Proceedings of the ASME 2011 14th International Conference on Environmental Remediation and Radioactive Waste Management
ICEM2011
September 25-29, 2011, Reims, France

ICEM2011-59198

UNCERTAINTIES ON THE EXTENSION OF A POLLUTED ZONE

Chantal de Fouquet
Mines ParisTech
Ecole des Mines de Paris
Fontainebleau, France

Yves Benoit
IFPEN
Rueil-malmaison, France

Claire Carpentier
ARCADIS
Le Plessis Robinson, France

Bruno Fricaudet
now at : BFR EX POLL
Montgeroult, France

RESUME

Data collected during the sampling of polluted sites are mainly used

- through an exploratory and variographic analysis, to characterize to characterize the concentration level and the spatial variability;
- at fixed support, to estimate the concentrations in order to map the pollution. Kriging gives also the standard deviation of the estimation error, making it possible to delimit the zones in which the estimation is considered to lack in precision. If a proportional effect is present the map of error standard deviation has to be corrected to take into account the increase of spatial variability with the local concentration mean.

A confidence interval can be derived conventionally from the kriging estimation and the associated error standard deviation. For a fixed limit threshold, the polluted site can then be divided in three areas:

- the polluted zone and the not polluted zone, at a fixed statistical risk, not necessary the same for both sets;
- the "zone of uncertainty", in which the estimated concentrations are close to the threshold. Because of the estimation error, it is not possible to specify if the exact concentrations exceed or not the threshold.

For an hydrocarbon soil pollution, usual and geostatistical forecasts are compared. The effective consequences of these various forecasts on the quality of the site remediation are quantified and discussed

INTRODUCTION

The estimation of polluted volumes and their concentration is one of the key points for the economic balance of a remediation project. But because of spatial variability concentrations are never known exactly. The only available maps are maps of estimated concentrations, which differ from the real concentrations. Calculated from the samples these maps are tainted with uncertainties. Geostatistical methods allow to quantify and to reduce some of them.

But spatial variability should be modelled correctly in order to ensure that kriging is actually more precise than an empirical mapping. The preliminary steps of exploratory and variographic analysis are thus needful and important (1, 2 and 3).

The difference between the (unknown) real concentrations and the estimated concentrations comes from the estimation error. The error mean is null (within a probabilistic model) and kriging gives the error variance. The comparison of the concentrations with a limit threshold can then take the estimation error into account, in more or less sophisticated ways. The simple approximation of a Gaussian error is convenient but not always pertinent. The methods of non-linear geostatistics, which are less simple but more rigorous where developed specifically to solve the selection problem.

SOME THEORY

To specify the situation, let us note Z the exact block concentration and Z* its estimation. The selection errors come from a selection made on the estimated value Z* instead of the unknown block concentration Z. Let us denote *t*, the selection

Copyright © 2011 by ASME

"threshold". When the estimation uncertainty is not taken into account, one selects the blocks with $Z^*>t$ instead of selecting those with $Z>s$.

Kriging makes it possible to take the uncertainty into account during the selection. Indeed, the estimation error is modelled as a variable with null mean and with standard deviation given by the kriging standard deviation σ_K. One can write

$$Z = Z^* + \sigma_K.\delta \quad \text{(Eq. 1)}$$

where δ is a variable with null mean and with unit standard deviation.

The *conventional calculation* consists in supposing the reduced error δ Gaussian, and in neglecting the correlation between δ and the estimated concentration Z^*. let us introduce two statistical risks, ζ and η, that the variable δ (and then the concentration Z) be respectively "very low" or "very large", i.e. respectively lower than the percentile q_ζ or greater than the percentile $q_{1-\eta}$: $\zeta = P(\delta < q_\zeta)$ and $\eta = P(\delta \geq q_{1-\eta})$. As the statistical risks are low, $q_\zeta \leq 0$ and $q_{1-\eta} \geq 0$. Up to the risk $\zeta + \eta$, the unknown real concentration lies in the confidence interval:

$$Z^* + \sigma_K.q_\zeta \leq Z < Z^* + \sigma_K.q_{1-\eta} \quad \text{(Eq. 2)}$$

For example, if the two risks are equal to 10%, the associated percentiles are close to ± 1.3, and the confidence interval is

$$\left[Z^* - 1.3\ \sigma_K, Z^* + 1.3\ \sigma_K \right] \quad \text{(Eq. 3)}$$

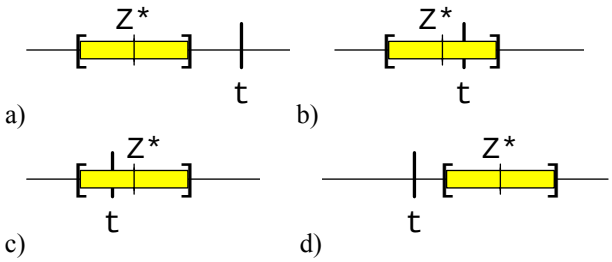

Figure 1. Confidence interval around the estimated concentration Z*, and comparison with the concentration quality threshold t. a) No pollution case and d) pollution case, up to a fixed statistical risk. b) and c) uncertainty case, when the threshold falls in the confidence interval. In (b) the estimated concentration is lower (b) and in (c) greater than the threshold.

