large as those at the surface (Table 1). Linking histograms and scatter diagrams with the location map (Figure 4) gives further insights. For both depths the histograms of the logarithms are bimodal, but the value and the frequency of the modes vary with depth. For the surface concentrations, the small-values mode is larger, whereas at depth the two modes have similar frequencies. At some points concentration is larger at depth and surface sampling may therefore not detect underlying 'hot spots'. This observation has some general consequences for remediation: in a zone where no pollution is detected at the greatest depth sampled, this does not necessarily indicate that there is an absence of pollution at greater depth.

*Example 2: regular grid site LOQUAS 2.* The scatter diagram of concentration plotted against depth shows that the empirical mean concentration slightly decreased from 0.5 to 1.5 m depth, and then increased (Figure 5). The variability of concentration is larger at 2.5 and 3.5 m. Compared with the concentrations at 0.5-m depth, the distribution of the 1.5-m concentrations moved towards smaller values (with the exception of the maximum). The largest values appear to be greater for the two greater depths. From the 25 regularly located drill-holes, the variation of concentrations with depth provides information for the area sampled (about 400 m$^2$). However, what happens at greater depth is not known and a complete assessment of contaminated volumes makes a full subsurface survey necessary.

### *Application 2: proportional effect*

Figure 6 shows the scatter diagram for the mean plotted against the standard deviation calculated for the five-point reference sampling patterns (site LOQUAS 0). For each pattern the calculated variance is the dispersion variance of a point in the pattern, in other words the variance of one point randomly taken from the five. This calculation is theoretically sound, though the number of points per pattern might seem small. Clearly, variability increases with the size of the mean (Figure 6), a phenomenon commonly observed for soil pollution and at different scales.

This proportional effect is a relationship between local variability and local mean (Matheron, 1974): local variability increases with the local mean. In the neighbourhood $W_{\mathbf{x}}$ of point $\mathbf{x}$ the variogram is written $\gamma_{\mathbf{x}}(\mathbf{h}) = f(m_{\mathbf{x}})\gamma_0(\mathbf{h})$, $m_{\mathbf{x}}$ denoting the local mean in $W_{\mathbf{x}}$, $f(m_{\mathbf{x}})$ a function with spatial mean equal to 1 for the whole studied domain $W$ and $\gamma_0$ the global variogram of $W$. The proportional effect is present for lognormal distributions (Chilès & Delfiner, 1999) and, in the presence of a proportional effect, transformation of the data can be convenient but it is not always needed.

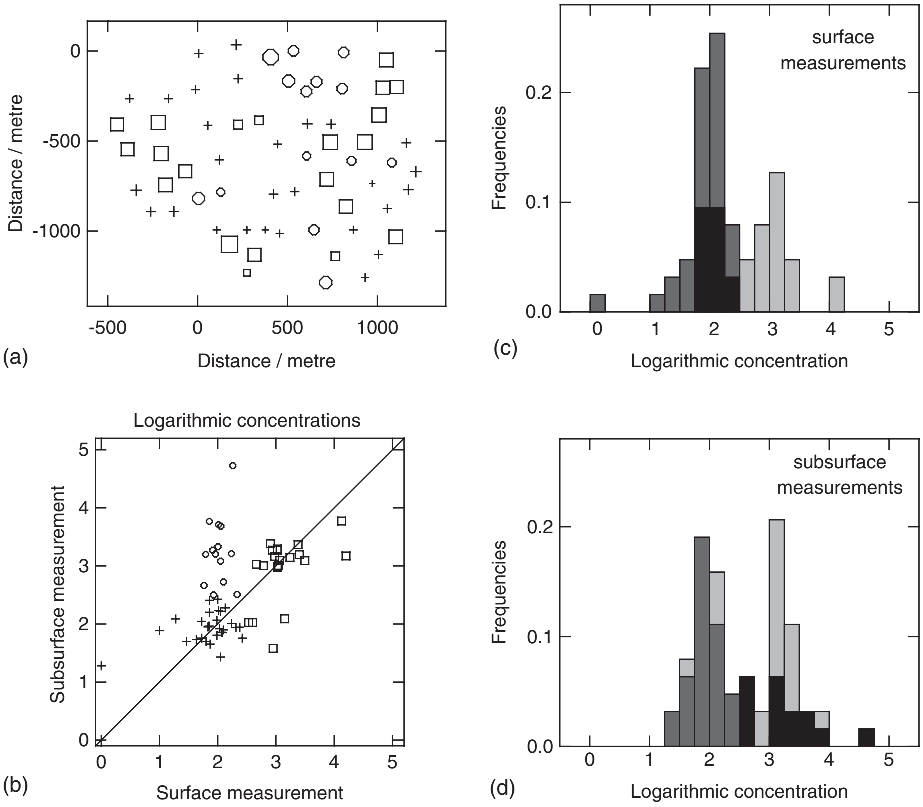


**Figure 4** Former petrochemical complex, large grid sampling. (a) Location of samples, (b) scatter diagram plotted between decimal logarithmic surface (abscissa) and subsurface (ordinates) concentrations at the same location, (c) histogram of decimal logarithmic surface concentrations and (d) histogram of decimal logarithmic subsurface concentrations. Symbols or colour of histogram classes correspond to the same points on all graphics: light grey or □ for large surface values, dark grey or + for small surface and subsurface values, and black or ○ for small surface and large subsurface values.

At a fixed support $v$ and domain $V$ the dispersion variance varies in the same way. If $D_W^2(v|V)$ denotes the dispersion variance calculated using the global variogram, the local dispersion variance is $D_{\mathbf{x}}^2(v|V) = f(m_{\mathbf{x}})D_W^2(v|V)$. This property is used to detect the proportional effect. Local mean and local standard deviation are calculated for a moving neighbourhood (ideally on a regular grid in order to have the same configuration of data location for all calculations) and their scatter diagram is plotted. The function $f$ is then fitted from the empirical regression of local standard deviation on local mean.

