GEOSTATISTICS FOR HIGHLIGHTING UNCERTAINTIES IN SOIL OR GROUNDWATER CONTAMINATION MANAGEMENT

Chantal de Fouquet

Mines ParisTech, centre de géosciences - géostatistique
35, rue saint-Honoré. 77305 Fontainebleau, France
+33 1 64 69 47 61
chantal.de_fouquet@ensmp.fr

ABSTRACT

The management of soil or groundwater contamination is based on an incomplete knowledge of the extension and of the concentration level. Indeed concentration maps are built from a limited number of data.

Exploratory data analysis helps to delineate the effectively recognized domain and to detect the survey gaps. For example in polluted soils a sampled depth restricted to a few meters does not always reach the vertical extension of the pollution. The "source term" of groundwater transfer remains thus often inaccessible.

In groundwater, large pesticide concentrations can be "censored" (wells for drinkable water are deleted if the quality threshold is exceeded), and a large proportion of micropollutant concentration in soils or water is usually indicated as "lower than the quantification limit". Attention has then to be paid when interpreting statistical results, namely in order to characterize an evolution.

The measurements are used to establish concentration maps. From the same survey plan the accuracy of the estimation depends on the spatial variability of the studied substance. Furthermore the spatial (or temporal) variability increases generally with the concentration. This “proportional effect” has to be taken into account in order to deduce a realistic value for the standard deviation of the estimation error.

The passage of the (estimated) concentration map to the demarcation of a polluted zone requires taking into account the estimation error, which is unknown but which expectation and variance can be modeled. Several non-linear estimation methods are available. A simplified conventional approach, based on a hypothesis of normal distribution of the estimation error, can supply non realistic bounds for a confidence interval. Indicator kriging or cokriging sometimes supplies inconsistent results. Disjunctive kriging (Matheron, 1973) or conditional expectation methods are a little more complex but give more consistent results, when the underlying hypotheses are valid. With these techniques a partition of the site in three zones with regard to a quality threshold is established, up to fixed statistical risks:

- the polluted zone;
- the not polluted zone;
- the "zone of uncertainty", in which it is not possible to specify if the real concentration exceeds or not the threshold.

Finally, the question of the "support» for the comparison of concentrations to a quality threshold is discussed.

INTRODUCTION

The management of a soil or groundwater pollution is based on an incomplete knowledge of its extension or of the concentration level. Indeed concentration maps are built from a limited number of samples taken in drillings or wells. These maps are thus never exact.

The use of geostatistics [Matheron, 1965] in order to improve the characterization of former industrial soils or groundwater pollution and to calculate the associated uncertainty is not new [de Fouquet et al., 2004; Demougeot-Renard et al., 2004]. The mapping of concentrations estimated by kriging was quickly completed by the calculation of the risk that concentration exceeds a fixed quality threshold. This requires implementing non linear estimators.

The study of micropollutants brings new difficulties. For each substance an important proportion of data, or even a large majority, is indicated as lower than a "quantification limit". Large pesticide values in groundwater can also be "censored»: when the concentration in a well for drinking water supply exceeds a policy value, the well is destructed for sanitary reasons. How can these characteristics of the data be taken into account in order to estimate the concentrations or their evolution (in groundwater)?

Among many other issues, the following points are examined in the paper:

- How to check if the sampling is appropriate to characterize the (initial) state of pollution?
- How to deal in a pragmatic way with imprecise data, for example values lower than a quantification limit?
- How to characterize the evolution of concentration between two dates?
- How to go from the concentration estimation to a confidence interval?

SOIL SURVEY OF FORMER INDUSTRIAL SITES

We examine the case of polluted soils, after the systematic site survey

Vertical Extension of Pollution

For site survey, drillholes are often limited to the first few meters, which means that the concentration remains unknown down. The vertical extent of pollution is therefore rarely known: to know it, drillholes should exceed the "pollution limit" in depth. For risk studies of groundwater transfer, the source term remains poorly characterized when the vertical extension of the pollution is unknown.

