Evaluation of an Automatic Procedure Based on Geostatistical Methods for the Characterization of Contaminated Sediments

G. Raspa, C. Innocenti, F. Marconi, E. Mumelter and A. Salmeri

Abstract This chapter describes a specific procedure that the Italian Central Institute for Marine Research (ICRAM) is researching on, in collaboration with the University of Rome La Sapienza, for the characterization of contaminated sediments. In the first part a description of the procedures developed in these years by ICRAM for a systematic and scientific approach to the characterization of contaminated sediments is provided. In particular it is illustrated how data analysis is performed by means of geostatistics in order to evaluate sediment volumes to be removed. Then, attention is focused on the need to develop an automatic procedure to estimate sediment contamination and the proposed procedure is described in detail. At last, results about applications of the procedure for some case studies are reported and the procedure's future development and progress are discussed.

Introduction

The Italian Ministry of Environment has made the Italian Central Institute for Marine Research (ICRAM) in charge of the environmental characterization of marine and brackish areas located within the contaminated sites of national interest. In compliance with its institutional assignment, ICRAM has defined guidelines and procedural models for a systematic and scientific approach to the characterization strategy, sampling and analytical methodologies, the processing of the characterization data, the evaluation of sediment quality and contaminated sediment management options.

Sediment characterization is realized by applying a site-specific sampling scheme defined on the basis of the information collected about the area and based on the conceptual model of contaminant transfer. As per the area dimension and morphology, sampling stations are disposed on a regular or an irregular grid, uniformly distributed all over the area; for regular sampling, grid size varies from 450 m × 450 m to 150 m × 150 m in more critical areas and is generally reduced to 50 m × 50 m

G. Raspa
University of Rome "La Sapienza", Dept. ICMMPM, via Eudossiana, 18 - 00184 Roma, Italy
e-mail: giuseppe.raspa@uniroma1.it

A. Soares et al. (eds.), *geoENV VI – Geostatistics for Environmental Applications*, 421–441

when specific operations have to take place (such as dredging, building of new docks, construction of confined disposed facilities). For irregular sampling, density continues to be chosen as indicated above, but sampling stations are placed based on an optimization process that permits to minimize the distance between any arbitrary point of the study area and its nearest sampling location (Van Groenigen and Stein 1998; Van Groenigen et al. 2000; Bação et al. 2004).

On the base of the defined sampling grid, surveys longer than or equal to 2 m are realized and core samples are taken approximately of the following sections, starting from the core top: 0–20 cm, 30–50 cm, 100–120 cm, 180–200 cm, another 20-cm section for each core's linear metre beyond 2 m (e.g. 280–300 cm, 380–400 cm) and the section relative to the last core's 20 cm.

The data analysis phase represents a very important step in the process towards the definition of the need for emergency and remediation interventions. Therefore a specialized team of experts has been created. This team, with the support of ISATIS, uses geostatistical methods to estimate the volumes of contaminated sediments.

Thanks to the experience gained in the last few years in the field of sediment characterization, ICRAM has been able to outline some peculiar aspects of a concentration's spatial variability. Sediment contamination generally shows histograms characterized by a strong and systematic asymmetry (Figs. 1 and 2). Due the complexity of the process of contaminant accumulation in the sediment matrix, experimental variograms are generally very irregular (Figs. 3–6) and very difficult to be adapted to a theoretical model; also, it is very hard to see variability structures

