THE APPLICATION OF MULTIVARIATE GEOSTATISTICAL ANALYSIS IN THE STUDY OF CONTAMINATED SITES

Michele Castore[1], Claudia Cherubini[2] , Concetta I. Giasi[2], Mauro Molinari[3]

Abstract

In an environmental concern it happens frequently to have to manage such a broad number of data that their analysis and elaboration are extended up to levels useful to catch a large amount of information; this will allow not only the improvement of the whole knowledge about the process object of study but also the optimization of further investigations by achieving significant reduction of the costs.
The oriented application of a multivariate geostatistical approach could represent a valid instrument of decisional support in various cases; for example it helps to formulate hypothesis concerning the individuation of the possible source of contamination and their relative causes, it allows to get to know whether the sampling distance could catch the spatial variability of the contamination, it helps to individuate better the area or the areas in which to intervene with a "fitting" of the sampling procedure, it allows to establish the most suitable position for the monitoring of the area.
The application of such techniques in a broad industrial area allows to put into evidence the benefits of this approach. In this work the above mentioned procedure has been applied to a contaminated site of national interest whose analytical data has been given by the Italian Ministry of the Environment.
For the whole study concern and for its results it has not been necessary to report and to elaborate specific data of pollution of the area that at the present state is still active; in fact the study has been carried out on contaminants present at low concentrations or anyway having values lower than the limits of table 1 of the D.M 471/99.
The elaboration has showed that in this area, in the case in which another sampling would be carried out or an efficient monitoring would be defined, it is necessary to take into account the existent spatial correlation and in particular of the ratio between correlation at short range and dimensions of the sampling grid used during the site characterization, parameter that assumes a fundamental rule in this concern. In particular the geometrical setting of the new boreholes or of the whole monitoring should be realized by trying to increase successfully this ratio.

Introduction

The processes of propagation and transport of contaminants in soil and groundwater are mostly influenced by spatial variability of the physical and hydraulic properties of the medium. Anyway, even if the variability is high, it is often possible to highlight a sort of correlation in the spatial distribution, whose interpretation enable to recognize a spatial structure that permits to consider those parameters as regional variables.
The existence of correlation between the registered values in sampled points could lead to a significant reduction in the number of observations. In fact, the knowledge of the structure of the variability and of the variance of the estimation provides a help to improve the campaign of characterization of soil in terms of costs of sampling and analysis.
In previous papers (Giasi e Masi., 2001; Giasi e Masi, 2002; Giasi, 2005) it has been pointed out how the use of geostatistics, preceded by a correct design of the sampling mesh that would follows optimization schemes, allowed significant reduction in the number of boreholes in comparison with

[1] Professional engineer
[2] Polytechnic of Bari
[3] ENI Refining & Marketing

the by the DM 471/99 prescribed number, without losing any significance in the results. The spatial distribution of the contamination is the result of the influence and interaction of a multiplicity of hydrogeological factors of the environmental setting and also of the nature and environmental behavior of the contaminants. It is therefore important to catch the different scales at which the phenomenon of the contamination takes place in order to detect the probable and different causes.
Strictly linked to this problem is the necessity of a predefined sampling scheme based on an accurate and exhaustive knowledge of the spatial and volumetric distribution of the contamination.
Many times in the practice, during the phase of characterization, the sampling procedure bases itself on a rigid mesh having a fixed lag or whose dimensions depend on a preliminary conceptual model. Anyway, usually this procedure proves to be ineffective in the description of the real state of contamination of the area, just because it is not able to catch the spatial variability of the contamination.
A badly defined sampling may bring to adjunctive, undue costs, on the one hand because the number of sampled points results higher than necessary, on the other hand because, in case of environmental pollution, an inaccurate prevision increases costs for the phase of remediation or bring to unacceptable risks for the collectivity (Stein et al., 1995).
Moreover the number of points and their location in space influence noticeably the reliability of the results (Cochran, 1977; Muller, 1998), therefore a careful planning could lead to a considerable saving of time and money, together with an increased precision of the estimation.
In this setting a multivariate geostatistical approach can provide an interpretative key of the phenomenon, as it utilizes completely all the information present in the available datasets.
In this paper the quantitative variables object of study have been subjected to multivariate geostatistical analysis, by defining their simple and cross variograms and adapting a linear model of coregionalization that includes the nugget effects, the short range and the long range structures.
The application of cokriging has allowed to estimate and map the variables in study, correlated among themselves whereas the fonts of variability at the three different spatial scales of the linear model have been subjected to factorial kriging in order to be synthesized and mapped in terms of regionalized factors.