Kriging can be done by using the global variogram $\gamma_0$. If $\sigma_{K,0}^2(\mathbf{x})$ is the kriging variance calculated by using the global

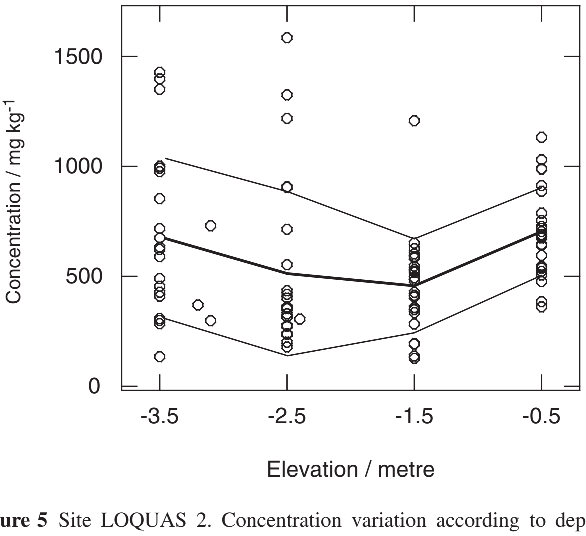


**Figure 5** Site LOQUAS 2. Concentration variation according to depth, equal to the elevation negatively counted (the surface corresponds to elevation 0). Twenty-five vertical drillings of the horizontal 5-m grid. Each data point is the empirical mean of three analyses on point-support taken on homogenized cores. Mean per class is indicated as a thick line with ± one standard deviation interval.

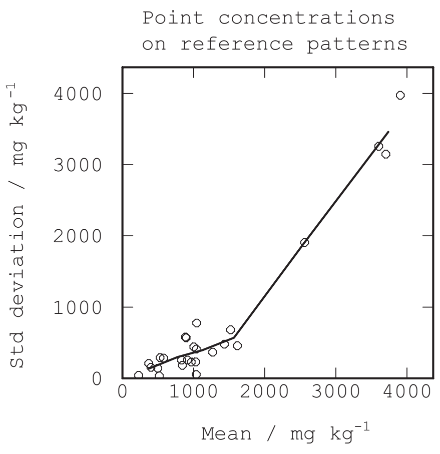


**Figure 6** Site LOQUAS 0. Scatter diagram between mean and standard deviation calculated on each five-point reference pattern of the multi-scale grids. The empirical regression line is shown.

variogram $\gamma_0$, the local kriging variance $\sigma^2_{K,\mathbf{x}}$ is corrected according to the local mean: $\sigma^2_{K,\mathbf{x}} = f(m_{\mathbf{x}})\sigma^2_{K,0}(\mathbf{x})$ (Chilès & Delfiner, 1999). At a fixed configuration of the data and the block (or point) to be estimated, the estimation variance depends on the local mean.

This exploratory analysis provides answers to several important questions, including whether (i) the sampling is suitable for the purpose of the study, (ii) some data seem aberrant and if so, an explanation is available, (iii) the site can be considered as homogenous and if not, with which criteria the heterogeneity must be treated, (iv) there is a proportional effect present and (v) there is information on the variation of pollutant concentrations with depth.

## Variogram analysis

### *Stages and tools*

Variogram analysis aims to characterize and quantify spatial variability. Different types of variogram are available (Chilès & Delfiner, 1999). These include variograms of concentrations or of transformed variables (logarithm, Gaussian, indicator at different values) and also those of different orders: $Z$ denoting a random function, the variogram of order $\nu$ is defined as $\frac{1}{2}$ E$[|Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})|^{\nu}]$. In addition to the usual variogram (order 2), the variogram of order 1 (madogram) is also used.

In complex situations (with a few very large values for example) these different variogram calculations help us to decide if a spatial structure is present or not. In addition to the variographic map (see later) they help to detect and to characterize any anisotropy. The 'variogram cloud' is the scatter diagram between the quadratic deviation $\frac{1}{2}\,(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x}))^2$ and the distance $|\mathbf{h}|$. Interactively linked with the data location map, it shows those data that contribute most to spatial variability, including those that may be spurious, so helping to detect anomalies.

If there are no marked differences between directional variograms the assumption of isotropy seems realistic. The 'variographic map' (Pannatier, 1997) represents directional variograms as a two-dimensional surface for lag vectors, centred at zero. It facilitates the detection of anisotropies and their characterization such as type (geometric, factorization of the covariance, etc.) and main directions, which can vary between the different spatial components. In the examples presented, variographic maps confirmed the absence of clear anisotropy in the horizontal plane.

*Example 1: former petrochemical complex.* In the area of former production units, the variogram cloud (Figure 7) of subsurface concentrations shows that the spatial variability at short distances results mainly from the contrast between two very large values (maximum larger than 30 000 mg kg$^{-1}$ with a median of 210 mg kg$^{-1}$, Table 1) and the surrounding smaller values.

The influence of these two large concentrations on the sample variogram is reduced by transformation to logarithms (Figure 8).