Three cases appear when comparing the concentration with a quality threshold t (Fig. 1):

- the not polluted blocks, up to the statistical risk η, for which the upper bound of the confidence interval is lower than the threshold t : $Z^* + \sigma_K.q_{1-\eta} < t$
- the polluted blocks, up to the statistical risk ζ, for which the lower bound of the confidence interval is greater than the threshold: $t < Z^* + \sigma_K.q_\zeta$
- the zone of uncertainty; for these blocks, the threshold falls within the confidence interval. In this case in addition to Eq. 2, we have $Z^* + \sigma_K.q_\zeta \leq t < Z^* + \sigma_K.q_{1-\eta}$, but we don't known if $t < Z$ (not polluted block), or $Z < t$ (polluted block).

In order to reduce pragmatically the risk of no detection of a polluted block (with concentration higher than the threshold), one can first consider a block as polluted if its estimated concentration Z^* is greater than the threshold. This is the same as putting $q_\zeta = 0$, i.e. $\zeta = 50\%$ in the Gaussian case (or when the error distribution is symmetrical), and restricting the uncertainty zone to the interval $\left[Z^*, Z^* + \sigma_K.q_{1-\eta} \right]$.

Second, including the whole uncertainty zone in the area to remediate is the same as considering a block polluted if $t \leq Z^* + \sigma_K.q_{1-\eta}$, or equivalently if $Z^* \geq t - \sigma_K.q_{1-\eta}$. On the *estimated* concentrations one applies the *corrected cut-off*

$$t' = t - \sigma_K q_{1-\eta} \quad \text{(Eq. 4)}$$

lower than the threshold *t. Taking the uncertainty into account necessarily increases the volume to be excavated*, because on the estimated concentration Z^* the selection is made at the corrected cut-off *t'* lower than the fixed threshold *t*.

The corrected cut-off depends on the kriging standard deviation, according to the joint localization of the block and the samples. If a proportional effect is present it is necessary to take into account the relationship between local concentrations mean and local variability, in order to avoid a too large extension of the uncertainty interval for low concentrations, and a too restricted interval for large concentrations areas.

When the uncertainty zone is wide, an additional sampling can be useful in order to reduce the estimation error standard deviation and thus the confidence interval. An economic calculation is needed to compare the additional sampling (and geostatistical study) cost with that of a systematic remediation of the whole uncertainty zone (4). If the spatial structure is little marked if any (low spatial correlation) and the threshold correspond to an intermediate percentile of the concentration distribution, an additional sampling does not always remove the uncertainties.

EXAMPLE

The LOQUAS project (LOCALIZATION and QUANTIFICATION of an organic Soil pollution) associated research institutes and companies. On former industrial sites polluted by hydrocarbons, a systematic survey was implemented, associating Gas Phase Chromatography and Pollut-Eval® pyrolysis measurements which allowed an exceptional systematic sampling, from decimetre to decametre scale (5, 6). These data made a detailed geostatistical analysis of the spatial variability possible, resulting in a large number of results, methodological or of practical interest.

Copyright © 2011 by ASME

The consequences of uncertainties related to the concentrations on the technical dimensioning of a remediation project are examined by simulation. A numerical model of samples and blocks concentrations, representing the spatial variability of a spot pollution, is built in reference to an experimental site. To simplify only a "unit layer" with 1 m thickness is considered.

The “fictitious” site of approximately 4 ha is supposed to be recognized according to a regular 32m-grid, in conformity with the practice of the offices in consideration of the site dimensions. The data location is presented on Fig. 2. The statistics of the 32 samples are the following: minimum 135 mg/kg, maximum 34750 mg/kg, mean 6900 mg/kg and coefficient of variation (ratio of the standard deviation to the mean) of 165%. The median (950mg/kg) is much lower than the mean, because of the distribution asymmetry.