The case study

In order to understand it better, the proposed methodology has been applied on a real site and in particular on a still working industrial area that cannot be mentioned for reasons of discretion.
In order to exclusively illustrate the above mentioned methodology it has been decided to carry out the study not on the contaminants, but on substances that are present in the area at very low concentrations, always lower than the limits of the Table 1 of the D.M. 471/99.
The available data of the studied area consist in chemical analyses carried out on 237 samples obtained from a 50x50 mesh. The investigated substances are total chromium, nickel, vanadium, lead and organic carbon. For each examined variable the relating base map has been plotted in which the measured points are represented; this representation enables not only to visualize globally the georeferenced data, but also to evaluate the homogeneity of the distribution and the possible presence of not sampled areas. Table 1 reports the descriptive statistics of all the variables.

Variable	Mean	Min	Max	Standard Deviation	Skewness	Kurtosis
Organic Carbon	0.25	0.01	1.24	0.21	1.72	6.64
total Cr (mg Kg-1)	7.99	0.8	46.10	8.28	2.73	11.23
Ni (mg Kg-1)	7.77	0.70	67.50	10.75	3.15	14.05
Pb (mg Kg-1)	6.10	0.22	34.40	5.62	2.33	9.38
Va (mg Kg-1)	11.49	1.10	38.10	6.00	1.21	5.08

Table 1- Descriptive statistics of all the variables

From the inspection of the table, we can notice high shifts of skewness and kurtosis from 0 and 3, respectively, which are the characteristic values of normal distribution. Therefore, the variables generally exhibit non symmetric distributions, with long tails and several outliers. The variables have been then normalised and standardised to 0 mean and unit variance through Gaussian transformation (Cherubini et al 2005). This procedure has made it possible to obtain the standardized and normalized histograms of relative frequency for each variable (fig 1 a-e).

