UNIVERSITÀ DEGLI STUDI DI NAPOLI "FEDERICO II"

in consorzio con
SECONDA UNIVERSITÀ DI NAPOLI
UNIVERSITÀ "PARTHENOPE" NAPOLI
in convenzione con
ISTITUTO PER L'AMBIENTE MARINO COSTIERO – C.N.R.
STAZIONE ZOOLOGICA "ANTON DOHRN"

Dottorato in Scienze ed Ingegneria del Mare
XVII ciclo

Tesi di Dottorato

GEOSTATISTICAL MODELS FOR ENVIRONMENTAL DATASETS

Candidato: Dott. Simone Sammartino

Relatore: Dott. Mario Sprovieri
Co-Relatore: Dott. Ennio Marsella

Il Coordinatore del Dottorato: Prof. Bruno D'Argenio

ANNO 2005

CONTENTS

INTRODUZIONE

Nelle scienze ambientali il modo in cui ogni singolo dato viene elaborato e interpretato riveste un ruolo importantissimo per la comprensione della complessità dei fenomeni naturali. Il valore misurato di una variabile ambientale è il risultato di una miscela di fattori tanto complessa, che può essere ragionevolmente considerata come derivante da un processo casuale e quindi modellizzabile in un contenuto stocastico. Le metodologie afferenti all'ambito deterministico non sono in grado di rappresentare in modo completo il reale meccanismo responsabile della distribuzione spaziale di una certa variabile, limitandosi a modellizzare soltanto la variabilità intrinseca del sistema, attraverso vincoli matematici (Isaaks and Srivastava 1989).

L'approccio geostatistico è basato sull'assunzione che la variabile è il risultato di un processo stocastico, e per questo la definisce come variabile aleatoria. Il suo valore misurato è valutato come una possibile realizzazione di una certa funzione aleatoria in un certo punto. Ognuna di queste realizzazioni è una variabile aleatoria e l'insieme di tali variabili aleatorie, valutate nell'intero dominio spaziale, è definita come la funzione aleatoria generante.

L'analisi variografica e la stima con kriging rappresentano i concetti base dell'approccio geostatistico. Una applicazione corretta di tali strumenti, e un'accurata interpretazione dei risultati, permettono di comprendere la reale struttura di variabilità del sistema e realizzare ogni genere di analisi spaziale, finalizzata a simulare il suo andamento e a costruire modelli spaziali adeguati.

In questo lavoro, oltre ad una discussione dettagliata degli algoritmi che saranno utilizzati nelle applicazioni, vengono proposti alcuni esempi di corretto utilizzo degli strumenti geostatistici. La teoria presentata è selezionata da numerose fonti bibliografiche ed è riorganizzata al fine di fornire un'immagine chiara e più completa possibile del significato matematico delle diverse procedure utilizzate. Alcuni tra gli algoritmi più noti sono discussi in dettaglio, con la ricostruzione completa degli sviluppi matematici. Per alcuni di questi sviluppi sono implementate alcune varianti specifiche, come la proposta dell'utilizzo di un intorno unico, indipendentemente dalla conformazione della matrice di stima, e della valutazione

contemporanea di varianza nugget e varianza di piccola scala nel processo di ottimizzazione delle strategie di campionamento.

L'implementazione dettagliata della variografia e della stima con kriging viene prospettata come applicazione sul dataset geochimico relativo ai campioni di sedimento marino prelevati nella zona meridionale della piattaforma campana. Una analisi variografica accurata ha permesso di discriminare due gruppi principali di variabili, tra i metalli in tracce; il primo caratterizzato da una variabilità di tipo mono-strutturale e l'altro da una variabilità di tipo nidificata. Il kriging delle componenti spaziali viene sfruttato per separare le due componenti spaziali e utilizzato come strumento ausiliario per la definizione dei valori di background naturale delle singole variabili selezionate.

La variografia viene successivamente implementata nell'ambito delle procedure di ottimizzazione, finalizzate al perfezionamento delle strategie di campionamento. Dopo una breve descrizione delle metodologie di campionamento già note, il variogramma viene presentato come uno strumento innovativo per l'ottimizzazione di un nuovo piano di campionamento e il miglioramento di uno già esistente. Le simulazioni geostatistiche vengono ampiamente sfruttate, in questo ambito, per modellizzare l'incertezza spaziale, mentre viene proposta la combinazione delle procedure di valutazione della varianza di nugget con la valutazione di tale incertezza. Un esempio concreto di tale approccio deriva dall'applicazione ai dati geochimici della 'Nuova Darsena' nel porto di Napoli.

Infine, il kriging delle componenti spaziali viene applicato, come tecnica di filtraggio delle immagini, ad un esempio di mosaico acustico side scan sonar. La componente nugget viene filtrata al fine di rimuovere il rumore *sale e pepe*, responsabile di una interpretazione e classificazione del segnale non corrette.

Il pacchetto software di applicazioni geostatistiche ISATIS, della società francese Geovariances (Bleines, Deraisme et al. 2004), è ampiamente usato come strumento principe per l'implementazione di analisi spaziali avanzate, in ambito stazionario e non stazionario, per il calcolo di variogramma direzionali e omnidirezionali, per le stime con kriging e per realizzazione di simulazioni iterative a diverse scale spaziali, attraverso le tecniche sequenziali gaussiane. Tutti i dataset

utilizzati per le applicazioni provengono dall'archivio della sede centrale di Napoli dell'Istituto per l'Ambiente Marino Costiero (I.A.M.C. - C.N.R.).

Gli esempi di corrette applicazioni geostatistiche presentati dimostrano quanto sia importante la messa a punto degli strumenti geostatistici al fine di onorare le caratteristiche di variabilità strutturale del dataset. La gran parte dei software di elaborazione dei dati spaziali oggi include l'analisi variografica e la stima con kriging, ma molti di essi prevedono soltanto applicazioni di tipo "black box" che presumono un impiego automatico delle procedure. L'attenzione verso il corretto utilizzo di tali strumenti dovrebbe rappresentarne, invece, l'aspetto cruciale. La grande sensibilità e dipendenza di tali metodologie rispetto alle caratteristiche dei dati, implicano la necessità di un uso attento ed esperto di tali applicazioni.

La geostatistica è una disciplina dalle grandi potenzialità, soprattutto perché è basata sulla modellizzazione della variabile in una prospettiva probabilistica e in questo modo rispetta la complessità intrinseca del sistema ambientale. Essa rappresenta un valido supporto non solo per stime o simulazioni spaziali, ma anche come strumento per l'ottimizzazione dei processi pre-analitici, che hanno una influenza critica sul risultato finale. Nella maggior parte dei casi, gli algoritmi di interpolazione vengono utilizzati per la produzione di mappe soggette a complesse interpretazioni, analisi di sistemi ambientali e decision making. E' evidente quindi come un controllo affidabile del processo di trattamento sia fondamentale per dare forza ai risultati.

INTRODUCTION

In environmental sciences the importance of how the single data is processed and interpreted is a crucial aspect for the correct understanding of the complexity of natural phenomena. The measured value of an environmental variable is the result of such a complex mixture of different factors, that it can be reasonably considered as deriving from a random process and thus it can be modeled in a probabilistic framework. Deterministic methodologies can poorly catch the real mechanism responsible for the spatial distribution of the variable, limiting to model, only by mathematical constraints, the intrinsic variability of the field (Isaaks and Srivastava 1989).

Geostatistical approach is based on the probabilistic concept that assumes the variable as the result of some stochastic processes, and thus regards it as a random variable. The measured value of a certain variable is seen as one realization of a certain random function in one specific point. Each of this realization is just one random variable and the whole of random variables in the entire spatial domain is regarded as the parent random function.

Variographic analysis and kriging estimation are the fundamentals of the geostatistical approach. A correct application of such instruments and an accurate interpretation of results allow to understand the real variability structure of the field and compute any kind of spatial analysis, aimed to simulate its behaviour and build adequate spatial models.

In this work, after a detailed discussion of the algorithms will be used in the applications, some examples of the right use of geostatistical instruments are proposed. The presented theory is collected from the numerous literary sources and restructured, in order to provide a clearest and most complete image of the mathematical meaning of the different practices. Some main algorithms are discussed in detail, with the rebuilt complete mathematical develop. Some personal concerns are presented; such as the proposal of the use of a unique neighbourhood, independently from the set of the estimation grid and of the contemporary evaluation of nugget variance and small scale variance during the optimization of sampling strategies.

The detailed implementation of variography and kriging estimation is presented as applied to the geochemical dataset deriving from marine sediment samples collected in the southern Campanian continental shelf. An accurate variographic analysis has allowed to discriminate two main groups of variables, among trace metals; one characterized by a mono-structural variability and one with a nested ones. Kriging of spatial components has been exploited to separate the two spatial components and used as an indicative tool to define the natural background values for the single selected variables.

Moreover, variography has been implemented as an optimizing procedure aimed to improve the efficiency of sampling plan strategies. After a brief discussion of existing different kinds of sampling strategies, variogram is presented as innovative tool to optimize a new sampling plan and to improve the efficacy of an existing one. Geostatistics simulations are widely exploited to model the spatial uncertainness and an innovative joins between such methodology and nugget variance assessment is proposed. A concrete example is applied to one of the inspected dock of the port of Naples, the 'Nuova Darsena'.

Eventually, kriging of spatial components is applied, as image filtering technique, to an example of side scan sonar acoustic mosaic image. Nugget effect component is filtered out in order to remove the *salt and pepper* noise responsible for incorrect interpretation and classification of signal.

The ISATIS geostatistical package, from the french Geovariances society (Bleines, Deraisme et al. 2004), has been used as a master tool to implement advanced spatial analysis in stationary and non stationary frameworks, to compute omnidirectional and directional variograms, kriging estimations and iterative computation of different scale sampling simulations, via sequential gaussian technique. All the datasets used for applications come from the collection of the Neapolitan center of the Coastal Environment Institute of the National Research Council (I.A.M.C. C.N.R.).

The shown examples of correct application of geostatistics demonstrate how it is important to tune the powerful instruments of such discipline, in order to meet "dataset demands". The majority of mapping software now include variogram analysis and kriging estimation, but most of them take into account only the so

called “black-box” practices, whereas the attention due to the setting of these instruments is a crucial aspect. The susceptibility and the so deep dependence of such methodologies on data behaviour, imply a very careful and expert use of such applications.

Geostatistics is a very powerful method, just because it is based on a modeling of variable from a probabilistic perspective and it respects the intrinsic complexity of the environmental system. It is a valid issue not only for estimation or simulation classical processes, but also for several pre-processing optimization tools, as sampling strategies definition, that have a critical influence on final results. In most cases, mapping algorithms are used to produce maps that are subject to tricky interpretation practices, environmental systems assessments and decision making. It is evident how a full and robust control of the processing pathway is fundamental to optimize and make reliable such results.

CHAPTER 1
THEORY FUNDAMENTALS

Here a detailed description of the theory applied in the rest of the work will be presented. Statistical inference, variographic analysis, the fundamentals of geostatistical inference and kriging will be discussed, with particular attention to the mathematical develop of the algorithms. All the algorithms described here have been collected from the most known bibliographic sources and rebuilt in a organic structure that can be regarded as a valid reference for the reader.

1. Statistical inference

Statistical inference concerns the problem of gain information about a certain population, from a set of samples generated by it (SurfStat Australia). When we want to obtain detailed description of a population, instead to work directly on it (it is often impossible), we can compute the statistical parameter of a certain set of samples, selected from the same population. Such operation leads to the estimation of the peculiar characteristics of the population, that is usually the more similar to reality the more the number of samples is close to the totality. Inference is, thus, exactly the operation of approximating a statistical description of a certain population, from the computation of statistical quantities of a defined sample (Soliani 2005). The detailed processes are:

1. sample selection
2. calculation of statistics of sample
3. estimation of population parameters based on sample statistics

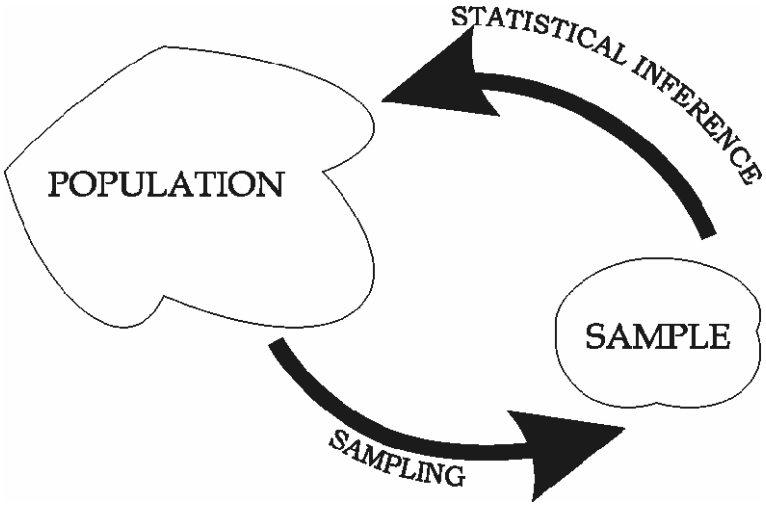


Figure 1.1 – Conceptual scheme of statistical inference.
Figura 1.1 – Schema concettuale dell'inferenza statistica.

Sampling is a selection process that can be repeated iteratively and can return any time a different result. If randomly repeated, this process gives origin to a random variable, studied to infer the population who generated it. In this case population can be regarded to be virtual and infinite because there are infinite random samples potentially extracted from it. In the statistical inference process, it is just the stochastic process that generated that certain sample to be studied and that process is exactly the population with its unknown and estimated statistical behaviour. Statistical inference is also finalized to quantify the correctness of such estimation. If we were able to extract some different subsets (samples) from the same population, we could try to estimate its statistics from a statistical treatment of all the samples. For example, if we want to estimate the mean value of a population we can extract several samples from it and examine the statistical distribution of the variable *mean value*. Such distribution is regarded as *sampling distribution*.

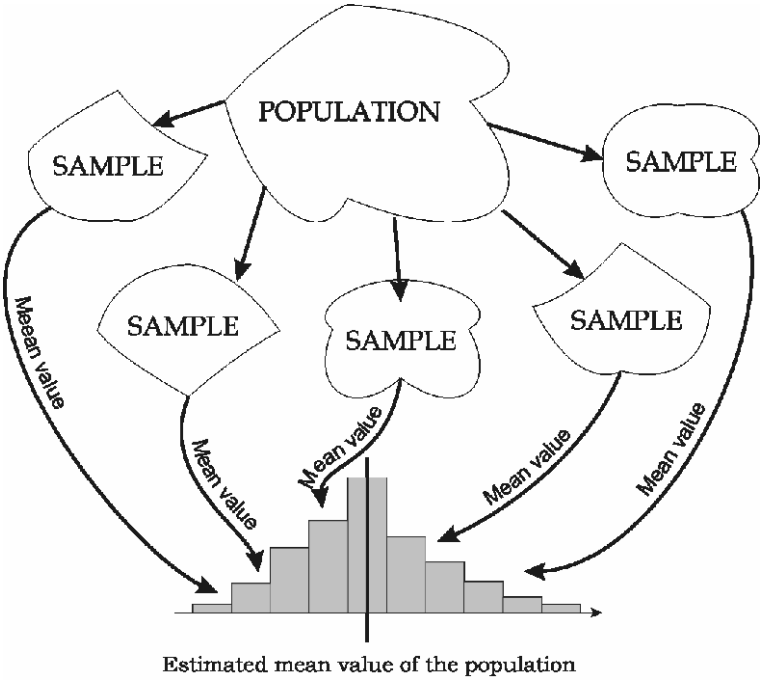


Figure 1.2 – The sampling distribution.
Figura 1.2 – La distribuzione del campione.

The estimated mean value of the population is the mean value of such distribution and the difference from such expected value of the samples and the real mean value of the population (usually unknown) is defined as the *bias*.

2. Geostatistics – an overview

Geostatistics is a relatively new branch of applied statistics that had origin from mining industries, where it has been used to affine techniques of computation of ore reserves. Both the African mining engineer DG Krige, with his practical applications, and the French mathematician George Matheron, with his theorization of the *Regionalized Variable* model, contributed to the born of such an important discipline, that has been applied on most of science and industry fields. Geostatistics is based on the study of the spatial behaviour of variables (Isaaks and Srivastava 1989). It concerns the modeling of such characteristics, with the aim of identifying the underlying structure, in order to implement different operations on data (estimation, simulation). The main difference between the geostatistical and the classical statistical approach to data analysis is exactly the spatial reference of datasets: all the statistical parameters are related to the spatial location of the single

information. Even the concept of variable is converted in its spatial context as the *regionalized variable* (Armstrong 1998; Journel and Huijbregts 2004).

The model of the regionalized random variable is the basis principle of such kind of science.

From a probabilistic point of view, a certain sampled value $z(x)$, measured in some specific location, is the outcome of a random process that generated it from a random variable $Z(x)$. In any x point of the region of interest, $Z(x_\alpha)$ can have a different property and thus generates different value $z(x_\alpha)$. In this way, all the infinite measured values $z(x)$ that can be collected in the entire spatial domain, can be viewed as a single outcome of all the infinite random variables involved.

Each single set of random variables is called *random function*. Each sample is a random variable and a regionalized variable respectively from a random and a regional point of view. Both of the concepts yield to the model of random function.

In practice, the regionalized variable is one realization of the random variable in each point of the spatial domain (Wackernagel 2003).

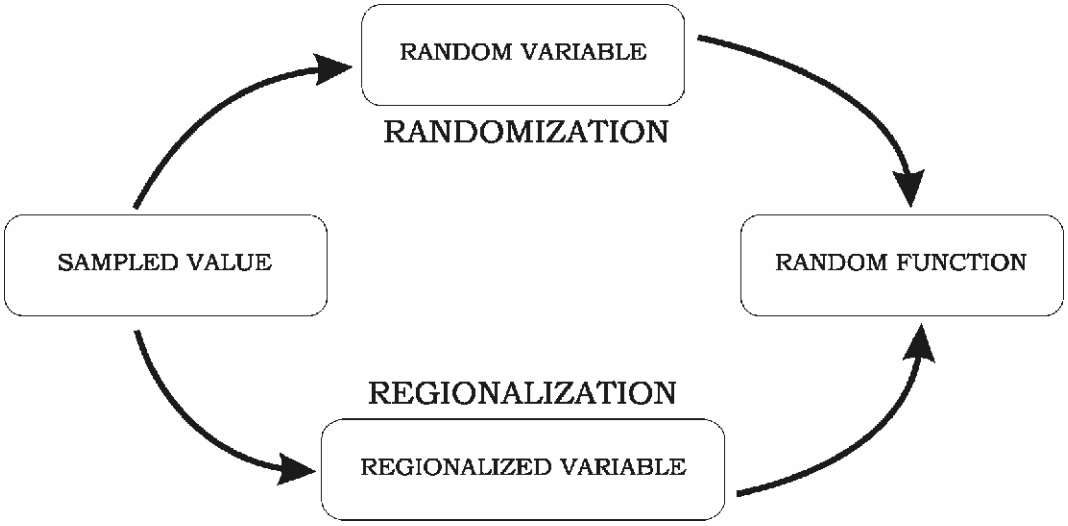


Figure 1.3 – Conceptual scheme of the Theory of Regionalized Function.
Figura 1.3 – Schema concettuale della Teoria della Funzione Regionalizzata.

The basic concept of Geostatistics is the probabilistic approach to the modeling of physical mechanism (Lee el. source). Each single value of a regionalized variable, and above all its spatial variability structure, is assumed to be described efficiently in a stochastic framework. Data values are regarded as the result of a

miscellaneous of complex factors and can be reasonably viewed as possible outcomes of a random process; in general, such randomness can be poorly described in a deterministic framework.

A probabilistic model can better represent the structural variability of the field and be the basis to implement any spatial data processing. Often, the complete knowledge of the model is not required and it is sufficient to specify only some parameters of the whole random process.

3. Structural analysis

The structural analysis concerns all the methodologies aimed to investigate the spatial structure of data and exploit it to build reasonable spatial models. As discussed before, the main target is to represent exhaustively the random function model.

Now, the problem concerns our knowledge of such random function. We can barely hope to make any assumption on it, knowing only one realization of such random function in few points only (one value for each sample). One solution is to reduce the study of the spatial law to the first two moments. Nevertheless, in order to be able to use such spatial law as a model of the spatial variability of the variable, we must take some other assumption. Such hypothesis regards the stationarity of the variable, that influences the validity of the law trough the spatial domain.

There are two main degree of stationarity: i) the second order one, that assumes both the mean value (first moment) and the covariance (second moment) are stationary, and ii) the intrinsic one that assumes the covariance function does not exist and introduces the use of the variogram function. The crucial aspect of such assumptions is just the existence of the covariance/variogram function and its dependence only on the distance between pairs of samples and not on the absolute position of samples. Such important statement is the core of the main branch of geostatistics, the *stationary geostatistics*, that is largely used in environmental data processing.