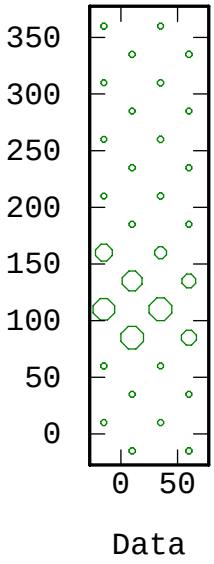

Figure 2. Data location, on 32m-grid. The symbol size is proportional to the sample concentration.

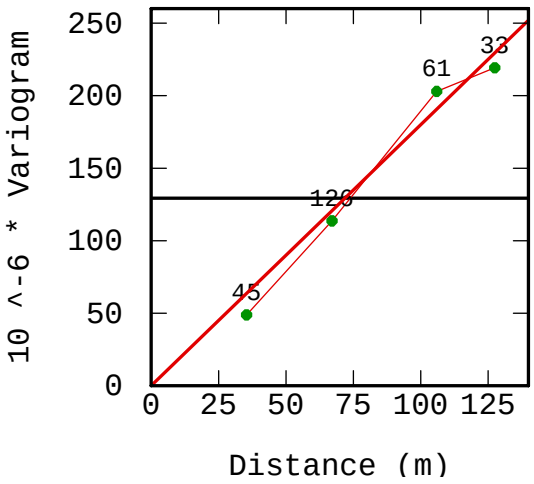

Figure 3. Sample variogram and fitted isotropic linear model.

The sample variogram is fitted with an isotropic linear model (Figure 3).

From the samples, the concentrations are estimated on 25m^3-blocks according to two methods:

- by “nearest data” (polygon of influence), a process rather close to those usually applied by the offices;

- by block kriging.

“Nearest data” interpolation and block-kriging are quite different (Fig. 4). The map of kriging error standard deviation takes into account the proportional effect model (Fig. 5). In spite of the regular sampling grid, the estimation uncertainty is larger on the pollution spot.

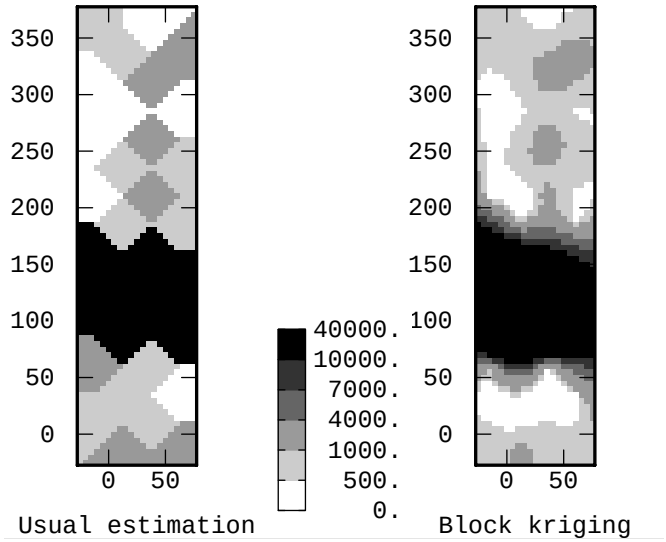

Figure 4. Estimation usuelle et krigeage de la teneur de bloc.

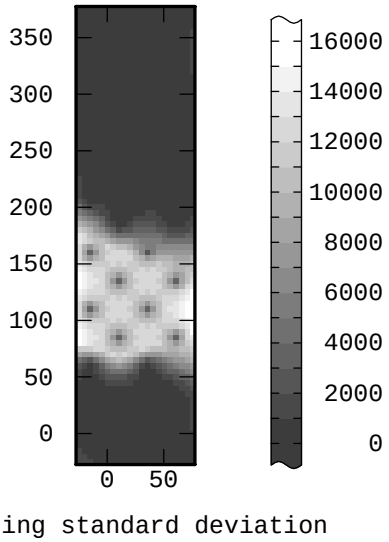

Figure 5. Standard deviation map of the block-kriging error. The proportional effect is taken into account.

Copyright © 2011 by ASME

The decision threshold is fixed at 7000 mg/kg of soil, according to a possible future use of the site. This value is close to the sample mean as well as to the 75% percentile (7130mg/kg).

The chosen statistical risks are symmetrical and equal to 10% ($\zeta = 10\%, \eta = 10\%$). Figure 6 shows the area considered as « not polluted » (zone A, 68% of the site surface), the area considered as polluted (zone C 13% of the site) and the uncertainty area divided in two parts. In one part the estimated concentration is lower than the threshold (B1, 4% of the site) and in the other one the estimated concentration is larger than the threshold (B2, 15%). The whole uncertainty area (19% of the site surface) represents more than twice the surface of the spot on estimated concentration (where estimated concentration is larger than the 7000mg/kg threshold).