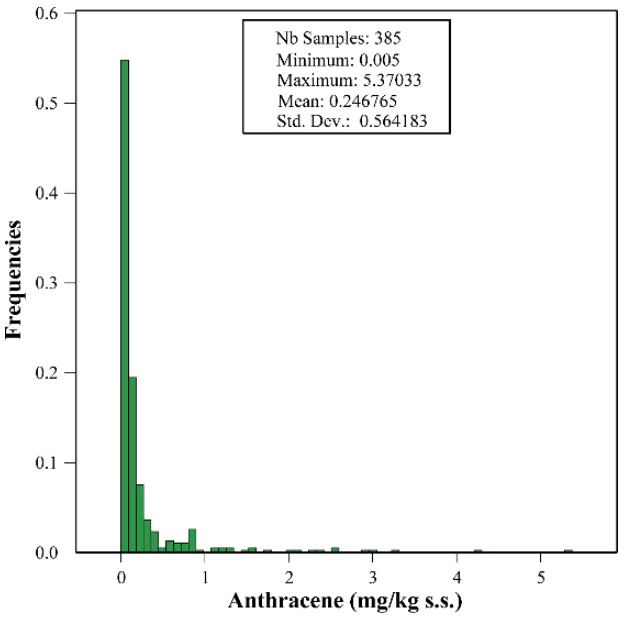

Fig. 1 Anthracene histogram

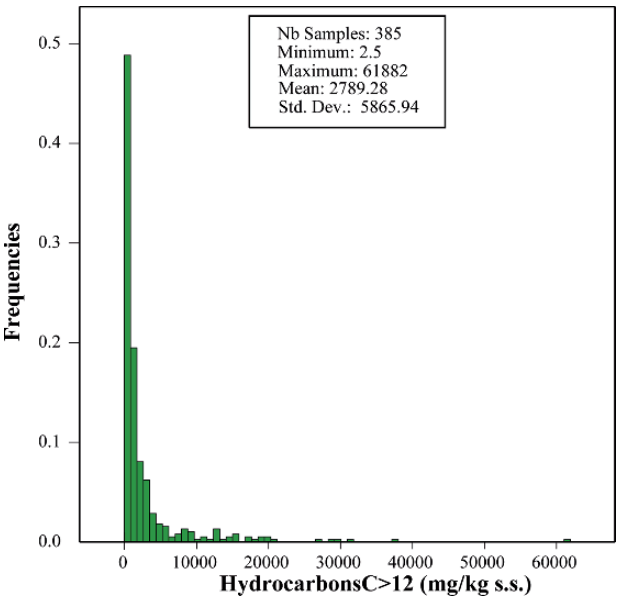

Fig. 2 Hydrocarbons C > 12 histogram

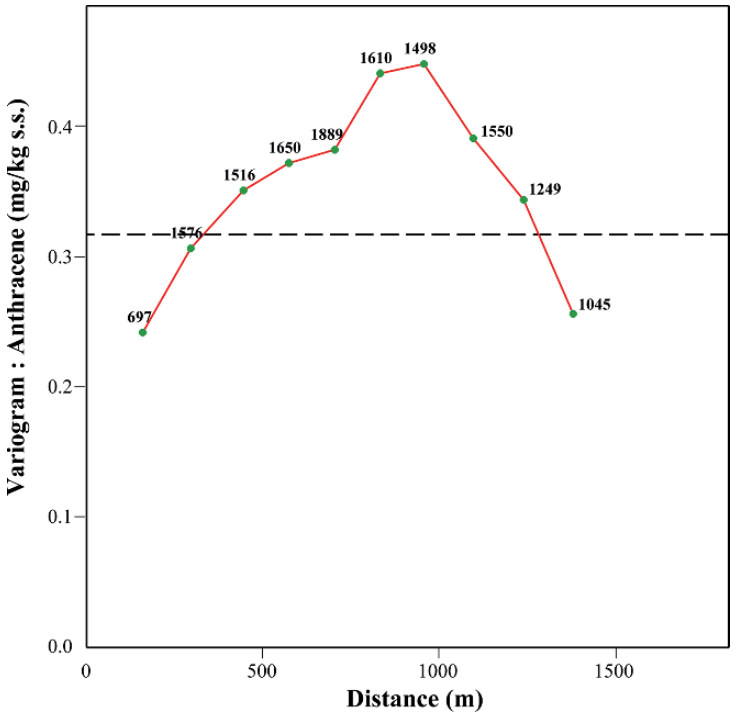

Fig. 3 Anthracene variogram in the horizontal plane

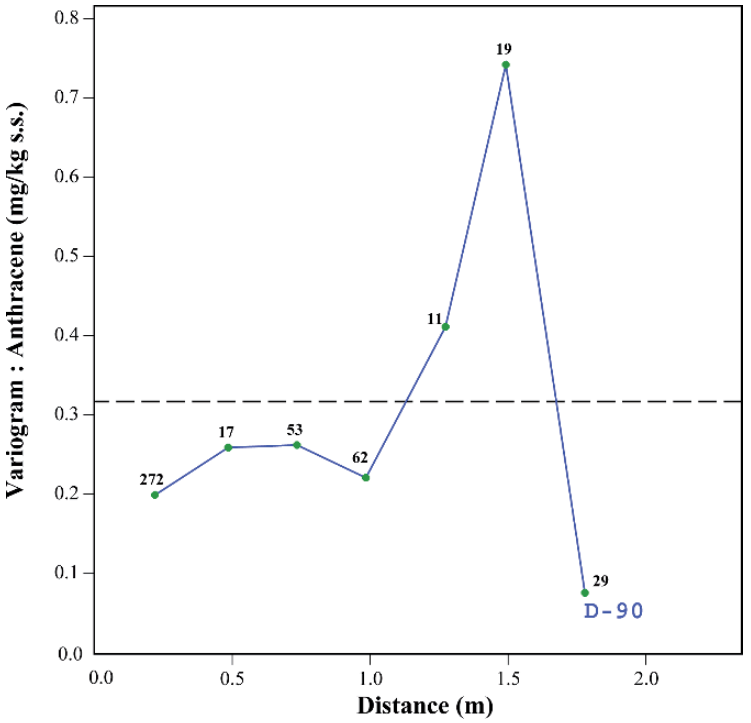

Fig. 4 Anthracene variogram in the vertical direction

by working on data logarithms. Moreover, experimental variograms usually show the existence of considerable low-scale spatial variability. Besides, in reality the contamination's spatial variability depends on space direction, and, usually, the variability in the horizontal plane is different from that in the vertical direction.

Variograms are usually modelled by means of a quasi-stationary nested model, typically composed of a nugget and two spherical variograms, one depending on the vertical component of distance and the other depending on the horizontal one:

$$\gamma(h) = \gamma_0(h) + \gamma_1(h_{xy}) + \gamma_2(h_z)$$

The horizontal component can be isotropic or anisotropic. A model with more than three structures or a too small nugget, due to the presence of many hot spots, may produce a lot of negative estimates, with many resulting problems in the interpretation process.

Data analysis permits the evaluation of sediment quality by producing 3D maps for sediment contamination; by comparing estimated maps with concentration limits, sediment volumes to be remediated are selected and management options are evaluated.