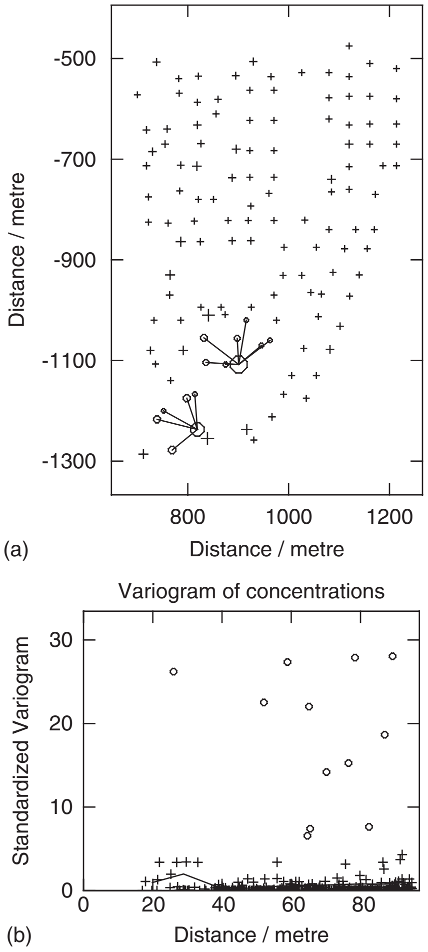


**Figure 7** Former petrochemical complex, subsurface concentrations. (a) Location of the data points on the area of former production units. (b) Standardized variogram cloud of concentrations: the halved quadratic increments are divided by the empirical variance. The circles on the variographic cloud correspond to pairs of sample points between two larger concentrations and the surrounding smaller concentrations. On the variogram cloud the broken line represents the sample mean variogram.

Directional variograms (Figure 8a) show that the assumption of isotropy is admissible for the log-transformed data. Their mean omnidirectional variogram (Figure 8b) shows the presence of nested spatial structures; the variability between 0 and 35 m is large and represents half of the sill. The mean variogram of concentrations appears to be more erratic because of the influence of the few larger concentrations (Figure 8c). In this case, non-linear geostatistics is more convenient than linear geostatistics in order to avoid an excessive influence of a few large values on the map of estimated concentrations.

*Example 2: site LOQAS 0, spatial variability at different scales.* On variograms calculated from the different grids (Figure 9) the spatial variability is much larger for the 0.50-m grid, which has the larger mean (Table 2). The sampling scheme, with a common distance for successive nested grids (Figure 2), allows the variograms

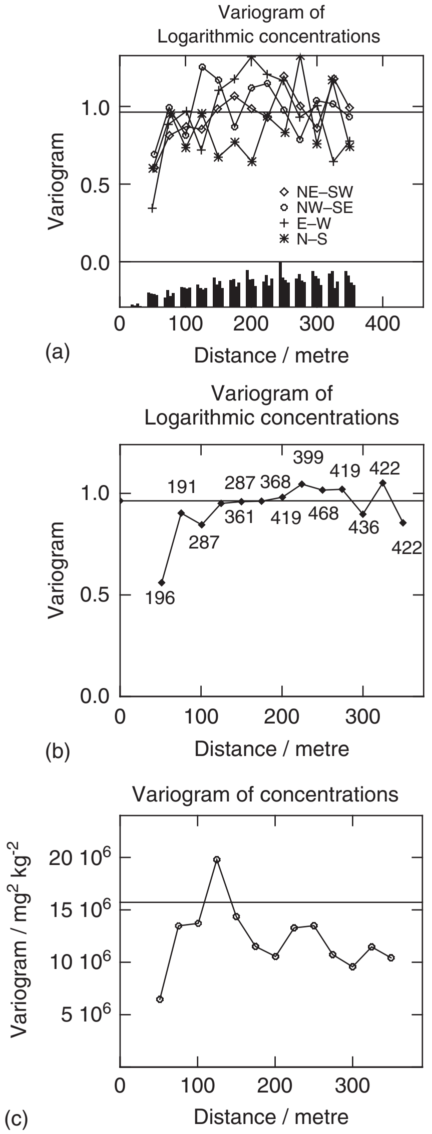


**Figure 8** Former petrochemical complex. Area of former production units. (a) Directional variograms, (b) mean omnidirectional variogram of the logarithmic transforms and (c) mean omnidirectional variogram of the concentrations. On (a) the histograms indicate the number of pairs per distance class for the four directions. On (b) the numbers indicated are the same for the mean variograms (b) and (c).

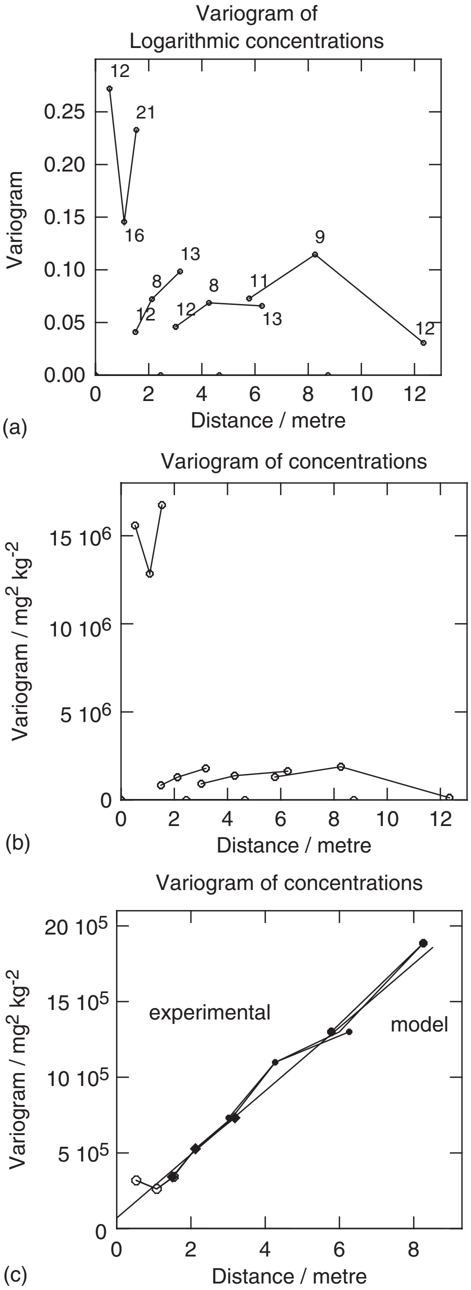


**Figure 9** Site LOQUAS 0, sample variograms of the empirical mean concentration of the five-point reference pattern at the different scales. (a) Variograms of concentration logarithm, (b) associated concentration variograms and (c) realignment of variograms with 6-m grid reference and fitted model, combining a nugget effect and a linear component. The number of pairs indicated on (a) is the same for all sample variograms.

to be rescaled. The 6-m grid was chosen as a reference because it samples the whole plot regularly.