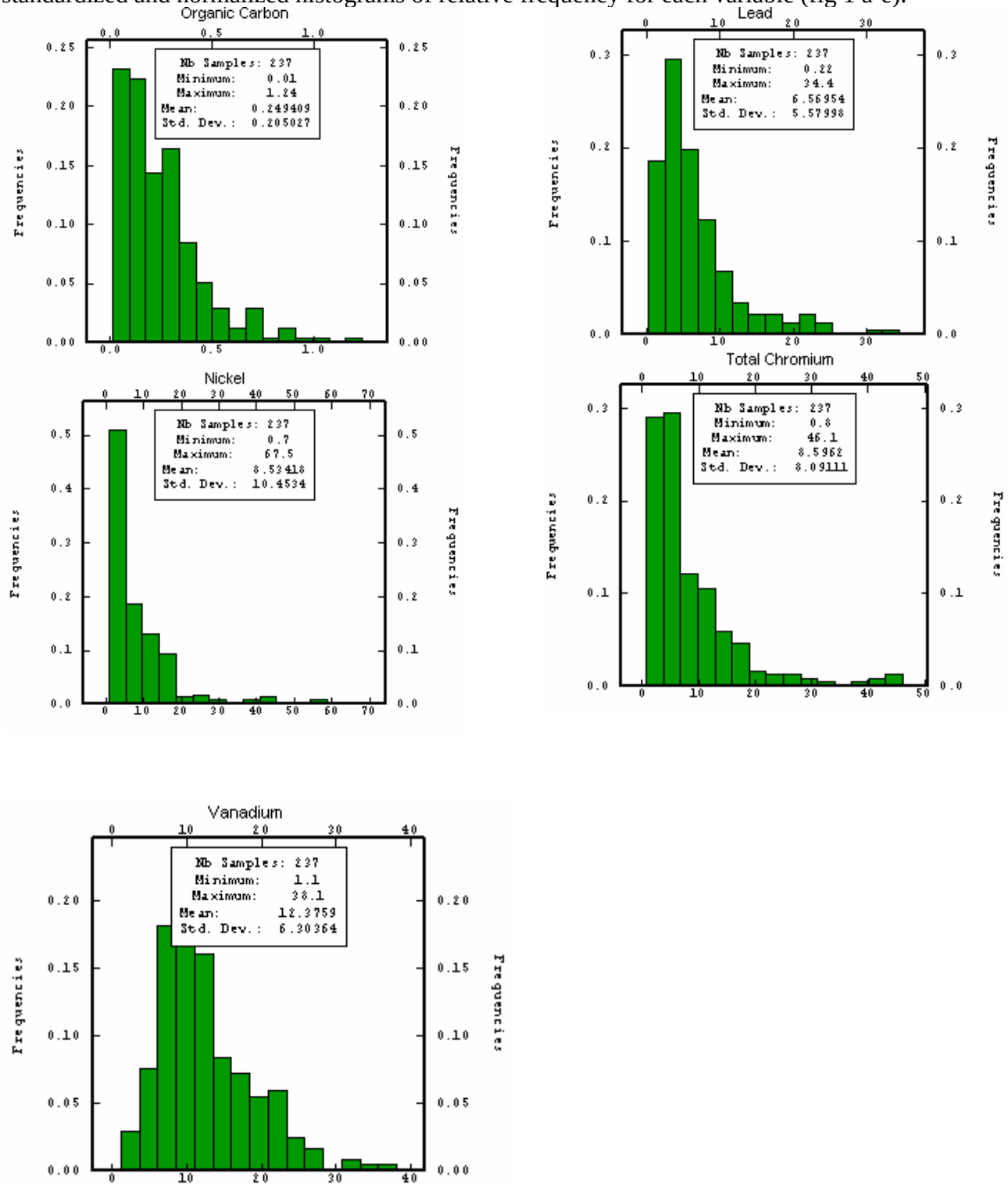

Fig.1e – *Histogram of the relative frequencies* of the georeferenced concentration data of organic carbon, lead, nickel, total chromium and vanadium.

The visual inspection of the variogram maps, (not shown) did not reveal any significant anisotropy in chemicals distribution, therefore an isotropic model of variogram was assumed.
The above mentioned descriptive statistics has been associated to principal component analysis, procedure that has permitted to extrapolate two factors able to describe and synthesize the behaviour of the examined elements, by explaining, summing up both contributes, almost 80- 85% of the element's variance.
The next step consists in the construction of the matrix of the experimental variograms; the set of direct and cross experimental semivariograms should be interpreted and opportunely modeled through the use of theoretical functions.
A Linear Model of Coregionalization (not shown) was fitted to the set of the 15 direct and cross-variograms including 3 basic spatial structures: 1) a nugget effect; 2) a short range spherical model (range = 249.58 m); 3) a long range exponential model (range = 1300.00 m). The nugget constitutes the not explained part of the semivariance, imputed to casual variability. More correctly it should be spoken of *relative nugget effect*, expressed as a percentage of the total sill; the higher the nugget effect, the smaller is the zone of spatial correlation between the samples and consequently the less effective would be the interpolation. The maps of the estimated values of organic carbon, total chromium, nickel, lead and vanadium concentration (Cherubini et al 2005) have been obtained after implementing the *back transformation*: for each estimated value, the cokriging has allowed to calculate the variance of the estimation error associated to it, providing a measure of the reliability of the interpolation and giving consequently the possibility to decide about the opportunity, in relation to specified goals, to operate a fitting in the sampling mesh.
The obtained maps show how total chromium and nickel on one side and lead and vanadium on the other side have similar spatial distribution, whereas the organic carbon behaves differently from the other elements.