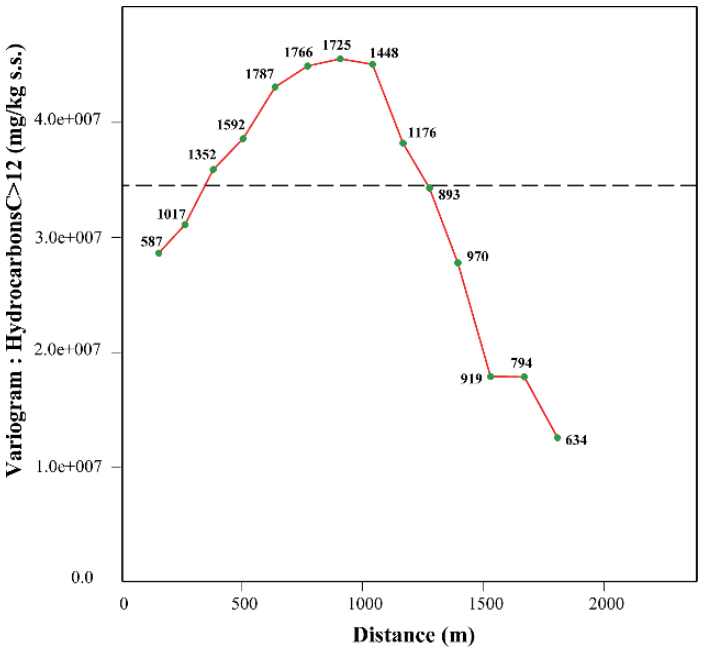

Fig. 5 Hydrocarbons C > 12 variogram in the horizontal plane

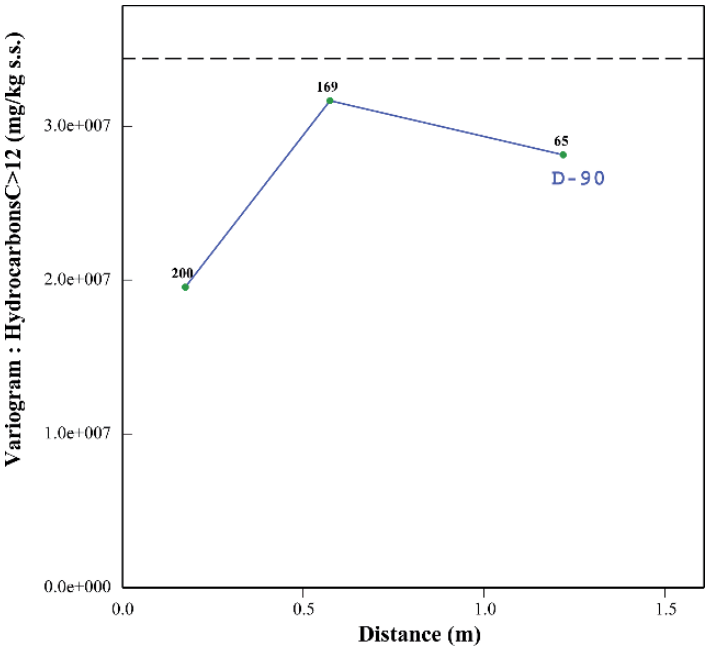

Fig. 6 Hydrocarbons C > 12 variogram in the vertical direction

Automatization of the Estimate Procedure

Generally, the modelling process for spatial variability of contaminated sediments is a very difficult task, in which some amount of subjectivity is unavoidable. However, this should be avoided as much as possible as it involves environmental and sanitary risks. Furthermore, most of the subjects in charge of the definition of the remediation interventions at specific sites (local administrations, port authorities, local agencies, etc.) do not have the geostatistical background or experience in estimating locations and quantities of sediments to be remediated. Therefore ICRAM has decided to develop a specific estimation tool, based on an automatic procedure that eliminates subjectivity but also takes into account the experience gained in these years by ICRAM's geostatistical working team.

This procedure allows the estimation of the quantity of contaminated sediments, starting from the information acquired by means of the above-described sampling strategy; this estimation is realized by dividing the site in 3D blocks and selecting those whose mean concentration exceeds a specific contamination threshold. Blocks mean concentrations are estimated by means of ordinary block kriging (Deutsch 1992).

Parameters

The estimate methodology has been chosen taking into consideration some operative aspects of the problem and, in particular, the necessity of modelling both vertical and horizontal anisotropies. The selected model of spatial variability is composed of four nested structures that allow the modelling of the nugget effect and of both horizontal and vertical anisotropies. In particular, one nugget structure and three 1D linear structures with zonal anisotropy (Chilès and Delfiner 1999) have been used:

$$\begin{aligned}\gamma(h) &= \gamma_0(h) + \gamma_1(h_{xy}) + \gamma_2(h_{xy}\cos\varphi) + \gamma_3(h_z)\\ &= \gamma_0(h) + m_1 h_{xy} + m_2 h_{xy}\cos\varphi + m_3 h_z\end{aligned}$$

where:

h is the module of $\vec{h}$, the vector connecting the estimate point to the sampled data.

$\gamma_0(h)$ is the the nugget structure, with sill C_0 that must be >0 in order to avoid models composed only of the sum of zonal components (Chilès and Delfiner 1999, pp. 96).

$\gamma_1(h_{xy})$ is the linear structure, with slope m_1; it depends on the horizontal component of $\vec{h}$.

$\gamma_2(h_{xy}\cdot\cos\varphi)$ is the linear structure, with slope m_2; it depends on the projection of h_{xy} on the direction forming an angle φ with the east–west direction (counterclockwise from E to W).

$\gamma_3(h_z)$ is the linear structure with slope m_3; it depends on the vertical component of $\vec{h}$.

A linear model is defined by only one parameter (slope), thereby allowing an easier automatic identification with respect to other variogram models (Chilès and Delfiner 1999). As indicated above, parameters defining the variogram model are five: C_0, m_1, m_2, m_3, φ; moreover, it is defined by another parameter that characterizes the estimation neighbourhood, that is the number of sample data (n) used to estimate a point. These six parameters jointly define the estimation model, that is the variogram model and the estimation neighbourhood characteristics that permit the calculation of the estimate. Parameters are obtained from an optimization procedure, based on the minimization of the cross-validation errors. With respect to parameter values obtained from the optimization process, a variogram can be isotropic or anisotropic; in the latter case, it can compete with a model exhibiting geometric anisotropy (Chilès and Delfiner 1999).

A mobile neighbourhood has been used in the estimation procedure; it moves all over the site and its shape and dimension are related to the variogram model and to the number of estimation points. Samples used to estimate a point x_0 are the n ones with the higher weights, with respect to the variogram; this corresponds to the use of a mobile neighbourhood with a shape defined by an iso-variogram surface and dimension so as to include all the n estimation samples.

By expressing $\gamma(h)$ with respect to h and taking Fig. 7 into consideration the following is obtained:

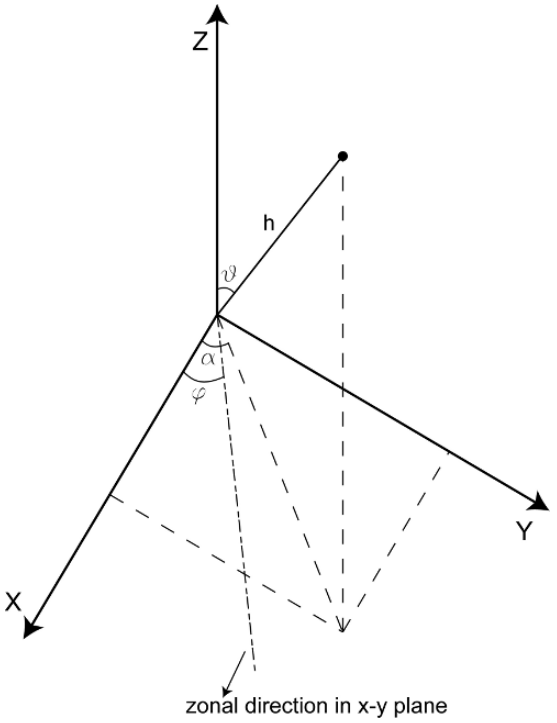

Fig. 7 Variogram components in space

$$\gamma(h) = \gamma_0(h) + m_1 h \sin\vartheta + m_2 h \sin\vartheta \cos(\alpha - \varphi) + m_3 h \cos\theta$$

and by considering $\gamma_0(h) = C_0[1 - \delta(h)]$ with $\delta(h) = \begin{cases} 1 & for h = 0 \\ 0 & for h \neq 0 \end{cases}$

it results:

$$h = \frac{\gamma(h) - C_0}{m_1 \sin\vartheta + m_2 \sin\vartheta \cos(\alpha - \varphi) + m_3 \cos\theta} \quad \forall h \neq 0$$

where

h is the "structural distance", that is the distance, along $\vec{h}$, between the point to be estimated and the surface on which the variogram has the defined value γ(h).

All measures disposed on an iso-variogram surface have the same weight in the estimation of $z(x_0)$, because their contribution does not depend on the distance from x_o.

In Figs. 8 and 9, the estimation neighbourhood centred in the $x_0$point is represented. In Fig. 8 the neighbourhood section on the horizontal plane with $z = 0$ is shown, while in Fig. 9 the surface is shown; variogram parameters used to build Figs. 8 and 9 are $\gamma(h) = 30$, $\gamma_0 = 3$, $m_1 = 8$, $m_2 = 10$, $\varphi = 45$Ł, $m_3 = 4$.