In this case, the economic balance between systematic remediation and additional sampling in and around the large concentration spot should be examined.

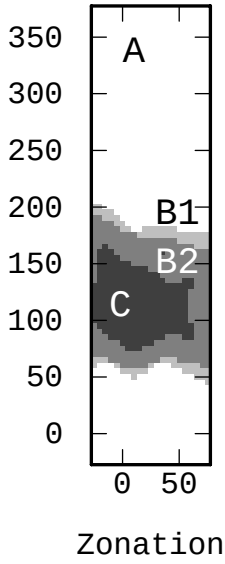

Figure 6. Zonation of the site in the not polluted zone (A), polluted zone (C), and uncertainty zone with estimated concentration lower than the threshold (B1) or greater (B2). Threshold at 7000 mg/kg of soil and statistical risks at 10% each.

SELECTION ERRORS

For each of the two estimation methods, the forecasts for volume of polluted soils are made by comparing the estimated concentration with the fixed threshold. For kriging, zones B2 and C are thus selected.

In the case of kriging, a third calculation is made by applying the "corrected threshold" $t' = t - 1.3\sigma_K$ to the estimated concentrations in order to take the uncertainty into account. The whole "uncertainty zone" is then supposed to be extracted: zones B1, B2 and C are thus selected.

In the real practice, the exact concentration of the blocks is not known. Only a (traditional or geostatistical) estimation is available, on which the forecasted result is calculated: volume of polluted soil, average concentration. But the selection is operated on real blocks. On a simulation, the exact concentration of the blocks (which were simulated jointly with the samples) is available. One can thus calculate the result actually obtained, by considering the exact concentration of the blocks which were selected according to their estimated concentration. Two types of error can thus be evaluated:

- the blocks whose exact (unknown) concentration exceeds the threshold, but which are regarded as not polluted, because their estimated concentration is lower than the threshold; these blocks are wrongly "abandoned";

- the blocks whose exact (unknown) concentration is lower than the threshold, but which are wrongly considered as polluted, because their estimated concentration is higher than the threshold.

The first selection error induces a risk, because remediation is insufficient; the second error induces an over cost, because of useless remediation (treatment of the soils, storage).

Lastly, it is possible on a simulation to calculate the result of the "ideal" selection, carried out according to the exact concentration of the blocks. In practice, this ideal selection is obviously impossible. On a simulation, the forecasts and the actual result can be confronted with the result of this "ideal" selection.

Following results are compared:

- the forecasted result, calculated on Z*;

- the actual result, calculated while selecting according to Z*, while returning then to the exact block concentration Z;

- the "ideal result", calculated on Z (with selection according to Z).

In addition to the selected volume (or tonnage), and the average concentration of these blocks, one is interested in the product Q of the volume by the concentration. Up to a factor (supposing the soil density independent of the concentration) this product Q is equal to the mass of pollutant in the contaminated blocks. Considering the selected blocks whose exact concentration is higher than the threshold, the ratio of the corresponding mass Q_e to the ideally selected mass Q_{id} expresses the percentage of pollution actually recovered in the polluted zone. In the ideal selection, this ratio is equal to 100%.

The results are recapitulated in table 1 For the two estimation methods (usual "nearest data" and kriging), the forecasts calculated according to the estimated concentrations, as well as the effective result, calculated according to the exact blocks concentration are given. Table 2 presents the result of the ideal selection carried out according to the exact blocks concentration in the absence of any estimation error (this is obviously impossible in practice).

Copyright © 2011 by ASME

At first sight kriging appears less favourable than the usual estimation, because it leads to excavate more volume. But with the usual estimation a part of the pollution is not detected and 22% of the excavated blocks are lower than the threshold. The deviation between the forecasted and the actual concentration of the excavated blocks is important, and could lead to wrong technical choices for soil stocking or remediation.

Selection on kriging without taking uncertainties into account (zones B2 and C) increases the tonnage in comparison with the previous usual method. 23% of the excavated blocks are lower than the threshold and only a small part of the pollution is not detected. The deviation between the forecasted and the actual concentration of the excavated blocks is here lower than in the previous case.

		excavated volume in $25m^3$ blocks	mean concentration of selected blocks mg kg^{-1}	recovered pollution Q / Q_{ideal}
usual estimation	forecast	420	25 185.	
	actual result	+92 wrongly -35 not detected	18 405.	93%
block-kriging	forecast	473	21 400	
	actual result	+111 wrongly -1 not detected	17 520	99%
block-kriging with uncertainty zone	forecast	543	19 312.	
	actual result	+180 wrongly 0 not detected	15 365.	100%

Table 1. Comparison of forecast made on estimated concentration with actual result calculated on the exact concentration of the selected blocks. Selection on usual estimation by « nearest data », on block-kriging and on block-kriging at "corrected" threshold to take uncertainties into account. Threshold: 7000 mg/kg of soil.