Check process

The precision of the estimation $z^*(\mathbf{x}_0)$ is described by the square root of the variance of kriging σ^2, that gives a measurement of the error of the estimation. As σ^2 is determined on all the study domain, it has been possible to represent the error maps that permit to detect areas characterized by high values of the variance, requiring a more accurate characterization (not shown).
An interesting and useful property of the error maps is that they depend uniquely on the adopted variogram model and on the relative disposition of the sampled points, and not on the absolute values of them (Cressie 1991); so it is possible, once chosen the variogram model, to define optimal sampling strategies that minimize the estimation error. This would permit to know more accurately the spatial and volumetric distribution of the contamination.
The estimation error depends, more specifically, on the following properties:

1. the variogram form (linear, spherical, etc); the presence of anisotropies or gaps from the normal distribution; the entity of the nugget variance;
2. the number of samples used for the estimation;
3. the geometric configuration of the samples (irregular or regular sampling and the last one with a triangular or square grid);

McBratney et al. (1981) developed some procedures to optimize the sampling grid in terms of the minimization of the variance of kriging, being assumed the variogram model.
Yfantis et al. (1987) found that, in case of variogram with a relatively low nugget, equilateral triangular grids produced more reliable kriging estimations than hexagonal or square ones.
In case of relatively high nugget and low density of sampling, the same authors found that hexagonal grids produced the least variances of kriging.
Sacks and Schiller (1988) introduced numerous algorithms based on the annealing simulations to optimize the sampling scheme in case of a small grid with only some predefined points. They

distinguished various optimization criteria, among which the minimization of the mean or of the highest variance of kriging.
A valid help in the topic of sampling could be also the stochastic annealing simulation implemented in such a way as to take into account physical barriers and previous measurements.

Individuation of the source points

In addition to providing a valid help in the cases in which there is to fit a sampling mesh in specified areas to get a more accurate estimation of the variable in study, the geostatistical analysis could be used also when the goal is to identify the source of the contamination, in all the cases in which the study is referred to pollution and not simple substances.
In fact in those issues it is useful and opportune to carry out a multivariate geostatistical analysis through the implementation of Factorial Kriging (FKA) that would enable to separate the different sources of variation in function of the spatial scale at which they are operating.
This analysis, aiming at describing the structure of a multivariate structure of spatial data, consists in three phases:

a) modeling the coregionalization of n variables according to a linear model (LMC);
b) Principal Component Analysis carried out on each coregionalization matrix;
c) Estimation of the regionalized factors through cokriging and graphical restitution in form of spatial maps.

In synthesis it could be pointed out that the FKA consists essentially in a principal component analysis that, instead of being carried out just on one correlation matrix, is applied to each correlation matrix that corresponds to a specified spatial scale.
In our case the decomposition of the variance- covariance matrix has been done by making use of a linear model of coregionalization made up by three structures:

1) A nugget effect
2) A spherical structure with a range of 249.58 m
3) An exponential structure with a range of 1300m

This model has been fitted to the set of the direct and cross- variograms of the examined variables.
Each coregionalization matrix obtained describes the relation between the 5 variables at a specified spatial scale defined by the variogram function. Afterwards a principal component analysis has been implemented for each coregionalization matrix in order to synthesize the relations among the variables (Wackernagel,1989). The nugget component has been discarded, being linked to the random variability of soils and also the error variance associated to field and lab measurements.
Once the linear model of coregionalization has been fitted it has been possible to extract the 2 Regionalized Factors, that, at the cost of an acceptable loss of information, provide a global and sufficiently exhaustive description of the process in study at the selected spatial scales.
As far as the present study is concerned, the necessity has emerged to carry out an analysis through a mathematical instrument able to synthesize the synergic behavior of the investigated substances with respect to geological factors and their own physical and chemical characteristics.
The tables 2 and 3 show the structure of the existing correlation between the regionalized factors and the variables.