The sample variograms of the concentrations (empirical mean on the five-point patterns) and of their logarithms have a similar shape (Figure 9a,b). They confirm the presence of a spatial structure at the decimetre to decametre scales. The decrease in the variogram between 6 and 12 m results from the larger concentrations in the centre of the site: the mean square deviations between patterns 12 m apart correspond to data pairs located on opposite edges, whereas those for patterns 6 m apart are calculated between central large concentrations and smaller concentrations on the edges. The last variogram lag will not be considered for modelling because it is too large relative to the study area.

After rescaling, the global variogram is fitted by a combination of a nugget effect and a linear variogram (Figure 9c). The validity of the model is limited to about $6\sqrt{2} \approx 8.5$ m, the last valid lag of the sample variogram.

Up to one half of the variogram is the variance of the estimation error of the point-support concentration at location $x + h$ from point-support concentration at location $\mathbf{x}$: $\gamma(\mathbf{h}) = \frac{1}{2}\mathrm{var}(Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x}))$. The variogram (Figure 9c) shows that the standard deviation of the difference between concentrations located 6 m apart is $\sqrt{2\gamma(6m)}$, about $\sqrt{2 \times 1.2510^6} \approx 1580$ mg kg$^{-1}$. As an example, this large spatial variability will result in poor accuracy of the estimates of 6-m square blocks from a 6-m sampling grid. During remediation, selection will then be made from an imprecise estimated map of concentrations. As a consequence, the spatial variability has to be taken into account to control the quality of any remediation. For a remediation threshold equal to 1000 mg kg$^{-1}$ and with the above fitted variogram, the standard deviation of the difference between concentrations 6 m apart is larger than the threshold. A control sample with concentration a little larger than the threshold does not prove that its neighbourhood (to be more precisely defined) is polluted, and a control sample smaller than the threshold does not prove that this neighbourhood is not polluted. A more pertinent approach is to evaluate the probability that at fixed support the concentrations exceed the threshold.

*Example 3: site LOQUAS 2, comparison between horizontal and vertical variability.* The vertical change in the mean concentrations (Figure 5) contributes to the vertical spatial variability and the vertical variogram calculated from the local 1-m grid appears to be linear; vertically the concentrations are not stationary.

Following the results of the exploratory analysis, horizontal variograms are calculated separately for the upper and lower depth intervals (0.5 and 1.5 m, and 2.5 and 3.5 m, respectively, Figure 10). The greater spatial variability of the lower depth interval is consistent with the observed larger dispersion (Figure 5). The global horizontal variogram would be the mean of these two variograms. Within a global model, the precision of the estimation would not be differentiated according to the specific variability of each depth; the kriging variance would be over-evaluated for the upper depth interval, and under-evaluated for the lower ones. At a depth of 1 m, vertical and horizontal variabilities are similar (Figure 10). The common intuition of a greater horizontal continuity of pollution thus breaks down at small distances.

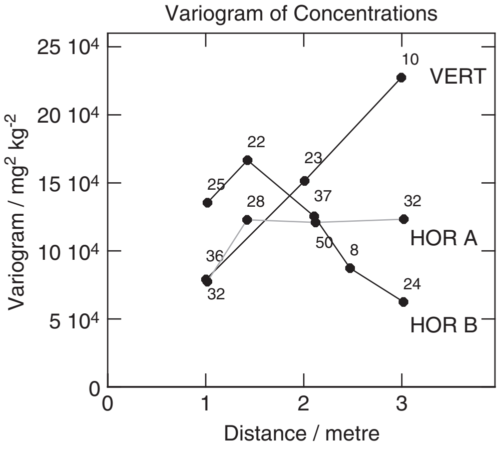


**Figure 10** Site LOQUAS 2. Vertical (VERT) and horizontal (HOR) variograms, calculated on the local, denser grid, with a 1-m lag. Data are the empirical means of four Pollut-Eval® measurements made at the centre of the cores. (a) Mean variogram of the two upper depths, in the fillings, and (b) mean variogram of the two lower depths. The number of pairs for each variogram point is indicated: this differs between the two horizontal variograms, because of the slight irregularities of the sampling design.

### *Quantifying the variance of measurement errors*

For one small section of site LOQUAS 0, 15 point-support samples were measured twice on two successive days. The scatter diagram indicates good reproducibility of the measurement, with a correlation coefficient of 0.8 (Figure 11a). If we assume that each pair of samples has the same concentration, the difference between the results comes from sampling and measurement errors (Gy, 1975). If $Y$ denotes the common 'true' concentration and $(\varepsilon_1, \varepsilon_2)$ the measurement errors the measurement results are $Z_1 = Y + \varepsilon_1$ and $Z_2 = Y + \varepsilon_2$. As the means of the two sets are very close, the expectations of both measurement errors are assumed to be identical. The empirical variances of both sets are also close, with respective standard deviations $\sigma_1 = 405$ and $\sigma_2 = 378$ mg kg$^{-1}$. The variances of the two errors will be assumed to be identical.