The exploratory data analysis allows a simple control of the depth reached by the survey. The histogram of the sampled depth indicates which levels are more or less densely sampled (Figure 1a). The scatter diagram between concentration and depth, with the plotting of "average per class" (empirical regression) and dispersion by class, shows the variation of concentration with depth which is not necessarily monotonic (Figure 1b). "Remediated" sites after decommissioning of installations sometimes show a superficial level with low concentration because of the added fills, and significantly larger concentration in depth in undisturbed soils. Several published cases show a concentration increase with depth [de Fouquet, 2011]. If the scatter diagram does not show decrease of concentration with depth (at same survey effort as in superficial levels), then the vertical extension of the pollution is unknown (except in specific cases, for example the presence of a formation stopping the pollution transfer and which top depth is known with a good accuracy). The maximum depth of the survey often depends on criteria other than the search of the vertical extension of pollution, namely the excavation depth for a new construction, or the survey cost.

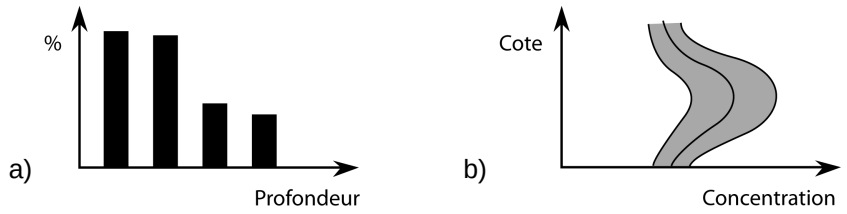

Figure 1. Depth of survey. a) Histogram of sample depth (*profondeur*), showing a decrease of sample number with depth; b) Scatter diagram between concentration and elevation (*cote*).

Drillhole spacing

The initial survey scheme of a site is sometimes guided by the following principle: the drillholes (assumed to be vertical) should be spaced enough to "decorrelate" concentrations. This approach contradicts the intuition that the average concentration is better known from a fine survey mesh. This paradox is explained by an inadequate application of an exact statistical result [de Fouquet, 2012]. Indeed, let us consider a sample of *n* data Z_i, supposed to arise from the same "theoretical" distribution, with variance S^2. The estimation variance of the "expectation" parameter *m* of this distribution from the data is

$$Var\ m^* = \frac{1}{n^2}\sum_{i=1}^{n} Var\ Z_i + \frac{1}{n^2}\sum_{i\neq j} Cov(Z_i, Z_j)$$
$$= \frac{1}{n}S^2 + \frac{1}{n^2}\sum_{i\neq j} Cov(Z_i, Z_j)$$

where *Cov* denotes the covariance. If the data are positively correlated then this estimation variance is minimal when the second term is zero, that is when there is no correlation between data. This is the reason why samples are sometimes spaced in agronomy and soil science. But the relevant quantity is generally not the probabilistic expectation *m*, but rather the spatial average on an area or a volume V: $Z_V = \frac{1}{V}\int_V Z(x)dx$. The associated estimation is then Z^*_V. Roughly speaking *t*he variance of the estimation error $Z^*_V - Z_V$ decreases with increasing density of sampling.

Spatial variability

An exceptionally dense sampling of hydrocarbon pollution obtained thanks to the LOQUAS project [Benoit et al., 2008] has shown that, on the studied sites, the main scales of spatial variability corresponded to a range (correlation distance) lower than 15m. However the survey used to build a remediation project is generally designed with a much larger distance between drillholes. This makes it impossible to map concentrations with a good accuracy.

This large spatial variability is detected by the sample variogram. It directly impacts the accuracy of the estimation.

In the case of an irregular sampling scheme, the map of the kriging error variance (even based on an approximate fit of the sample variogram) can be used to detect gaps in the survey. It provides an indication of the lateral extension of the area where estimation is effectively possible.

