Published in Journal of Environmental Quality (JEQ) 33, 1694-1702, 2004
which should be used for any reference to this work

Forecasting the Number of Soil Samples Required to Reduce Remediation Cost Uncertainty

Hélène Demougeot-Renard,* Chantal de Fouquet, and Philippe Renard

ABSTRACT

Sampling scheme design is an important step in the management of polluted sites. It largely controls the accuracy of remediation cost estimates. In practice, however, sampling is seldom designed to comply with a given level of remediation cost uncertainty. In this paper, we present a new technique that allows one to estimate of the number of samples that should be taken at a given stage of investigation to reach a forecasted level of accuracy. The uncertainty is expressed both in terms of volume of polluted soil and overall cost of remediation. This technique provides a flexible tool for decision makers to define the amount of investigation worth conducting from an environmental and financial perspective. The technique is based on nonlinear geostatistics (conditional simulations) to estimate the volume of soil that requires remediation and excavation and on a function allowing estimation of the total cost of remediation (including investigations). The geostatistical estimation accounts for support effect, information effect, and sampling errors. The cost calculation includes mainly investigation, excavation, remediation, and transportation. The application of the technique on a former smelting work site (lead pollution) demonstrates how the tool can be used. In this example, the forecasted volumetric uncertainty decreases rapidly for a relatively small number of samples (20–50) and then reaches a plateau (after 100 samples). The uncertainty related to the total remediation cost decreases while the expected total cost increases. Based on these forecasts, we show how a risk-prone decision maker would probably decide to take 50 additional samples while a risk-averse decision maker would take 100 samples.

Managing industrial polluted sites requires an assessment not only of the volume of polluted soil but also of the corresponding uncertainty. Indeed, uncertain estimates result in environmental and financial risks. On one side, polluted soil may stay in place or be discovered unexpectedly, while on the other side, clean soil may be excavated for no reason. The uncertainty results from the small quantity of information available as compared with the complexity of the spatial patterns of pollutant concentration in the soil. Geostatistical techniques allow modeling of the uncertainty based on available data. As a consequence, it is now becoming more and more frequent to use geostatistical techniques to map soil pollution, estimate the volumes exceeding a given concentration threshold, or improve the process of selective remediation (Flatman and Yfantis, 1984; Garcia and Froidevaux, 1996; von Steiger et al., 1996; Hendriks et al., 1996; Demougeot-Renard and de Fouquet, unpublished data, 2003; Saito and Goovaerts, 2003).

The consulting practice shows, however, that high volumetric and cost uncertainties are still common when designing a remediation scheme. One of the reasons is that at the beginning of a contaminated site investigation, sampling is mainly designed to answer the specific questions raised by the risk assessment: evaluation of the toxicological and ecotoxicological properties of pollutants, or investigation of the risk of migration. Sampling is subsequently used to map pollutant concentrations and assess remediation volumes. Sampling in this phase is thus generally not designed for geostatistical interpolations and additional sampling becomes necessary. It may be questioned, therefore, at a given stage of a site investigation, whether it is possible to use all available information to forecast the optimal number of additional samples to sufficiently reduce the volumetric and cost uncertainty.

Because of its high practical interest, optimal sampling procedure is a topic widely studied in the literature. Within the geostatistical framework, a classical issue is the definition of rules to locate the additional samples when the variogram is known. The most usual approach is to locate subsequent samples where the interpolated concentrations are the most uncertain (i.e., where the kriging standard deviation is high) (Burgess et al., 1981). An improved approach is to identify the points where the probability to have a high concentration is above a threshold (Van Tooren and Mosselman, 1996; Van Groeningen et al., 1997). Johnson (1996) follows the same concepts but within a combined geostatistical–Bayesian approach. Barabàs et al. (2001) used indicator kriging within a transformed space (to follow the geometry of an estuarine river) to delineate contaminated areas and to forecast additional sampling needs. One of the most recent tools to optimize sampling locations is the simulated annealing technique, which allows searching for optimal locations while accounting for constraints such as building locations (Van Groeningen et al., 2000). However, it was soon recognized that optimal sampling should also account for economical factors

H. Demougeot-Renard, Eidgenössische Technische Hochschule Zürich, Institut für Raumplanung und Landschaftsentwicklung, Hönggerberg, CH-8093 Zürich, Switzerland. Current address: University of Neuchâtel, Centre for Hydrogeology, 11 rue Emile Argand, CH-2007 Neuchâtel, Switzerland. C. de Fouquet, Ecole Nationale Supérieure des Mines de Paris, Centre de Géostatistique, 35 rue Saint Honoré, 77305 Fontainebleau, France. P. Renard, Université de Neuchâtel, Centre d'Hydrogéologie de Neuchâtel, 11 rue Emile Argand, CH-2007 Neuchâtel, Switzerland. Sponsoring organizations: Agence De l'Environnement et de la Maîtrise de l'Energie (ADEME), 2 Square Lafayette, BP 406, 49004 Angers, Cedex 1, France. Gaz De France (GDF), 361 avenue du président Wilson, 93211 Saint Denis, France. Received 7 Oct. 2003. *Corresponding author (helene.demougeot@unine.ch).

Abbreviations: C_c, cleanup cost; C_i, investigation cost; C_r, overall remediation cost; C_u, uncertain cost; S, remediation cutoff; V_e, excavated volume; V_{lr}, low-risk volume; V_c, cleanup volume; V_s, stored volume; V_u, uncertain volume; α, inferior probability threshold; β, superior probability threshold.

(Srivastava, 1987). Along this line, Englund and Heravi (1993, 1994) proposed a strategy to minimize remediation costs, including sampling costs and misclassification error costs. Okx and Stein (2000) combine statistical decision trees and indicator kriging to assess the economic validity of an additional sampling stage.

All of these methods present one or several drawbacks. Often, the probabilities of exceeding a concentration cutoff are calculated approximatively, on the basis of ordinary kriging (Burgess et al., 1981; Englund and Heravi, 1993, 1994) or residual kriging (Van Tooren and Mosselman, 1996). Some authors use indicator kriging, but in that case they reduce the pollution phenomenon, often described as a continuous spatial variable, to a discrete phenomenon (Johnson, 1996; Van Groeningen et al., 1997; Okx and Stein, 2000; Barabàs et al., 2001). Another issue is that the sampling strategy is based on estimations, but not on forecasts (Van Tooren and Mosselman, 1996; Van Groeningen et al., 1997). An additional data sampling is advised on the basis of probabilities calculated using real data. Probabilities that may be calculated taking into account additional data that are still not sampled are not forecasted. Furthermore, sampling is usually not optimized with respect to economical criteria (Van Tooren and Mosselman, 1996; Johnson, 1996; Van Groeningen et al., 1997). Finally, most of these approaches aim at reducing the environmental and financial risks due to soil misclassification using probability mapping, while the decision makers think instead in terms of overall amounts of soil requiring remediation.