The cross-variogram of the measurements is defined as

$$\gamma_{12}(\mathbf{h}) = \frac{1}{2}\mathrm{E}[\{Z_1(\mathbf{x}+\mathbf{h}) - Z_1(\mathbf{x})\}.\{Z_2(\mathbf{x}+\mathbf{h}) - Z_2(\mathbf{x})\}]. \quad (1)$$

If we denote $\gamma_{AB}$ as the cross-variogram between the random functions $A$ and $B$, it is easy to show (Faucheux *et al.*, 2008) that

$$\gamma_{12}(\mathbf{h}) = \gamma_Y(\mathbf{h}) + \gamma_{Y\varepsilon_1}(\mathbf{h}) + \gamma_{Y\varepsilon_2}(\mathbf{h}) + \gamma_{\varepsilon_1\varepsilon_2}(\mathbf{h}). \quad (2)$$

Under the assumption of absence of mutual correlation between both errors and the true concentration $Y$, all terms are null except the first one: $\gamma_{12}(\mathbf{h}) = \gamma_Y(\mathbf{h})$. Although true concentrations are inaccessible, their variogram becomes accessible because of the measurement cross-variogram, provided that the assumptions are justified.

The mean of the two measurements is $Z_m = \mathrm{Y} + \frac{1}{2}(\varepsilon_1 + \varepsilon_2)$. Under the preceding assumptions and because both measurement error variances are equal, its variogram is

$$\gamma_{Z_m}(\mathbf{h}) = \frac{\sigma^2}{2} + \gamma_Y(\mathbf{h}). \quad (3)$$

The variogram of the mean is then the sum of the concentration variogram and a constant equal to the half of the measurement error variance. Figure 11(b) shows that the difference between the sample variogram of the measurements' mean and their cross-variogram is nearly constant, giving an empirical standard deviation of the measurement error of about 180 mg kg$^{-1}$.

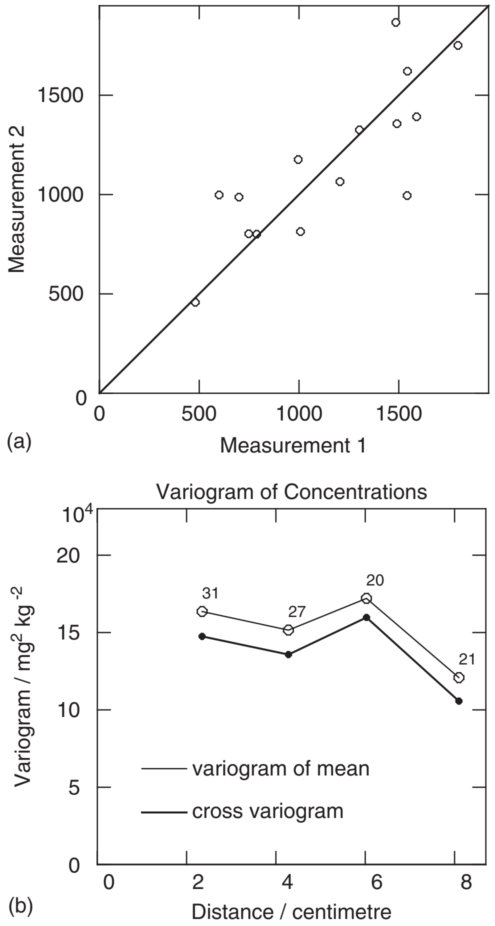


**Figure 11** Site LOQUAS 0. Repeatability of Pollut-Eval® measurements of hydrocarbon concentrations. (a) Scatter diagram of the duplicated measurements denoted 1 and 2 (units are mg kg$^{-1}$). (b) Variogram of the mean of both measurements (upper line) and of their cross-variogram (lower line). The indicated number of pairs is the same for the simple and the cross-variogram.

The error variance can also be calculated directly from the measurement differences. Under the weaker assumption of no mutual correlation of measurement errors and concentration at the same point, the variance of the difference $Z_1 - Z_2 = \varepsilon_1 - \varepsilon_2$ is twice the variance of the measurement error. This calculation gives a standard deviation of 170 mg kg$^{-1}$. Compared with the measurement mean of about 1155 mg kg$^{-1}$ both evaluations are consistent. From

$$\mathrm{var} Z_m = \frac{\sigma^2}{2} + \mathrm{var}\ Y, \tag{4}$$

the standard deviation of $Y$ is about 350 mg kg$^{-1}$.

On the sample variogram (Figure 11b) the nugget effect is much larger than the variance of measurement errors. It reflects the spatial variability at distances less than 2 cm and the sampling errors that arise when the material that is finally analysed is subsampled from the original core (Gy, 1975).

The assumption of independence between measurement errors and concentrations could break down if soil properties vary according to the concentrations. On site LOQUAS 2 the order of magnitude of the measurement error is close to that obtained on site LOQUAS 0, but independence cannot be assumed for all datasets: some cross-variograms are no longer parallel to the associated variograms of measurement means. ‘Redundant’ measurements are thus very useful for assessing some statistical characteristics of the ‘sampling errors’.

## Estimation

Before applying well-established methods such as kriging, it is necessary to look back at the physical relevance of the modelling.

### *Remarks on the variables and the support*

*Additivity property.* Additive variables represent ‘extensive’ quantities, such as length or mass: the mass of a mixture of several samples is the sum of their individual masses. Processing with additive quantities must take into account the possible differences of support between the different data or datasets or those between the data and the volume to estimate. In 2-D modelling, the concentration of pollutant in a layer with a variable thickness is estimated with the help of the amount accumulated, which is a product of the concentration and the layer thickness (Chilès & Delfiner, 1999).

In soil studies, concentrations (of clay for example) are sometimes expressed as gravimetric quantities measured for specific particle size classes; information on the proportion or density of the different classes is sometimes lacking, making it impossible to derive the total quantity at data points.

‘Intensive’ variables describe properties that are independent of the quantity of material present. These variables (for example, permeability or electrical resistivity), which are derived from physical laws, are not additive. Many soil variables such as hydraulic conductivity or pH are therefore non-additive.