	Carbon	Chromium	Nickel	Lead	Vanadium	EigenVal.	Var.Perc.
Factor 1	0.3105	0.5755	0.4479	0.4370	0.4252	1.2688	88.49
Factor 2	0.7476	-0.1462	-0.0580	0.2844	-0.5793	0.1461	10.19

Table 2 - Spherical - Range = 249.58m

	Carbon	Chromium	Nickel	Lead	Vanadium	Eigen Val.	Var. Perc.
Factor 1	0.1525	0.4944	0.7316	0.2496	0.3670	0.9182	72.42
Factor 2	0.1092	0.0347	-0.5379	0.6223	0.5570	0.2423	19.11

Table 3 - Exponential - Scale = 1300 m

The first two factors explain most variance both at short and long range (98.68%, and 91.51%, respectively). The short-range component of the first factor (F1) explains 88.49% of the variance and is mostly correlated with chromium (0.5755) and in smaller measure with the other variables, whereas the long-range component of F1 explains the 72.42% of the variance and is mainly correlated to nickel (0.7316) and in smaller measure to chromium (0.4944).
The second factor F2 at short range explains the 10.19% of variance and is strongly correlated with organic carbon (0.7476) whereas at long range it explains the 19.11% of the variance and is positively correlated with lead (0.6223) and vanadium (0.5570) and less with the others (negatively with nickel) (Cherubini et al 2005).
From the carried out analysis it emerges that the behavior of heavy metals doesn't seem to be influenced by the presence of organic carbon; in fact the first substances, showing a similar environmental behavior are mostly affected by the factor F1; whereas the organic carbon, being mainly linked to the characteristics of the soil and the possible presence of organic substances is affected in a stronger way by the second factor. His short range variability could be ascribed to the presence of anthropic activity.
As far as heavy metals are concerned it should be pointed out that the long range model doesn't detect a similarity in the behavior of them; therefore a successive integrative investigation should be performed by operating at a short range scale; from now on just this scale will be taken into account.

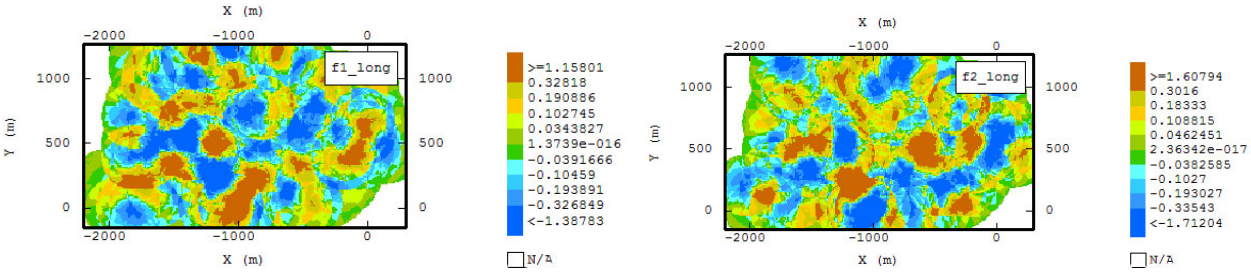

Fig.2 – Maps of the F1 and F2 Factors at long range scale.

Discussion of the results

Starting from a number of samples of 237 (located according to a 50x50 m mesh), the study has detected a short range variability of the phenomenon equal to 250 m, dimension able to express the spatial structure of the studied variables by synthesizing with enough completeness their behavior.
The Italian ministerial decree law 471/99 plants, for sites having dimension in the range between 50.000 and 250.000 m^2, from 15 to 60 sampling points; even performing the maximum scheduled by this legislation, that corresponds for the studied area to a 200x200 m grid, it would have been possible to detect the scale of the variability of the phenomenon, avoiding a redundancy in the information, with consequential high saving in terms of characterization times and economical resources.
However, the range of 250 m constitutes an upper bound because in order to describe the caught variability itself there is the need to investigate the process at a lower scale.
A proper methodology would be therefore to proceed hierarchically by starting from a sampling mesh of 100x100 m (as suggested by the legislation for industrial sites): in this case this dimension represents, in relation to the physical- chemical properties and the geological characteristics of the