The aim of this paper is to propose a new methodology to forecast the number of samples required to reach an acceptable uncertainty level in terms of overall volumes and costs. The methodology is applied during an iterative sampling campaign. The calculations are based on conditional simulations, which permit taking nonlinearity into account as well as modeling the specific features related to remediation activities. The basic principle of the calculation is to simulate the addition of new samples within the geostatistical model and then to forecast the evolution of the volume, the cost, and their respective uncertainties. The methodology includes (with great care) the possible effects that may bias the estimations (sampling errors, support effect, information effect) and follows in detail the procedure, which is used in the field to excavate a polluted soil. Finally, the method is illustrated on a former smelting work, which polluted the underlying soils with lead.

MATERIALS AND METHODS

We assume that preliminary site investigation and risk assessment have shown that part of the soil requires remediation. We also assume that we are in the phase of site investigation, the aim being to estimate the amount of polluted soil and to estimate a remediation cost. Thus, we are not yet in the remediation phase. It is further assumed that soils with pollutant concentrations greater than the remediation cutoff (S) will be treated either on site or ex situ, since nowadays, these are the most frequently applied remediation techniques. Soils will thus be selected and excavated before being sent to the treatment unit (Fig. 1). We consider that volumes and costs are "estimated" when they are calculated using real data. We consider that volumes and costs are "forecasted" when they are calculated using both real data and "simulated" data of an additional "simulated" sampling stage. We use "calculations" when both estimates and forecasts are considered.

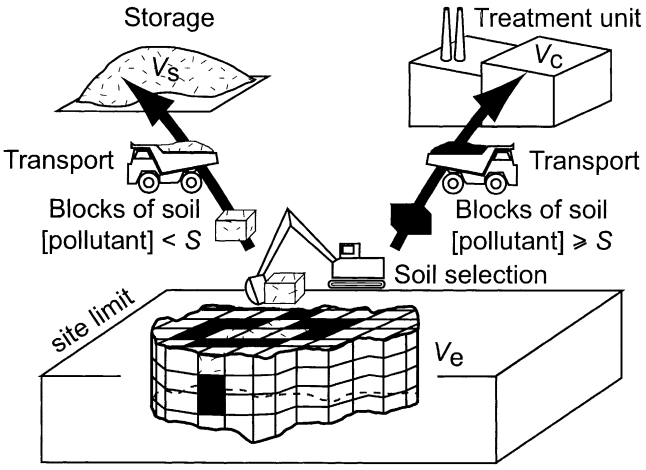

Fig. 1. The different steps of a common remediation works (soil selection, excavation, transport, storage, and treatment), and the corresponding soil volumes (V_e, excavated volume; V_s, stored volume; V_c, cleanup volume).

Principle

The methodology is integrated into an iterative sampling campaign. After a sampling stage j, the aim is to model the impact of the future resampling (stage $j + 1$) on the uncertainty. Five main steps are involved in the calculation:

Step 1. Remediation volumes and uncertainty are estimated from the available investigation data.
Step 2. Remediation costs and uncertainty are estimated.
Step 3. A conditional simulation is used to simulate the sampling of additional data in those parts of the site where the volumetric and financial accuracy is insufficient.
Step 4. Remediation volumes and uncertainty are forecasted using both simulated and real data.
Step 5. Remediation costs and uncertainty are forecasted.

Step 3 to 5 are repeated for a series of increasing numbers of additional simulated data. It allows forecasting of the evolution of uncertainty with the number of samples. The result is then used as a decision aid tool in the design of the next sampling stage. After the sampling stage $j + 1$, the volume and cost estimates may be different from the forecasts calculated at stage j. Such differences may occur, especially at the beginning of the investigation when few data are available. Estimations at stage $j + 1$ are, as a matter of fact, based on the available information on the spatial structure at stage $j + 1$ while forecasts at stage $j + 1$ are based only on the available information at stage j. The sampling design procedure can then be applied to a supplementary "simulated" sampling stage $j + 2$. The iterative sampling is stopped when the volumetric and financial uncertainty estimations are acceptable. The different steps of the procedure are detailed in the following sections.

Calculation of Remediation Volumes and Uncertainties Using Geostatistics (Steps 1 and 4)

Both the estimation and forecast of the volumes are obtained using conditional simulations of pollutant concentra-

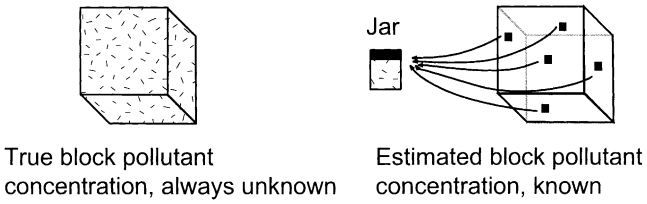

Fig. 2. Information effect: the true block concentration, always unknown, is estimated by the concentration of a composite of samples collected in the block.

tions. A simulation $T(x)$ of a random function $[Z(x): x \in R^n]$ is defined as a realization of $Z(x)$, randomly selected from the set of all possible realizations (Matheron, 1973; Journel, 1974). The term $T(x)$ follows the same spatial distribution, in particular the same expectation, covariance, and histogram as $Z(x)$. It does not minimize the estimation variance, so it is nonbiased, but not an optimal estimator of $Z(x)$. Because $T(x)$ reproduces the spatial distribution of $Z(x)$, the application of a cutoff S to $T(x)$: $[T(x) \geq S]$ provides a nonbiased estimator of $[Z(x) \geq S]$. Moreover, a conditional simulation $T(x|n)$ honors the experimental data at the n sampling points. A diffusive model is selected here because it permits simulation of continuous variables such as pollutant concentrations. Conditioning data are real data for volumetric estimations, while they include both real and "simulated" data for volumetric forecasts. The cumulative distribution function $F(S)$ of $Z(x)$ defines the probability that $Z(x)$ is greater than the cutoff S: $F(S) = P[Z(x) > S|n]$. This probability can be estimated using a set of K conditional simulations. At each location x, the probability estimate is calculated as the ratio of the number of simulated values $T_i(x|n)$ that exceed S, to the total number K of simulated values:

$$P[T(x) > S|n] = \frac{1}{K}\sum_{i=1}^{K}\{I[T_i(x|n);S]\} \qquad [1]$$

where $I[T_i(x|n);S) = 1$ when $T_i(x|n) > S$, and $I[T_i(x|n);S) = 0$ otherwise. In addition, the remediation volumes can only be correctly assessed if the conditions of soil selection for cleanup are modeled. Three effects have to be accounted for:

(i) Change of support. In Europe, the remediation volume is usually estimated using site investigation data measured in boreholes. The investigation samples are thus rather small (order of magnitude: dm^3). They can be considered as points, while soils will be excavated for cleanup in blocks (remediation blocks) of large dimensions (order of magnitude: m^3). The remediation volume resulting from applying S on a distribution of point pollutant concentrations differs from the volume resulting from applying S on a distribution of block pollutant concentrations (Rivoirard, 1994).

(ii) Information effect (Fig. 2). During the remediation stage, blocks of soil are selected for cleanup using pollutant concentrations measured in composite samples. One composite sample is composed of a set of L small samples collected systematically or randomly in each block. However, there is a difference between the pollutant concentration measured in the composite and the true yet unknown block concentration (Rivoirard, 1994). Applying S on the distribution of concentrations measured in composite samples thus provides a different remediation volume estimate from the volume estimate resulting from applying S on the distribution of the true block concentrations.

(iii) Sampling and analytical errors. They depend on the laboratory equipment, field equipment, and the experience of the operators, but they always affect real data.

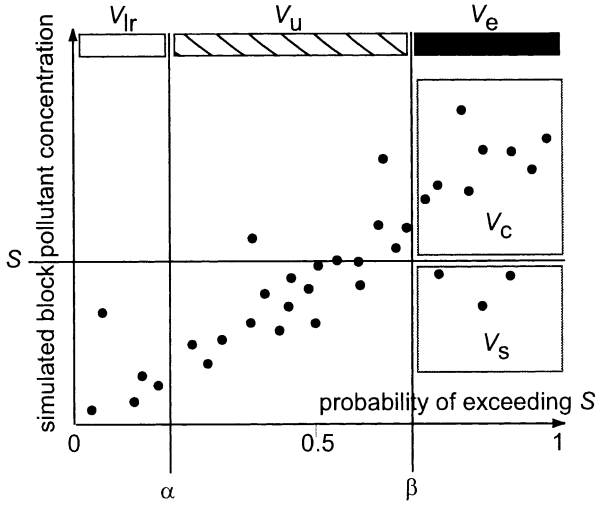

Fig. 3. Theoretical graph of block concentration as a function of block probabilities of exceeding the remediation cutoff *S*. It illustrates the meaning of the low-risk volume V_{lr} (blocks of soils with probabilities less than α), the uncertain volume V_u (blocks of soils with probabilities greater than α and less than β), and the excavated volume V_e (blocks of soils with probabilities greater than β). The excavated volume is composed of the cleanup volume V_c (blocks of soils with pollutant concentrations greater than *S*) and the stored volume V_s (blocks of soils with pollutant concentrations less than *S*).

To account for these soil selection conditions, we performed conditional simulations in the following manner. A series of K point simulations is generated on a fine grid discretizing the blocks of the remediation grid. To account for the information effect, each block-simulated concentration is subsequently estimated as the mean of L simulated point concentrations:

$$T_v(x|n) = \frac{1}{L}\sum_{j=1}^{L} T_{pj}(x|n) \qquad [2]$$

where T_v is a block simulation, T_p a point simulation, and L the number of point samples used to make the composite sample. To model the presence of sampling and analysis errors, a variable $\varepsilon(x)$ is added as follows:

$$H_V(x|n) = T_v(x|n)[1 + \varepsilon(x)] \qquad [3]$$

The variable $\varepsilon(x)$ is a relative error chosen with a uniform distribution between -1 and $+1$, and independent of $T_v(x|n)$. The resulting variable $H_V(x|n)$ models realistic block-simulated values, which are used to calculate the probabilities of exceeding S in the blocks. At each block, the probability is the ratio of the number of block-simulated values exceeding S to the total number K of simulations. The remediation volumes are calculated using the resulting set of block conditional simulations and block probabilities (Fig. 1 and Fig. 3), accounting for the two phases that usually lead to a refined volumetric estimation:

(i) At the end of the site investigation, the envelope V_e for the remediation volume is estimated.

(ii) During the remediation stage, the remediation volume is more finely investigated. Within V_e, pollutant concentrations are measured systematically in the blocks of the remediation grid (Fig. 1). Due to the remaining uncertainty, V_e still includes nonpolluted blocks. Consequently, V_e is made of both the cleanup volume V_c com-

posed of blocks of soils with concentrations exceeding S, and the stored volume V_s composed of nonpolluted blocks of soils to be excavated to access the polluted blocks.

The first remediation volume estimate V_e is defined as the set of blocks with probabilities higher than a maximal "acceptable" probability β:

$$P[H_V(x) > S|n] \geq \beta \quad [4]$$

The probability threshold β corresponds to the maximal acceptable risk of soil misclassification, which is either a risk of including nonpolluted soils in V_e, or a risk of excluding polluted soils of V_e. The major difficulty is to define β taking into account the risk assessment study, the possible site use following restoration, and the general context in which restoration takes place. Within V_e, the volume of soil $V_c(i)$ where pollutant concentrations exceed S is calculated for each block conditional simulation (index i). The term $V_c(i)$ is calculated as the sum of the unit volumes V_b of the t block-simulated values exceeding S included in V_e:

$$V_c(i) = \sum_t V_b \, I[H_{Vit}(x|n); S] \quad [5]$$

Similarly, the volume of soil $V_s(i)$ in V_e where the pollutant concentration falls below S is calculated for each block conditional simulation i. The term $V_s(i)$ is calculated as the sum of the unit volumes V_b of the t block-simulated values falling below S:

$$V_s(i) = \sum_t V_b \, \{1 - I[H_{Vit}(x|n); S]\} \quad [6]$$

These calculations lead to two distributions of volumes $V_c(i)$ and $V_s(i)$. Depending on the skewness of the distributions, their mean or their median provides an estimate of the volumes V_c and V_s excavated from V_e. Uncertainties of the volumetric calculations are modeled inside and outside V_e. Inside V_e, the volumetric uncertainty is modeled by the dispersion of the distributions $V_c(i)$ and $V_s(i)$. This may be quantified by the interquartile range $[Q_{25\%} - Q_{75\%}]$ or by the coefficient of variation σ/μ. Outside V_e, the volumetric uncertainty can be modeled providing the definition of a second probability threshold α ($\alpha < \beta$). Environmental risks may be considered as nonsignificant outside V_e for blocks with a probability smaller than α:

$$P[H_V(x) > S|n] \leq \alpha \quad [7]$$

These blocks can remain without any remediation or additional monitoring. They make up the low-risk volume V_{lr}. Environmental and financial risks are still significant in the blocks where probabilities exceed α and fall below β:

$$\alpha < P[H_V(x) > S|n] < \beta \quad [8]$$

Blocks with pollutant concentrations exceeding S may remain untreated while blocks with pollutant concentrations below S may be excavated for remediation, although unnecessary. These blocks make up the uncertain volume V_u, which represents the nonacceptable uncertainty remaining outside V_e.

Calculation of the Overall Remediation Cost and Uncertainty (Steps 2 and 5)

Cost estimates and cost forecasts are calculated using the block simulations and block probabilities. Cleanup costs include both the costs of remediating the polluted soils within V_e and the costs due to the excavation and the storage of nonpolluted soils within V_e. The cost function has been developed to calculate a frequency distribution of cleanup costs $C_c(i)$ from the volumetric estimates V_e, $V_c(i)$, and $V_s(i)$. The cost function (Renard-Demougeot, 2002) is parameterized to fit various pollution scenarios and different commercial and technical proposals. The general cost model is the following:

If $V_C < V_{threshold}$:

$$C_c = A/V_e + BV_{sort} + \sum_{q=1}^{nc} \left\{ \sum_{j=1}^{np} [C_1(q,j)V_c(q,j)] + D(q)V_c(q) \right\} + EV_s + FV_c \quad [9]$$

If not:

$$C_c = A/V_e + BV_{sort} + \sum_{q=1}^{nc} \left\{ \sum_{j=1}^{np} [C_2(q,j)V_c(q,j)] + C_3(q,j)] + D(q)V_c(q) \right\} + EV_s + FV_c$$

where A, B, C_1, C_2, C_3, D, E, and F are composite unit prices allowing inclusion of planning, cleanup, transport, excavation, storage, filling up, and soil screening. Because in practice, different remediation techniques may be applied simultaneously in V_e depending on the pollutant concentrations and the granulometric classes of the soil, the cleanup volume V_c has to be partitioned in granulometric and concentration classes. The notation is $V_c = \sum_{q=1}^{nc} \sum_{j=1}^{np} V_c(q,j) = \sum_{q=1}^{nc} V_c(q)$, where nc and np are respectively the number of concentration and granulometric classes. The total cleanup cost C_c is the sum of the costs of these different remediation techniques. Furthermore, a technical proposal may provide different cleanup unit prices. Unit prices may be higher for the volume of soil exceeding the remediation volume $V_{threshold}$ that has been declared before starting remediation.

Cost uncertainty outside V_e is defined as the cost of unexpected discovery of soils requiring remediation in V_u. A distribution of cost uncertainty $C_u(i)$ is calculated applying the same cost function (Eq. [9]) to the uncertain volume V_u and to the distributions of volumes of polluted soils $V_cu(i)$ and nonpolluted soils $V_su(i)$ included in V_u.

Sampling optimization requires balancing the increase of volume and cost accuracy, with the increase of investigation cost. Investigation costs C_i are calculated applying a parameterized investigation cost function to the number of samples. Finally, for a given number of samples collected at investigation stage j, an overall remediation cost C_r is defined as the sum of investigation cost C_i, cleanup cost C_c, and cost uncertainty C_u:

$$C_r = C_i + C_c + C_u \quad [10]$$

A distribution $C_r(i)$ is obtained from C_i, $C_c(i)$, and $C_u(i)$. Depending on the skewness of the distribution, the mean or the median of this distribution provides an estimate of the overall cost.

Modeling Sampling of Various Numbers of Additional Data (Step 3)

If at stage j, the remaining uncertainties on the volume and cost estimates are too high for decision making, we propose to forecast the uncertainties on volumes and costs that could be estimated if additional samples were collected at stage $j + 1$ as follows:

Step 1. A point conditional simulation of pollutant concentrations, generated at stage j, is selected randomly. It is consid-

ered as the reference of the state of pollution of the industrial site.

Step 2. The N_{j+1} point-simulated values are selected on the reference. They are taken in the uncertain volume V_u, at the nodes of a regular rectangular grid. If necessary, the axes of the grid are oriented according to the anisotropy of the spatial structure. Such a sampling scheme is commonly recommended for inferring the spatial structure of a phenomenon (Flatman et al., 1988).

Step 3. A new set of block simulations is generated (Eq. [2] and [3]), conditioned by the N_j real site investigation data and the N_{j+1} "simulated" additional data, using the variogram model fitted at stage j.

Step 4. Volumes V_e, $V_c(i)$, $V_s(i)$, V_u, and V_{lr} are recalculated with that new set of simulations (Eq. [4]–[8]).

Step 5. Costs $C_c(i)$ and $C_u(i)$ are recalculated for these volumes (Eq. [9]). The investigation cost C_i is recalculated accounting for N_{j+1} additional data. The overall remediation cost $C_r(i)$ is subsequently calculated (Eq. [10]).

Steps 2 to 5 are repeated for a series of successive values of N_{j+1}. The results are represented in the form of two graphs: a volume forecast graph representing V_e and V_u for stage $j + 1$ as a function of N_{j+1}, and a cost forecast graph representing C_i, $C_c(i)$, and $C_u(i)$ for stage $j + 1$ as a function of N_{j+1}. These graphs are then used as decision criteria in the design of the following sampling phase.