As a linear estimator, point-support kriging (kriging a quantity on the same support as the data) is consistent but this is not necessarily the case for block-support kriging. Block-kriging is defined as the mean for the whole block of all point-support estimations. It is thus an estimator of the arithmetic mean over the block; as a block-support estimator, it is suitable only if the arithmetic mean effectively characterizes the block-support value.

Let biodiversity be quantified by the number of species per unit area. If $n_1$ and $n_2$ denote the number of species present in two different plots, the number of species present in their union lies between the maximum of $n_1$ and $n_2$ and the sum $n_1 + n_2$, depending on whether the species of the less diversified plot are present in the other plot. Estimating the number of species by block-kriging on a surface $S$ greater in area than the sampling plots would dramatically under-estimate the result. Furthermore, the number of species estimated in this way on $S$ could be less than that recorded on the most diverse plot sampled inside

$S$. Block-kriging actually provides an estimation of the mean of the number of species on the plot-support within $S$. Block-support estimation needs more information, namely the species recorded by plot. As for any estimation method, uncritical application of kriging without considering the properties of the variables studied can therefore provide inconsistent results.

*Choosing the support.* The definition of a polluted volume raises the important question of the support and of whether the quality threshold is related to sample-support concentrations (resulting from the practice and the sampling constraints) or to block-support concentrations, for blocks about tens of m$^3$ or more such as the units selected during remediation. This choice may have important consequences for the delimitation of polluted areas. When the support increases, the concentration histogram becomes more tightly distributed around the mean. The proportion of values above a given threshold, and the surface considered as polluted, thus vary with the support (Rivoirard, 1994; Chilès & Delfiner, 1999).

### Linear or non-linear estimation

Webster & Oliver (1990) and Chilès & Delfiner (1999) give more details of kriging. The kriging estimation of the block-support concentration $Z_V$ (or of the point-support concentration $Z(x)$) is a linear combination of data at sample points $x_\eta$:

$$Z_V^* = \sum_\eta \lambda_V^\eta z(\mathbf{x}_\eta). \tag{5}$$

Weights $\lambda_V^\eta$ are chosen so that the estimate is unbiased (on average, the estimation error is zero) and the precision optimal (the variance of the estimation error is minimized). The weights depend on the variogram. Kriging provides the variance of the estimation error $Z_V - Z_V^*$, which quantifies the accuracy of the estimation. Different kriging variants allow different types of auxiliary information to be taken into account, a measurement uncertainty to be assigned to data, and different assumptions on stationarity to be made.

The map of the estimated concentrations $Z_{V,i}^*$ should not be confused with the map of the true concentrations $Z_{V,i}$, which is in practice unattainable. The difference between the two is the map of estimation errors $Z_{V,i} - Z_{V,i}^*$, for which only the mean (equal to zero) and the variance are known.

To delineate the polluted zone with reference to a threshold $s$, the searched area is the set of blocks $i$ so that $Z_{V,i} \geq s$. Applying the threshold to the estimated map corresponds to the condition $Z_{V,i}^* \geq s$. The associated areas may differ significantly:

$$\begin{aligned} Z_{V,i} \geq s &\Leftrightarrow Z_{V,i}^* + (Z_{V,i} - Z_{V,i}^*) \geq s \\ &\Leftrightarrow Z_{V,i}^* \geq s - (Z_{V,i} - Z_{V,i}^*). \end{aligned} \tag{6}$$

and the difference increases with the amplitude of the estimation error. Two approaches are used to solve this issue.

The rigorous approach is based on non-linear geostatistics (disjunctive kriging or conditional expectation), generally within the framework of the anamorphosed Gaussian model (Rivoirard, 1994). The point-support concentration $Z(\mathbf{x})$ is modelled as the transform by an increasing function (anamorphosis) $\varphi$ of a random function $Y(\mathbf{x})$ with Gaussian spatial distribution. Various criteria are available to check the validity of this model (Rivoirard, 1994). The change of support can be made within the discretized Gaussian model to determine the block-support anamorphosis $\varphi_V$ between block-support concentrations $Z_V$ and associated Gaussian transforms $Y_V$. When the anamorphosis is bijective with a function that is both one-to-one (injective) and onto (surjective), the issue of exceeding a threshold is very simple to formulate because of the properties of the Gaussian distribution: conditionally to the observed data $Z(\mathbf{x}_\alpha) = \varphi(Y(\mathbf{x}_\alpha))$

$$\begin{aligned} Z_{V,i} \geq s &\Leftrightarrow \varphi_V(Y_{V,i}) \geq s \\ &\Leftrightarrow Y_{V,i}^* + \sigma_{K,i} R_i \geq \varphi_V^{-1}(s) \\ &\Leftrightarrow R_i \geq \frac{\varphi_V^{-1}(s) - Y_{V,i}^*}{\sigma_{K,i}}, \end{aligned} \tag{7}$$

where $Y_{V,i}^*$ and $\sigma_{K,i}$ denote the kriging and the kriging standard deviation, respectively, on Gaussian transforms, and $R$ a Gaussian random function spatially independent of $Y$. The correlation between ordinary kriging and associated error can be taken into account for the lognormal case (Rivoirard, 1994) and the general case (Emery, 2006). Within this model, probability maps or confidence interval are easily derived.

Because of its relatively easy implementation, indicator kriging has been widely applied (Goovaerts *et al.*, 1997), without checking the validity of underlying assumptions. This method dramatically reduces the information provided by concentration measurements, because only two values (0 or 1) are considered. Webster & Rivoirard (1991) present a very instructive application of non-linear geostatistics applied to copper and cobalt deficiency in pastures. The study included a check of the hypotheses on the spatial bivariate distribution within the anamorphosed Gaussian model using indicator variograms and cross-variograms, a comparison between kriging and disjunctive kriging for block-support estimation of concentrations and a comparison between disjunctive kriging and conditional expectation for mapping the probability that the block-support concentration exceeds a threshold.