The greater m_3 is, that is the more the variability along the z-axis, the more flattened is the iso-variogram surface on the x-y plane.

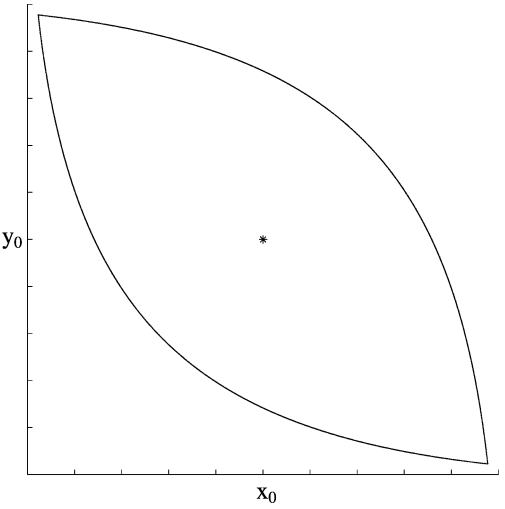

Fig. 8 Neighbourhood section

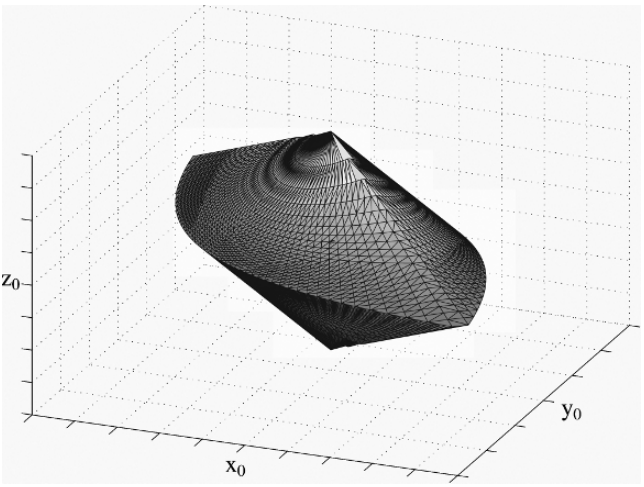

Fig. 9 Neighbourhood axonometric view

Objective Function

The automatic procedure's aim is to select blocks whose mean concentration exceeds the threshold value. The values of the six estimation parameters described above are obtained by an optimization process that minimizes selection errors; since these cannot be calculated for each grid point, where the true value is not known, the selection is evaluated based on the cross-validation data.

Figure 10 shows what is obtained by mapping on a Cartesian plane the couples $\{z_i^*, z_i\}$ relative to all measure points. The cloud of points is more scattered the greater the estimation error. If z_s is the threshold concentration above which the point is evaluated as contaminated, the estimation error produces a selection error. Particularly, as per statistical tests, two types of errors might be present, that is:

- Type I error: It occurs when $z > z_s$ and $z^* < z_s$, that is when a point that should be selected is left on place. This might cause an environmental damage.
- Type II error: It occurs when $z < z_s$ e $z^* > z_s$, that is when a point that should be left in place, is selected. This leads to higher remediation costs.

In this scenario, it is defined as an objective function whose value depends on errors of type I and type II, weighted in a different way, since the possible consequences are different. Moreover, the extracted estimation parameters might produce negative weights in ordinary block kriging. This occurs when data close to the location that is being estimated contain outlying values. When negative weights are applied to high data values, this might lead to negative contaminant estimates (Deutsch 1996). Experience shows that negative values might indeed be produced when estimating

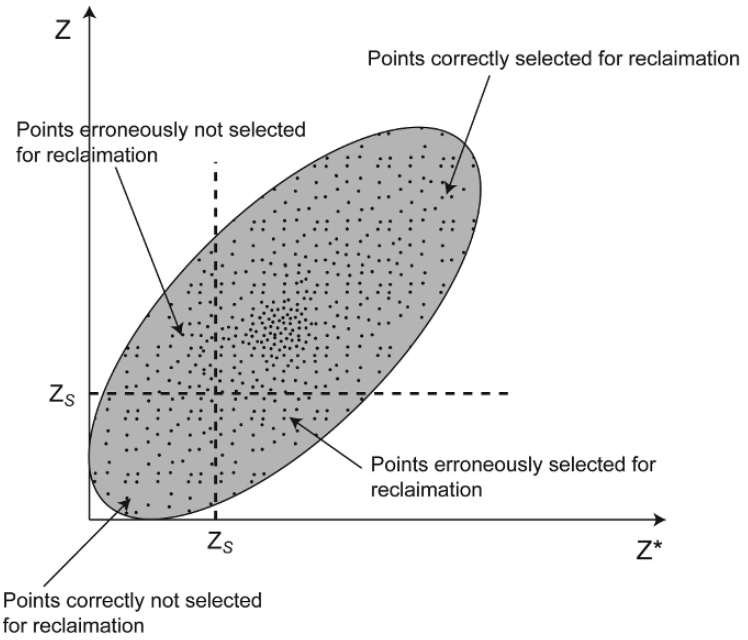

Fig. 10 Scatterplot estimations–measures

contamination in sediments, due to its spatial distribution, which is complex and heterogeneous. In order to avoid negative estimates, the proposed objective function contains a specific term that produces an exponential increase of the function value as the negative estimates increase.

The proposed objective function is:

$$f = \sum_{s=1}^{n_s} \left(\left(r \sum_{i=1}^{n_I} \frac{z_i - z_i^*}{z_i} + \sum_{i=1}^{n_{II}} \frac{z_i^* - z_i}{z_i} \right) e^{\frac{N_{neg}}{N} 100} \right)$$

where

r is the parameter whose value is greater than 1; it assigns higher weight to type I errors than to type II errors; its value must be determined experimentally, by evaluating its influence on the cross-validation results.

n_I number of type I errors.
n_{II} number of type II errors.
n_s number of thresholds $z_s$respect to which the selection is realized.
N_{neg} number of points for which the estimated value is < 0.

The objective function is obtained by summing the selection errors for different thresholds. In order to optimize the objective function with respect to a wide range of limits and to avoid applied thresholds that are lower than the minimum concentration

value or higher than the maximum one, the selection errors for three thresholds are calculated, obtained from the 25-, 50- and 75-quantiles of the contaminants' experimental distribution. In order to minimize the selection errors, the objective function has to be optimized by finding its minimum value.

Optimization

The six model parameters (C_0, m_1, m_2, φ, m_3, n) are calculated by minimizing the objective function, using an optimization procedure based on genetic algorithms (GA). Two different genetic optimization procedures have been developed and tested.

Genetic Algorithms

GA were created by J. Holland in the 1960s and were inspired by Darwin's theory of evolution. GA have been used to solve a wide variety of optimization problems, including those for which the objective function is discontinuous, not derivable, stochastic or strongly not linear.