The three main feature of the variogram function are:

- the behavior at small spatial scales (nugget variance);
- the way it grows and reaches a certain threshold (authorized model type);

- the threshold value it reaches (sill variance).

These are the crucial clues to mould the model (our simplification of the spatial law) in order to reproduce the complex variability of the variable. The basic process of the variographic approach is based on: i) the computation of the experimental variogram – that is the primitive calculation directly deriving from the real data, and ii) its fitting with a deterministic function (chosen among a family of authorized models), in order to synthesize the variability of the variable and express it with some few parameters.

3.1 Empirical approach

A first and easy kind of approach to better understand the application of geostatistical methodology is the empirical one (Raspa 2000; Raspa 2004). Let's consider a certain spatial variable $Z(x_i)$, measured in some location x_i with $i = 1...n$ and some distance vector h, oriented in a certain direction, with an associated tolerance ε. Let's consider the groups of pairs of samples that have a distance kh with $k \in \aleph$, along that direction, within $\pm\varepsilon$ tolerance, and calculate for each group the scatterplot and the correlation coefficient.

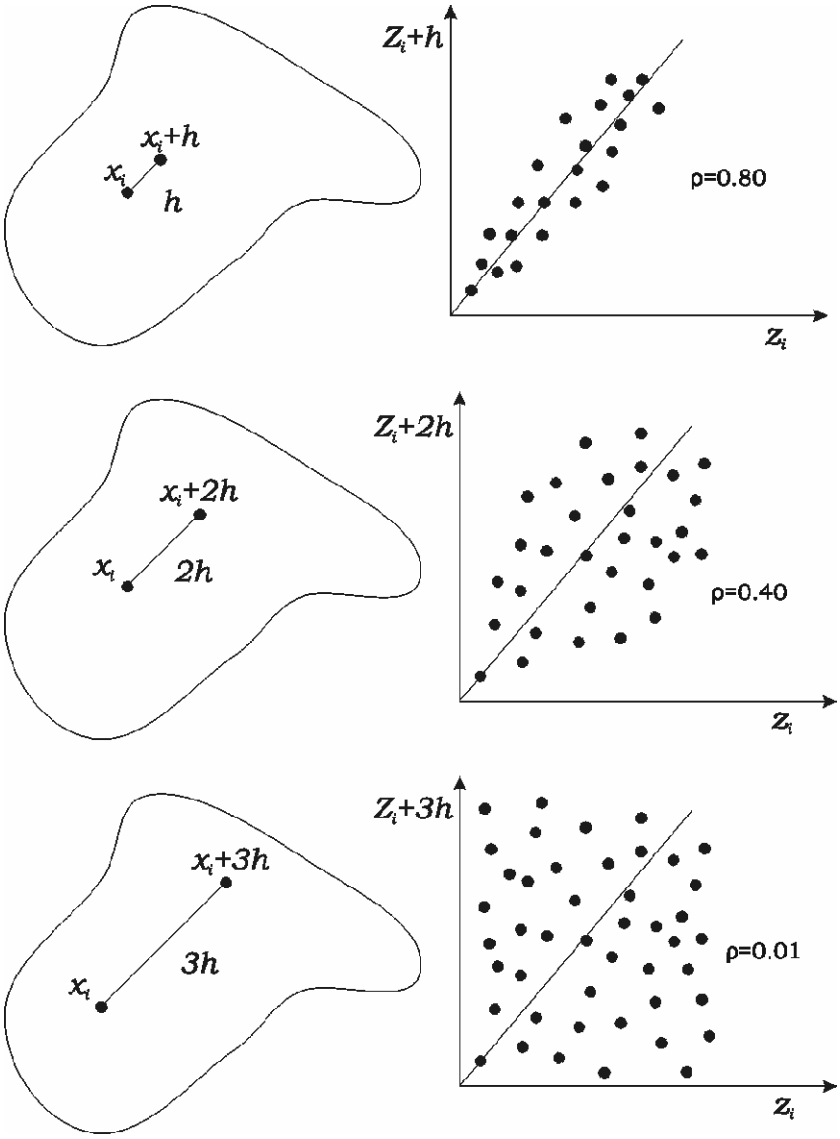


Figure 1.4 – Empirical evaluation of covariance function.
Figura 1.4 – Valutazione empirica della funzione covarianza.

If we plot the values of correlation coefficient against the distance h, we can obtain a synthetic form to explain the spatial variability of variable $Z(x_i)$. We'll observe that ρ, that is exactly one for zero distance (when x_i and $x_i + h$ are coincident), decreases with h, and it tends asymptotically to null value with the increasing distance among pairs.

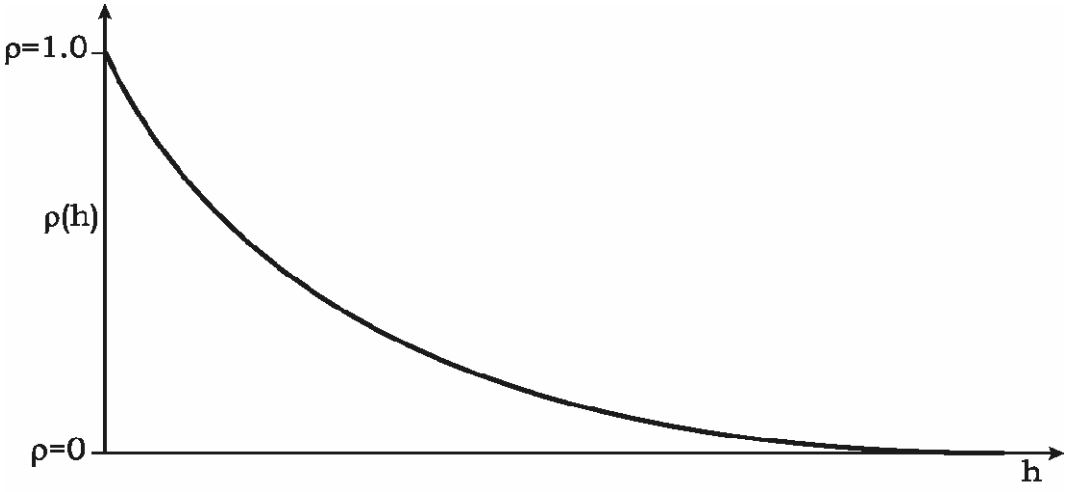


Figure 1.5 – Trend of correlation coefficient with distance.
Figura 1.5 – Andamento della funzione coefficiente di correlazione.

Now, if we substitute the standardized covariance (correlation coefficient), with the variance of histograms of pairs for each class distance, we obtain an analogue function, that show some peculiar differences.

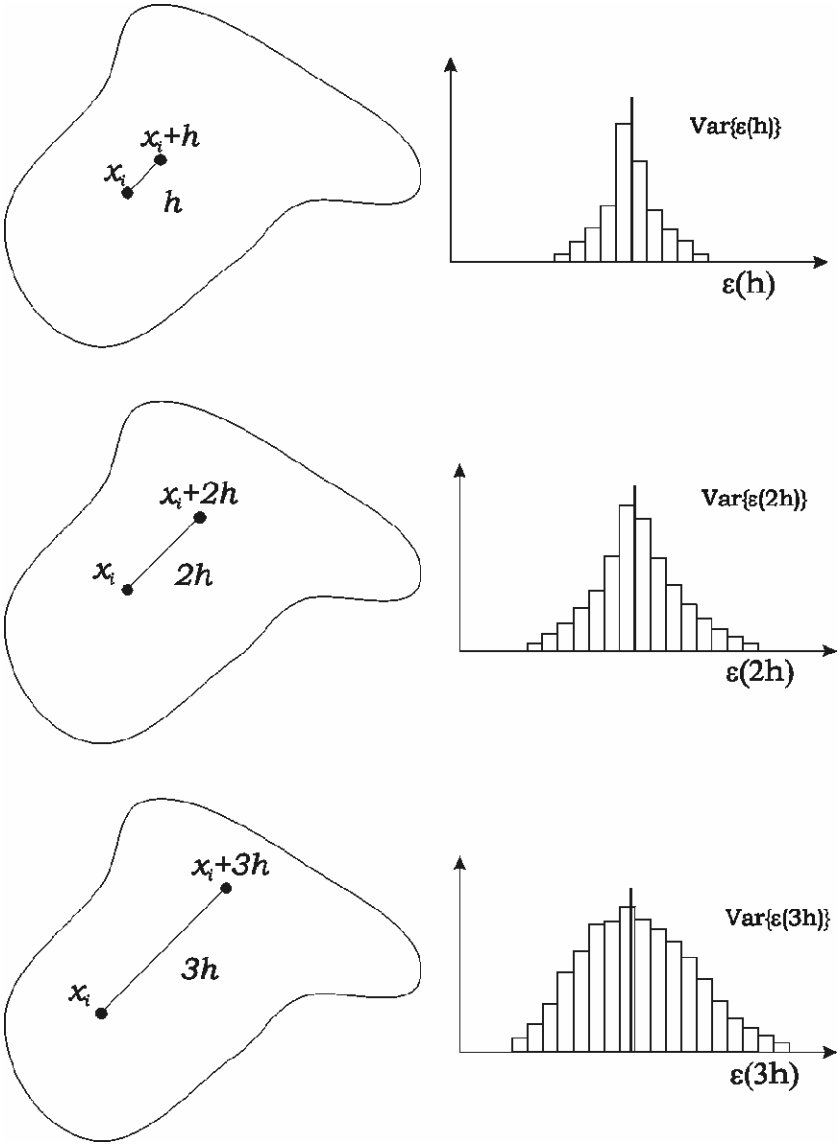


Figure 1.6 – Empirical evaluation of variance.
Figura 1.6 – Valutazione empirica della varianza.

Variance of increments, that can be regarded as *variogram* function (more precisely as twice the variogram function), presents a mirrored shape of covariance one. It is zero at zero distance and increases with increasing distance tending asymptotically to a certain value that is usually comparable to the sample variance of the whole dataset.

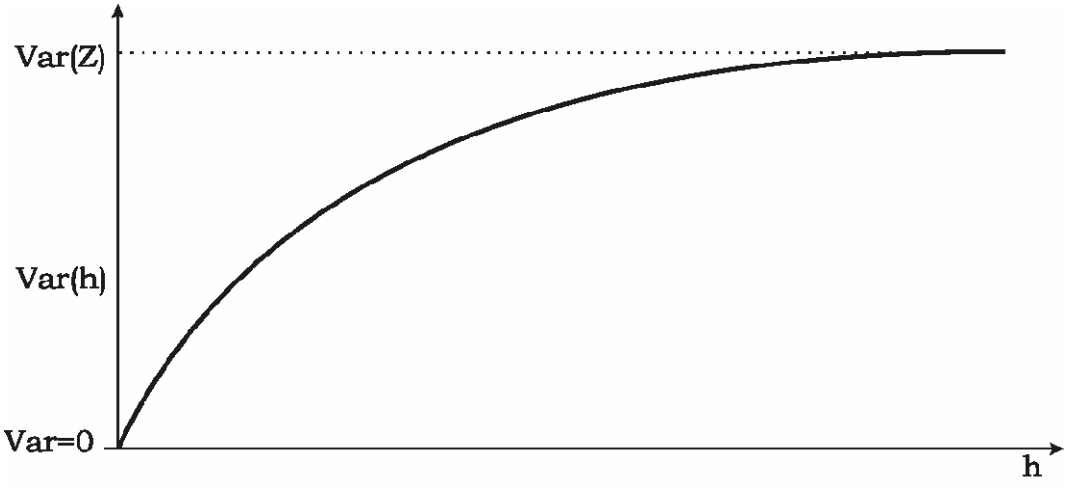


Figure 1.7 – Trend of variance with distance.
Figura 1.7 – Andamento della funzione varianza.

Variogram function is one of the most important instruments of the structural analysis and just the transposition of such empirical approach to a more formal mathematical one, represents the basis of the probabilistic theory from which Geostatistics takes origin.

3.2 Probabilistic approach

In any location of the entire spatial domain, the sampled value can be regarded as the realization of a regionalized random function. It means this values are one possible outcome of a function that generates them following a certain *regionalized probability density function* (Deutsch 2002; Dutter 2003).

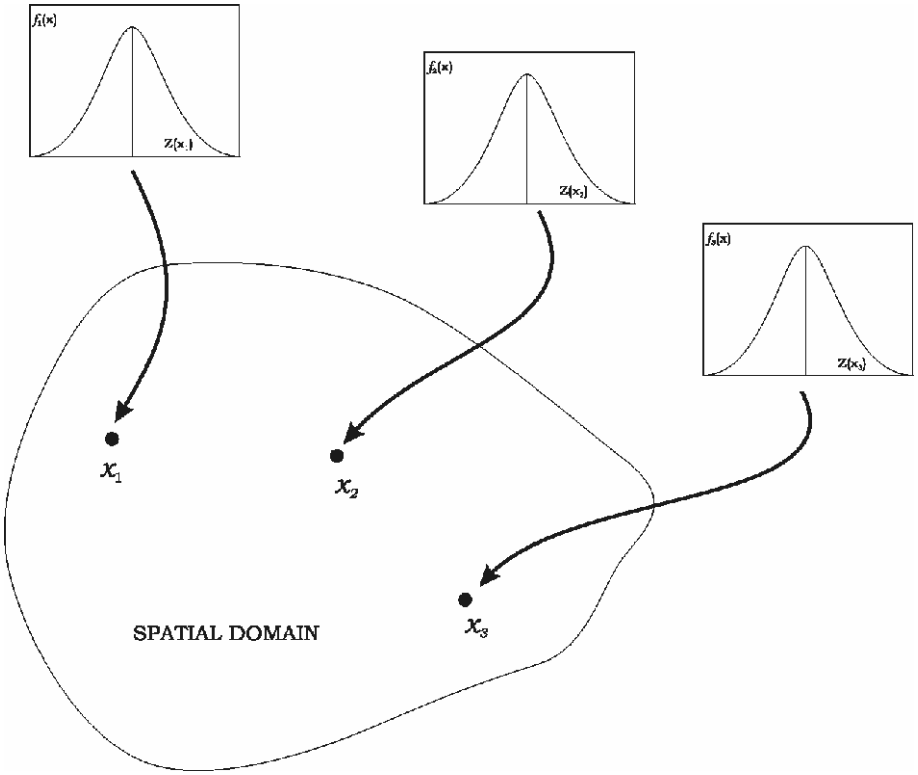


Figure 1.8 – The measured value in each point is one of the possible outcome of one realization of the random function.
Figura 1.8 – Il valore misurato in ogni punto è uno dei possibili risultati di una realizzazione della funzione aleatoria.

Considering that all the possible realizations are represented by the same random function, it must be necessary related to some spatial law. Such law is represented by the sum of the joint distribution functions (cumulative distribution functions) of the *n* random variables:

$$F_{x_1 \ldots x_n}(z_1,\ldots z_n) = pr\{Z(x_1) \le z_1,\ldots, Z(x_n) \le z_n\}.$$

Such spatial law is an example of an extreme generalization of the stochastic modelling of any universal natural process, and if we knew it in details, we could presume to be able to reproduce any feature of such phenomenon. Unfortunately it is not possible, because our knowledge of the phenomenon is based on only *one* realization of the random function, moreover available only in *some* points, of the entire spatial domain.

We need a model to implement statistical inference on sample, but the theorized one (the spatial law) is too generic and cannot be explicated. We then need some simplifications of such model, based on some statistical hypothesis.

The most important assumption is the *stationarity* of the random function (strict stationarity), where the vectors of random variables $\{Z(x_1),...,Z(x_n)\}$ and $\{Z(x_1+h),...,Z(x_n+h)\}$ have the same distribution function. That is, any spatial distribution of samples (if any shift is applied to any samples configuration) presents the same distribution function.

Because of the limitation of the amount of available samples, such condition is reasonably acceptable only up to a certain degree; that is only some degrees of stationarity are acceptable (Deutsch 2002). Thus a wide range of degrees of stationarity are defined, going from the closest to the ideal description, to the farthest one.

As most of geostatistical processes (linear estimations) need the involvement of only the first two moments (mean and variance) of the entire spatial law, the assumption of stationarity can be limited only to such two moments.

3.3 Spatial models

Spatial probabilistic models are needed to model the random function and to implement geostatistical inference to spatial variables. Once analyzed the spatial structure of the dataset, we can assume to build different models following the degree of stationarity examined. Starting from the pure stationary model and ending to the different models of non stationary functions, we can be able to model any kind of spatial variability.

3.3.1 Stationary models

Stationarity hypothesis is essential to simplify the probabilistic model and to allow to implement statistical inference on the sample (Isaaks and Srivastava 1989). We already admitted that strict stationarity, applied to the whole random function, is unreliable, firstly because we lack a complete description of the phenomenon and then because we often do not need such generalization. For most cases, especially if variable have a Gaussian-like distribution, stationarity hypothesis can be reduced to the first two moments. In such case, in fact, distribution function can be reasonably represented by the two first moments.

Second order stationarity

Condition 1:

The first moment exists and it is translation invariant

$E[Z(x_\alpha)] = m$ for all the x.

Condition 2:

The second moment (covariance) exists and does not depend on the position but only on the distance h

$$Cov(x_1, x_2) = Cov(x, x+h) = E[Z(x) \cdot Z(x+h)] - m^2 = Cov(h).$$

Assumed the shift invariance of the first moment it is reasonable to accept that the spatial variability function is the same for the whole spatial domain, once fixed the module and direction of h.

The assumption of second order stationary should include also the condition of existence of the variance and its shift invariance. Nevertheless, it is implicit in the second condition, because at $h = 0$ covariance corresponds with variance.

$$Cov\{Z(x), Z(x+h)\} = Cov\{Z(x), Z(x)\} = Cov(0) = Var\{Z(x)\} \quad \text{for } h = 0.$$

Thus, even the variance exists and it is translation invariant for the whole spatial domain:

$$Var\{Z(x)\} \equiv Var\{Z(x+h)\}.$$

When both the conditions are satisfied and the stationarity is of the second order, $Cov(h)$ or $C(h)$ is called *covariance function*, that describes the variability of the auto-correlation of the increments with distance h:

$$Cov(h) = E[Z(x) \cdot Z(x+h)] - m^2.$$

Covariance function encloses some important features:

1. it is an *even* function: $Cov(h) = Cov(-h)$
2. it always assumes positive values at $h = 0$. In fact $Cov(0) = Var\{z(x)\} \geq 0$
3. Schwartz inequality is worth: $|Cov(h)| \leq Cov(0)$

Variogram function

Variogram is the most common function in geostatistics and it often substitutes covariance function because of its lack in non stationarity cases (see chapter 3.3.2). Let's consider a new random function that explicates the difference between the quantities $Z(x)$ and $Z(x+h)$, within a certain distance $h \pm$ its tolerance ε, and call it *increment* $[Z(x) - Z(x+h)]$.

The variance of such random variable is:

$$Var[Z(x) - Z(x+h)] = E\{[Z(x) - Z(x+h)] - E[Z(x) - Z(x+h)]\}^2 =$$

$= E\{[Z(x) - Z(x+h)]^2\} - \{E[Z(x) - Z(x+h)]\}^2$, according to covariance function of the same variable $[Z(x) - Z(x+h)]$;

in a second order stationary case, if generally $E[Z(x)] = E[Z(x+h)]$, then $E[Z(x) - Z(x+h)] = 0$.

Thus $Var[Z(x) - Z(x+h)] = E\{[Z(x) - Z(x+h)]^2\}$.

Variogram is defined as half of variance of increment, that is half the expectation value of the square difference between $Z(x)$ and $Z(x+h)$:

$$\gamma(x,h) = \frac{1}{2} E\{[Z(x) - Z(x+h)]^2\} =$$

$$= \frac{1}{2}\left\{E[Z(x)]^2 + E[Z(x+h)]^2 - 2E[Z(x)\cdot Z(x+h)]\right\} =$$

$$= \frac{1}{2}\left\{Var[Z(x)] + m_x{}^2 + Var[Z(x+h)] + m_{x+h}{}^2 - 2Cov[Z(x), Z(x+h)] - 2m_x m_{x+h}\right\}$$

$$= \frac{1}{2}\left\{Var[Z(x)] + Var[Z(x+h)] - 2Cov[Z(x), Z(x+h)] + [m_x - m_{x+h}]^2\right\}.$$

In a second order stationary case, when $m_x \equiv m_{x+h}$,

$Var\{Z(x)\} \equiv Var\{Z(x+h)\} \equiv Cov(0)$ and $Cov\{Z(x), Z(x+h)\} = Cov(h)$

$\gamma(h) = \frac{1}{2}\{2Cov(0) - 2Cov(h)\}$;

that is:

$\gamma(h) = Cov(0) - Cov(h)$.