ESTIMATION AND DEMARCATION OF THE AREA WHERE A THRESHOLD IS EXCEEDED

Kriging provides a map of estimated concentrations, as well as the associated accuracy.

Kriging requires specifying the "support" of the quantity to be mapped. Is it the same as for the data (the analyzed samples) or is it for example a "block" of 25mx25mx1.5m (more than 930m^3 of soil), or in the case of groundwater, an area of 250mx250m (6.25 ha)? In the case of groundwater, an information as

important as the presence of a strainer and its location is not always available. The estimation refers then to a level which is not specified, with depth and thickness supposed to be constant; data are supposed to represent the averaged concentration of a sample homogenized on all the thickness. If data on depth or thickness of the water sample or of the groundwater table are available, it is obviously possible to take them into account for the estimation.

The accuracy of the estimation can be improved by means of variables which describe the environment, or (even qualitative) observations linked to concentrations [Jeannée et al., 2003; de Fouquet et al., 2007]. To be realistic, the standard-deviation map of the estimation error has to be corrected in order to take the proportional effect into account: indeed variability is larger where concentration is larger. As a result, with the same survey mesh, the precision is better where the concentration is lower, and it gets worse where concentration increases. Generally the proportional effect is taken into account in a simple way. Because the kriging estimation remains unchanged if the variogram is multiplied by a constant, the estimation is performed using a global variogram. The error standard deviation is then multiplicatively corrected by a factor calculated locally and depending on the estimated concentration [Chilès & Delfiner, 1999].

The accurate delineation of a polluted area cannot be obtained by simply comparing the map of estimated concentration with the quality threshold. Indeed, it is necessary to take into account the estimation error, which remains unknown but which variance is known in the geostatistical model.

The idea is then to build a map of the probability that the real concentration exceeds the quality threshold, and to derive a partition into three zones [de Fouquet et al., 2011]:

- (A) the zone supposed to be unpolluted, up to a fixed statistical risk;
- (B) the "zone of uncertainty", in which it is not possible to specify whether or not the actual concentration exceeds the threshold, because of the uncertainty on the estimation;
- (C) the zone supposed to be polluted, up to a(nother) fixed statistical risk;

The zone of uncertainty can itself be divided into two parts (Figures 2 and Figure 3):

- (B1) the estimated concentration is a little less than the threshold, and because of the uncertainty, the probability that the actual concentration exceeds the threshold is not negligible.
- (B2) the estimated concentration exceeds a little the threshold and because of the uncertainty, the probability that the actual concentration is below the threshold is not negligible.

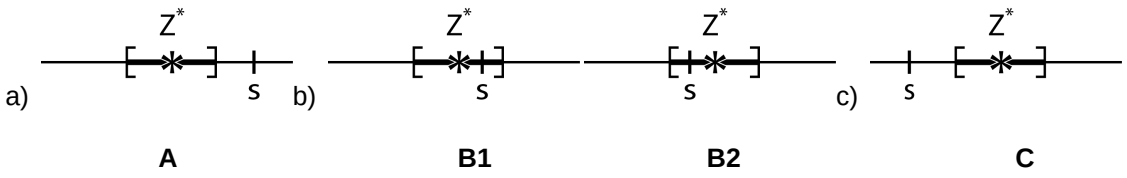

Figure 2. Comparizon of the concentration with a threshold *s*. Up to statistical risk on law and large values, a confidence interval is build around the estimated concentration *Z. a) absence of pollution, b) uncertainty (with cases B1 and B2, see text) and c) pollution.**

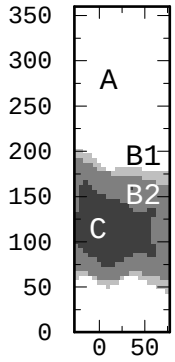

Figure 3. Delineation of pollution, showing the "uncertainty zone" (from [de Fouquet et al., 2011])

Economic calculation and consideration of practical constraints of remediation works will allow to decide whether the "uncertainty zone" should be added to the area where the threshold is exceeded, or if a preferential additional survey of this uncertainty zone (with measurements inside and around) is possible.