The basic idea is to select the best solutions and to combine them so that they evolve towards an optimal point. The function to be optimized is called the objective function and the variables on which it depends are called genes; the specific sequence of genes forms a chromosome, i.e. an individual that represents a possible solution. Initially many individual solutions are randomly generated to form a starting population. During each successive stop, a proportion of the existing population is selected to breed a new generation. Individual solutions are selected through a *fitness-based* process, where fitter solutions (as measured by a fitness function) are typically more likely to be selected.

The next step is to generate a second-generation population of solutions from those selected through genetic operators: crossover (implemented by combining genes of two chromosomes to form a new individual) and mutation (random change of some genes). In order to obtain a population with a fitness function whose best value doesn't get worse, it is necessary to retain the best individual generated so far from generation to generation; this technique is called elitism. This generational process is repeated until a stop condition is reached.

As already mentioned, two different genetic optimization procedures have been developed and tested; they differ principally in the use of diverse crossover algorithms. In particular:

Procedure (a): the used selection is a rank selection (proposed by James Baker); the individuals' selection probabilities are assigned according to the individuals' rank, which is based on the objective function values. This method avoids excessive preference in the selection of a few individuals with the best fitness, reducing selection pressure when variance on population fitness is high.

Procedure (b): the used selection is a tournament selection (proposed by J. Haataja); it is one of the several existing methods of selection that runs a "tournament" among a few individuals, chosen at random from the population, and selects the winner (the one with the best fitness) for crossover. Selection pressure can be easily adjusted by changing the tournament size. If the tournament size is larger, weak individuals have a smaller chance to be selected.

In order to choose the two algorithms' parameters, a sensitivity analysis on different test functions has been realized; test functions have been obtained from the literature (Winter et al. 1995) and are characterized by a very irregular behaviour. This analysis has led to the choice of the following values: crossover probability equal to 0.8 and mutation rate equal to 0.02. Moreover, elitism is applied, so that the best individual of the population is always carried to the next generation.

Hereafter are reported the results of the application of the two algorithms to the maximization of a test function characterized by some local maximums (Figs. 11–14).

Test function:

$$z = f(x, y) = 10+3(1-x)^2 e^{(-x^2-y^2)} - 10(\frac{x}{5} - x^3 - y^5)e^{(-x^2-y^2)} - \frac{1}{3}e^{(-(x+1)^2-y^2)}$$

The realized tests have shown that the two procedures converge and that they can find with good approximation the optimum of very irregular functions, characterized by many local maximums. Therefore they have been used for the optimization of the objective function defined in the automatic procedure.

In the procedure, a solution is represented by a six-component vector, composed of a numerical value for each of the variables that define the variogram model and

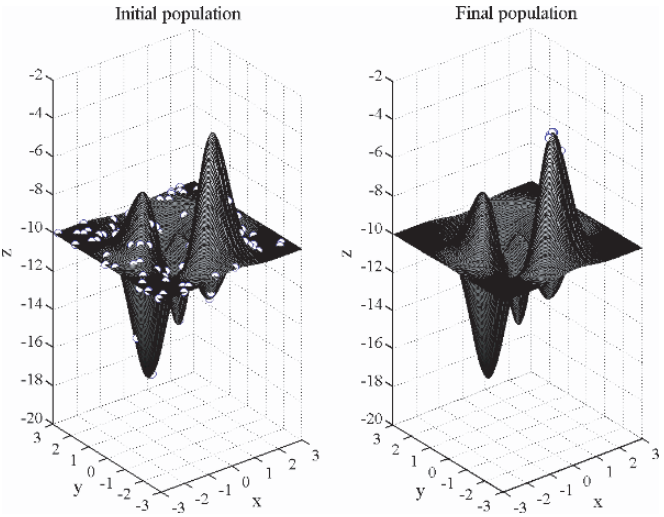

Fig. 11 Function plot and population values (100 individuals); each circle represents an individual (procedure a). Optimum value −3.4494

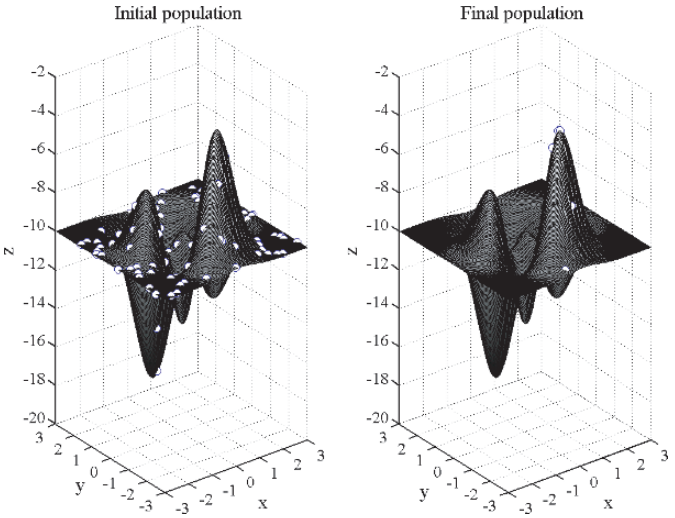

Fig. 12 Function plot and population values (100 individuals); each circle represents an individual (procedure b). Optimum value −3.4489

the neighbourhood. Nugget component varies between a fraction of experimental variance and the variance itself. Parameters m_1, m_2 and m_3 vary between 0 and $p/2$; φ varies between 0 and p, while n varies between 4 and 20 points. The fitness function matches with the above-defined objective function. A population of individuals