area, a plausible dimension of the sampling mesh: the spatial variability of the investigated substances could have been caught through a lower number of samples than the ones actually realized.
A possible fitting of the sampling mesh could be performed locally in those areas where the variance of the estimation error shows high vales; this could be ascribed by a shorter range local variability of the phenomenon. Just in this case the allocation of new economical resources for the sampling procedure would be justified.
The carried out study has also made it possible to formulate hypothesis concerning the kind of emission source of the substances.
The distribution of the two factors both at short and at long range looks as “pepper and salt” type with a high component of erraticity. This puts into evidence that the points of emission for the examined inorganic chemicals *are not concentrated in fixed locations:* more precisely we could assert that in the study case we are in presence of *more than one points of emission*, jointly working, acting intermittently and being ascribed to causes of anthropic origin. The areas in which the factors analyzed at short range show higher values are the ones where to investigate about any cycle of production operating there and where to carry out a more accurate sampling in the case in which there is the need of detecting the source of a situation of contamination.
The realized study has made it possible to point out that the ratio between the spatial correlation scale and the dimension of the sampling grid is a fundamental parameter that should be taken into account in planning further investigations aimed at fitting the existing mesh; in this hypothesis there is to increase the above mentioned ratio (in our case equal to 250/50): that would lead to the reduction of the sampling lag and consequently of the nugget effect.
Similar suggestions could be useful also in planning new sampling points for the test phase of any remediation action and furthermore for the monitoring procedure performed afterwards.

References

Giasi C.I., Masi P. “Considerazioni sulle modalità di campionamento nella caratterizzazione dei siti contaminati.” Siti Contaminati n.5/2001, pp.17-21. Ranieri Editore Milano.
Giasi C.I., Masi P. “Cost-effectiveness in polluted site sampling campaign" Brownfields 2002" in Brownfield Sites Assessment, Rehabilitation & Development.pp.209-217. Editors:C.A. Brebbia, D.Almorza & H. Klapperich. WITpress Southampton .
Castrignano A., Cherubini C., Giasi C.I.,. Castore M., Di Mucci G., Molinari M.: Using Multivariate Geostatistics for Describing Spatial Relationships Among Some Soil Properties. ISTRO (International Soil Tillage Research Organization) Soil – agriculture, environment, landscape Proceedings of International Conference Brno 2005 - Czech republic June 29 – July1 2005.
Chiles, J.P., Guillen, A., 1984. Variogrammes et krigeages pour gravimétrie et le magnétisme. Sciences de la Terre, Série Informatique 20, pp. 455-468.
Goovaerts, P., Webster, R.,1994. Scale.dependent correlation beetween topsoil copper and cobalt concentrations in scotland. Eur. J. Soil Sci. 45, 79-95.
Goulard, M., Voltz, M.,1992. Linear coregionalization model : tools for estimation and choice of cross-variogram matrix. Math. Geol. 24 (3), 269-286.
Matheron, G., 1982. Pour une analyse krigeante des données regionalisées in Report 732. Centre de Geostatistique, Fontainebleau.
Annamaria Castrignanò Gabriele Buttafuoco Maggio 2004 Analisi spaziale mediante tecniche geostatistiche Internal report
Cochran, W.G., 1977. Sampling technics. Wiley & Sons, 428 pp.
Cressie, N.A.C., 1992. Statistics for Spatial Data. John Wiley and Sons, New York, 900 pp.

McBratney, A.B., Webster, A., Burgess, T.M., 1981. The design of optimal sampling schemes for local estimation and mapping of regionalized variables: 1. Theory and method. Comput. Geosci., 7, 331-334.
Sacks, J., Schiller, S. 1988. Spatial Designs, in Statistical Decision Theory and Related Topics IV, Volume 2, 385-395.
Wackernagel, H. 2003. Multivariate Geostatistics: an introduction with Applications. Springer-Verlag, Berlin, 3rd ed., 388 pp.
Yfantis E. A Flatman G. T Behar J. V. (1987) Efficiency of kriging estimation for square, triangular and hexagonal grids, Mathematical Geology 19, 183- 205.