The proportion of the concentrations which exceed a fixed threshold depends on the considered support, and thus the "polluted", "not polluted", or "uncertainty area" also depends on the support. The extension and the average concentration of these areas vary considerably depending on whether the selection is made on sample support (corresponding to the support of the data) or on 'blocks" actually selected during the site remediation. Prediction of polluted volumes has to be carried out on a realistic support.

CALCULATION OF THE RISK THAT CONCENTRATION EXCEEDS THE QUALITY THRESHOLD

Several non-linear estimation methods allow calculating the probability that concentration exceeds a quality threshold. Papritz et al. (1999) show an instructive comparison of methods.

The simplified conventional approach is based on an assumption of Gaussian (i.e. normal) estimation error. Let the support be chosen, for example a "block" of fixed dimensions for a site remediation. Z denotes the unknown real concentration, and Z^* its estimation. The estimation error $Z-Z^*$ is unknown, but the geostatistical model provides its variance (in the model, the probabilistic mean of the estimation error is supposed to be zero; in practice, the model has to be adapted to each site particularities to keep this assumption realistic). The approach is as follows: the estimated value is known and equal to z^*, and the real concentration is supposed to be Gaussian with mean (expectation) z^* and variance equal to the kriging variance σ^2_K (the calculation can be refined to take the correlation between kriging and kriging error into account). Let R denotes a reduced Gaussian variable. The concentration is written $Z = z^* + \sigma_K R$. The distribution of Z is expressed from the classical distribution of a reduced Gaussian variable, which allows to calculate the probability that the real concentration exceeds a fixed quality threshold s, or to construct a "confidence interval" (it could be called an "uncertainty interval") around the estimated concentration. It is imperative that this calculation takes the possible proportional effect into account, because otherwise, the uncertainty interval may be unrealistic: therefore areas with a large "risk" to exceed the threshold are not detected, or areas with very low concentration are added to the "uncertainty zone". But this model is very simplistic: the "distribution" of the concentration is supposed to be symmetric around z^*, and can take negative values (Figure 4a).

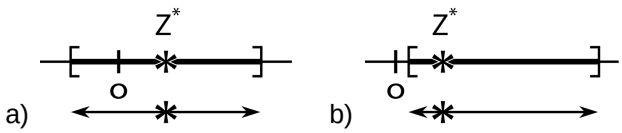

Figure 4. Uncertainty interval around the estimated concentration Z*. a) Simplified model, assuming a Gaussian error; b) classical transformed Gaussian model, with dissymmetric interval around the estimated concentration.

Another seemingly simple method is to estimate the indicator associated with the fixed threshold (this variable is 0 if the data concentration is lower than the threshold and 1 if it is greater). But this encoding does not take into account the measured concentration with accuracy: the coding is identical for a concentration much larger or only a little larger than the threshold. Furthermore it is not possible this way to adequately deal with the change of support between sample and block to be estimated.

Non-linear methods developed for mining estimation can be transposed to the environmental context. The usual classic model assumes that the concentration Z is the transform of a Gaussian variable Y, which is written $Z = \varphi(Y)$, with assumptions on the "spatial distribution" of Y. These assumptions are easy to test using some sample variograms [Rivoirard, 1994]. If these assumptions are valid, the "distribution" of the concentration is computable and can be mapped even with change of support (block model). This distribution is no more necessarily symmetric around z* (Figure 4b). The risk that the concentration exceeds a fixed threshold or an "uncertainty interval" for the real concentration is directly derived from this distribution [Rivoirard, 1994].

Desnoyers (2010) and Musci (2011) show examples of testing the model hypotheses. When the hypotheses of the "transformed Gaussian model" are not valid then other models have to be tested.