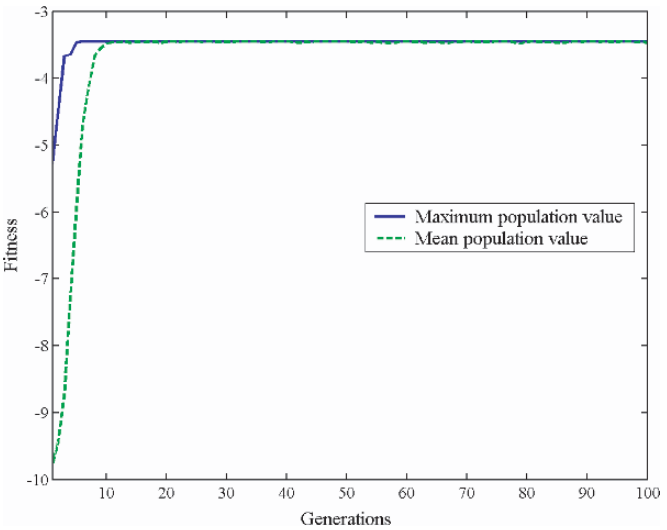

Fig. 13 Fitness versus generations. Procedure a

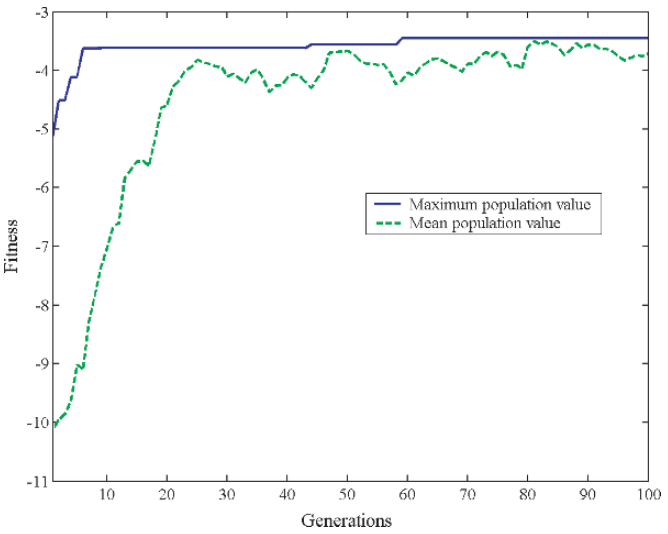

Fig. 14 Fitness versus generations. Procedure b

evolves until a defined end criterion is reached; on the basis of experimental results it has been decided that when the fitness decrement is below 5% for 20 consecutive generations, the procedure be stopped.

Once the end criterion has been defined, the number of individuals comprising the population is determined by taking into consideration two opposite matters: on the one hand, a large population should determine a result nearer to the optimum, but, on the other hand, it requires a longer calculation time (in fact, a cross-validation is realized for each individual). The number of cross-validations executed for each contaminant is equal to the individuals' number, multiplied by the number of generations and by the number of samples. In order to determine the number of individuals to be used, the procedure was run for 28 contaminants, selected from a case study with about 400 samples.

In Fig. 15 the progress of CPU time mean (AMD Athlon XP 2800+) and fitness mean versus number of individuals are reported. For each number of individuals used, the mean of the calculation time that is necessary to obtain the objective function's minimum is reported, as determined for 28 contaminants; the mean of fitness function's value is also reported, calculated for 28 contaminants and normalized to 1.

The data obtained show that fitness decreases rapidly up to 70 individuals, and then more slowly, while CPU time increases almost linearly. Between 100 and 170 individuals, a slight fitness increment can be observed, due to the extreme irregularity of the function whose behaviour does not depend linearly on the GA parameters.

Figure 15 has been obtained for a number of samples equal to 400. In order to define the number of individuals to be used in a more general case, it is necessary to perform some experiments with different number of samples and to express the

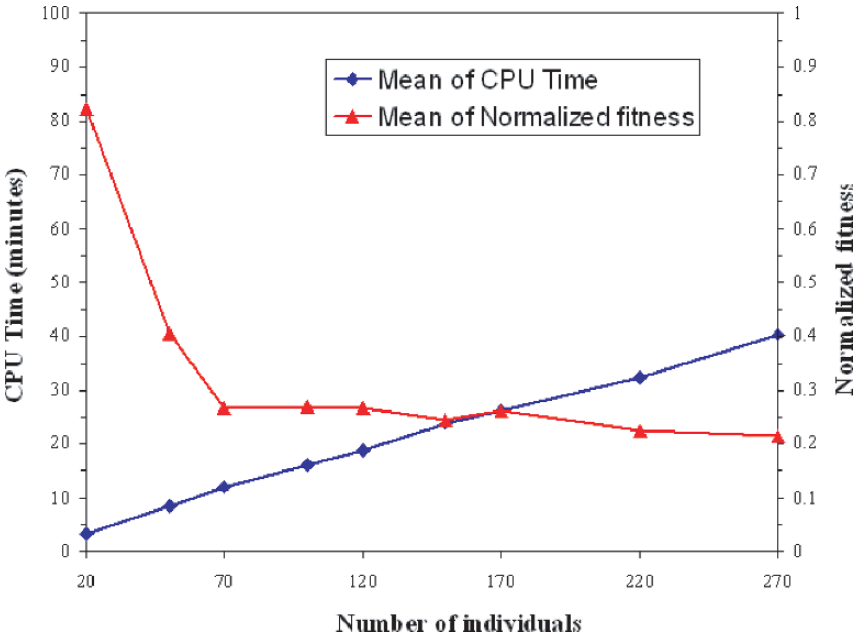

Fig. 15 CPU Time and normalized fitness versus number of individuals

CPU time and the fitness increment with respect to the number of samples and individuals.

In Figs. 16 and 17, the progress of the variance and the mean of cross-validation experimental errors are reported, versus the number of individuals, for 2 of the

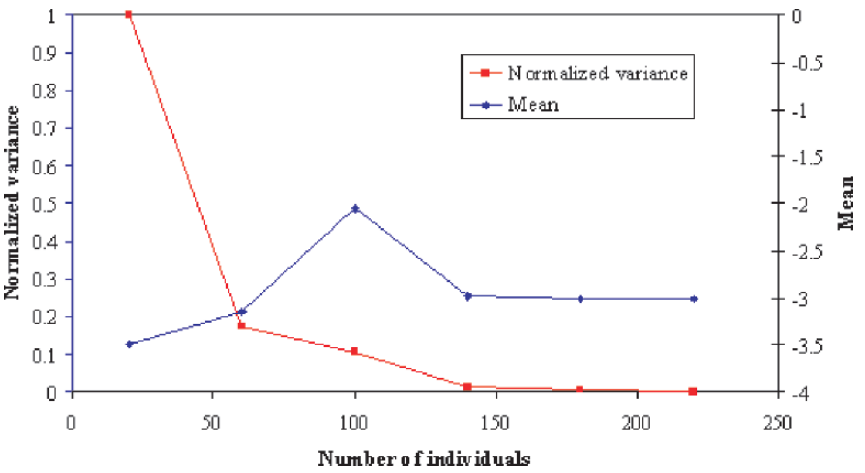

Fig. 16 Normalized variance versus number of individuals (zinc)