It shows that variogram function depends on covariance and consequently, in stationary conditions, it is translation invariant too. If covariance exists and Schwartz inequality is worth, variogram is limited to the variance of the original dataset (called *a priori* variance) (Raspa 2000).

$|Cov(h)| \le Cov(0)$ that is $|\gamma(h) - Cov(0)| \le Cov(0)$ that is $\gamma(h) \le 2Cov(0)$ that is $\gamma(h) \le 2Var\{Z(x)\}$

That is the reason why variogram is half the variance; in this way it tends asymptotically to the a priori variance and not twice this value.

Unfortunately, most of times, such stationarity hypothesis are too hard to be applied to sampled variables and real life is much more complicated than our theoretical assumptions would hope to treat with. We have to simplify our assumption to some weaker form of stationarity.

Quasi-stationarity

Sometimes, the mean of the random function is not exactly constant, but it varies weakly within a certain spatial sub-domain. The so called *quasi-stationary* model assumes that the function can be considered as a second order stationary one, within the said neighbourhood.

In a more general description of the random function, its variogram will be:

$$2\gamma(x,h) = E\left\{\left[Z(x) - Z(x+h)\right]^2\right\} - \left[m_x - m_{x+h}\right]^2,$$

that is exactly correspondent to the classical formulation of the second order stationary model, when $m_x \equiv m_{x+h}$.

Now, let's consider a new variable, said residual, with zero mean, that represents, in any point, the difference between the variable $Z(x)$ and it's local mean $m(x)$:

$$\mathrm{Y}(x) = Z(x) - m(x).$$

We admit that such residual is second order stationary, that is its covariance exists and its variogram is dependent only on h and it is up limited. Its variogram is:

$$\gamma(h) = \frac{1}{2}E\left\{\left[\mathrm{Y}(x) - \mathrm{Y}(x+h)\right]^2\right\} - \left\{E\left[\mathrm{Y}(x)\right] - E\left[\mathrm{Y}(x+h)\right]\right\}^2,$$

that is

$$\gamma(h) = \frac{1}{2}E\left\{\left[\mathrm{Y}(x) - \mathrm{Y}(x+h)\right]^2\right\}$$ because its mean is zero everywhere.

If we substitute $\mathrm{Y}(x) = Z(x) - m(x)$ in the last and rearrange the equation

$$2\gamma(h) = E\left\{\left[Z(x) - m(x) - Z(x+h) + m(x+h)\right]^2\right\} =$$
$$E\left\{\left[\left[Z(x) - Z(x+h)\right] - \left[m(x) - m(x+h)\right]\right]^2\right\} =$$
$$= E\left\{\left[Z(x) - Z(x+h)\right]^2 + \left[m(x) - m(x+h)\right]^2 - 2\left[Z(x) - Z(x+h)\right]\left[m(x) - m(x+h)\right]\right\}$$

and expand the average

$$2\gamma(h) = E\left\{\left[Z(x) - Z(x+h)\right]^2\right\} -$$
$$- 2\left[m(x) - m(x+h)\right]E\left\{\left[Z(x) - Z(x+h)\right]\right\} + \left[m(x) - m(x+h)\right]^2 =$$
$$= E\left\{\left[Z(x) - Z(x+h)\right]^2\right\} - 2\left[m(x) - m(x+h)\right]^2 + \left[m(x) - m(x+h)\right]^2 =$$
$$= E\left\{\left[Z(x) - Z(x+h)\right]^2\right\} - \left[m(x) - m(x+h)\right]^2$$

that is exactly correspondent with the general variogram of the function $Z(x)$ when it is not second order stationary.

Thus, studying variogram of non stationary function $Z(x)$, means studying variogram of its residual. In particular, such variogram is exactly correspondent to the stationary one, except for the second part that explains the trend component (the spatial variability law of the mean) of the global variability law of the random function.

Let's assume, now, that the trend is a linear function:

$$m(x) = a_0 + a_1 x,$$

and substitute it in the last formulation of variogram.

$$2\gamma(x,h) = E\left\{\left[Z(x) - Z(x+h)\right]^2\right\} - \left[a_0 + a_1 x - a_0 - a_1(x+h)\right]^2 =$$
$$= \gamma(h) - \left[a_1(x - x - h)\right]^2 = \gamma(h) + {a_1}^2 h^2$$

The last equation explains the shape of the variogram of a non stationary function. It is the linear combination of the "true" variogram of the variable and its trend, that is represented by a parabolic shape (deterministic). It is just the presence

of a parabolic behaviour of experimental variogram that allows us to realize we are treating a non stationary random function (Clark 2000).

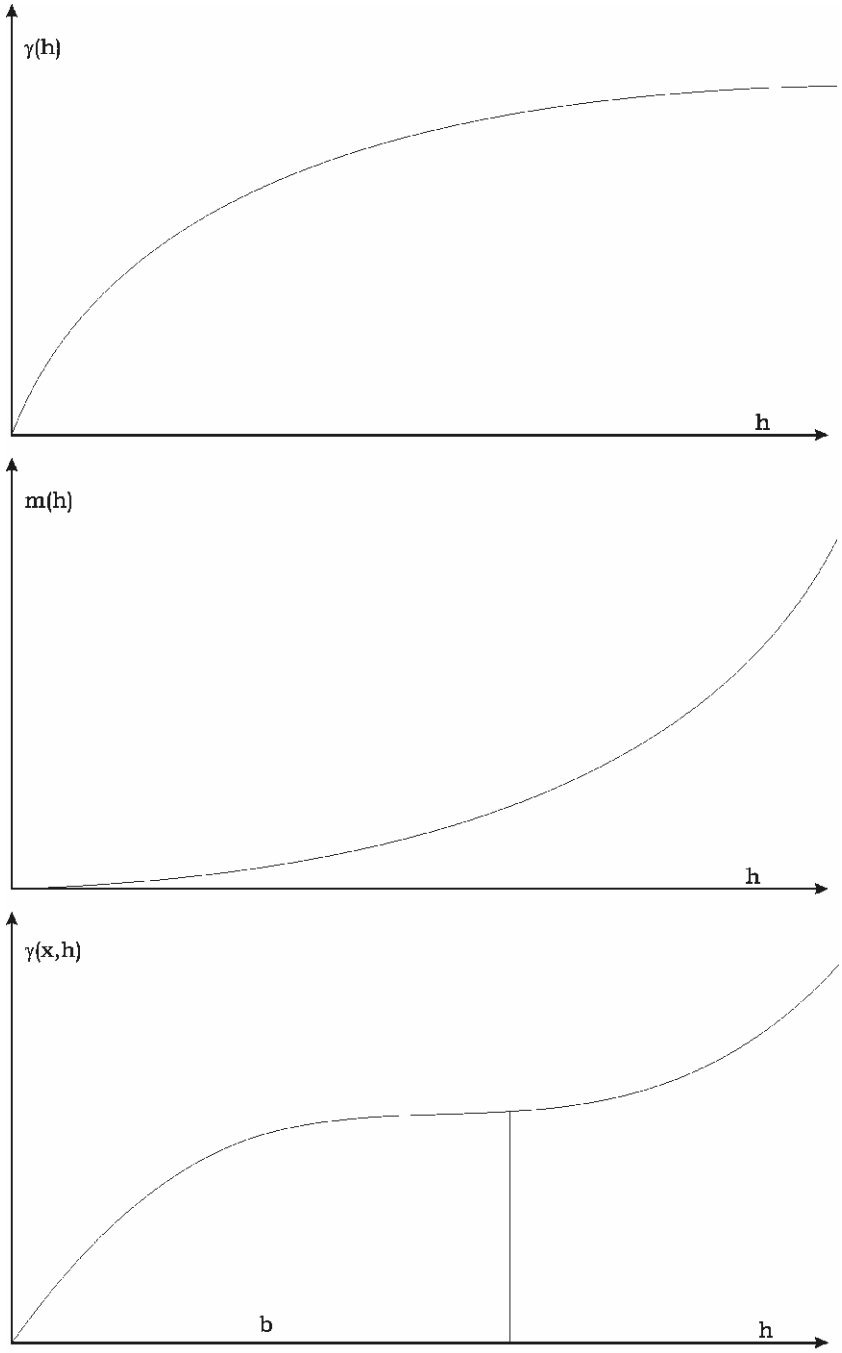


Figure 1.9 – The quasi stationary model is the sum of a bounded model and a parabolic component.
Figura 1.9 – Il modello quasi stazionario è la somma di un modello limitato e di una componente parabolica.

Such parabolic drift depends on coefficient a_1. Thus, there will be a certain range b within which the parabolic component is negligible and it does not influence the true variogram shape. In practice, if we consider a_1 such that for $h \leq b$, $a_1{}^2 h^2 \cong 0$, we can consider the function as a second order stationary one (Raspa 2000). It means that if we limit our neighbourhood to b we can treat with a stationary random function, even if it is not exactly so (Armstrong 1998; Wackernagel 2003; Journel and Huijbregts 2004; Raspa 2004).

Thus, its variogram would be simply:

$$\gamma(h) = \frac{1}{2} E\left\{\left[\mathrm{Y}(x) - \mathrm{Y}(x+h)\right]^2\right\}.$$

When such range *b* cannot be reasonably defined and it is so short that the parabolic component is never negligible, we are forced to reformulate the problem in a non stationary framework.

Intrinsic Hypothesis

A even weaker hypothesis of stationarity is based on *stationarity of increment* and it is applied to the majority of conditions. If the random function presents a constant mean, but its variance varies with the increase of neighbourhood, we can apply second order stationarity assumption to the increment of the random function, instead of it own. One common example is the result of a Brownian random walk, that shows a variation of the variable with a constant mean and a varying variance, whose increment is second order stationary (Raspa 2000).

Thus, for the intrinsic hypothesis, the first moment of such increment is zero anywhere in the domain:

$$E\left[Z(x) - Z(x+h)\right] = 0$$

and its second moment (variance) exists and has a finite value $2\gamma(h)$, depending on the vector value h, but not on the position of itself:

$$Var[Z(x)-Z(x+h)] = E\{[Z(x)-Z(x+h)]^2\} - \{E[Z(x)-Z(x+h)]\}^2 =$$
$$= E\{[Z(x)-Z(x+h)]^2\} = 2\gamma(h)$$

It yields to the theoretical definition of the variogram:

$$\gamma(h) = \frac{1}{2} E\{[Z(x)-Z(x+h)]^2\}.$$

When the increment of a random function is second order stationary it is called *Intrinsic Random Function,* and the variogram is called exactly *intrinsic function* (Wackernagel 2003). Intrinsic stationarity can be extended to higher order increment; the simple increment can be regarded as the zero order increment that is able to filter out zero degree drift (constant ones). Higher order intrinsic functions can exist, able to filter out higher order trends (see chapter 4.3.5).

Intrinsic models are less demanding than second order stationary ones. Stationary models are always intrinsic (that is they admit stationarity of increment), while intrinsic ones are not second order stationary. In fact they are used to treat random function that do not admit covariance function. In particular the intrinsic models that are not stationary are called *strictly intrinsic*. Theoretical treatment of intrinsic models is exactly the same of second order ones, except that variogram is computed in substitution of covariance function (Raspa 2000; Raspa 2004).

3.3.2 Non stationary models

In order to treat some weak stationarity, we tried to reduce it to a stationary framework, making some assumptions about modelling variable or its increment. When the assumptions made before are not sufficient and the function is inevitably non stationary, we are forced to take into account the trend and integrate it in the structural analysis process. Nevertheless, we have to keep in mind that the stationarity is a feature of the stochastic model and not of the random function (Raspa 2004); besides it is dependent on the spatial scale of the studied phenomenon.

There are two main approaches to solve such problem:

- the *Dichotomy* or *Drift models*, related to Universal Kriging method;
- the *Intrinsic random function of order k*, related to Irf-k Kriging method.

Dichotomy – drift models

Sometimes, natural phenomena act following some well known physical mechanism, and modelling of spatial variability must take into account such feature. Very often, such mechanism determines a global trend variability of the field, added to the random one. At this point we can try to model such trends, taking into account the affordability of our estimate. The more the process is effectively driven by a known physical mechanism, and such mechanism is well known by the user, the more the reliability of our assumption is valid.

The dichotomy principle affirms that the measured value of a certain variable is the sum of a deterministic global trend and a stochastic local variability. Thus, these two components can be separated and the first one can be modelled in a pure deterministic framework. Nevertheless, such concept must be read following the scale of observation of the phenomenon and the authenticity of the deterministic behave of the trend (Raspa 2004). The scale of observation is very important, because it determines the border between the pure non stationary framework, that involves a deterministic modelling of trend variation, and the quasi-stationary one, for which, as we have already seen, a simple restriction of neighbourhood is sufficient to carry the model back to a stationary situation.

Moreover, the solidity of the hypothesis about the deterministic nature of the phenomenon is a crucial aspect. We must assume and apply a deterministic approach to modelling of trend, only if we have a clear idea about the nature of physical phenomenon which is based on. We can not use it as in black box, because deterministic trend analysis is very sensitive to the kind of chosen model. The variogram of residual is distorted proportionally to the correctness of the trend fit (Wackernagel 2003).

The variable is regarded as a composed quantity:

$Z(x) = \mathrm{Y}(x) + m(x)$, where $m(x)$ represents the trend deterministic component and $\mathrm{Y}(x)$ the random residual part.

Thus, drift component is approximated with a polynomial function $m(x)=\sum_l a_l f^l(x)$, where a_l are the coefficients of the polynomial and f^l are the degree l monomials.

The crucial aspect of dichotomy approach is just the determination of such trend component, that is, as said before, strictly dependent on the scale of observation and on the kind of model chosen. This aspect involves a subjective evaluation of the phenomenon and it is very critical.

The residual $\mathrm{Y}(x)$ is treated as a random function in a second order stationary framework or in a intrinsic one, depending on the existence of covariance function. The variogram of residuals is said "underlying", because its inference is not direct and it depends on the deterministic modelling of the trend. The variography of the drift model is the base of *Universal Kriging* estimation approach.

Intrinsic random function of order k – Irf-k

The Irf-k method is a generalization of the intrinsic random function model, in which higher order increments are computed to filter out some polynomial trend of a non stationary random function (Journel and Huijbregts 2004). The order of Irf-k is exactly the order of the trend it is able to filter out. The random function is regarded as the sum of a unknown varying mean and a residual stationary random function (Buttafuoco and Castrignanò 2005):

$Z(x)=m_x+\mathrm{Y}(x)$, where $m(x)=\sum_{l=0}^{K} a_l f_l(x)$ is a linear combination of monomials f_l with unknown coefficients a_l.

Such monomials are defined with the condition that their linear combination, evaluated in each x_i point is zero:

$$\sum_{i=0}^{n} \lambda_i f^l(x_i)=0 .$$

Once such conditions are satisfied, the linear combination $Z(\lambda)=\sum_{i=0}^{n}\lambda_i Z(x_i)$ is defined as the *authorized linear combination* or *increment* of order *k*. Such higher order increment is able to filter out higher order trend. Structural analysis computed on such transformed variable ($Z\lambda$) is similar to the classical stationary practice, except that covariance is computed with an automatic procedure and it is the sum of some authorized basic structures.

The nested so called *generalized covariance* is:

$K(h)=\sum_{p} b_p K^p(h)$, where the components b_p are generally polynomials and splines.

Non stationary methods like Universal Kriging or Irf-k both are sensible instruments that need a cautious use. Both are based on a decomposition of the variability into a deterministic low fluctuating component and a rapidly fluctuating random one that is investigated with some specific structural analysis tools. In the case of Universal Kriging, such global mean is fitted by a deterministic function, arbitrarily chosen by the user and the resulting variogram of residuals is computed traditionally (Wackernagel 2003).

Conversely, in Irf-K, the global mean is fitted with a deterministic combination of basic functions, such that the random function is transformed in a higher order increment that is stationary. In the UK, the most critical step is just the fit of the detrending function, that is chosen arbitrarily, often not knowing exactly its real nature. The resulting underlying variogram is necessarily biased. On the other hand, Irf-K approach, once chosen the degree of the detrending function, is an automatic method and, as such, it is poorly verifiable.

3.4 Geostatistical inference

The most critical step of geostatistical approach is the modeling of the experimental variogram. It is regarded as a synthetic form for explaining the structural variability of data. In order to be able to honour such information, in spatial modelling approach, we have to express the result of the structural analysis

with a continue and theoretical function. Such process needs a very careful attention of the user, in order to correctly represent the real model of the original dataset.

3.4.1 Variogram evaluation

Often, covariance function cannot be computed because of its lack, in the case the mean of the variable is not constant. Conversely, variogram of increment of the variable, is an universal good estimator of the spatial variability that can be correctly used also when the variable is intrinsic (Burrough and McDonnell 1997). Besides, the shape of variogram curve helps us to catch the non stationarity degree of the field, in order to choose the more appropriate method to process it. Once the variogram has been computed, if its shape is revealing enough stationary to proceed with classical methods, the next step is the fit of the theoretical variogram function (Clark 2000).

The raw data and their related experimental variograms cannot be exploited to obtain a complete knowledge of the variability structure of the phenomenon, and some species of continuous function is needed (Isaaks and Srivastava 1989). The inference of the experimental variogram is based on a least square fit of the experimental values of semivariance for each lag. Such fit can be made by a single or a series (nested variograms) of so called *authorized models*. The authorized functions are the ones that follow particular conditions, among which the main one is to be positive definite (Wackernagel 2003).

By fitting a continuous mathematical function on raw variogram we can exploit such powerful instrument in order to model the variability structure for the whole spatial domain (not only on the points we have measured values). Such precious informations can be used in estimation and simulation processes.

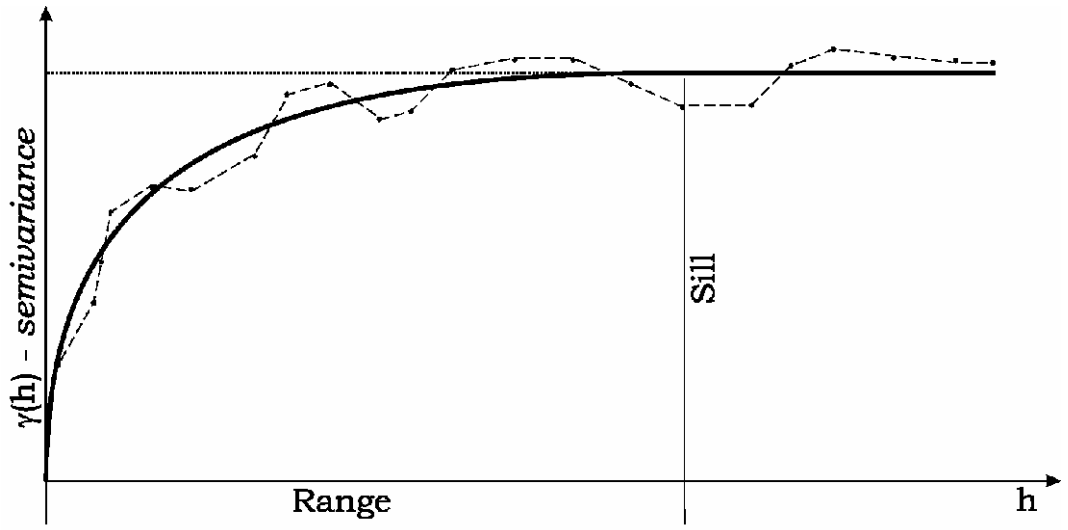


Figure 1.10 – The experimental variogram and its theoretical fitting function.
Figura 1.10 – Il variogramma sperimentale e la sua funzione approssimante.

The crucial aspect of the fitting operation is the choice of the model. In this step, the subjectivity of the process of adjusting the model and the way to associate weights to the different lags is the moment in which the experience and the smartness of the geostatistician have the greatest impact (Armstrong 1998). The technique is usually based on a dualistic process of manual adjusting and a statistical verification by a least square optimization procedure. However, the "try and catch" method should be always weighted on the knowledge of the phenomenon. The main parameters of a typical variogram function are: i) the *sill* (the variance value at which the function tends asymptotically) and ii) the *range* (the lag value at which the sill is reached). The most evident differentiation among the family of authorized models can be made between *bounded* and *unbounded* variograms.