APPLICATION TO A FORMER SMELTING WORK

Site Description

The former smelting work is located in France. The site covers an area of 3 ha. The water table lies at around 2 m below ground. The principal wind directions in this area are west to north-northeast. Historical data have shown that the chimney of the smelting work was responsible for dispersion of nontreated dust and smoke. Seventy five samples were collected from the smelting site and its neighborhood at depths of 0 to 4 m, in six sampling stages (Fig. 4). The sampling support was homogeneous, consisting of soils collected along boreholes on a height of 25 cm. Lead concentrations were measured in all the samples (Row 1 of Table 1). A detailed risk assessment study has shown that soil with Pb concentration >300 mg kg^{-1} involves a risk for human health via dust inhalation. The 300 mg kg^{-1} value was thus chosen as the regulatory remediation cutoff S, without specifying at which sampling support S had to be applied. Soil was remediated by soil washing on site. The zone that was excavated for remediation had been delineated without using geostatistics. This zone was sorted according to 212 Pb concentrations measured in blocks of a selective remediation grid. The Pb concentrations were measured on composite samples made of four small samples taken at the corners and one small sample taken at the center of each block (Fig. 2). One block had a 10-m side length and a 0.30-m height. The soil was excavated in three layers and either sent to the washing unit or a storage area.

Estimating Volumes and Costs at the Sixth Investigation Stage (Steps 1 and 2)

The remediation volumes and the overall costs were estimated using the 75 real site investigation data points. Because the density of the data on the site is variable, a declustering algorithm was applied. The statistics of the declustered data are presented in Row 2 of Table 1. The Gaussian transformed data have shown an anisotropic spatial structure, which has been modeled with a combination of a nugget effect 0.1 and two anisotropic spherical models of sills 0.5 and 0.65:

$$\gamma(h) = 0.1 + 0.5\mathrm{Sph}(50\ m_{NE}, 70\ m_{NW}, 0.7\ m_{vert}) + 0.65\ \mathrm{Sph}(140\ m_{NE}, 1000\ m_{NW}, 4\ m_{vert}) \quad [11]$$

where NE is northeast, NW is northwest, vert is vertical, h is a vector of distance, and Sph is the spherical model. The anisotropy axes of the variogram model are consistent with the main wind directions.

A total of $K = 200$ conditional simulations of point Pb concentrations were generated on a fine grid, in the framework of a multigaussian model, in a large neighborhood so that the 75 available data points were used for conditioning each simulated value. Each fine grid cell had a 1.43-m side length and a 0.30-m height, so that 49 point-simulated values were included in one block of the remediation grid. The grid was oriented according to the anisotropy axes of the variogram model. The block simulations were calculated using the point simulations, accounting for:

(i) The change of support and the information effect. Block-simulated Pb concentrations were considered as the mean of five point-simulated values taken in each block (Eq. [2]), by analogy with the real block concentrations measured during remediation (Fig. 2).

(ii) The sampling and analysis error. It has been shown that large errors have affected the real block Pb concentrations (Renard-Demougeot, 2002). These large errors were modeled by the variable $\varepsilon(x)$ taken as a uniform

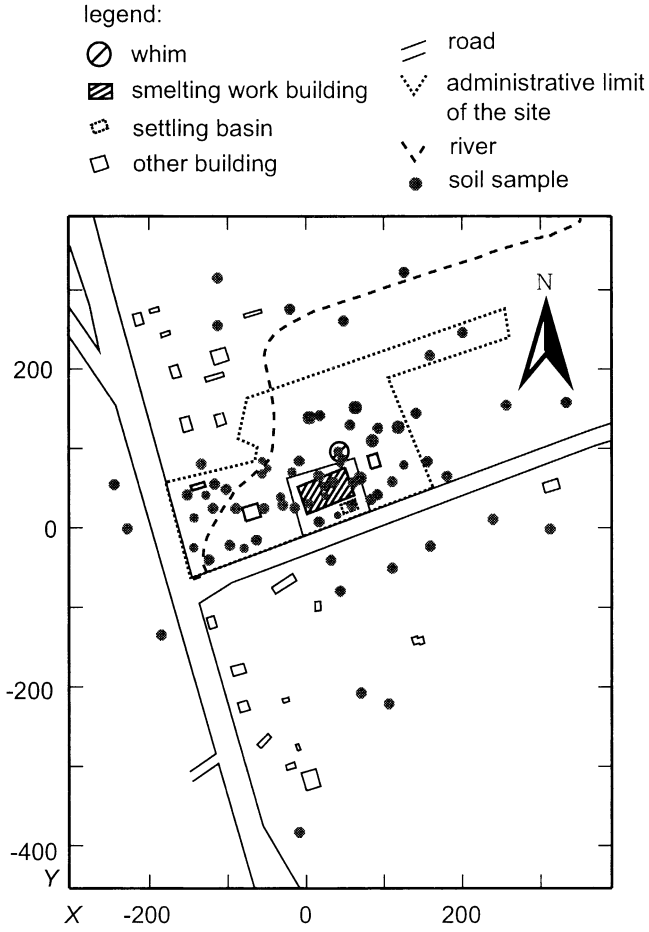

Fig. 4. Location of the 75 samples, collected in six investigation stages, on the map of the former smelting work.

Table 1. Statistics of soil Pb concentrations measured on boreholes (homogeneous sampling support) in the former smelting work (unity: mg kg^{-1}).

Data	Number	Minimum	Maximum	Mean	Standard deviation	Coefficient of variation
Not declustered	75	7	53 338	1 864	6 590	3.54
Declustered	75	7	53 338	1 293	5 416	4.19

distribution on the interval $[-1; +1]$, whose variance is large, equal to 0.33 (Eq. [3]).

For each block, the probability that block Pb concentration exceeds 300 mg kg^{-1} was subsequently calculated as the ratio of the number of block-simulated values exceeding 300 mg kg^{-1} to the total number of simulated values ($K = 200$).