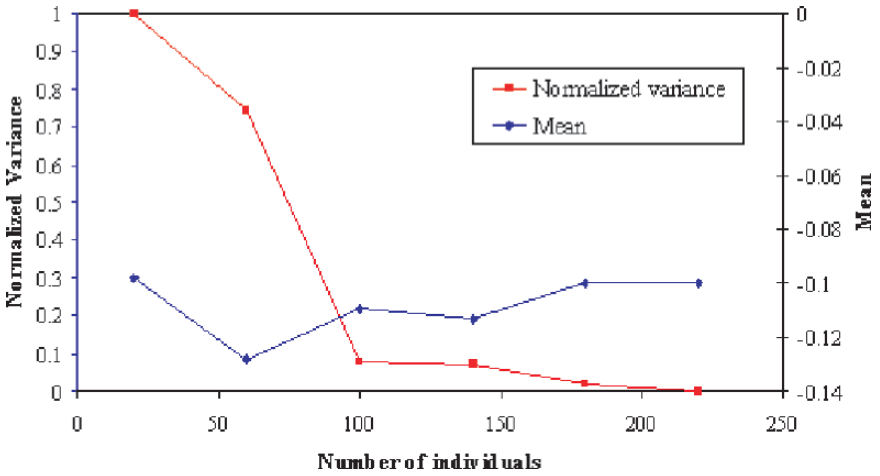

Fig. 17 Normalized variance versus number of individuals (arsenic)

28 contaminants taken into account in the above-described case study; variance has been normalized to 1. In order to obtain the minimum value, for each step the procedure was run 12 times, each time with a different initial configuration. For both contaminants, the experimental variance slowly decreases when the number of individuals increases, even though the objective function is based on the minimization of the selection errors and not on the minimization of the cross-validation experimental variance.

Procedure Steps

Ordinary block kriging method has been implemented to automate the process of reconstruction of 3D contamination, by means of a procedure that consists of the following steps:

1. Procedure starts from a randomly generated population of solutions $(C_0, m_1, m_2, \varphi, m_3, n)$.
2. For each individual (solution) the cross-validation is realized and the objective function is calculated; an objective function value is obtained for each individual; the best one is that corresponding to the minimum.
3. A new population of solutions is generated starting from the old one.
4. Point 3 is repeated until convergence criteria is reached.
5. Parameters obtained from the optimization process are used to estimate contaminant concentration on a 3D grid; blocks that need to be remediated are selected on

the basis of the estimated values, by comparing them with the thresholds defined by the user.

6. The procedure exposed is repeated for all contaminants.

Comparison with Interactive Procedure

In this section the results of the cross-validation are reported, obtained for the two contaminants extracted from the above-described case study. A population composed of 100 individuals was used. As indicated above, the procedure's aim is to minimize the fitness function (Figs. 18 and 21).

In the case of zinc, the algorithm stops after 53 generations, with a value of 5962.6. For arsenic, the algorithm stops after 44 generations, with a value of 6202.6. In the above-described cases, the cross-validation results of the automatic procedure are comparable to those obtained with the interactive procedure (Table 1–2 and Figs. 19–20, 22–23).

Conclusions

A specific automatic procedure has been implemented and it has been run for different cases, showing converging results; moreover, cross-validation results are similar to those obtained with the interactive procedure. Before adopting the proposed

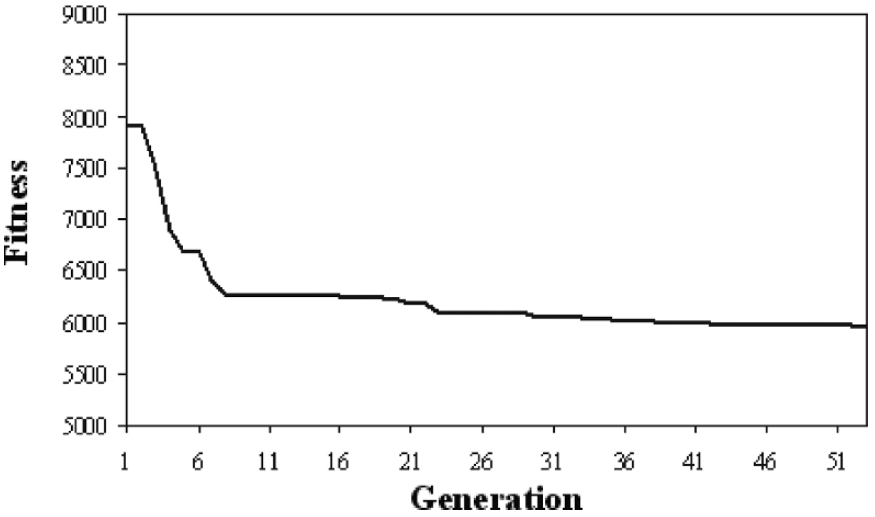

Fig. 18 Fitness value versus number of generations (zinc)

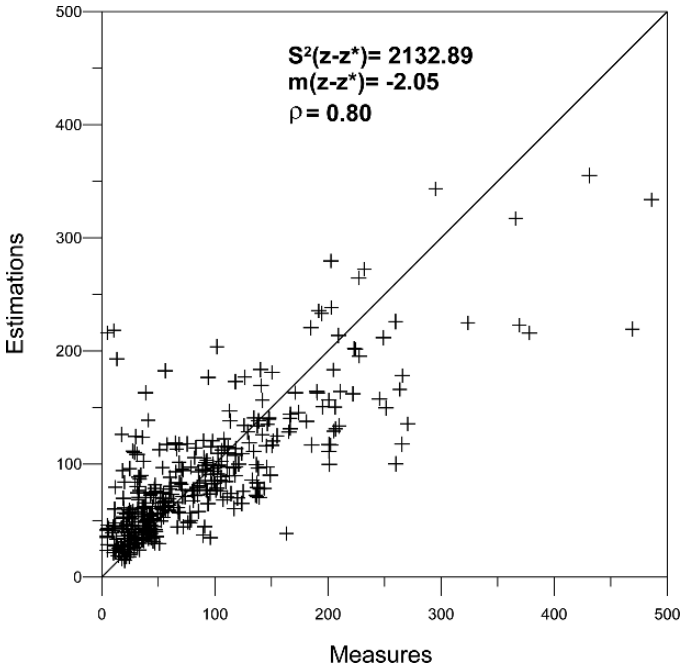

Fig. 19 Scatterplot Zn–Zn* (Interactive procedure)

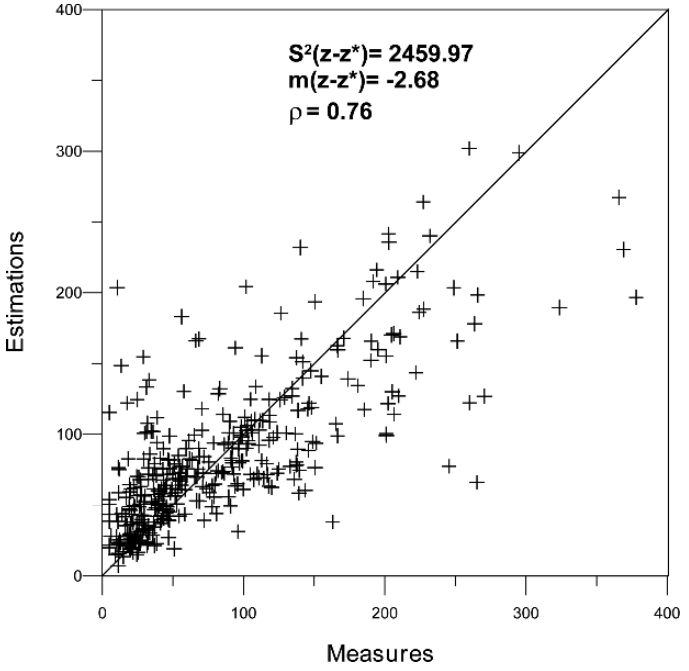

Fig. 20 Scatterplot Zn–Zn* (automatic procedure)