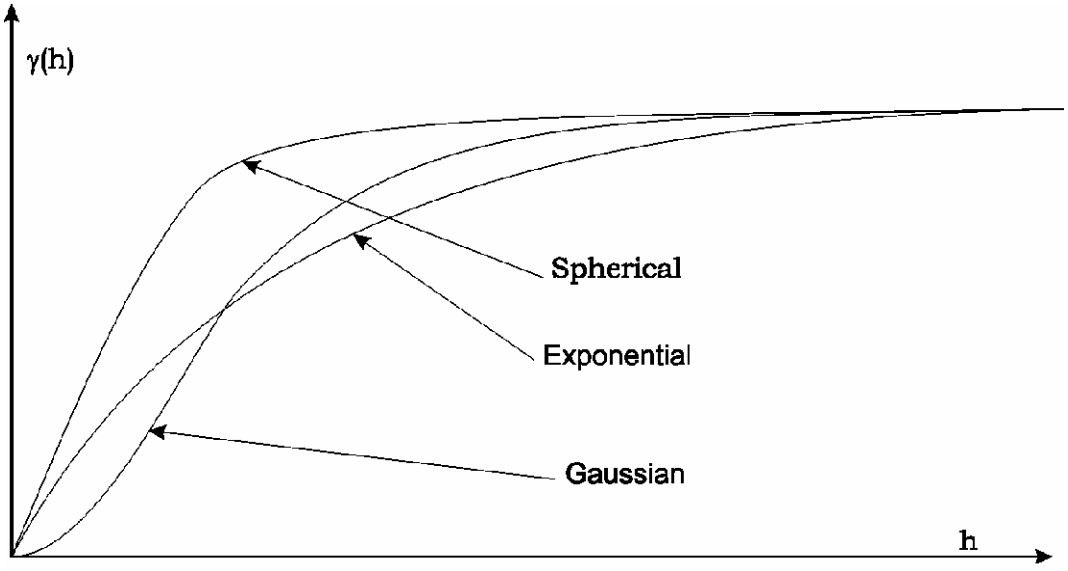


Figure 1.11 – The three most known example of bounded variogram models.
Figura 1.11 – Tre esempi, tra i più noti, di modelli di variogramma limitati.

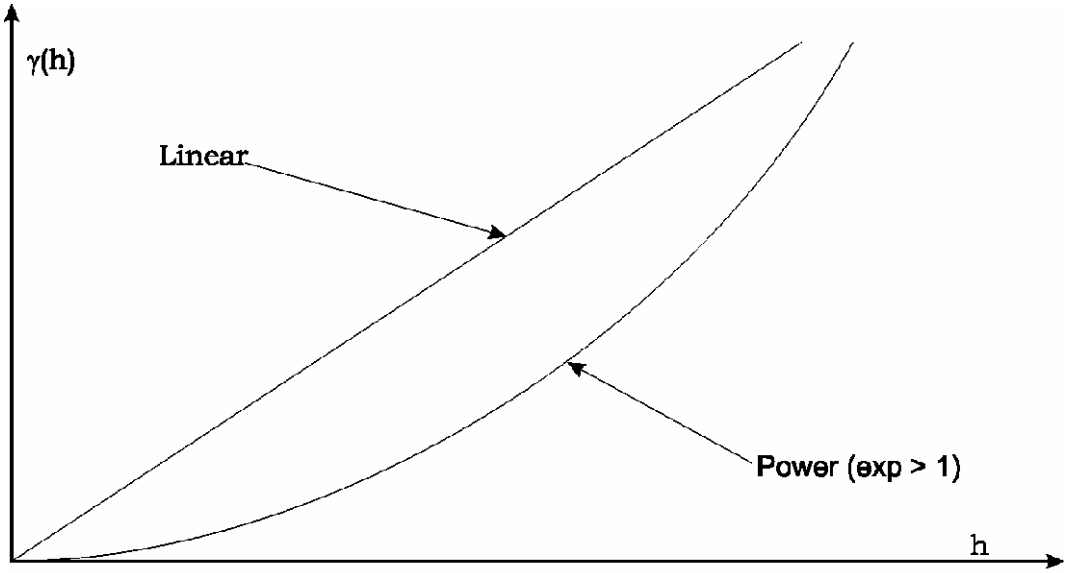


Figure 1.12 – Two examples of unbounded variogram models.
Figura 1.12 – Due esempi di variogrammi non limitati.

Those in figure are some of the most common authorized variogram models. The bounded ones (spherical, exponential and gaussian) show an asymptotic tendency to attest on the sill value and, although they have different behaviour at the origin, they represent a stationary random function. Contrarily, the unbounded ones, the ones that continue increasing not reaching a sill, characterize some form of non

stationarity: an intrinsic stationarity for the linear model and a marked non stationarity for the power model.

3.4.1.1 Nested structures

The so high complexity of natural phenomena is almost always related to the amount of factors that cooperate to generate a certain spatial distribution. The variability structure is the result of a mix of factors contributing each with a different weight and at a different spatial scale. Such situation needs to be quantified in the stochastic interpretation of the phenomenon and must not be ignored in the process of modelling of the random variable.

Let's assume that the stationary variable is sum of different independent variables:

$$Z(x)=\sum_{u=0}^{S-1} Z_u(x) \text{ and } Z(x+h)=\sum_{u=0}^{S-1} Z_u(x+h).$$

Its variogram is (Raspa 2004):

$$\gamma(h)=\frac{1}{2}E\left\{\left[Z(x)-Z(x+h)\right]^2\right\}=\frac{1}{2}E\left\{\left[\sum_{u=0}^{S-1} Z_u(x)-\sum_{u=0}^{S-1} Z_u(x+h)\right]^2\right\}=$$

$$=\frac{1}{2}\left\{\left[\sum_{u=0}^{S-1}\left[Z_u(x)-Z_u(x+h)\right]\right]^2\right\}=$$

$$=\sum_{u=0}^{S-1}\sum_{u=0}^{S-1} E\left\{\left[Z_u(x)-Z_u(x+h)\right]\left[Z_u(x)-Z_u(x+h)\right]\right\}.$$

Expanding the square:

$$2\gamma(h)=2\sum_{u=0}^{S-1} E\left\{\left[Z_u(x)-Z_u(x+h)\right]^2\right\}+2\sum_{u=0}^{S-1}\sum_{u=0}^{S-1} E\left\{\left[Z_u(x)-Z_u(x+h)\right]\left[Z_u(x)-Z_u(x+h)\right]\right\}=$$

$$=\sum_{u=o}^{S-1} E\left\{\left[Z_u(x)-Z_u(x+h)\right]^2\right\}+\sum_{u=0}^{S-1}\sum_{u=0}^{S-1} Cov\left(\left[Z_u(x)-Z_u(x+h)\right],\left[Z_u(x)-Z_u(x+h)\right]\right)$$

As $Z(x)$ and $Z(x+h)$ are independent, their covariance is null and:

$\gamma(h) = \sum_{u=0}^{S-1} \left\{ [Z(x) - Z(x+h)]^2 \right\} = \sum_{u=0}^{S-1} \gamma_u(h)$, that is, the variogram of $Z(x)$, is the sum of the variograms of all its components. Such components are regarded as *spatial components* and each of their contribute to the total variability (the weights) is represented by each own sill. The global variogram is called *nested variogram*.

One typical example of nested variogram is the one in figure.

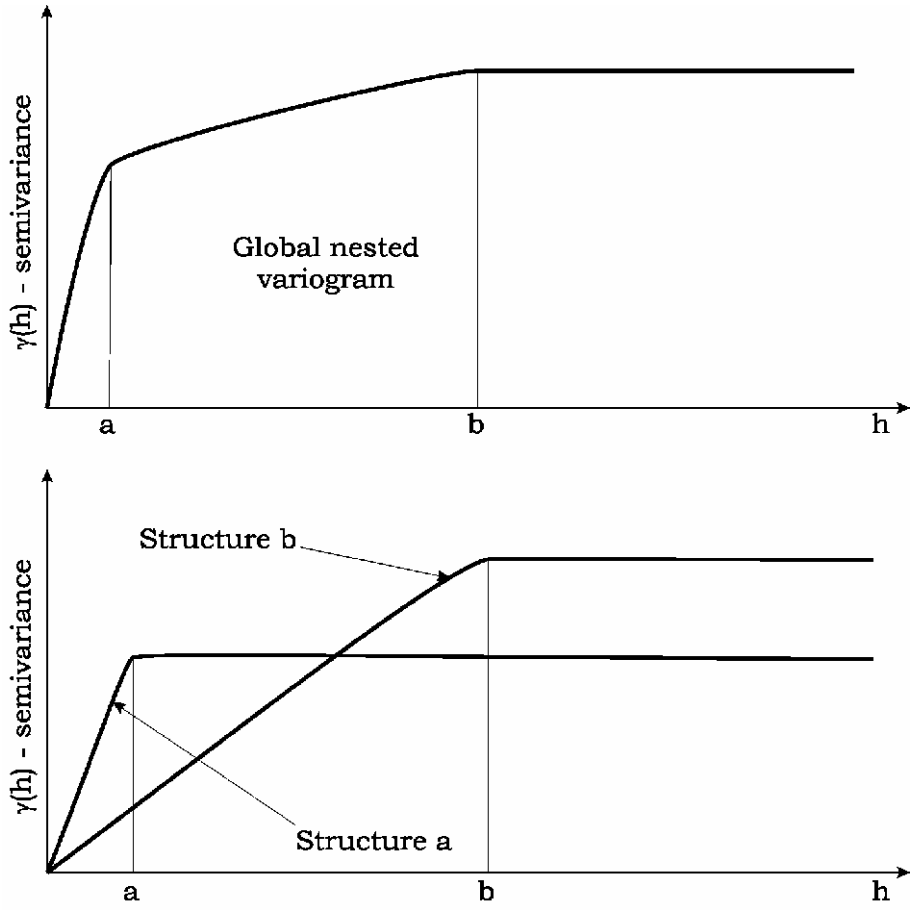


Figure 1.13 – Top: the global nested structure. Down: the exploded structure with the two spatial model components.
Figura 1.13 – In alto: la struttura nidificata globale. In basso: la stessa struttura scomposta nelle sue due componenti spaziali.

Two independent components (γ_a *and* γ_b) are contributing to the global structure of variability, each with its range *a* and *b*, that give an indication of the

spatial scale at which such structure is acting, and with its sill, that represents the magnitude of each contribute (Isaaks and Srivastava 1989).

3.4.1.2 The nugget effect

One of the most important variogram models, for joining the classical structural analysis with the study of small scale variability and optimizing of sampling strategy is the *nugget model*. Variogram mathematical formalism assumes the semivariance value is zero at zero lag, that is, two points have exactly the same value if their reciprocal distance is zero.

With the distance increasing it is assumed that the differences among pairs increase with the said law. Nevertheless, with the distance tending to zero such difference must tend to zero. Very often, it happens that the high dissimilarity of very close pairs show non zero values at the smallest lags. As said before experimental variograms are computed directly from raw data and it is unreal we can have different measures realized in the same point.

Moreover, even if it was real, it would hard to obtain exactly the same value for two or more replicas of the measure. The inevitable measurement error are around the corner and its minimization is limited by human and machine accuracy.

Anyhow, more often, information at lag values less than sampling size are not available and small scale variability are only estimated by known values.

Such situation leads to discontinuous experimental variogram showing non zero value at lag = zero, revealing a general marked small scale variability.

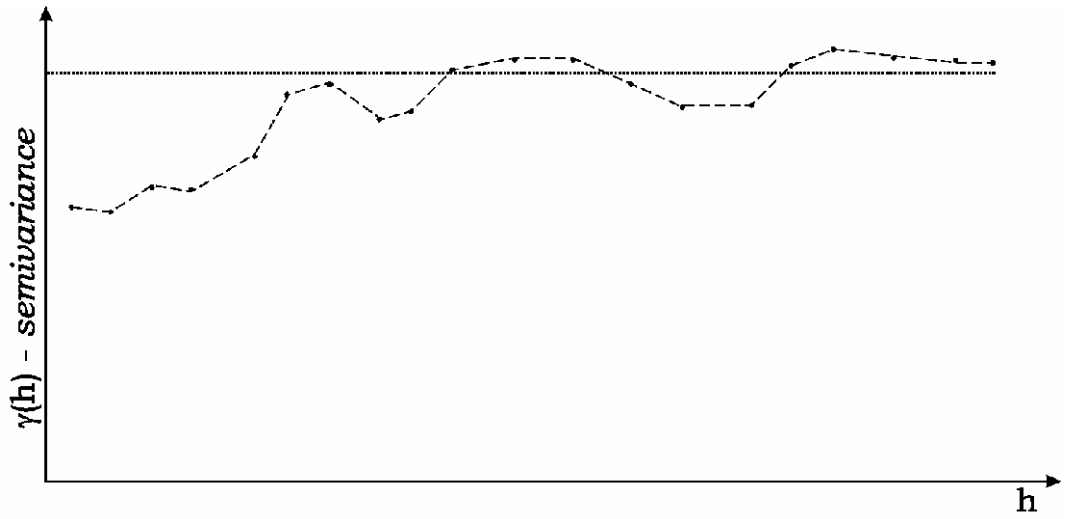


Figure 1.14 – An example of experimental variogram with a marked nugget effect.
Figura 1.14 – Un esempio di variogramma sperimentale con una marcata struttura nugget.

In general, environmental variables are assumed to have a regular behaviour at small scale. Scaled with the scale of investigation, it is unreal that two point very close each other can show very different values of altitude, pollutant concentration or moisture. High dissimilarity in pairs, at lags tending to zero, must inevitably be attributed to other causes. Such causes can be i) the ratio between the scale of the phenomenon we are interested in, and the scale of investigation and ii) the analytic errors. As said before, if the minimum sampling size is very larger than the scale of the phenomenon we are observing, we are not able to extract the correct small scale information. Moreover if the analytic error are comparable with the a sample global variance, the small scale variability is deeply influenced by this uncertainness.

In order to be able to model such small scale variability we can use a special variogram model (the nugget model) that has formally infinite range and the sill equal to the estimated intercept on the semivariance axis.

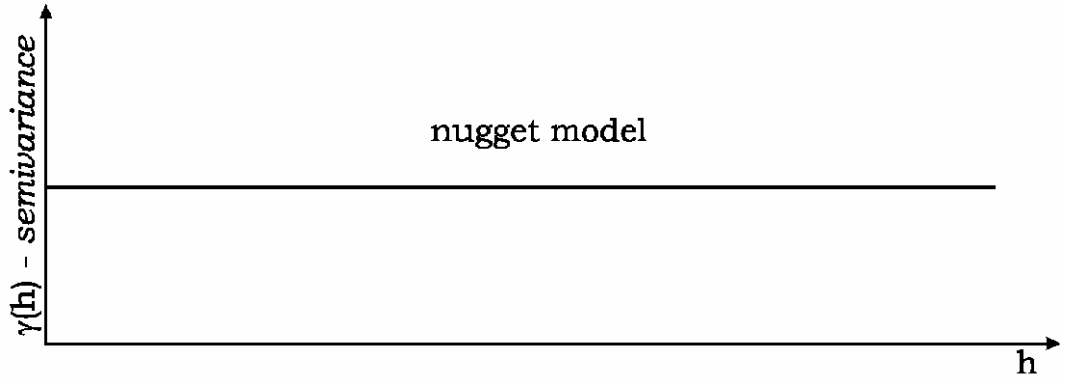


Figure 1.15 – The pure nugget model.
Figura 1.15 – Il modello nugget puro.

Almost always, the nugget model is used as spatial component in a nested variogram model. The small scale variability is just modelled by nugget model, that represents the translation of the main model/s to non zero semivariance values at the smallest lags. Such translation is called *nugget variance*.

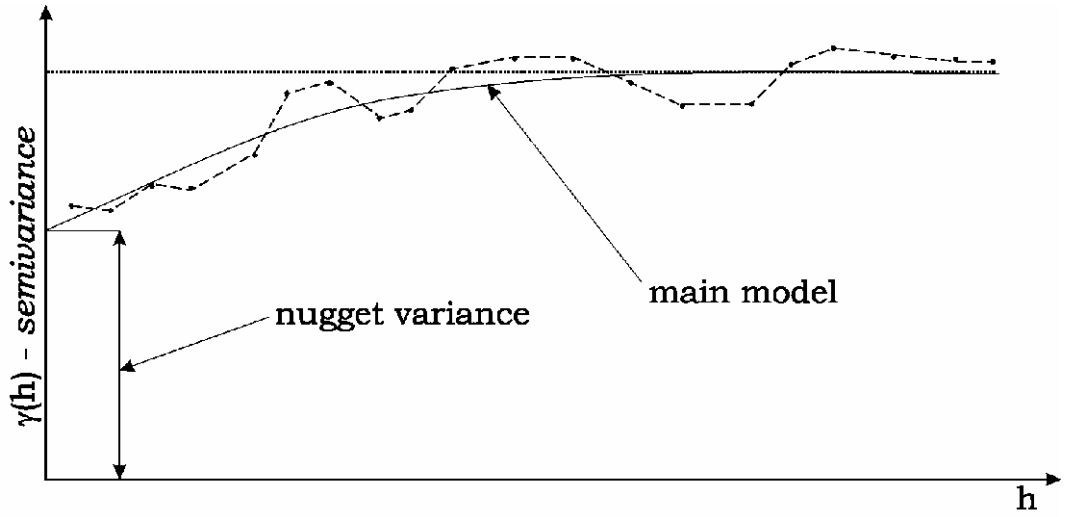


Figure 1.16 – The example of figure 1.14 with its modelled theoretical variogram model. The nugget effect component is evident.
Figura 1.16 – Lo stesso esempio della figura 1.14, per il quale è stato stimato il modello approssimante. La componente nugget è evidente.

Eventually, in some cases data show the complete absence of auto-correlation and their model is called *pure nugget*.

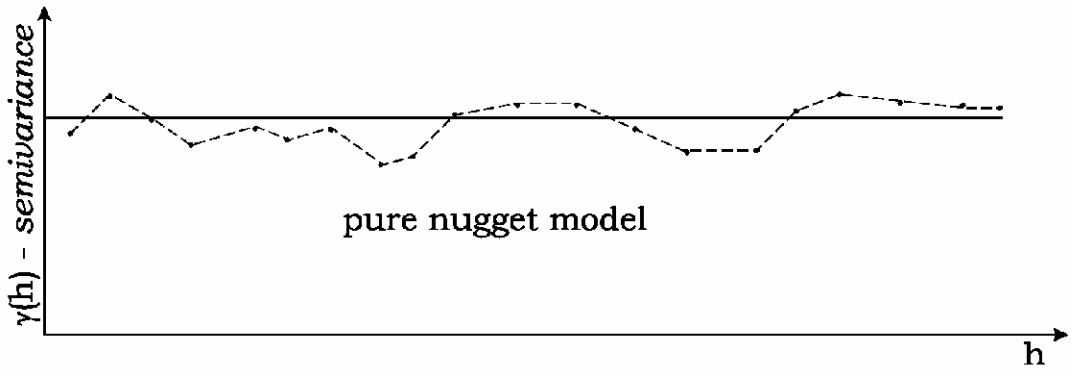


Figure 1.17 – An example of pure nugget experimental variogram and its theoretical model .
Figura 1.17 – Un esempio di variogramma sperimentale puramente nugget e del suo modello approssimante.

Such cases are not conveniently processed by a geostatistical approach, because the main assumption, that is the presence of a well designed spatial structure, comes to missing.

4. Estimations

One of the main target of the spatial analysis of environmental data is the so called *spatial estimation*. It is clear that *in loco* measures can be realized only for some points; it would be unreliable to imagine to obtain a complete knowledge of the field, measuring it in any point of the spatial domain. Moreover the distance between two samples could be infinitely reduced, giving origin to the request of an infinite amount of samples.

Taking into account the poor availability of information about the spatial distribution of a certain variable, is clear that the method that allows us to pass from a *discrete information* to a *continue description* of the phenomenon must be assigned the maximum importance. Another limitation is represented by the fact that we have often scattered data, that is the samples are irregularly distributed trough the spatial domain, while we need a regular distribution of information in order to obtain a continue descriptive surface.

Moreover, we are often interested to know the value of a certain variable in a point out of the convex hull (the convex polygon that contains all the points) of known data (Isaaks and Srivastava 1989).

All these needs are reassumed under the definition of *estimation process*, that is the process with which we can estimate the value of a certain variable where it is not known, knowing only few information, often scattered in the spatial domain and, sometimes, not surrounding completely the target point. The main target is to obtain a regular matrix of information, on a dense regular grid with a certain resolution, with which we can build any vector or raster graphic representation of the estimated spatial distribution of the variable.

4.1 The linear estimator

The main instrument to deal with the estimation process is the *linear estimator* (Raspa 2004). For each target point the linear estimator is the linear combination of the known points:

$$\tilde{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i);$$

where $\tilde{Z}(x_0)$ is the estimated value in the target point, $Z(x_i)$ are the known values and λ_i are the weights of such values. In practice, the estimated target point value is the result of a weighted average of a certain amount of known values. The term "certain" is well used in this case, because it is unreal that all the point of the whole spatial domain can influence the value of the target point. Thus there must be a certain area of influence within which the target point value depends on the surrounding known values. Such area is called neighbourhood and its extent depends on the method chosen for the estimation process.