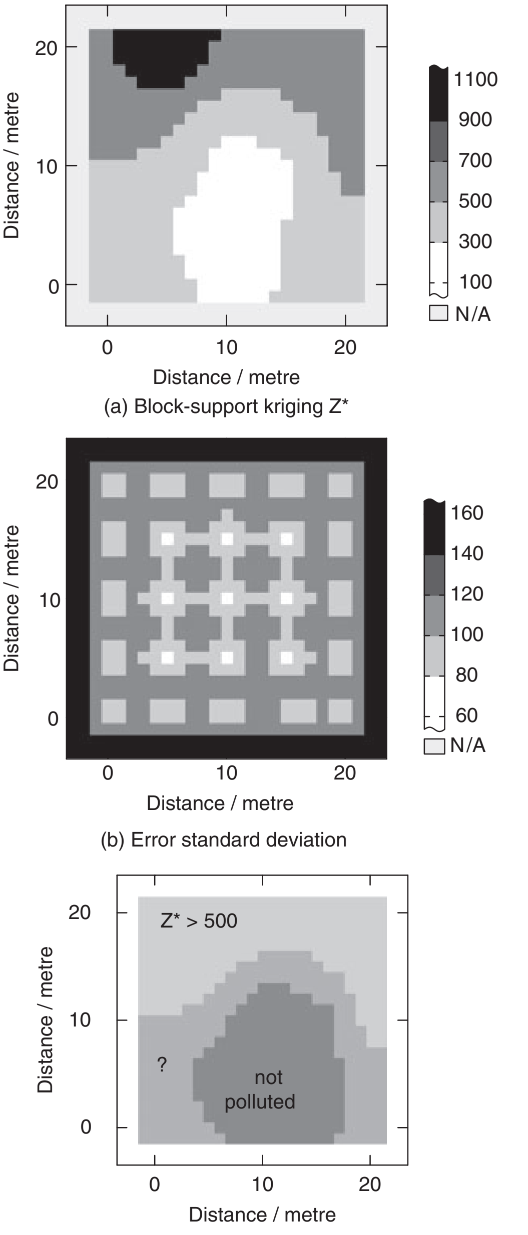
In a less rigorous but pragmatic approach, the estimation errors are conventionally considered to be Gaussian, making it possible to derive conditional confidence intervals. From these confidence intervals, three zones can be delineated to compare the unknown concentrations with a fixed threshold (Renard-Demougeot, 2002; Bobbia *et al.*, 2008). First the ‘unpolluted zone’, where the concentrations can be assumed to be smaller than the threshold, up to a given statistical risk; second, the ‘uncertain zone’, where it is not clear whether the true concentrations are greater or smaller than the threshold. This uncertain zone includes concentrations close to the threshold and depends on the local precision of the estimation. Finally, there is the zone where the estimated concentrations exceed the threshold.

To obtain a confidence interval the Gaussian hypothesis is not really necessary. Chilès & Delfiner (1999) quote general results with a very mild assumption on the error distribution, but at a given statistical risk, the width of the associated confidence interval is larger.

*Example, site LOQUAS 2*

Figure 12 presents the kriged map (with unknown mean) for 1-m$^3$ blocks (Figure 12a), calculated by using the mean variogram of the depth interval. The associated standard deviation map (Figure 12b) shows the small irregularities of the sampling grid. The uncertainty is considered to be too large when the error standard deviation exceeds 120 mg kg$^{-1}$ and the associated estimated values are omitted (shown in light grey on Figure 12a). In 194 blocks the concentration is estimated to be larger than the threshold of 500 mg kg$^{-1}$.

At a conventional statistical risk of 2.5%, the 'uncertainty' zone corresponds to the blocks with $Z^*_{V,i} \leq s \leq Z^*_{V,i} + 2\sigma_{K,i}$. To take into account a proportional effect, the error standard deviation is corrected using a linear relationship between local mean and (dispersion) standard deviation. As a consequence, the uncertainty increases with the estimated concentrations. The uncertainty zone thus includes 170 additional blocks (Figure 12c). The penalty for uncertainty is a large increase of the zone to be remediated.



(a) Block-support kriging Z*

(b) Error standard deviation

(c) delimitation of the 'not polluted' and 'uncertainty' zones

**Figure 12** Site LOQUAS 2. (a) Block-support kriging of hydrocarbon concentration; (b) associated estimation standard deviation calculated from the global variogram; and (c) delimitation of the 'unpolluted' zone with a conventional statistical risk of 2.5%. The 'uncertainty zone' (see text) is denoted as ?

## Conclusion

The soil is a complex environment for which probabilistic modelling has proved its worth (Webster, 2000; Heuvelink & Webster, 2001; Webster, 2007). However, remediation project managers or soil scientists do not always make proper use of statistics (Webster, 2001).

Exploratory and variographic analyses help us to understand how soil properties or pollution are spatially organized on the site. For the examples used here, the variability at 'short distances' is large but spatial correlation is present at the metre to decametre scales. On former industrial sites the survey is not always intensive enough to detect spatial correlation where it is present or the short-range components which then occur on the sample variogram as a nugget component. As a consequence the estimation will be imprecise, and at a fixed remediation threshold the 'uncertainty zone' can cover an important part of the site. For the economic balance of benefits from a remediation project, the uncertainties have to be taken into account from the very beginning (Benoit *et al.*, 2008).

Soil pollution was presented here in a univariate context. However, soil science is essentially multivariate. Webster & Oliver (1985) and Oliver & Webster (1989) addressed the classification as precursors from a statistical and geostatistical point of view. There are still many problems within soil science. New random models are being developed (Milne *et al.*, 2010). To the current open questions (Heuvelink & Webster, 2001) one can add the wide class of change of scale problems (Matheron, 1993) because most soil properties are not additive.