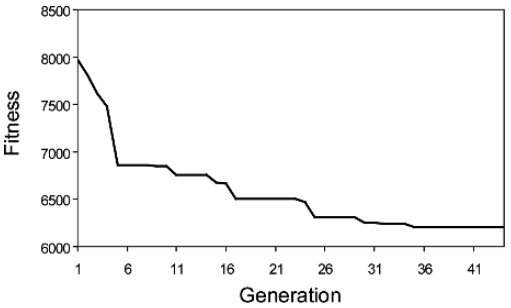

Fig. 21 Fitness value versus number of generations (arsenic)

Table 1 Cross-validation selection errors with respect to three thresholds (Zinc). Interactive procedure (up value) versus automatic procedure (down value)

Limit (mg/kg)	Type I error (%)	Type II error (%)	Correct selection (%)
31.2	2.86	14.81	82.34
	1.04	16.1	82.86
56.2	2.86	19.74	77.40
	2.34	16.62	81.04
114.31	8.83	5.195	85.97
	7.27	5.97	86.75

Table 2 Cross-validation selection errors with respect to three thresholds (arsenic). Interactive procedure (up value) versus automatic procedure (down value)

Limit (mg/kg)	Type I error (%)	Type II error (%)	Correct selection (%)
15.08	4.16	12.47	83.38
	4.94	7.27	87.79
18.64	9.09	15.58	75.33
	8.31	13.77	77.92
23.06	12.99	11.43	75.58
	8.31	9.09	82.6

automatic procedure, some other tests need to be performed with different case studies and different contaminants, taking into account not only the cross-validation results, but also the estimation results and their consequences on the selection of sediments to be remediated, with respect to the results of the interactive procedure. In the next few months, depending on the results from experimentations, possible variants of the procedure will be studied, with respect to both the objective function and the variogram parametrization.

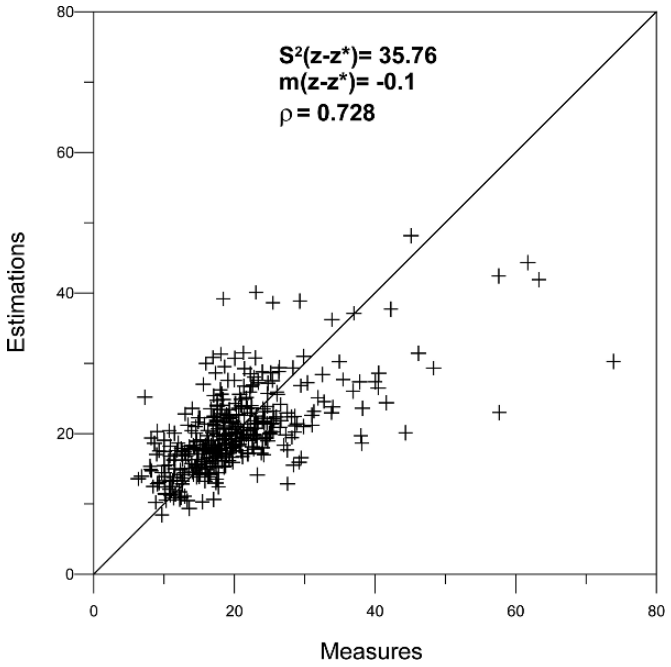

Fig. 22 Scatterplot As–As* (interactive procedure)

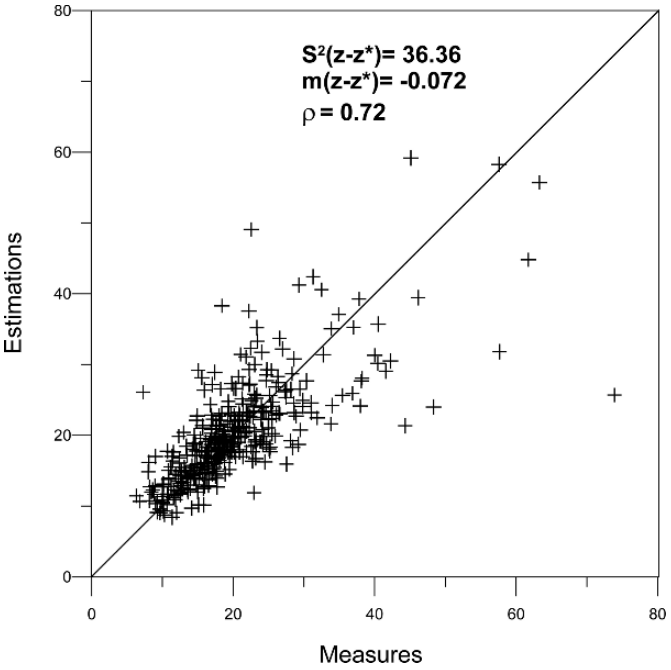

Fig. 23 Scatterplot As–As* (automatic procedure)

References

Bação F, Caeiro S, Painho M, Goovaerts P, Costa MH (2004) Delineation of estuarine management units: evaluation of an automatic procedure. geoENV 2004 Neuchâtel, Pre-Proceedings

Chilès J-P, Delfiner P (1999) Geostatistics. Modeling spatial uncertainty. Wiley series in probability and statistics, pp 95–97

Deutsch CV (1996) Correcting for negative weights in ordinary kriging. Comput Geosci, Elsevier Sci Ltd 22(7):765–773

Deutsch CV, Journel A (1992) GSLIB: Geostatistical software library and user's guide. Oxford University Press, New York

Mitchell M (1996) An Introduction to genetic Algorithms. MIT Press

Van Groenigen JW, Stein A (1998) Constrained optimization of spatial sampling using continuous simulated annealing. J Environ Qual 27:1078–1086

Van Groenigen JW, Pieters G, Stein A (2000) Optimizing spatial sampling for multivariate contamination in urban areas. Environmetrics 11:227–244

Winter G, P'eriaux J, Gal'an M, Cuesta P (1995) Genetic algorithms in engineering and computer science. Wiley series in computational methods in applied sciences