Main features of linear estimators

I – Unbiasedness of the estimator – the estimator is said *unbiased* when its average error is null. That is, the estimator makes as many overestimations as underestimations of the real unknown value. The error is the difference between the real and the estimated value:

$$\varepsilon(x_0) = Z(x_0) - \sum_{i=1}^{n} \lambda_i Z(x_i).$$

If $Z(x)$ is a random variable, the error $\varepsilon(x)$ will be too. As said before, the unbiased estimator is the one that shows a null averaged error:

$$E\left[Z(x_0) - \sum_{i=1}^{n} \lambda_i Z(x_i) \right] = 0 \text{ that is } E[Z(x_0)] - \sum_{i=1}^{n} \lambda_i E[Z(x_i)] = 0 .$$

If we assume that $Z(x)$ is stationary, that is $E[Z(x)] = m$, the equation becomes:

$m - m\sum_{i=1}^{n} \lambda_i = 0$ that is $m\left[1 - \sum_{i=1}^{n} \lambda_i \right] = 0$ that leads to the condition:

$$\sum_{i=1}^{n} \lambda_i = 1 .$$

Such condition is the one that guarantees for the unbiasedness of the estimator.

II – Maximum precision of the estimator – in order to obtain the best estimation, we must ask the estimator to return the smallest dispersion of the estimation. If we could observe the hypothetical probability distribution function of the estimated value, we would check the width of the curve (Wackernagel 2003; Raspa 2004). If the estimator were unbiased, such curve would be centred on zero and it would have the smallest width. Unfortunately we cannot compute the *pdf* of the estimate, but we can calculate its variance:

$$\sigma^2{}_s = E\left\{ \left[Z(x_0) - \sum_{i=1}^{n} \lambda_i Z(x_i) \right]^2 \right\}$$

expanding the square and dividing the *i* index in *a* and *b* indices:

$$= E\left\{ Z^2(x_0) - \sum_a \sum_b \lambda_a \lambda_b Z(x_a) Z(x_b) - 2\sum_a \lambda_a Z(x_a) Z(x_0) \right\}$$

$$= Cov(0) + \sum_a \sum_b \lambda_a \lambda_b Cov(x_a - x_b) - 2\sum_a \lambda_a Cov(x_a - x_0);$$

that is $\sigma^2{}_s = \sigma^2 + \sum_a \sum_b \lambda_a \lambda_b Cov_{a,b} - 2\sum_a \lambda_a Cov_{a,0}$.

Such estimation variance must be minimized in order to assure the optimization of estimation process. We have an expression of estimation variance based on sample variance and covariance of known and target points. The minimization of such expression depends only on the weights of linear estimator.

4.2 Minimization of error variance

In order to minimize the estimation error variance we should set its derivative to zero. Such derivative would be computed with respect to the *n* weights, thus obtaining *n* equations for *n* unknowns. Nevertheless, taking into account the condition for the unbiasedness of the estimator $\sum_{i=1}^{n} \lambda_i = 1$, we have *n+1* equations for only *n* unknowns.

The Lagrange parameter

In order to introduce the $n+1^{th}$ variable we can invoke the so called *Lagrange parameter*:

$$2\mu\left(\sum_{i=1}^{n} \lambda_i - 1 \right).$$

Such new equation must not affect the equality:

$$\sigma^2{}_s = \sigma^2 + \sum_a \sum_b \lambda_a \lambda_b Cov_{a,b} - 2\sum_a \lambda_a Cov_{a,0} + 2\mu\left(\sum_a \lambda_1 - 1 \right).$$

In fact, if we apply the condition of unbiasedness to such term:

$\sum_{i=1}^{n} \lambda_i = 1$, that is $\sum_{i=1}^{n} \lambda_i - 1 = 0$.

Thus $2\mu\left(\sum_{i=1}^{n} \lambda_i - 1 \right) = 0$.

In this way we obtain a set of *n+1* equations with *n + Lagrange parameter* unknowns.

Minimization of the estimation error variance

The introduction of Lagrange parameter as additional equation allows us to express the unbiasedness condition directly trough the minimization derivative. In fact, setting the derivative of the variance to zero with respect to μ, we obtain exactly the unbiasedness condition and we do not need it anymore (Isaaks and Srivastava 1989):

$$\frac{\partial\left(\sigma^2{}_s\right)}{\partial\mu} = \frac{\partial\left(2\mu\left(\sum_a \lambda_a - 1 \right) \right)}{\partial\mu} = 2\sum_a \lambda_a - 2 .$$

Setting it to zero: $2\sum_a \lambda_a - 2 = 0$, we obtain exactly the unbiasedness condition:

$\sum_{i=1}^{n} \lambda_i = 1$.

Since the unbiasedness condition is included in minimization equations, the minimization procedure leads to find the weights that minimize the estimation error variance and contemporarily respect the unbiasedness condition (Isaaks and Srivastava 1989). The estimator that follows this kind of condition is the *best unbiased*.

Going on with the other derivative, we can make an example with the weight λ_1:

$$\frac{\partial\left(\sigma^2{}_s\right)}{\partial\lambda_1}=\frac{\partial\left\{\sigma^2+\sum_a\sum_b\lambda_a\lambda_b Cov_{a,b}-2\sum_a\lambda_a Cov_{a,0}+2\mu\left(\sum_a\lambda_1-1\right)\right\}}{\partial\lambda_1}=0,$$

spreading the equation in four terms:

$$\frac{\partial\sigma^2{}_s}{\partial\lambda_1}=\frac{\partial\sigma^2}{\partial\lambda_1}+\frac{\partial\left(\sum_a\sum_b\lambda_a\lambda_b Cov_{a,b}\right)}{\partial\lambda_1}-2\frac{\partial\left(\sum_a\lambda_a Cov_{a,0}\right)}{\partial\lambda_1}+2\frac{\partial\left[\mu\left(\sum_a\lambda_1-1\right)\right]}{\partial\lambda_1}=0$$

where:

- the first term does not depend on λ_1, that is $\frac{\partial\sigma^2}{\lambda_1}=0$;

- the second term can be expanded separating the contribute for λ_1 from the other ones:

$$\frac{\partial\left(\sum_a\sum_b\lambda_a\lambda_b Cov_{a,b}\right)}{\partial\lambda_1}=\frac{\partial\left(\lambda_1{}^2 Cov_{1,1}+2\lambda_1\sum_{b=2}^{n}\lambda_b Cov_{1,b}\right)}{\partial\lambda_1}=$$

$$=2\lambda_1 Cov_{1,1}+2\sum_{b=2}^{n}\lambda_b Cov_{1,b}=2\sum_{b=1}^{n}\lambda_b Cov_{1,b}$$

- the third term contains only one term in which λ_1 is involved:

$$\frac{\partial\left(\sum_a\lambda_a Cov_{a,0}\right)}{\partial\lambda_1}=\frac{\partial\left(\lambda_1 Cov_{1,0}\right)}{\partial\lambda_1}=Cov_{1,0};$$

- the fourth term also contains only one term involving λ_1:

$$\frac{\partial\left[\mu\left(\sum_a \lambda_1 - 1\right)\right]}{\partial\lambda_1} = \frac{\partial(\mu\lambda_1)}{\partial\lambda_1} = \mu\,.$$

The equation becomes:

$$\frac{\partial\sigma^2{}_s}{\partial\lambda_1} = 2\sum_{b=1}^{n}\lambda_b Cov_{1,b} - 2Cov_{1,0} + 2\mu = 0\,,\ \text{that is}$$

$$\sum_{b=1}^{n}\lambda_b Cov_{1,b} + \mu = Cov_{1,0}\,.$$

Such equation is the result of the derivative of the estimation error variance to zero with respect to λ_1. We can obtain the same result, implementing the same method with the other weights.

Thus, more in general, the join of unbiasedness condition and minimization of error variance of the linear estimator leads to the global system:

$$\begin{cases}\sum_{b=1}^{n}\lambda_b Cov_{a,b} + \mu = Cov_{a,0} \\ \sum_{a=1}^{n}\lambda_i = 1\end{cases}$$

This is the system often referred as *Kriging system*.

4.3 Kriging

In 50's the south african engineer Daniel G. Krige was the first to apply covariance function analysis to model the spatial continuity in order to quantify the reciprocal autocorrelation of the variable in linear spatial estimation. Gold mines exploitation and petroleum reservoirs design were the first application fields that obtained great profits from such applications. In particular, the dependence of the

estimation on the spatial continuity laws defined in structural analysis gave the chance to calibrate the model in order to honour the measured data, and the possibility of extracting the information about the estimation error variance revealed to be a very useful feature for the intent (Burrough el. source).

The word “kriging” takes origin just from the name of the engineer that developed it.

It consist in a linear estimation of known values, based on the unbiasedness and the greatest precision of the results (Isaaks and Srivastava 1989). Its peculiarity is based on the use of covariance and variogram as instruments for asses such main features. Kriging is regarded as B.L.U.E. (Best Linear Unbiased Estimation), that is, *best*, because its error variance is guaranteed to be the minimum, *unbiased* because the average error is guaranteed to be null, *linear estimation*.

4.3.1 The Ordinary Kriging system

As defined before, the kriging system is:

$$\begin{cases} \sum_{b=1}^{n} \lambda_b Cov_{a,b} + \mu = Cov_{a,0} \\ \sum_{a=1}^{n} \lambda_i = 1 \end{cases}$$

that, written in matrix notation, is:

$$\begin{vmatrix} Cov_{1,1} & \cdots & Cov_{1,n} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ Cov_{n,1} & \cdots & Cov_{n,n} & 1 \\ 1 & \cdots & 1 & 0 \end{vmatrix} \cdot \begin{vmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu \end{vmatrix} = \begin{vmatrix} Cov_{1,0} \\ \vdots \\ Cov_{n,0} \\ 1 \end{vmatrix}.$$

We have seen how the variogram is complementary to variogram function. Thus we can express the kriging system with variogram function.

Taking into account the equation:

$\gamma(h) = Cov(0) - Cov(h)$, that can be expressed as $\gamma_{a,b} = \sigma^2 - Cov_{a,b}$, the system becomes:

$$\begin{cases} \sum_{b=1}^{n} \lambda_b \left(\sigma^2 - \gamma_{a,b}\right) + \mu = \sigma^2 - \gamma_{a,0} \\ \sum_{a=1}^{n} \lambda_i = 1 \end{cases}$$; erasing common terms:

$$\begin{cases} \sum_{b=1}^{n} \lambda_b \gamma_{a,b} - \mu = \gamma_{a,0} \\ \sum_{a=1}^{n} \lambda_i = 1 \end{cases}$$ the classical expression of kriging system explicited with variogram.

4.3.2 Estimation error variance

One of the most important feature of kriging is the possibility to associate to the estimated value, the measure of its statistical reliability (Burrough and McDonnell 1997). Such additive parameter is just the estimation error variance that will be regarded as *kriging variance*. The value of such variance must be as lower as possible as assumed before.

Tacking into account the kriging system, and in particular the first term $\sum_{b=1}^{n} \lambda_b Cov_{a,b} + \mu = Cov_{a,0}$, we can multiply each term by λ_a and obtain:

$$\lambda_a \left(\sum_{b=1}^{n} \lambda_b Cov_{a,b} + \mu \right) = \lambda_a Cov_{a,0}.$$

Expanding it:

$$\sum_{a=1}^{n} \lambda_a \sum_{b=1}^{n} \lambda_b Cov_{a,b} + \sum_{a=1}^{n} \lambda_a \mu = \sum_{a=1}^{n} \lambda_a Cov_{a,0},$$

that is:

$$\sum_{a=1}^{n}\lambda_a\sum_{b=1}^{n}\lambda_b Cov_{a,b}=\sum_{a=1}^{n}\lambda_a Cov_{a,0}-\sum_{a=1}^{n}\lambda_a\mu,$$

and considering the unbiasedness condition $\sum_{a=1}^{n}\lambda_a=1$ the last term is only μ and the equation becomes:

$$\sum_{a=1}^{n}\sum_{b=1}^{n}\lambda_a\lambda_b Cov_{a,b}=\sum_{a=1}^{n}Cov_{a,0}-\mu.$$

Substituting it into the general equation of error variance

$$\sigma^2{}_s=\sigma^2+\sum_a\sum_b\lambda_a\lambda_b Cov_{a,b}-2\sum_a\lambda_a Cov_{a,0},$$

it becomes $\sigma^2{}_s=\sigma^2+\sum_{a=1}^{n}Cov_{a,0}-\mu-2\sum_{a=1}^{n}\lambda_a Cov_{a,0}$,

that is $\sigma^2{}_s=\sigma^2-\left(\sum_{a=1}^{n}\lambda_a Cov_{a,0}+\mu\right)$.

Taking into account again $\gamma_{a,b}=\sigma^2-Cov_{a,b}$, the equation becomes $\sigma^2{}_s=\sigma^2-\left(\sum_{a=1}^{n}\left(\lambda_a\sigma^2-\gamma_{a,0}\right)+\mu\right)$, that is:

$$\sigma^2{}_s=\sum_{a=1}^{n}\lambda_a\gamma_{a,0}+\mu.$$

From this equation is clear that the value of estimation variance depends on the variogram function value. The more is the proximity of the target point from a

known one, the more is the reliability of the estimate. In particular its smallest value depends on the presence of the nugget effect. Excluding the case in which the target point coincides with a known one, for which the variance is exactly zero (the estimated value is exactly the known one), the nugget variance is the lower limit for the variance of estimation.

4.3.3 Main features of kriging

I – *exactness of interpolation* – as said before if the target point coincides with a known one, its estimated value will be exactly the known one. Its variance is zero, because the variogram function is a discontinuous function and it shows zero semivariance value at zero lag. Moreover, if we apply the system in order to estimate the target point, we will obtain zero weights for each sample but for the known target one, and the two values will coincide.

II – *effect of range* – in the linear combination of known samples we affirmed that there is a certain area of self correlation among pairs within which the involvement of information takes sense. Beyond such threshold samples are too far away and their influence is assumed to be negligible. The variogram shape offers a way to quantify such limit. In a classical case of bounded variogram, in fact, the semivariance values tend to the samples variance value (the *a priori* variance) and indicate the lack of self correlation among pairs. It allows the kriging system to associate very low weights to samples at distances greater than range. In this way the system is able to auto-tune itself functionally to range of self correlation of real data (Raspa 2004). Moreover, the range is a precious indication to quantify the neighbourhood of the estimates, in order to reduce the computational complexity of the algorithm.

III – *the screen effect* – the unbiasedness condition does not assure the presence of positive weights. In some cases, when some extreme values appear, they are associated with negative weights. When two or more samples are aligned in one direction, the closest known point is assigned the highest weight and consequently the one behind has a very low one. In some cases such low weights are also negative; such effect is called *screen effect*, because closest points make a screen to shadow the most distant ones. The presence of negative weights can introduce,

sometimes, some negative estimates, that are often not coherent with natural variables. In this cases the user can artificially set such estimates to zero (Isaaks and Srivastava 1989).

IV – *effect of small scale behaviour of variogram* – the shape of variogram near the origin is a very important indication for the understanding of the continuity and regularity of the variable. The more is the steepness of the function the more is the discontinuity and the small scale variability of the variable. To better understand it, we can think that the pure nugget model can be regarded as a common model with a very high discontinuity at zero lag and infinite range. Conversely, when the variable is very regular, the variogram shows a very slow increase with lag and a parabolic behaviour at small lags. Such different cases act on the setting of kriging weights in different ways. The more the variogram is regular, the more the weights decrease slowly with distance and estimates are more similar. Contrarily, if the increase of semivariance values is rapid, the weights decrease quickly and estimates tend to produce the so called bull's eyes effect.

4.3.4 Simple Kriging

In the case the variable is known and second order stationary, the kriging system is submitted to some modification. The mean of the variable is known:

$E[Z(x)] = m$ and it is constant for the whole spatial domain.

The estimation, expressed with respect to the residual $\mathrm{Y}(x) = Z(x) - m$ is defined as:

$\tilde{Z}(x_0) - m(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) - m(x_i)$, that is:

$$\tilde{Z}(x_0) = m + \sum_{i=1}^{n} \lambda_i [Z(x_i) - m].$$

As *m* is the same for the whole spatial domain, $E[Z(x)] \equiv E[\tilde{Z}(x)]$ for any point and the condition of unbiasedness that assumed that $E[Z(x) - \tilde{Z}(x)] = 0$ is useless. Thus we have not any condition on the weights and the kriging system is simply:

$$\sum_{b=1}^{n} \lambda_b Cov_{a,b} = Cov_{a,0},$$

with estimation error variance (kriging variance):

$$\sigma^2{}_s = Cov_0 - \sum_{a=1}^{n} Cov_{a,0}.$$

Note that the system is expressed with respect to covariance function, because the second order stationarity guarantees for its existence.

4.3.5 Kriging with trend

I – Universal Kriging

When the variable cannot be assumed to be stationary, we have to modify the situation in order to lead it to a stationary case (Wackernagel 2003; Raspa 2004). The variogram is no more usable because it is able to filter only a constant mean (in intrinsic hypothesis) and we have to move the approach on the residuals:

$$\mathrm{Y}(x_i) = Z(x_i) - m(x_i).$$

The estimator $\tilde{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i)$ becomes

$$\tilde{Z}(x_0) - m(x_0) = \sum_{i=1}^{n} \lambda_i [Z(x_i) - m(x_i)], \text{ that is:}$$

$$\tilde{Z}(x_0) = \sum_{i=1}^{n} Z(x_i) + m(x_0) - \sum_{i=1}^{n} \lambda_i m(x_i).$$

Assuming $m(x_i) = \sum_l a_l f^l(x_i)$, the general expression for the estimator becomes:

$$\tilde{Z}(x_0) = \sum_{i=1}^{n} Z(x_i) + \sum_l a_l f^l(x_0) - \sum_{i=1}^{n} \lambda_i \sum_l a_l f^l(x_i), \text{ that is:}$$

$$\tilde{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) + \sum_a a_l \left[f^l(x_0) - \sum_{i=1}^{n} \lambda_i f^l(x_i) \right],$$

that coincides with the stationary form:

$$\tilde{Z}(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i), \text{ with the condition } \sum_{i=1}^{n} \lambda_i f^l(x_i) = f^l(x_0).$$

Such condition is the same that assures that the average error $E\{\tilde{Z}(x) - Z(x)\} = 0$.

With such condition and using the residual covariance (because we are not able to compute the direct variogram or covariance from raw data) the kriging (universal) system becomes:

$$\begin{cases} \sum_{b=1}^{n} \lambda_b Cov^R{}_{a,b} + \sum_l \mu_l(x_0) f^l(x_a) = Cov^R{}_{a,0} \\ \sum_{b=1}^{n} \lambda_b f^l(x_b) = f^l(x_0) \end{cases}$$

where Cov^R is the covariance of residuals,

with the minimized estimation variance (Universal Kriging variance):

$$\sigma^2{}_s = Cov^R{}_0 - \sum_{a=1}^{n} \lambda_a Cov^R{}_{a,0} - \sum_l \mu_l(x_0) f^l(x_0).$$

Such approach is called *Universal Kriging*, in which the trend is filtered by a deterministic function and the underlying covariance (or variogram) is computed in

order to implement the system. As seen before, the weights of known samples used for the linear combination are submitted to respect the unbiasedness condition and to minimize the estimation variance, but they have to filter the trend too. With such weights, the linear combination of the original variable coincides with the same linear combination of residuals. That is why they are able to filter the trend (Trevisani 2004).

The more is high the degree of the trend, the more severe are the conditions and less is the importance of randomness of the phenomenon. As seen before, the detrending approach is based on a polynomial fit of the surface and the underlying covariance is computed on the residuals. Such method is, thus, extremely sensible to the choice of the shape of the trend, that is rarely justified by a real knowledge of the physical meaning of the phenomenon. More often the fit is a deterministic result and the reliability of such results is often doubtful.

II – Kriging in Irf-k

The alternative method to Universal Kriging is the kriging estimation applied to the Irf-k technique. As seen before, this method is a generalization of the intrinsic stationarity approach in which the zero order increment (the simple increment) of the variable is able to filter out a constant trend. In this way, using the variogram function as the main instrument for the structural analysis, we can be sure to be able to model second order (zero mean), intrinsic (constant mean) or quasi stationary (intrinsic within a certain sub-domain) variables. Generalizing the concept of the intrinsic stationarity, the higher order increment is able to filter out higher order trend (Buttafuoco and Castrignanò 2005).

The further feature is the way to model the spatial continuity, for which a simplified form of covariance is used. Such function, the *generalized covariance*, is regarded as the sum of polynomials, plus some spline components.

Summing up, the Irf-k is a function for which a linear combination of increments exists, it is second order stationary, it is able to filter out higher order trend, and its spatial continuity is modelled by the generalized covariance (Trevisani 2004).