MICROPOLLUTANTS

For some organic substances in soils or water (surface or groundwater), a large proportion of values is indicated as lower than a "quantification limit", the exact concentration remaining unknown. Commonly, these values are reduced to half of the quantification limit (QL). This approach can induce bias: distributions of positive variables as concentrations (for example a lognormal distribution) typically show for low values an increasing histogram. The proportion of values between 0 and 1/2 QL is then less than between 1/2 QL and QL. Concentrations below QL have thus an average greater than 1/2 QL.

In addition, putting at the same value all concentrations lower than QL modifies the spatial variability of the concentrations.

The way to handle these variables depends among others on the proportion of values lower than the Quantification Limit, and of the ratio between Quantification Limit and Quality threshold.

When the Quantification Limit is low enough (a ratio of 1/10 between QL and quality threshold can be taken as order of magnitude) it may occurs that different ways of handling non-quantified values provide in practice very similar results [Musci, 2011]. As a precaution, it is propose to put all non-quantified values equal to the quantification limit, possibly adapting the variogram. One can also look for bounds around the result, putting all non-quantified values to 0 (lower bound) or to the Quantification Limit (upper bound) and making a sensitivity study on the variogram.

When the Quantification Limit is close to the quality threshold, or when one is also interested in the sum of several pesticides, a specific model is required.

Handling with "censored" large concentrations requires attention. Indeed, the destruction of a well after a too large measured concentration has as a consequence a truncated concentration time-series, with a preferential lack of large values for the considered substance. This induces a bias in the estimation and prevents to characterize the evolution of concentrations in time, or at least strongly complicates this characterization. Renard et al. (2007) introduced a "persistence" hypothesis, by completing the corresponding truncated time series by the last large value measured if it was above the quality threshold. This pragmatic method does not necessarily give an upper bound for the estimation, but partially corrects the preferential nature of the measurements.

CHARACTERIZATION OF CONCENTRATION EVOLUTION IN GROUNDWATER

Ideally, the characterization of the evolution of concentration in groundwater takes into account the hydrodynamics. But hydrodynamic modeling is imperfect because of the uncertainty on flow parameters, linked to their spatial variability [Delhomme, 1979; de Marsily et al., 2005; Delhomme et al., 2006; Mazuel et al., 2006; Renard, 2007].

The evolution of concentration in groundwater can be characterized by a multivariate kriging (cokriging). Concentrations at two different dates are considered as two different variables, which facilitates the modeling: cokriging concentrations at dates t_1 and t_2 takes into account the measurements available at each of these two dates, and ensures consistency in the estimations (the difference of estimated concentrations at dates t_1 and t_2 is then equal to the estimation of the difference). To know if the evolution is "significant" or not, it must be reported to the associated uncertainty. Cokriging can provide the variance of the estimation error on the concentration at each date and also on the difference. An accurate characterization of the evolution would require maintaining sampling at wells where large concentrations have been previously measured.

CONCLUSION

Methods for site diagnosis and characterization of pollution level in soil or groundwater can still be improved. For example micropollutants pose specific estimation problems. Besides, removing wells where large concentrations were previously measured is not compatible with an accurate characterization of the temporal evolution of the concentrations in groundwater. The permanent increase of substances to be controlled requires a reflection on the systematization of their survey, taking into consideration the costs of the analysis.

Geostatistics offers a range of methods for the assessment of concentrations, but the application to more difficult contexts requires continuing the development of new models. Finally, the dissemination of geostatistical methods in the environmental industry can be further improved.

ACKNOWLEDGEMENT

The author thanks cordially Gaëlle Le Loc'h for her attentive proofreading and Philippe Le Caër for the graphics.

REFERENCES

Matheron G. 1965. Les variables régionalisées et leur estimation. Masson, & C^{ie}, Ed.

Delhomme J. P. 1979. Spatial variability and uncertainty in groundwater flow parameters. *Wat. Res. Res.* 15 (2) 281-290.

Rivoirard J. 1994. Introduction to disjunctive kriging and non-linear geostatistics. Clarendon Press.

Chilès, J.P., Delfiner, P. 1999. Geostatistics: Modeling Spatial Uncertainty. Wiley, New York.