Furthermore, we assume that remediation is required for soils with a probability exceeding $\beta = 0.6$. This value $\beta = 0.6$ is considered as a reasonable threshold for that specific site. The volume V_e that requires excavation was thus evaluated at $V_e^* = 10\,912$ m^3. We further assume that soil can remain without any remediation or additional monitoring if the probability of Pb concentrations to exceed 300 mg kg^{-1} falls below $\alpha = 0.2$. The uncertain volume was thus evaluated at $V_u^* = 23\,327$ m^3. The investigation cost C_i was calculated at investigation Stage 6 with real unit prices: $C_i^* = 68.6$ kEuros. A distribution of soil washing costs $C_c(i)$ was calculated with actual market unit prices, using the volumetric estimates. An estimate of C_c is provided by the median of the distribution $C_c(i)$: $C_c^* = 984.8$ kEuros. A distribution of uncertain costs $C_u(i)$ was calculated applying the cleanup cost function to V_u. An estimate of C_u is provided by the median of this distribution: $C_u^* = 1295.8$ kEuros. In these conditions, the volumetric uncertainty V_u^* represents 214% of the excavated volume V_e^*. While the financial uncertainty C_u^* corresponds to 132% of the cleanup cost estimate C_c^*.

Forecasting Volumes and Costs at the Seventh Investigation Stage (Steps 3 to 5)

Additional sampling at Stage 7 was "simulated" according to the sampling design procedure described in the "Modeling Sampling of Various Numbers of Additional Data" section, above. Steps 2 to 5 were repeated for a series of numbers of additional Pb concentrations N_7: $N_7 \in (12, 25, 50, 100, 200)$. The location of the selected values is a trade-off between standing at the nodes of a regular grid and being within the uncertain volume V_u, made of several patches (Fig. 5). Figure 6 maps the forecasted probabilities for different tested numbers N_7. The volume forecast graph (Fig. 7) shows that the excavated volume forecast V_e increases while the uncertain volume forecast V_u decreases as N_7 increases. The cost forecast graph (Fig. 8) shows that the cleanup cost forecast C_c increases while the uncertain cost forecast C_u decreases as N_7 increases. The investigation cost forecast C_i increases very slightly as N_7 increases.

Decision Making: How to Design Sampling Using the Volume and Cost Forecasts Graphs?

An optimal number of samples that permits having an "acceptable" remediation volume and cost uncertainty level does not exist. Only a "best compromise" can be found. Criteria, goals, constraints, and decision makers' preferences, which are specific to each remediation context, determine this "best compromise." The volume forecast graph provides environmental decision criteria in terms of volumes intended for excavation and remediation, and their uncertainties. The cost forecast graph provides financial decision criteria in terms of

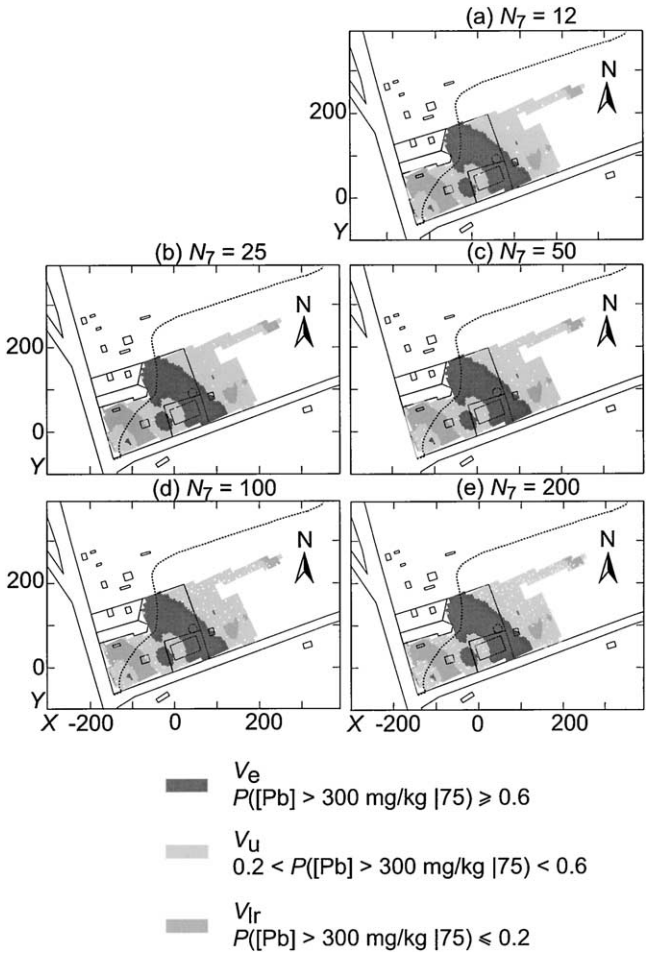

Fig. 5. Maps showing the location (white dots) of the simulated additional samples for the successive values of N_7. The points are superimposed on the simplified map of the probabilities where block concentration exceeds 300 mg kg^{-1} in the superficial layer (estimated at Stage 6 with the 75 real site investigation data points).

investigation costs, cleanup costs, and uncertainties on the cleanup costs, for various planned sampling numbers. To illustrate how the volume and cost forecast graphs can be used to choose a number of additional data to be collected, we assume that decision makers have four different goals and constraints. The first goal consists of minimizing the relative volumetric uncertainty:

$$\text{Min}(V_u/V_e) \quad [12]$$

The second goal consists of minimizing the relative financial uncertainty:

$$\text{Min}(C_u/C_c) \quad [13]$$

The third goal consists of minimizing the overall remediation cost:

$$\text{Min}(C_i + C_c + C_u) \quad [14]$$

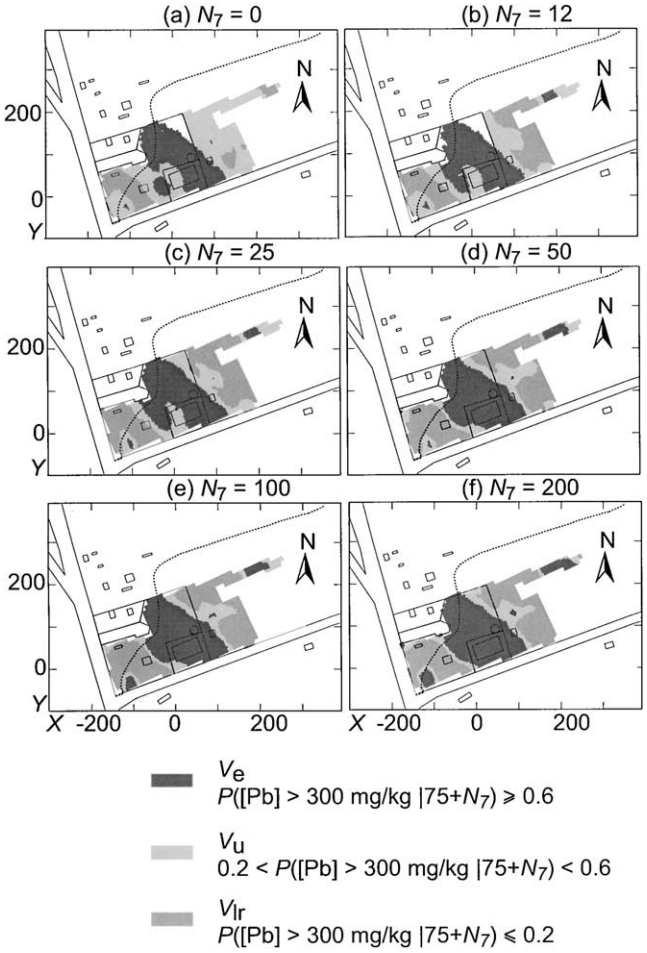

Fig. 6. Simplified maps of probabilities where block concentrations exceed 300 mg kg^{-1} in the superficial layer. These maps illustrate how the forecasted uncertain volume reduces when additional data are "simulated." Note that these maps are only an illustration of the uncertainty reduction but are not real forecasts as the additional data are simulated and not actual data.

The last criterion is a common financial constraint, consisting of limiting the relative sampling cost to a given percentage:

$$(C_i/C_c) \leq \text{percentage} \qquad [15]$$

The weights given to these objectives depend on the decision makers' profiles. We consider two decision makers' profiles. We call a risk-averse decision maker a person who is trying to avoid risks. He or she will probably assign heavier weights to Eq. [12] and [13] than to Eq. [14] and [15]. We call a risk-prone decision maker a person who is ready to take risks. He or she will probably assign heavier weights to Eq. [14] and [15] than to Eq. [12] and [13].

The volume forecast graph (Fig. 7) shows that the ratio (V_u/V_e) decreases rapidly as N_7 increases, and tends to stabilize for $N_7 \geq 100$. For $N_7 = 100$, V_u represents 50% of V_e, instead of the 214% estimated at Stage 6. Similarly, the cost forecast graph (Fig. 8) shows that the ratio (C_u/C_c) decreases rapidly as N_7 increases, and tends to stabilize for $N_7 \geq 100$. For $N_7 = 100$, C_u is forecasted at 35% of cleanup cost, instead of the 132% estimated at Stage 6. When $N_7 > 100$, the gain in accuracy on the forecasted volumes and costs is small. The forecasted cost $C_r = C_i + C_c + C_u$ (Table 2) is minimal for $N_7 = 50$. The ratio (C_i/C_c) slightly decreases as N_7 increases, up to

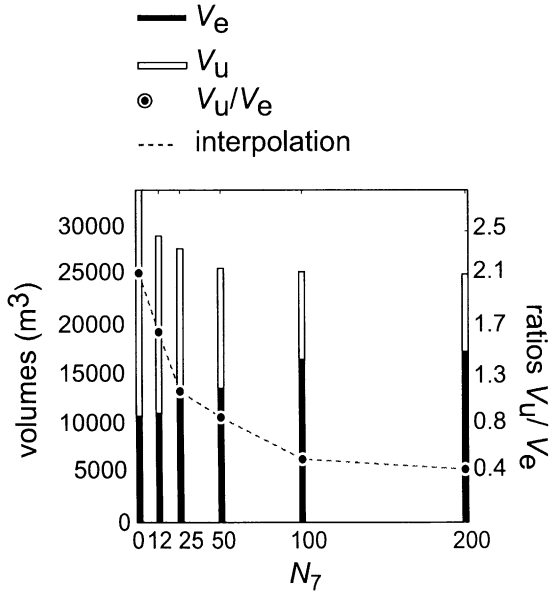

Fig. 7. Volume forecasts graph: excavated volumes V_e and uncertain volumes V_u as a function of the number N_7 of additional data whose sampling is "simulated" at Stage 7.

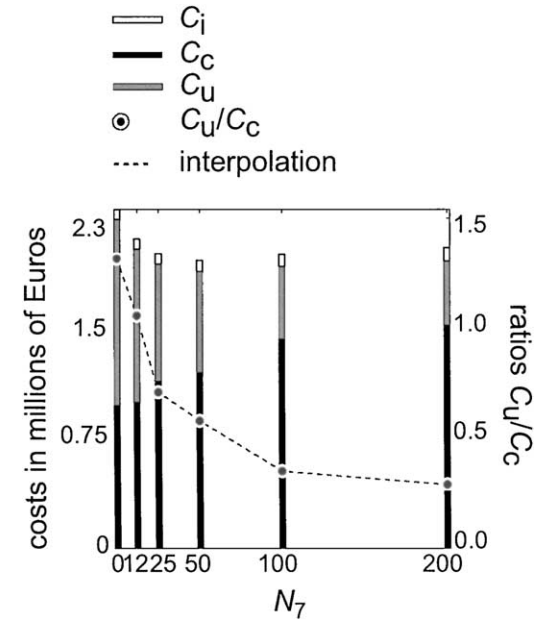

Fig. 8. Cost forecasts graph: investigation costs C_i, cleanup costs C_c, and uncertain costs C_u as a function of the number N_7 of additional data whose sampling is "simulated" at Stage 7.

Table 2. Forecasted investigation costs C_i, cleanup costs C_c, and uncertain costs C_u, and corresponding ratios according to the number N_7 of additional data whose sampling is "simulated" at Stage 7 (unity: $\times 10^5$ Euro).

	N_7					
	0	12	25	50	100	200
C_i	0.69	0.72	0.73	0.78	0.84	0.95
C_c	9.85	10.08	11.06	12.15	14.45	15.43
C_u	12.96	10.64	8.16	7.03	5.06	4.45
$C_i + C_c + C_u$	23.48	21.42	20.41	19.96	20.35	20.82
C_u/C_c (%)	132	106	71	58	35	29
C_i/C_c (%)	6.9	7.1	6.4	6.4	5.8	6.1

$N_7 = 100$. This ratio increases when $N_7 > 100$. The minimum ratio, calculated for $N_7 = 100$, is forecasted at 5.8%.

According to these forecasts, leaving aside any objectives or constraints other than the four defined above, it is reasonable to think that a risk-averse decision maker would choose to collect $N_7 = 100$ additional samples at Stage 7. Similarly, we can think that a risk-prone decision maker would prefer $N_7 = 50$, since, even though the corresponding (V_u/V_e) and (C_u/C_c) ratios are higher than those forecasted with $N_7 = 100$, this number of additional data minimizes C_r and maintains the ratio (C_i/C_c) at an average value (6.4%).