The kriging (Irf-k) system is the same of universal kriging, where covariance of residuals is substituted by generalized covariance:

$$\begin{cases} \sum_{b=1}^{n} \lambda_b K_{a,b} + \sum_{l} \mu_l(x_0) f^l(x_a) = K_{a,0} \\ \sum_{b=1}^{n} \lambda_b f^l(x_b) = f^l(x_0) \end{cases}$$, where K is the generalized covariance,

with the minimized estimation variance (Irf-k Kriging variance):

$$\sigma^2{}_s = K_0 - \sum_{a=1}^{n} \lambda_a K_{a,0} - \sum_{l} \mu_l(x_0) f^l(x_0).$$

4.3.6 Kriging of spatial components

When the structural analysis of a certain variable returns a nested variogram with some different spatial components, we can be interested to estimate the unknown value of the variable regarding only one of such components (Bleines, Deraisme et al. 2004). For example, a nested variogram with two spherical models with two different ranges can be at the base of a multi-source phenomenon. Thus, the measured value can be the results of two different processes acting at two different scales and we could be interested to understand how one of the component had influenced the spatial distribution of the variable.

The estimator takes the form:

$\tilde{Z}(x_0) = \sum_{a=1}^{n} \lambda_a^u Z(x_a)$, where the apex *u* represents the u_{th} spatial component

and the kriging (of the component) system is the same of ordinary one:

$$\begin{cases} \sum_{b=1}^{n} \lambda_b^u \gamma_{a,b} - \mu^u = \gamma_{a,0}^u \\ \sum_{a=1}^{n} \lambda_a^u = 1 \end{cases}$$, where γ^u represents the u_{th} model component.

The sum of the estimated component with kriging of the component must be exactly equal to the estimated global value (with the global nested variogram) with ordinary kriging (Wackernagel 2003; Raspa 2004).

4.4 Cross validation

One of the main target of the geostatistician is to check the reliability of the estimation, to compare the results of different methods application and finally to evaluate the effects of setting the main parameters of each procedure.

One of the most used method to implement such tests is the so called *cross validation*. Known points represent the reference information and in general a comparison between they and their estimations is the approach used to verify the affordability of the method. Nevertheless, if kriging is asked to estimate a target point in a position coinciding with a known one, it returns exactly the same value.

Thus, in order to implement a concrete comparison, the *leaving one out* methodology suggests to remove one known point at once and to ask the method to estimate the value in the position of such removed point, using all the other informations (Isaaks and Srivastava 1989). Iterating the procedure for each known point, we can obtain two vectors of estimated and known information, that can be statistically compared with traditional techniques.

One of the most common one is the classical scatter plot between $Z(x)$ and $\tilde{Z}(x)$, in which the more the points cloud is stretched on a line, the higher is the correspondence between the two variables.

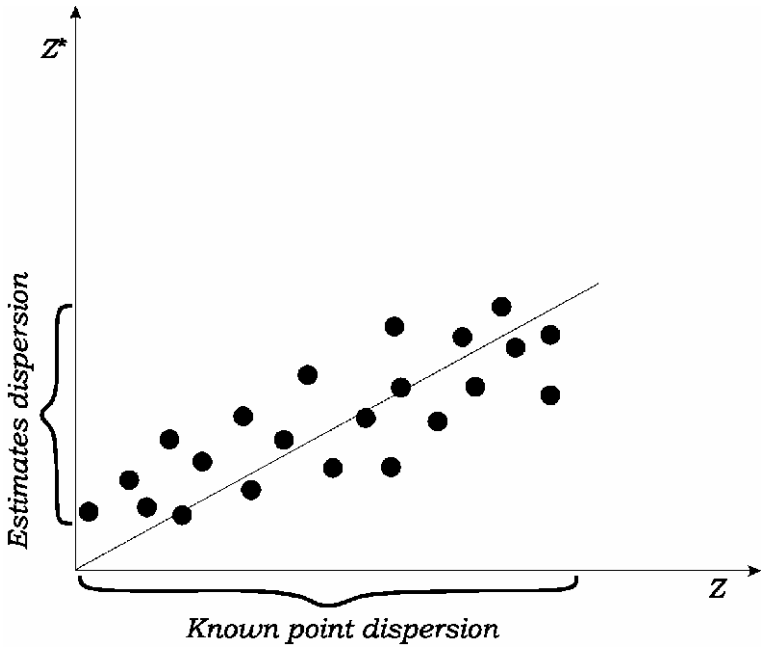


Figure 1.18 – The scatter plot of known points values and their estimates.
Figura 1.18 – Lo scatter plot dei punti noti e delle loro stime.

In general, the dispersion of estimated values is lower than the real ones, because their variance is minimized by the kriging procedure.

The histograms of the standardized error $\frac{Z(x)-\tilde{Z}(x)}{\sigma}$ is a very important tool to evaluate the unbiasedness of the estimation procedure and to quantify graphically the dispersion (the variance) of estimates.

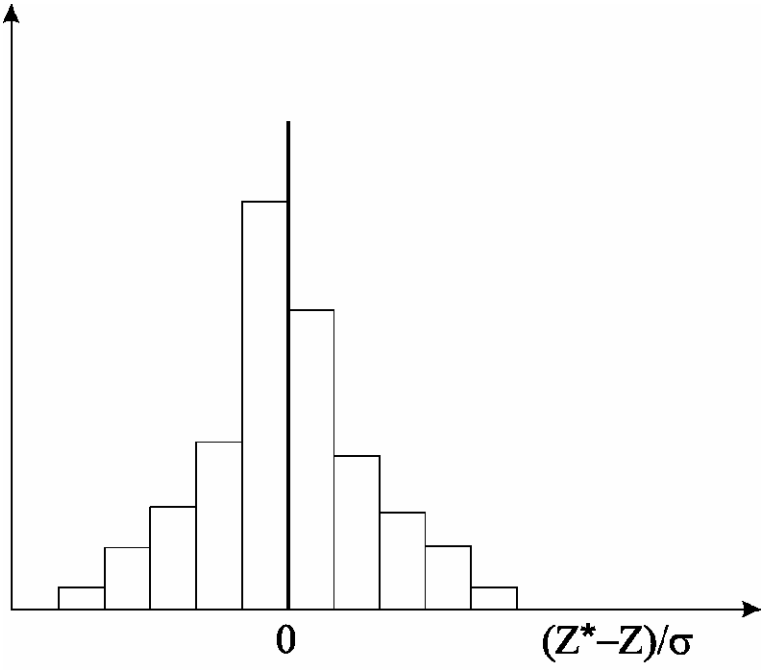


Figure 1.19 – Histogram of estimates.
Figura 1.19 – Istogramma delle stime.

Moreover, such representation is useful to focus on the outliers. The scatter plot between the estimates and their standardized errors is a powerful tool to locate under- and over-estimates. The reference line (zero error variance) is the border between these two kinds of errors.

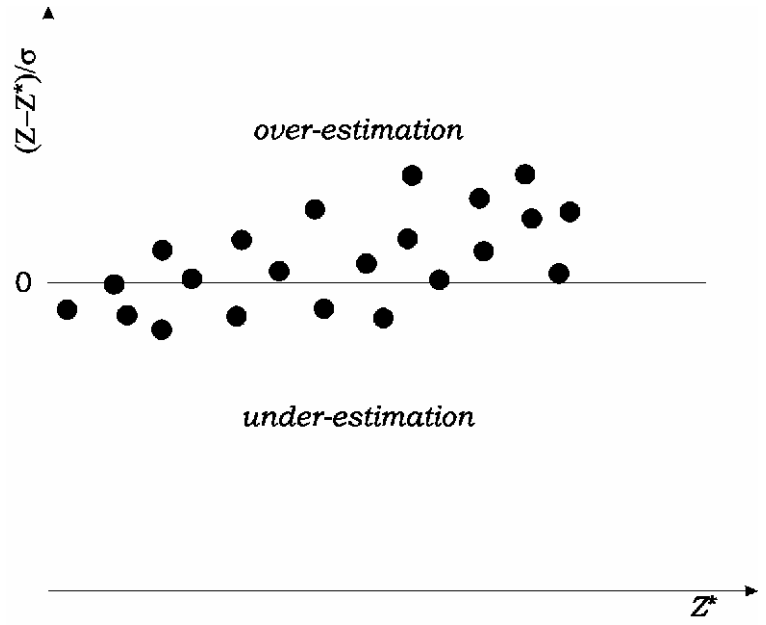


Figure 1.20 – The scatter plot of estimates and their standardized errors.
Figura 1.20 – Lo scatter plot delle stime e dei loro errori standardizzati.

The last parameter is very useful also to quantify univocally the reliability of the whole estimation procedure. The square mean of such standardized error $E\left\{\left[\frac{\tilde{Z}(x)-Z(x)}{\sigma}\right]^2\right\}$ should be one, for a perfect estimate. Such parameter is often used to compare results obtained with different set of the main factors in the estimation method definition, in order to check for the best combination of variogram model and neighbourhood size. Moreover it is useful for different method comparison.

5. Geostatistical simulations

As we have seen in previous sections, kriging algorithm produces the most probable estimation of a certain variable on a certain grid. Such estimation is unique because it follows some conditions of unbiasedness and minimization of error variance, defined in the procedure (Isaaks and Srivastava 1989; Wackernagel 2003; Bleines, Deraisme et al. 2004). Because of the structure of the algorithm, kriging estimations are smoothed, realistic and precise representations of the spatial distribution of a certain variable, following the spatial law estimated just by real data (Journel and Huijbregts 2004). In most cases kriging results are more affordable than other interpolation methods because the algorithm is able to mould itself agreeing with the variability structure of the field.

For many modelling application, such as flow models and quantitative manipulations of environmental variables, such unique representations of the field are not useful because they do not provide a quantitative measure of the variability of the system. Very often, in mathematical models applications, iterative versions of the random function distribution are very useful and geostatistical simulations provide it. Such methodologies return several reproductions of the function, such that they share the same histogram and variogram of the original raw data. The simulated realizations of the random function are know everywhere and are usually iteratively computed, for hundreds times. Such set of outputs can be exploited to compute the spatial variability of the system. One of the most widely used simulation method is the Sequential Gaussian Simulation (SGS), that is based on

the sequential simulation, the iterative simulation of each grid node, of the gaussian transformed variable.

A summary description of the method is supplied here, while, for more details it is reminded to cited literature (Bleines, Deraisme et al. 2004)

After a gaussian transformation of raw data (anamorphosis), finalized to give a gaussian distribution to raw data, we can apply simple kriging to them (the mean value is known and it is zero) in order to compute the value and the variance of each grid node.

The simulated value is:

$$Z_S = Z_{SK} + \sigma_{SK} U ,$$

where Z_{SK} and σ_{SK} are the values computed with simple kriging and U is a random normal function. Such procedure is usually repeated iteratively.

Assessing spatial uncertainty

Geostatistical simulations are conveniently used to compute spatial uncertainty. Having hundreds of realizations of the original random function, that share the same spatial law, we can combine them to compute some spatial statistics of the variability of the system. Local variability can be measured summarizing the n equally probable values of each grid node into some uncertainty indices. Such as the interquartile range, or the standard deviation of each grid node, can be used as measures of local dispersion. Such technique will be used to compute the local variability in order to locate the most sensible points for the infilling sampling plan.

CHAPTER 2
GEOCHEMISTRY OF SOUTHERN CAMPANIAN CONTINENTAL SHELF

In autumn 2004, the National Coastal Marine Environment Research Institute (I.A.M.C. Istituto per l'Ambiente Marino Costiero – C.N.R.) planned an oceanographic survey, focused on the geomorphologic research of the Campanian continental shelf. In this occasion, the Geochemistry Laboratory of the Institute, proposed to take part of the project in order to plan an intensive geochemical investigation of the area. The original sampling plan, focused on superficial sampling of the 0-200 mt. depth area, has been enriched by about 150 additional samples expressly devoted to geochemical analysis. The practical complexity of the survey, the economical affords and meteorological adversities caused a reduction of the planned amount of samples and the final set consisted on 104 samples.

The 104 geochemical samples have been analyzed for trace metals, grain size and Total Organic Carbon content and the results have been processed with a geostatistical approach, in order to estimate the spatial distribution of the variables and eventually make any assumption about the natural and anthropic origins of the different trace metals.

Structural analyses results, combined with them of classical statistics, allowed to classify the variables, dividing them in some group. A quasi stationary outline has been identified for almost all the variables (Myers 1989; Petitgas 1997). For the first group a well designed nested spatial structure has been recognized (Raty and Gilbert 1998). The two spatial components have been evaluated by a nested variogram and then used to implement kriging of spatial components (Morris 1999). Such methodology allowed to identify two different spatial scales distributions of the variables, one with high frequency rapid oscillations and one with some smoothed low frequency variations (Facchinelli, Sacchi et al. 2001; Franklin and Mills 2003). The second group showed mono structural variogram and have been treated with classical ordinary kriging, while log-normal kriging has been implemented on the last group in order to manage variables presenting high skewed distributions.

1. The sampling plan

As said before, the cruise was focused on the geological analysis of the area and the geochemical exploration represented only an infilling of the defined sampling plan. The first one was defined following a systematic strategy in which a regular grid was built and each node represented a potential sample. Later, the same resolution has been kept, but with samples located only on the bathymetric lines. In order to reduce the redundancy of information and to avoid the aliasing of structural analysis (see chapter 3) it has been decided to infill the existent plan with a stratified random one on a denser grid. Such infilling has been computed in three different areas, that are widely known as risk ones:

- *Sele* river estuary;
- Harbour of *Agropoli* (SA);
- City of *Sapri* (SA) – Golfo di Policastro.

The original grid resolution has been reduced to 700-800 mt. and 100-1500 mt., respectively for the area between 10-50 mt. depth and 50-200 mt. depth. Each cell of such grid has been sampled randomly.

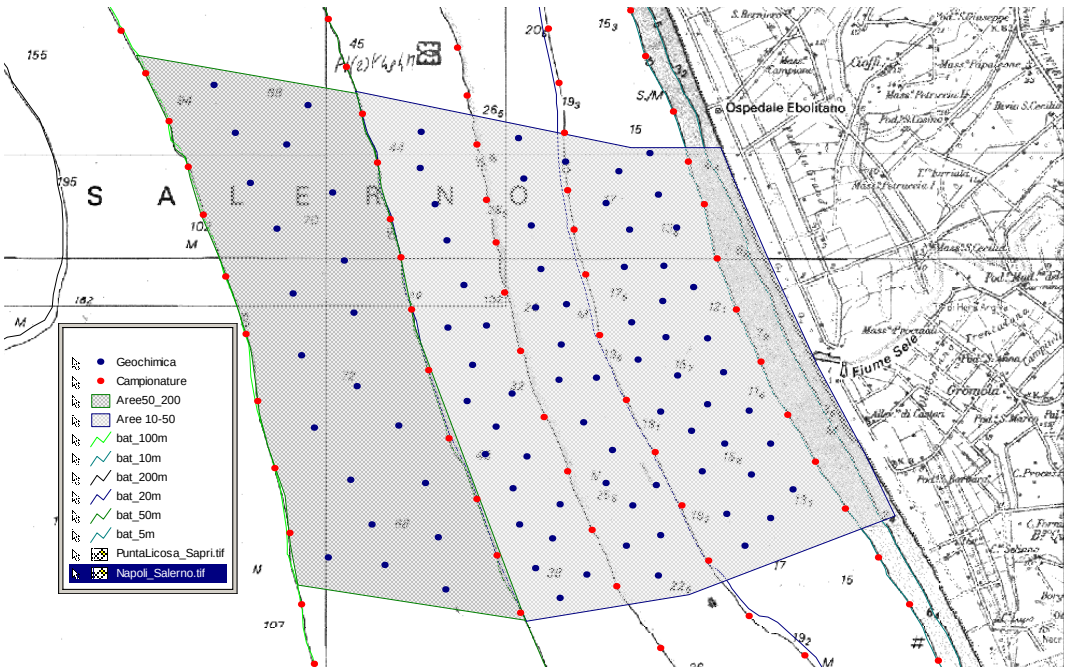


Figure 2.1 – Sampling plan of the area out of the estuary of Sele river.
Figura 2.1 – Piano di campionamento dell'area al largo di Foce Sele.

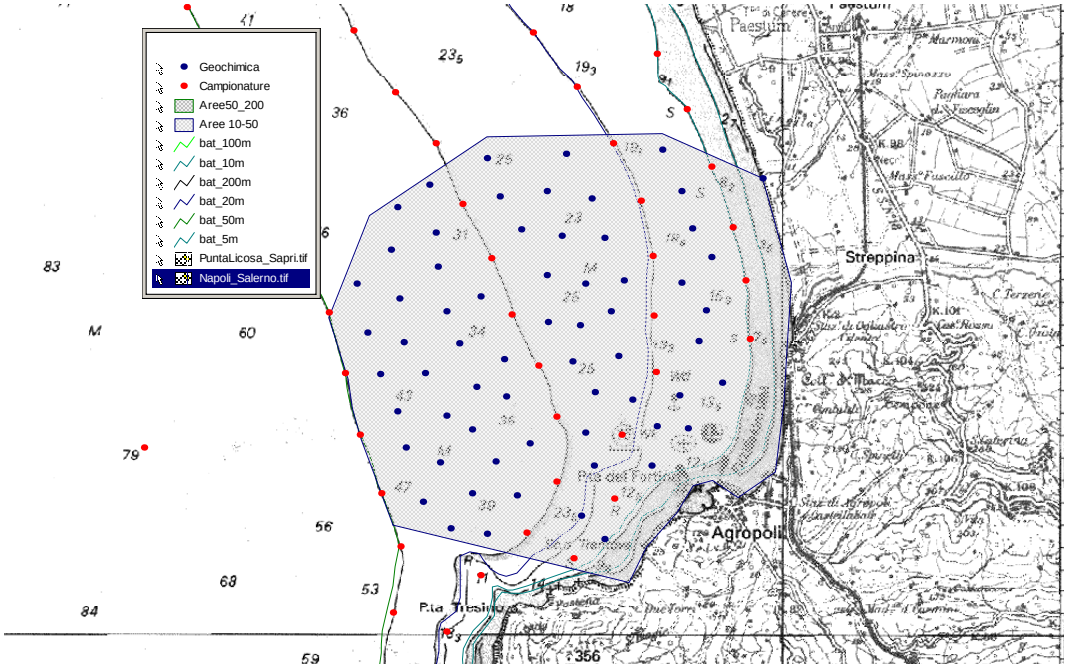


Figure 2.2 – Sampling plan of the area out of the bay of Agropoli.
Figura 2.2 – Piano di campionamento dell'area al largo della baia di Acropoli.

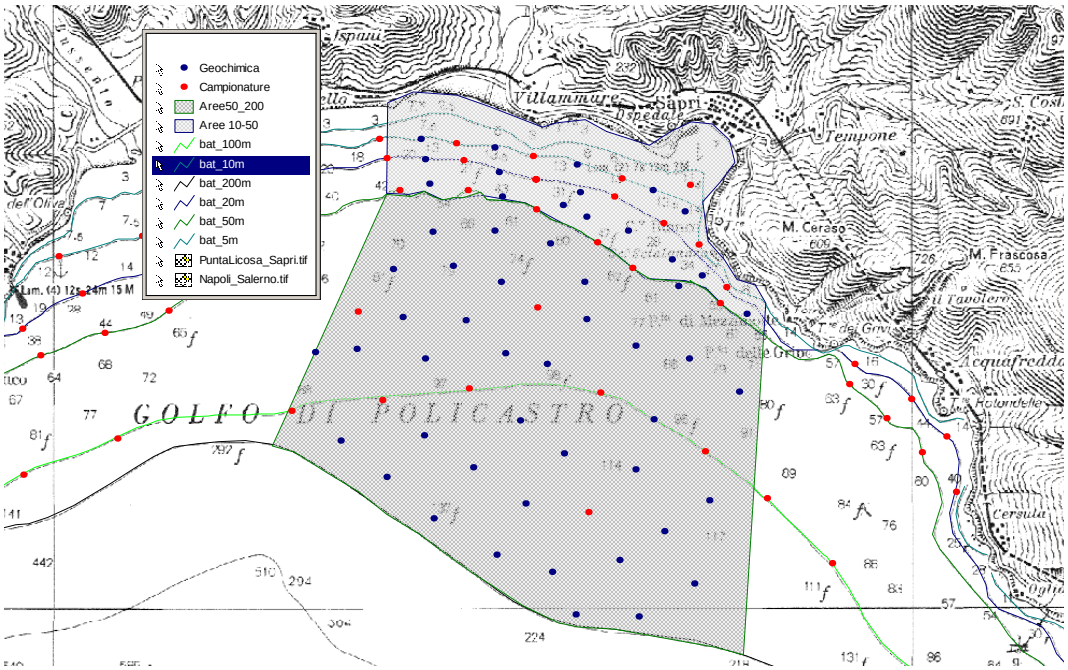


Figure 2.3 – Sampling plan of the area of Gulf of Policastro.
Figura 2.3 – Piano di campionamento dell'area del Golfo di Policastro.