Papritz A., Dubois J.-P. 1999. Mapping heavy metals in soil by (non-) linear kriging: an empirical validation. In Gomez-Hernandez J., Soares A. et al. (eds) geoENV II – Geostatistics for Environ. Appl. Kluwer.

Jeannée N., de Fouquet C. 2003. Apport d'informations qualitatives pour l'estimation des teneurs en milieux hétérogènes : cas d'une pollution de sols par des hydrocarbures aromatiques polycycliques (HAP). *Comptes rendus Geoscience*, 335 (5) 441-449.

Demougeot-Renard H., de Fouquet C. Renard Ph. 2004. Forecasting the number of soil samples required to reduce remediation cost uncertainty. *Journal of Environmental Quality.* 33 (3) 1694-1702.

de Fouquet C., Prechtel A., Setier J. C. 2004. Estimation de la teneur en hydrocarbures totaux du sol d'un ancien site pétrochimiques : étude méthodologique. *Oil&Gas Sci. & Techn. Rev. IFP* 59(3) 275-295

de Marsily G., Delay F., Gonçalvès J., Renard P., Teles V. & Violette S. 2005. Dealing with spatial heterogeneity. *Hydrogeology Journal*, 13, 1:161-183.

Delhomme J. P. & de Marsily G. 2006. Flow in porous media: an attempt to outline Georges Matheron's contributions. In M. Bilodeau, F. Meyer and M. Schmitt (eds). Space, Structure and Randomness: Contributions in Honor of Georges Matheron. Lecture Notes in Statistics Springer: 69-88.

Mazuel S., de Fouquet C., Chilès J.-P., Goblet P., Krimissa M. 2006. Geostatistical modelling for the quantification of uncertainties on the unsaturated zone and the groundwater transfer. *Groundwater hydraulics in complex environments.* IAHR, Toulouse.

Renard D., de Fouquet C. 2007. Characterization of pesticide concentrations in the *Craie du Nord* aquifer system. In Aquifer systems management : Darcy's legacy in a World of impending water shortage. Selected papers from the International association of hydrogeologists (IAH) Dijon symposium 2006. L. Chery, G. de Marsily (eds). Chap 38, 511-524.

de Fouquet C., Gallois D., Perron G. 2007. Geostatistical characterization of the nitrogen dioxyde concentration in an urban area. Part I: spatial variability and cartography of the annual concentration. *Atm. Env.* 41(2007) 6701-6714

Renard Ph. La fin des certitudes ? *Bulletin d'Hydrogéologie No 22 (2007). Centre d'Hydrogéologie, Université de Neuchâtel*. Editions Peter Lang

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

Desnoyers Y. 2010. Approche méthodologique pour la caractérisation géostatistique des contaminations radiologiques dans les installations nucléaires. Thèse de doctorat, Ecole Nat. Sup. des Mines de Paris.

de Fouquet C., Benoit Y., Carpentier C., Faucheux C., Fricaudet B. 2010. On site survey of organic pollutions: some results of the LOQUAS project. *Proceedings Consoil 2010*, Salzburg 22-24 septembre.

Musci F. 2011. Quantifying uncertainty for environmental pollution management in decision making. Scuola interpolitecnica di Dottorato. Final Dissertation. Politecnico di Bari.

de Fouquet C. 2011. From exploratory data analysis to geostatistical estimation: examples from the analysis of soils pollutants. *European Journal of Soil Science*, special issue: Pedometrics. 62(3) 454-466

de Fouquet C., Benoit Y., Carpentier C., Fricaudet B. 2011. Uncertainties on the extension of a polluted zone. Proceedings ASME 14th Int. Conf., ICEM2011-59198, 25-29 septembre 2011, Reims.

de Fouquet C. 2012. Environmental Statistics Revisited: Is the mean reliable? *Environ. Sci. Technol. 46(4) 1964-1970. DOI : 10.1021/es2024143*