## Acknowledgements

I thank Y. Benoit (IFP), manager, C. Faucheux and E. Polus-Lefebvre (Mines ParisTech), C. Carpentier and B. Fricaudet

(Arcadis) for their contribution to the LOQUAS project. Thanks to M. Lark and the reviewers for their remarks and linguistic help, to G. de Marsily and M. and P. Duplat for carefully reading the paper and Ph. Le Caer for help with the graphics.

## References

Benoit, Y., de Fouquet, C., Fricaudet, B., Carpentier, C., Gourry, J.-C., Haudidier, N. *et al*. 2008. Cross linked methodologies to assess the contamination extension of hydrocarbon polluted soil. *Proceedings Consoil*, Milano, 3–6 June 2008 (ISBN of Proceedings CD: 978-3-00-024598-5).

Blanchet, D., Benoit, Y. & Haeseler, F. 2005. Trace analyses of hydrocarbons to understand the fate of these contaminants in aquifers. Analyse des traces d'hydrocarbures pour la compréhension du devenir de ces contaminants dans les aquifères. *Oil & Gas Science and Technology – Revue IFP*, **60**, 913–922. doi: 10.2516/ogst: 2005064.

Bobbia, M., Cori, A. & de Fouquet, C. 2008. Représentativité spatiale d'une mesure de la pollution atmosphérique. *Pollution Atmosphérique*, **197**, 63–75.

Chilès, J.-P. & Delfiner, P. 1999. *Geostatistics: Modelling Spatial Uncertainty.* John Wiley & Sons, New York.

Emery, X. 2006. Ordinary multigaussian kriging for mapping conditional probabilities of soil properties. *Geoderma*, **132**, 75–88.

Faucheux, C., Lefebvre, E., de Fouquet, C., Benoit, Y., Fricaudet, B., Carpentier, C. *et al*. 2008. Characterisation of a hydrocarbon polluted soil by an intensive multi-scale sampling. In: *Geostats 2008, Proceedings of the 8th International Geostatistics Congress* (eds J.-M. Ortiz & X. Emery), pp. 961–970. GECAMIN Ltd, Santiago.

de Fouquet, C. 2006. *La modélisation géostatistique des milieux anthropisés. Habilitation à diriger des recherches, Mémoire des Sciences de la Terre n° 2006-13*. Université Pierre-et-Marie Curie, Paris.

de Fouquet, C., Prechtel, A. & Setier, J.-C. 2004. Estimation de la teneur en hydrocarbures totaux du sol d'un ancien site pétrochimique: étude méthodologique. *Oil & Gas Science & Technology: Revue IFP*, **59**, 275–295.

Goovaerts, P., Webster, R. & Dubois, J.-P. 1997. Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics. *Environmental & Ecological Statistics*, **4**, 31–48.

Gy, P. 1975. *Théorie et pratique de l'échantillonnage des matières morcelées*. P. Gy ed., Cannes.

Heuvelink, C.B.M. & Webster, R. 2001. Modelling soil variation: past, present, and future. *Geoderma*, **100**, 269–301.

Matheron, G. 1974. *Effet proportionnel et lognormalité ou le retour du serpent de mer*. Technical Report N-377, Ecole des mines de Paris, Fontainebleau.

Matheron, G. 1993. Quelques inégalités pour la perméabilité effective d'un milieu poreux hétérogène. In: *Cahiers de géostatistique*, *Compte-Rendu des journées de Géostatistique, fascicule 3* (ed. C. de Fouquet), pp. 1–20. Ecole des mines de Paris, Fontainebleau.

Milne, A.E., Webster, R. & Lark, R.M. 2010. Spectral and wavelet analysis of gilgai patterns from air photography. *Australian Journal of Soil research*, **48**, 309–325.

Oliver, M.A. & Webster, R. 1989. Geostatistically constrained multivariate classification. In: *Geostatistics*, *Proceedings of the 3rd International Geostatistics Congress, Avignon*, France (ed. M. Armstrong), pp. 383–395. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Pannatier, Y. 1997. Variowin-software for spatial data analysis in 2D. *Computational Statistics & Data Analysis*, **25**, 243–244.

Renard-Demougeot, H. 2002. *De la reconnaissance à la réhabilitation des sols pollués: estimations géostatistiques pour une optimisation multicritère*. Dissertation ETH n°. 14615. Doctorat en Sciences Naturelles, Ecole Polytechnique Fédérale de Zürich.

Rivoirard, J. 1994. *Introduction to Disjunctive Kriging and Non Linear Geostatistics.* Oxford University Press, Oxford.

Webster, R. 2000. Is soil variation random? *Geoderma*, **97**, 149–163.

Webster, R. 2001. Statistics to support soil research and their presentation. *European Journal of Soil Science*, **52**, 331–340.

Webster, R. 2007. Analysis of variance, inference, multiple comparisons and sampling effects in soil research. *European Journal of Soil Science*, **58**, 74–82.

Webster, R. & Oliver, M.A. 1985. Utilisation exploratoire de la géostatistique pour la cartographie du sol dans la forêt de Wyre, Angleterre. *Sciences de la Terre, n°24*, *Séminaire CFSG sur la géostatistique, 17-18 Juin 1985. Fontainebleau.* Fondation Scientifique de la Géologie et de ses Applications, Nancy.

Webster, R. & Oliver, M.A. 1990. *Statistical Methods in Soil and Land Resource Survey.* Oxford University Press, Oxford.

Webster, R. & Rivoirard, J. 1991. Copper and cobalt deficiency in soils: a study using disjunctive kriging. In: *Cahiers de Géostatistique*, *Compte-rendu des journees de Géostatistique, fascicule 1* (ed. C. de Fouquet), pp. 205–223. Ecole des Mines de Paris, Fontainebleau.