The total amount of potential samples has been defined to about 600 units. Such prospective was obviously optimistic, because of economic and temporal practical limitations.

At the end of the cruise, the amount of sample for geochemical analyses was 104 units, widely distributed along the southern coastal area (within 200 mt. depth).

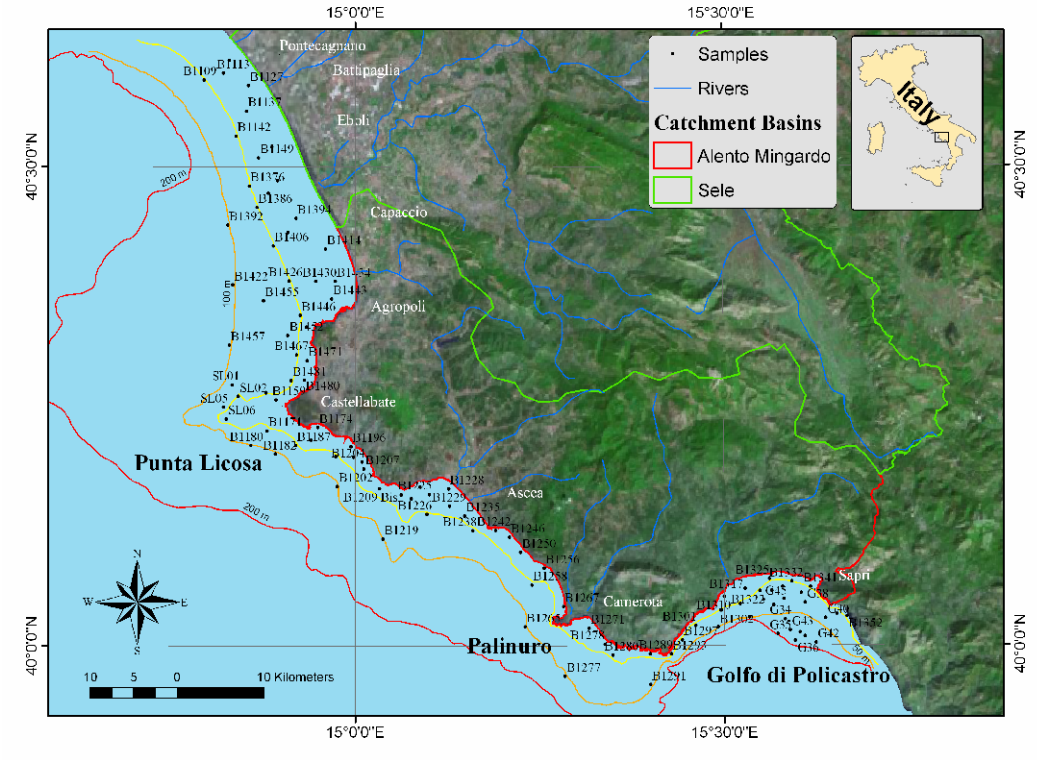


Figure 2.4 – The final sampling plan of the entire area.
Figura 2.4 – Il piano di campionamento definitivo dell'intera area.

Most of samples were located within the 100 mt. bathymetry, while only the ones around the deeper Golfo di Policastro overcame that quote, until 200 mt depth. The involved catchments basins are the Alento Mingardo river and Sele river.

2. The dataset

Geological description, together with sedimentological examination, has been executed onboard on the whole dataset, before to store them for the laboratory analyses. The spatial distribution of descriptive grain size values is conveniently correlated with bathymetry, where silt/clay samples are located mainly in the deepest Golfo di Policastro area and beyond the estuary of Sele river and coarser sand, with some zones with the presence of *Posidonia Oceanica*, are widely

distributed along the shoreline, concentrated in Punta Licosa area, where bottom is shallower (Russo 1990; Ferraro, Pescatore et al. 1997).

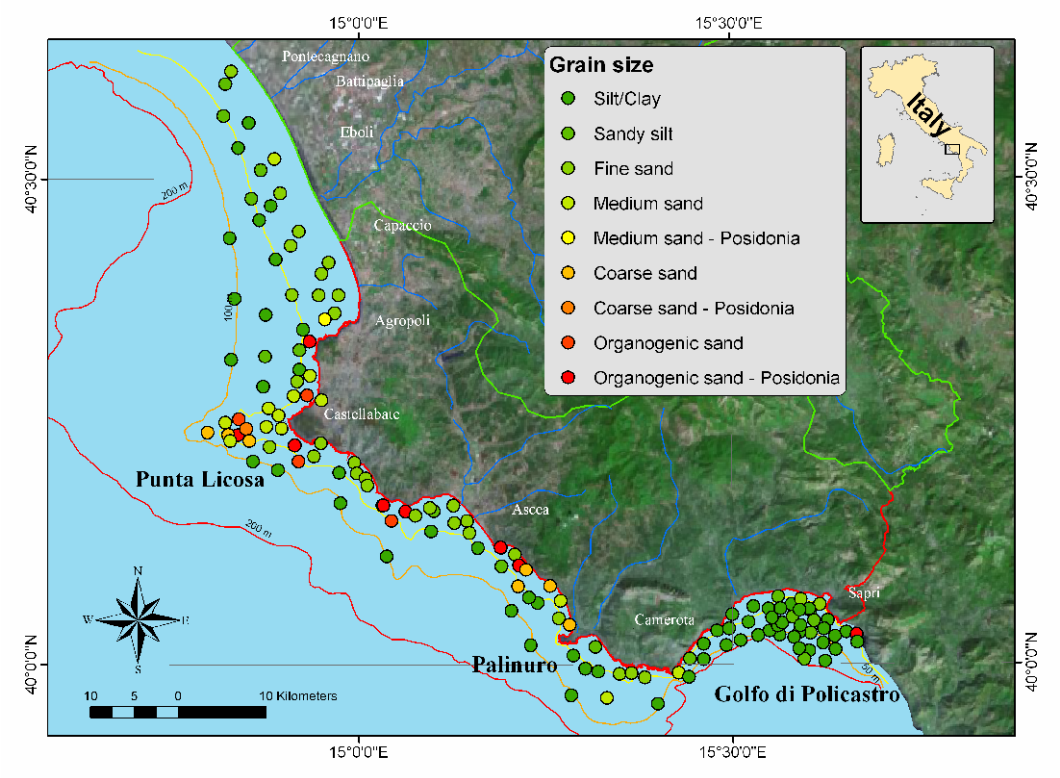


Figure 2.5 – Grain size analyses results on the samples.
Figura 2.5 – I risultati delle analisi granulometriche effettuate sui campioni.

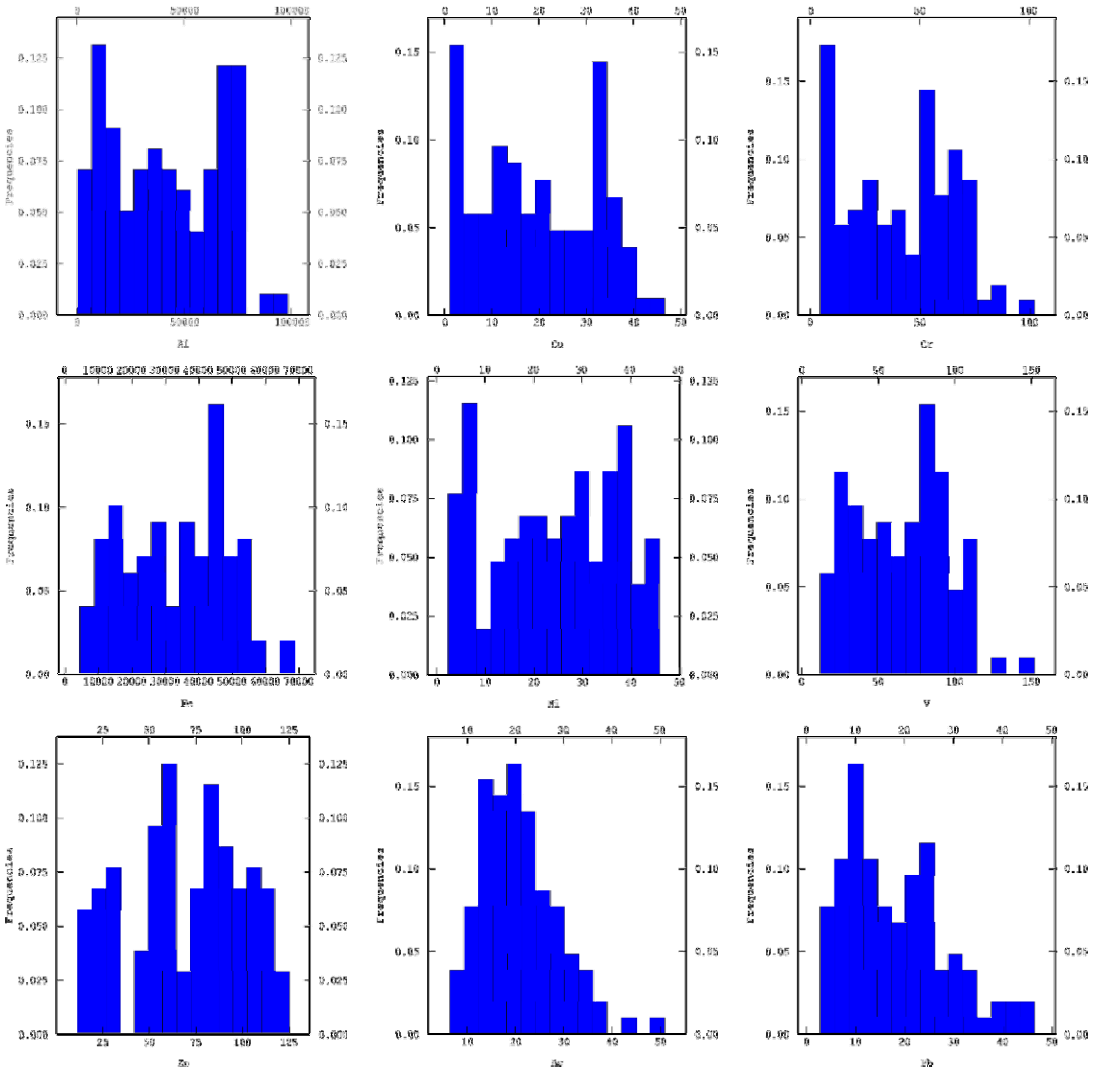
As we will analyze later, such grain size distribution will influence deeply the variables distribution. Eleven trace metals have been analyzed. Main basic statistics are presented in table.

	N	MIN	MAX	MEAN	MEDIAN	VARIANCE	STD DEV	SKEWNESS	KURTOSIS
Al	99	497	98158	41389	40696	6.7e08	26002	0.074	-1.30
Fe	99	4540	68519	33368	35550	2.4e08	15493	0.003	-1.01
Cr	104	4.46	101.76	41.16	42.37	592	24.32	0.061	-1.05
Cu	104	1.15	46.44	19.60	18.81	148	12.18	0.136	-1.24
Ni	104	2.49	45.82	24.05	25.36	166	12.87	-0.133	-1.19
V	104	12.04	151.45	65.24	68.31	900	30.00	0.128	-0.65
Zn	104	11.99	124.86	69.81	73.43	935	30.57	-0.221	-0.93
Pb	104	2.75	46.29	18.00	16.44	110	10.47	0.678	-0.13
Hg	102	0.002	0.35	0.06	0.04	4.0e-03	0.06	2.222	5.67
Cd	104	0.02	2.00	0.15	0.10	4.8e-02	0.22	6.243	50.58

As	104	6.47	50.77	20.85	20.18	63	7.96	0.859	1.26

Table 2.1 – Basic statistics of the analyzed trace metals.
Tabella 2.1 – Statistiche di base dei metalli in tracce analizzati.

Generally, the variables are quite normally distributed, while Arsenic and Pb present a rather skewed distribution. Cadmium and Hg are significantly unbalanced toward low values and they show some high outliers.



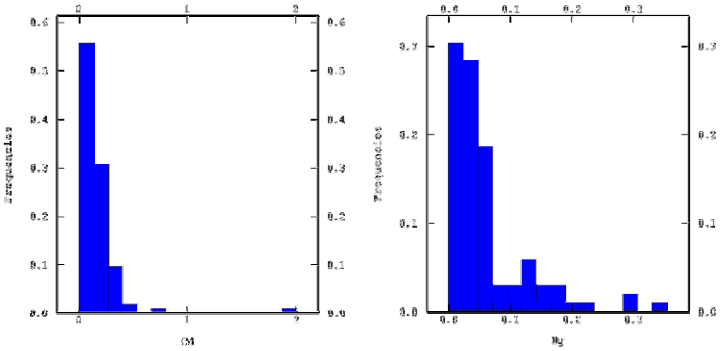


Figure 2.6 – Histogram of the analyzed trace metals.
Figura 2.6 – Istogramma dei metalli in tracce analizzati.

One of the most important aspect of multivariate datasets processing is the quantification of correlation among variables.

	AL	FE	CR	CU	NI	V	ZN	PB	HG	CD	AS
Al	1.00	0.95	0.95	0.92	0.92	0.92	0.91	0.81	0.38	0.21	0.19
Fe		1.00	0.89	0.89	0.92	0.85	0.90	0.76	0.39	0.23	0.18
Cr			1.00	0.96	0.96	0.96	0.95	0.85	0.37	0.19	0.19
Cu				1.00	0.97	0.93	0.96	0.88	0.46	0.16	0.22
Ni					1.00	0.90	0.97	0.81	0.41	0.16	0.14
V						1.00	0.91	0.89	0.32	0.26	0.33
Zn							1.00	0.83	0.39	0.20	0.15
Pb								1.00	0.32	0.32	0.43
Hg									1.00	-0.05	0.10
Cd										1.00	0.06
As											1.00

Table 2.2 – Correlation coefficient matrix of the analyzed trace metals. In red the values beyond 0.80.
Tabella 2.2 – Matrice di correlazione tra I metalli in tracce analizzati. In rosso i valori oltre lo 0.80.

The correlation coefficient table (computed on 99 valid samples) shows a general good correlation among the group Aluminium, Fe, Cr, Cu, Ni, V, Zn, while Lead seems to be slightly less concordant with them. Mercury, Cd and As are evidently uncorrelated. Such behaviour is confirmed by the analysis of distribution, that suggests the same gathering, and will be validated in structural analysis. The scatterplot between Aluminium and Fe, for example, reveals the typical high correlation between such metals, whose behaviour is deeply influenced by grain size distribution.

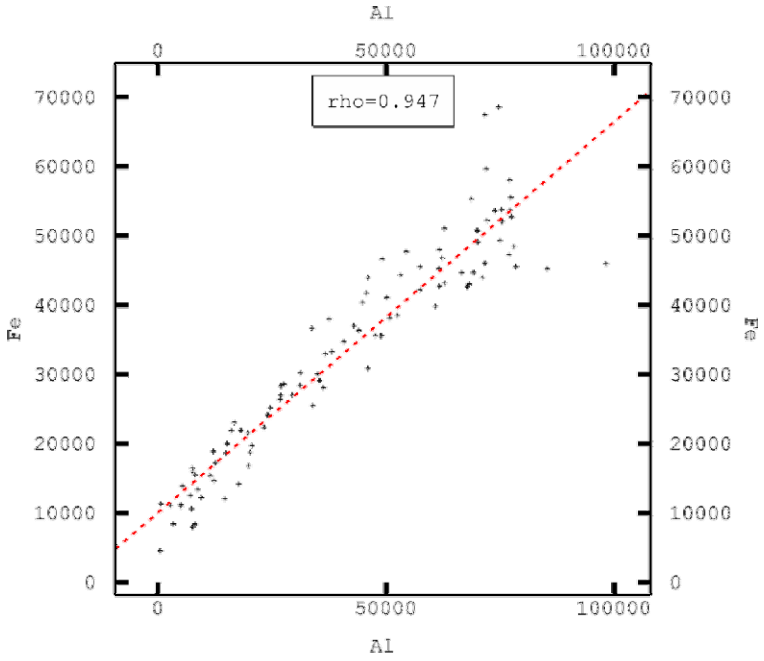


Figure 2.7 – Scatter plot of Aluminium and Iron.
Figura 2.7 – Scatter plot tra Alluminio e Ferro.

Conversely, the scatterplot between Arsenic and most of other variables (Iron, in the example figure) reveals the presence of two subpopulations that each shows a good correlation. Such situation will be analyzed in the structural analysis, when influence on variogram computation will be investigated.

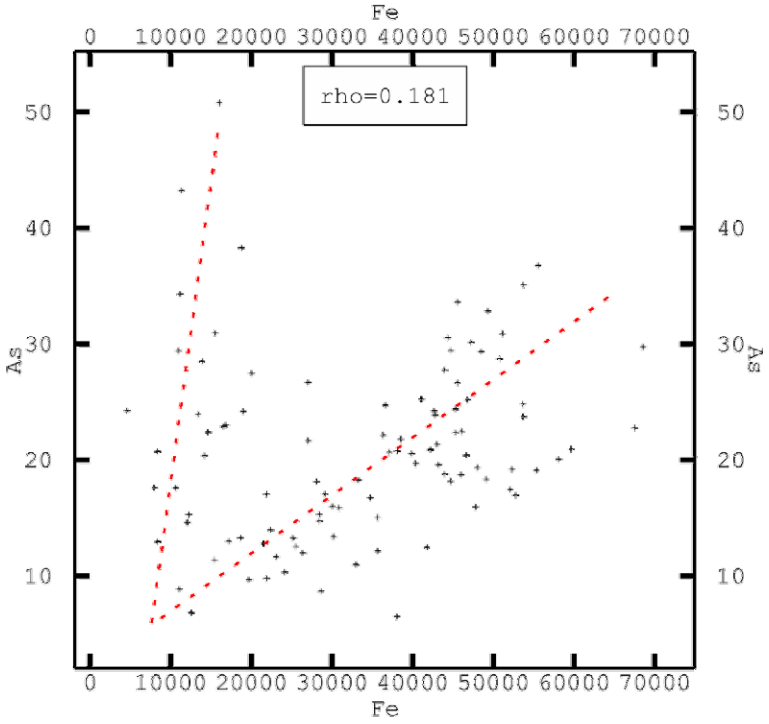


Figure 2.8 – Scatter plot of Arsenic and Iron.
Figura 2.8 – Scatter plot tra Arsenico e Ferro.

Principal component analysis of the multivariate dataset confirms the gathering made on the basis of the classical statistical parameters (Carlon, Critto et al. 2001; Reisa, Sousab et al. 2004).

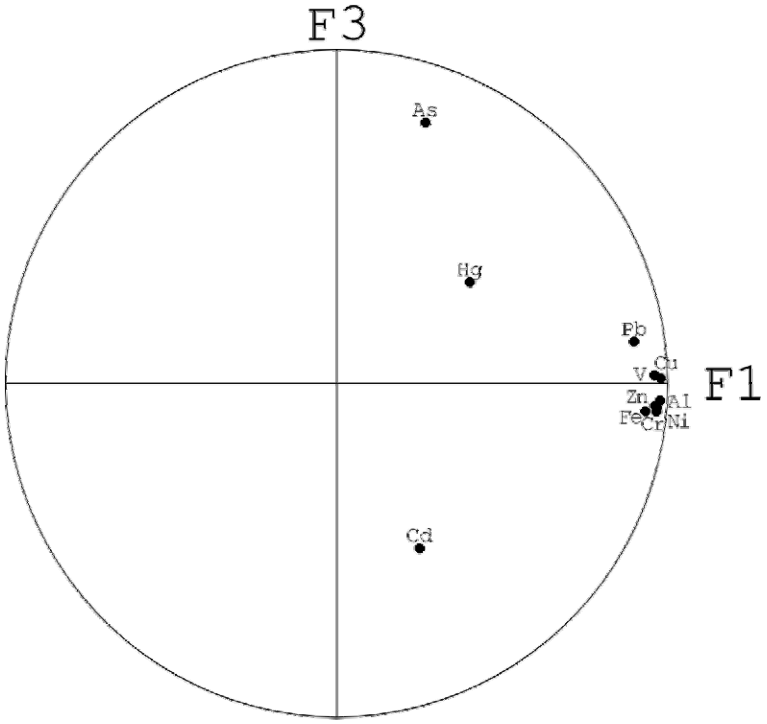


Figure 2.9 – Factors circle plot with factors 1 and 3.
Figura 2.9 – Grafico dei fattori della ACP. Fattori 1 e 3.

PCA correlation circle, between factors one and three, shows a very clear cluster made by Aluminium, Fe, Cr, Cu, Ni, V, Zn, Pb, along factor one, while arsenic, mercury and cadmium are spread along the factor three axis. Such statistical behaviour will be confirmed by structural analysis.