CONCLUSIONS

The methodology presented in this paper allows support of the design of the number of additional samples that would be worth taking within a framework of an iterative sampling procedure. The value of a data sample is evaluated in terms of volumes and costs uncertainty reduction. The methodology proposes to locate additional samples in zones where the uncertainty remains important, as it has already been suggested in the literature. As illustrated on a former smelting work, a "best compromise" of additional data number can be chosen at each investigation stage j, using volumes and costs forecasted for stage $j + 1$. The choice is not unique since it depends on criteria, objectives, and constraints other than those quantified by geostatistics, as well as on the decision makers' preferences.

The advantage of the methodology is its flexibility. The use of conditional simulations, the modeling of soil selection and excavation, the modeling of support effect, information effect, and sample bias, and the parameterized cost function allows it to be close to the reality of remediation works. Indeed, the comparison of volumes and costs estimated by this method has shown to be in good agreement with actual values determined after remediation.

Nevertheless, the method is not helpful when the pollution does not show a spatial structure. In the case of a pure nugget effect, no additional sampling will improve the existing uncertainty. A technical limitation of the method is the choice of a single conditional realization to simulate sampling of additional points. Because only one realization is used, the forecasted volumes and costs represent only samples from a distribution and not the expectation. It means that these values should not be taken as estimations even if the uncertainty estimates are correct. In any case, it is then important to (i) check the consistency of the reference with the probability map and (ii) apply the approach to various simulations taken as references, and to compare the results from multiple realizations. The methodology will then provide intervals for the forecasted volumes and costs, which are much more useful than expectations. Another limitation is that the volumes and costs are forecasted with the variogram model adjusted at stage j. The variogram model that will be adjusted to the real data at stage $j + 1$ may differ from that model, especially at the beginning of the investigations when only few data are available.

ACKNOWLEDGMENTS

The authors thank Martine Louvrier and Philippe Bégassat of the Agence De l'Environnement et de la Maîtrise de l'Energie (ADEME) for supplying the data and for their technical support, ADEME and Gaz de France for their financial support, and Alfred Stein for reviewing an early version of the article.

REFERENCES

Barabàs, N., P. Goovaerts, and P. Adriaens. 2001. Geostatistical assessment and validation of uncertainty for three-dimensional dioxin data from sediments in an estuarine river. Environ. Sci. Technol. 35:3294–3301.

Burgess, T., R. Webster, and A. MacBratney. 1981. Optimal interpolation and isarithmic mapping of soil properties. IV. Sampling strategy. J. Soil Sci. 32:643–659.

Englund, E., and N. Heravi. 1993. Conditional simulation: Practical application for sampling design optimisation. p. 613–624. *In* A. Soares (ed.) Geostatistics Troia '92. Kluwer Academic Publ., Dordrecht, the Netherlands.

Englund, E., and N. Heravi. 1994. Phased sampling for soil remediation. Environ. Ecol. Stat. 1:247–263.

Flatman, G.T., E.J. Englund, and A.A. Yfantis. 1988. Geostatistical approaches to the design of sampling regimes. p. 74–84. *In* L.H. Keith (ed.) Principles of environmental sampling. Am. Chem. Soc., Washington, DC.

Flatman, G.T., and A.A. Yfantis. 1984. Geostatistical strategy for soil sampling—The survey and the census. Environ. Monit. Assess. 4:335–349.

Garcia, M., and R. Froidevaux. 1996. Application of geostatistics to 3D modelling of contaminated sites: A case study. p. 309–325. *In* A. Soares, J. Gomez-Hernandez, and R. Froidevaux. geoENV I—Geostatistics for environmental applications. Kluwer Academic Publ., Dordrecht, the Netherlands.

Hendriks, L.A.M., H. Leummens, A. Stein, and P.J. De Bruijn. 1996. Use of soft data in a GIS to improve estimation of the volume of contaminated soil. Water Air Soil Pollut. 101:217–234.

Johnson, R. 1996. A bayesian/geostatistical approach to the design of adaptive sampling programs. p. 102–116. *In* R.M. Srivastava, S. Rouhani, M.V. Cromer, A.I. Johnson, and A.J. Desbarats (ed.) Geostatistics for environmental and geotechnical applications. ASTM9 STP 1283. Am. Soc. for Testing and Materials, West Conshohocken, PA.

Journel, A. 1974. Geostatistics for conditional simulation of ore bodies. Econ. Geol. 69:673–687.

Matheron, G. 1973. The intrinsic random functions and their applications. Adv. Appl. Probability 5:439–468.

Okx, J.P., and A. Stein. 2000. Use of decision trees to value investigation strategies for soil pollution problems. Environmetrics 11:315–325.

Renard-Demougeot, H. 2002. De la reconnaissance à la réhabilitation des sols industriels pollués: Estimations géostatistiques pour une optimisation multicritère. Ph.D. diss. no. 14615. Swiss Fed. Inst. of Technol. (ETHZ), Zürich.

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

Saito, H., and P. Goovaerts. 2003. Selective remediation of contaminated sites using a two-level multiphase strategy and geostatistics. Environ. Sci. Technol. 37:1912–1918.

Srivastava, R.M. 1987. Minimum variance of maximum profitability? CIM Bull. 80(901):63–68.

Van Groeningen, J.W., G. Pieters, and A. Stein. 2000. Optimizing sampling for multivariate contamination in urban areas. Environmetrics 11:227–244.

Van Groeningen, J.W., A. Stein, and R. Suurbier. 1997. Optimisation of environmental sampling using interactive GIS. Soil Technol. 10(2):83–97.

Van Tooren, C.F., and M. Mosselman. 1996. A framework for optimisation of soil sampling strategy and soil remediation scenario decisions using moving window kriging. p. 259–270. *In* A. Soares, J. Gomez-Hernandez, and R. Froidevaux (ed.) geoENV I—Geo-

statistics for environmental applications. Kluwer Academic Publ., Dordrecht, the Netherlands.
Von Steiger, B., R. Webster, R. Schulin, and R. Lehmann. 1996. Mapping heavy metals in polluted soil by disjunctive kriging. Environ. Pollut. 94:205–215.