The volume deviation between the two estimation methods could be different if the threshold is move very aside of the mean (7).

Taking uncertainties into account allows detecting the whole pollution, at the price in this example of an increase of the volume to be remediated.

The "ideal" but impossible selection (the exact block-concentration remains always unknown before excavation) corresponds here to the smallest volume to be excavated and to the highest associated mean concentration, because none of these blocks is wrongly selected (tableau 2).

In practice only forecasts are available for the determination of the remediation workings. One observes that the more "optimist" forecast based on the usual estimation method (with lower volume and larger concentration than with kriging) has an incomplete remediation as actual consequence. The balance of the remediation project can be questioned if additionally works are indispensable.

	excavated volume in $25m^3$ blocks	mean concentration of selected blocks mg kg^{-1}	recovered pollution Q_{ideal}
result on exact concentrations	363	29 650	$8.22\ 10^6$

Table 2. « Ideal » selection on exact blocks concentration. Threshold: 7000 mg/kg of soil.

TOWARD NON LINEAR GEOSTATISTICS

The conventional calculation based on Eq. 1 and on a Gaussian hypothesis for the estimation error is not always realistic. The Gaussian hypothesis is not necessary (8) but at fixed statistical risk the confidence interval becomes then wider (Eq. 2), or equivalently at same confidence interval the statistical risks are greater.

The bounds of the confidence interval (Eq. 2) derived from the direct modelling of the estimation error are not always pertinent. For example the lower bound can be negative. In a more rigorous modelling the concentration is considered as a transform $Z = \Phi(Y)$, were for example Y is a « Random Function with Gaussian spatial distribution », and the function Φ is defines the model for the concentration histogram. These non linear methods take the « change of support » into account, i.e. the different variability between the available sample concentration and the block concentration (to be modelled).

Copyright © 2011 by ASME

Used for mining estimation for decades, conditional expectation and disjunctive kriging are more and more applied for pollution estimation (2, 7, 9 and 10).

Lastly, in order to improve the precision, other information than the concentrations can be taken into account, for example organoleptic observations (11).

REMERCIEMENTS

The LOQUAS project was supported by the French ANR.

REFERENCES

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

(2) Hofer, C., and A. Papritz, 2010. "Predicting threshold exceedance by local block means in soil pollution surveys". Mathematical Geosciences, 42, (6), 631-656, doi: 10.1007/s11004-010-9287-4

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

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

(5) Faucheux C., Lefebvre E., de Fouquet C., Benoit Y., Fricaudet B., Carpentier C., and Gourry J.-C. 2008. « Characterisation of a hydrocarbon polluted soil by an intensive multi-scale sampling." In *Geostats 2008, proceedings of the 8th international geostatistics congress*, 1-5 dec. 2008, Santiago, Chile. Ortiz J.-M., Emery X. eds.

(6) Benoit Y., de Fouquet C., Polus-Lefebvre E., Faucheux C., Fricaudet B., Carpentier C.,Gourry J.-C., and Haudidier N. «LOQUAS : Combinaison des reconnaissances géophysiques et physico-chimiques pour l'estimation géostatistique de pollutions de sols par hydrocarbures. Deuxièmes rencontres nationales de la recherche sur les sites et sols pollués : pollutions locales et diffuses.» ADEME. Paris, 20-21 octobre 2009.

(7) Rivoirard J. 1994. Introduction to Disjunctive kriging and Non-linear geostatistics. Oxford University Press. Oxford.

(8) Chilès J.-P., Delfiner P. Geostatistics. Modeling spatial uncertainty. Wiley series in probability and statistics. 1999.

(9) Demougeot-Renard H., and de Fouquet C. 2004. Geostatistical approach for assessing soil volumes requiring remediation: validation using lead-polluted soils underlying a former smelting works. *Environmental science & technology.* 38 (19) 5120 – 5126.

(10) Desnoyer Y. 2010. « Approche méthodologique pour la caractérisation géostatistique des contaminations radiologiques dans les installations nucléaires ». Thèse de Doctorat, Ecole nationale Supérieure des Mines de Paris. Mines ParisTech.

(11) Jeannée N., and 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.

Copyright © 2011 by ASME