3. Structural analysis

Structural analysis is the crucial aspect of such kind of application and its use will represent a fundamental step to optimize the kriging estimation procedures will be applied on the dataset. A detailed variographic analysis will be computed on data, in order to evaluate the consistence of the spatial structure and the presence of eventual anisotropies and non-normal behaviours of variables (Rosenbaum and Soderstrom 1996).

Influence of lognormality and presence of outliers on variogram computation

As seen before, the whole spatial law of the random function must be simplified and reduced to the definition of the behaviour of the first two moments (mean and variance/covariance). Such assumption is based on the consideration that only a

normal-like variable can be completely described by the first two moments (Bleines, Deraismé et al. 2004). Consequently, a more general geostatistical approach assumes that the variable (and its modelled random function) follows a normal distribution. When it does not, log-normal kriging is used. Such technique is based on a first transformation of the raw variable in its logarithms and a back-transformation of the estimated values into their antilogarithms.

The original variable can be regarded as:

$$Z(x) = e^{Y(x)} - \beta ,$$

where the variable *Y* stands for the logarithm of *Z* and β is the shift that makes *Z* positive. Log-normal kriging is based on the estimation of *Y* and on the final back transformation of *Y* in the original form *Z*, following:

$$\tilde{Z}(x) = e^{\left[\tilde{Y}(x) + \frac{1}{2}\sigma_Y^2 + \mu\right]} - \beta ,$$

where the estimated variable $\tilde{Z}(x)$ derives from the back transformation of $\tilde{Y}(x)$, added with its estimation variance. The use of lognormal kriging, especially in the case of geochemical variables is somewhat tricky and its use must be carried on with care.

When some outliers are suspected to be present, a try-catch strategy is applied. Variogram is iteratively computed in order to evaluate the influence of eventual outlier on the conformity of spatial structure of the random function. If the apparent lack of structure is demonstrated to be due to some outlier they are assumed to be spike and rejected.

3.1 Choice of calculation parameters

The choice of variogram calculation parameters is a very important aspect, because the correct interpretation of the spatial structure of the random function depends on them.

The two-dimensional variogram map represents the semivariance values in a 360° polar diagram, in which each cell value is defined for a certain distance and direction. Such graph is very useful for the definition of lag parameters and the recognition of anisotropies (Deutsch 2002).

The spatial distribution of samples causes an anisotropic availability of pairs.

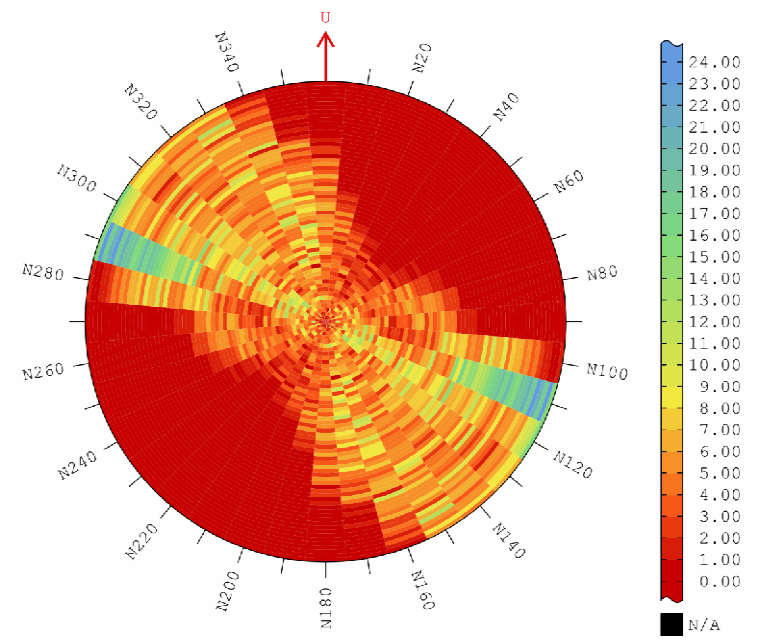


Figure 2.10 – Polar plot of pairs abundance. Lag value = 1000 mt. # of lags = 60.
Figura 2.10 – Grafico polare del numero di coppie. Lag = 1000 m. # di lags = 60.

Number of pairs is widely concentrated along the Northwest - Southeast direction, that is approximately the long-shore direction. Because of the main target of the cruise, in fact, the samples are distributed along the coastline and are concentrated within the band of 10-200 mt. depth. Considering that the estimation procedure must be computed isotropically in all the direction and that a too much big neighbourhood is often unnecessary, we can reduce the computation domain to the minor radius of the ellipse.

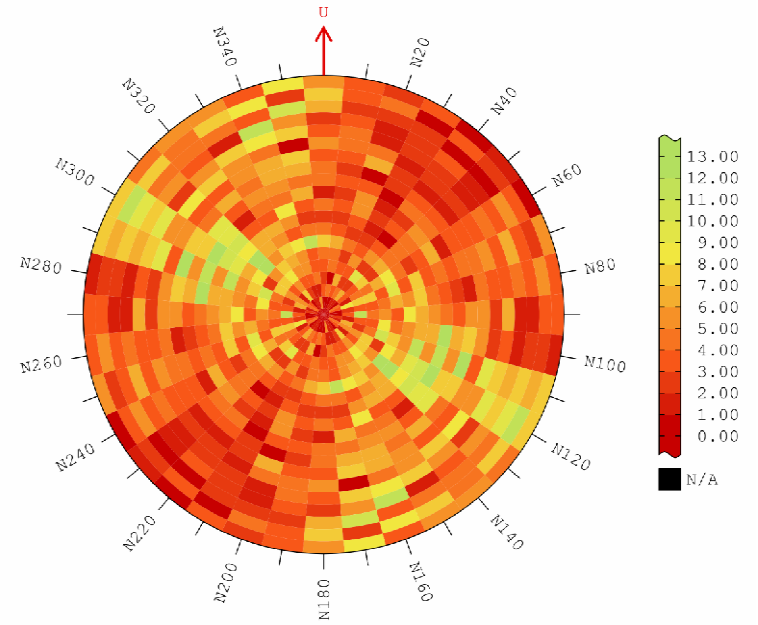


Figure 2.11 – Polar plot of pairs abundance. Lag value = 1000 mt.; # of lags = 20.
Figura 2.11 – Grafico polare del numero di coppie. Lag = 1000 m. # di lags = 20.

Reducing the computation domain to 20 lags (lag = 1000 mt.), the pairs availability decreases but it is suitable also in cross-shore direction.

The quasi-stationarity of data

If we observe the omnidirectional experimental variogram of Zinc, computed over 90 lags, we can note a rapid increase of variability around the 60 Km lag.

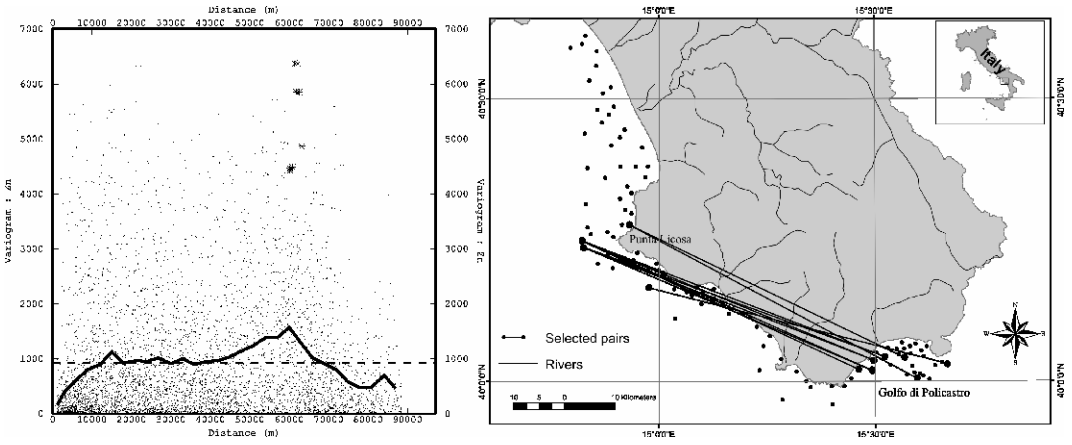


Figure 2.12 – Left: the variogram cloud and its fitted theoretical model. Right: location map of samples. In both the figure, the pairs responsible for the semivariance peak are highlighted.
Figura 2.12 – A sinistra: la nuvola variografica e la sua funzione approssimante. A destra: mappa di ubicazione dei campioni. In entrambe le figure le coppie responsabili

del picco di semivarianza sono evidenziate.

Such peak is related to the pairs belonging to Golfo di Policastro and Punta Licosa areas. It means that the most elevated dissimilarity among pairs is represented by the comparison between such two areas. As seen before, these two zones presents the most pronounced dissimilarity about the grain size, showing a broadly silt/clay sea-bottom in Golfo di Policastro and a coarse sandy one in Punta Licosa. Such dissimilarity is the indicator of a slight non stationary situation, in which the mean value of the variable (Zinc in this case) is varying trough the domain, showing very different averaged values between the two zones. We should chose a non stationary approach, in order to process such variable.

Nevertheless, taking into account the considerations made before, regarding the pairs availability trough all the directions of investigation, we can simply reduce the spatial domain of variogram computation, following the procedure of the so called quasi stationary approach. Moreover, as affirmed before it is unreliable that such a large neighbourhood (with the related variogram range) would be necessary.

In fact, the same experimental variogram, reduced to 30 lags, show a stationary well defined spatial structure.

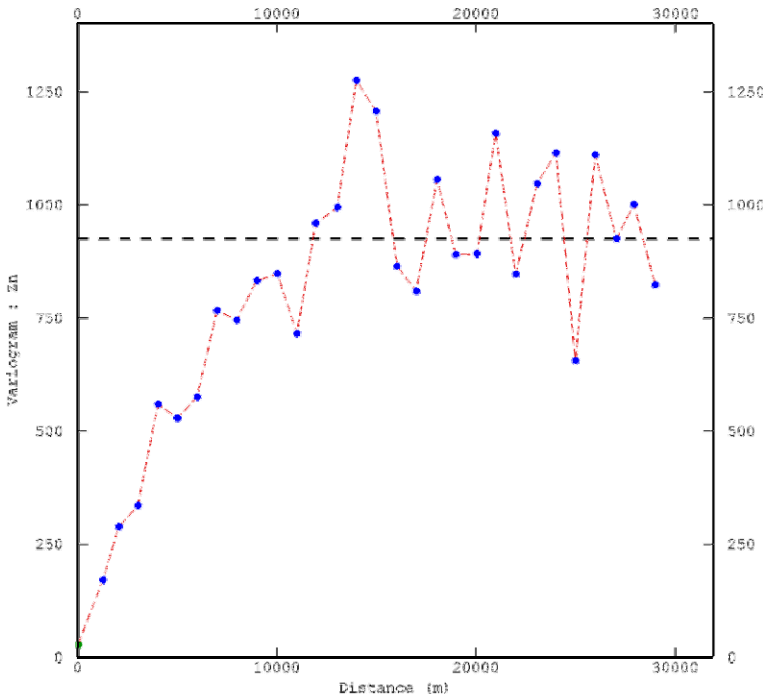


Figure 2.13 – Experimental variogram of Zinc. Lag value = 1000 mt.; # of lags = 30.

Figura 2.13 – Variogramma sperimentale dello Zinco. Lag = 1000 m. # di lags = 30.

In this case, as omnidirectionally computed, the pairs availability allows to extend the upper limit to 30 lags, instead of the 20 defined before. In order to evaluate eventual anisotropic behaviour of the variables, we can observe the variogram map (Copper in the example).

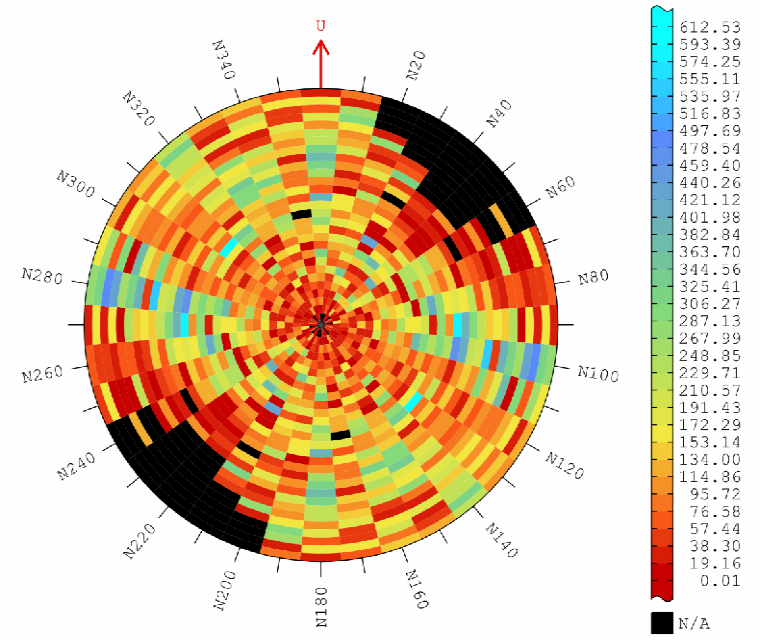


Figure 2.14 – Variogram map of Copper. Lag value = 1000 mt.; # of lags = 30.
Figura 2.14 – Mappa variografica del Rame. Lag = 1000 m. # di lags = 30.

The distribution of cell semivariance values does not reveals any evident anisotropy.

Nested structures

If we observe the experimental variogram of Aluminium, we can note how a fitting spherical model is not able to catch the behaviour of the first lags, while a exponential one is not able to model correctly the right sill.

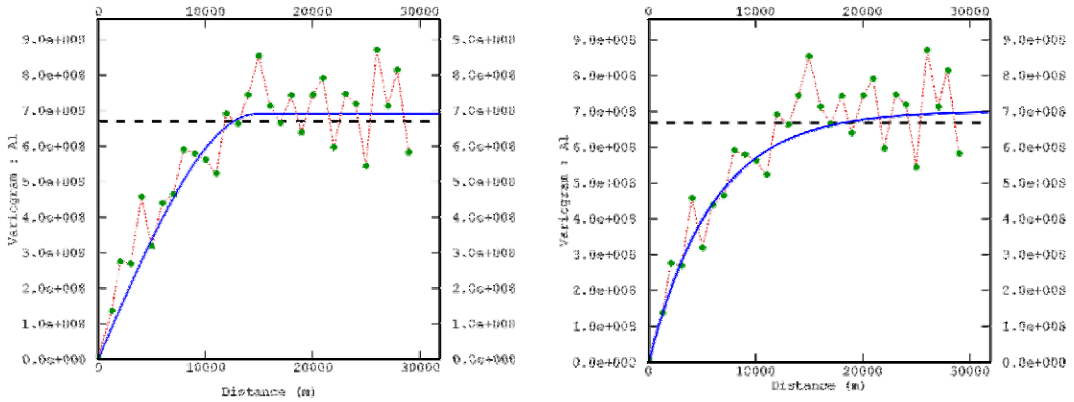


Figure 2.15 – Experimental variogram of Aluminium. Left: modelled with a spherical model (Sill: 6.90e008, range: 14.80 e003). Right: modelled with an exponential model (Sill: 7.01e008, range: 17.95e003).
Figura 2.15 – Variogramma sperimentale dell'Alluminio. A sinistra: approssimato con un modello sferico (Sill: 6.90e008, range: 14.80 e003). A destra: approssimato con un modello esponenziale (Sill: 7.01e008, range: 17.95e003).

The pairs availability amount keeps over 30 units for all the lags, and we can trust on their semivariance values, trying to fit the model at all the scales. One way to try to honour the data is to introduce another structure for the smaller scales.

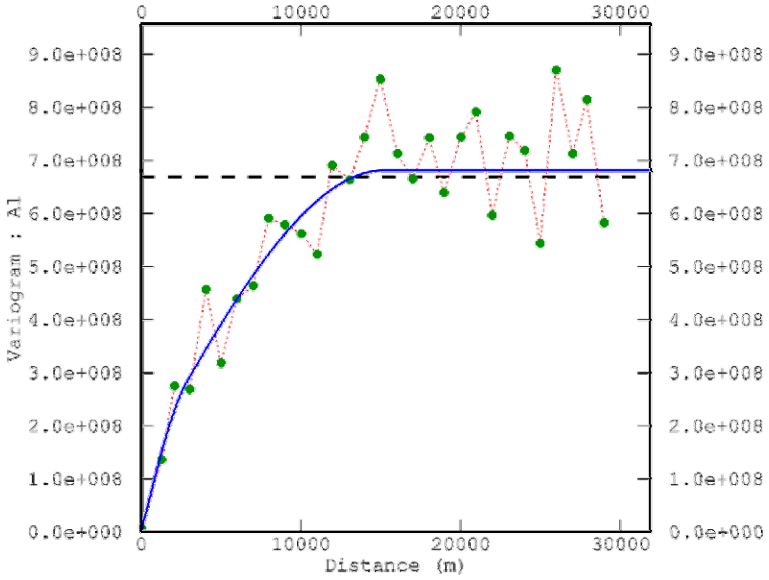


Figure 2.16 – Experimental variogram of Aluminium and its theoretical nested model: 1.29e008 * *sph* (2.79e003) + 5.52e008 * *sph* (15.23e003).
Figura 2.16 – Variogramma sperimentale dell'Alluminio e suo modello approssimante: 1.29e008 * sph (2.79e003) + 5.52e008 * sph (15.23e003).

Such choice leads to the definition of a nested model, where two spherical models are combined to represent the whole variability structure of the random

function. In this way, in fact, both the small scale and the large scale (the sill) of the variogram are correctly modelled. The range of the first component is about one fifth of the large scale one, while the sum of the two sills is nearby the sample variance trough which the values oscillate. As said before, such situation reveals the role of two main mechanism contributing to the spatial distribution of the variable (Bourennane, Salvador-Blanes et al. 2003).

3.2 The trace metals

At this point we can go on analyzing each variable at once, following the calculation parameters, as defined here.

Aluminium

As seen before, Aluminium is one of the variable that shows a nested variogram. Its histogram is sufficiently normal-like.

	N	MIN	MAX	MEAN	MEDIAN	VARIANCE	STD DEV	SKEWNESS	KURTOSIS
AL	99	497	98158	41389	40696	6.7e08	26002	0.074	-1.30

Table 2.3 – Basic statistics for Aluminium.
Tabella 2.3 Statistiche di base per l'Alluminio.

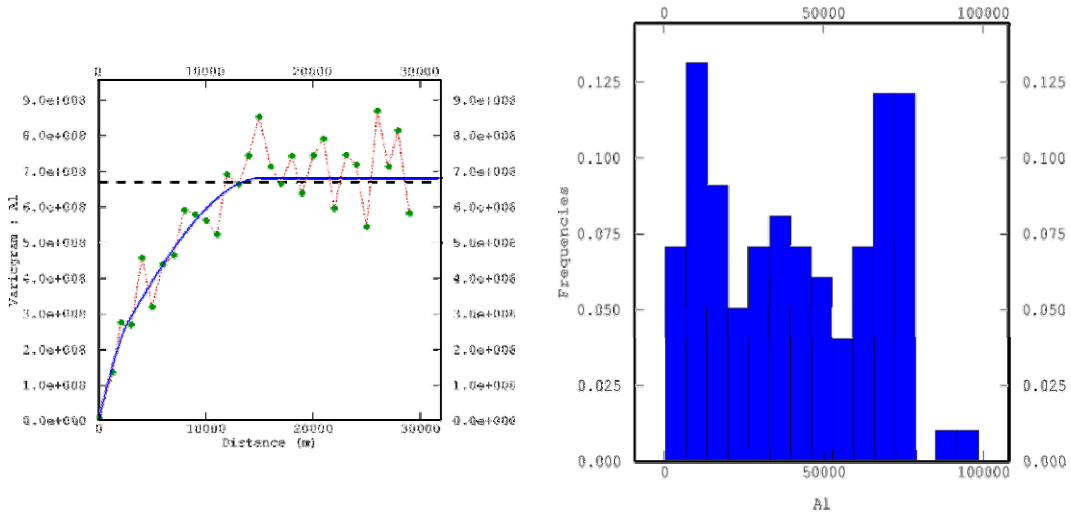


Figure 2.17 – Variogram and histogram of Aluminium.
Figura 2.17 – Variogramma e istogramma dell'Alluminio.

With the theoretical nested variogram model:

$$\gamma(h) = 1.29e008 * sph^{(2.79e003)} + 5.52e008 * sph^{(15.23e003)}.$$

Iron

Variogram of iron presents the same nested structure, but with an even more regular small scale variability. The short range model, in fact, is a gaussian one, that reveals a homogeneous variation of variable values at small scales. Its histogram is close to be normal-like.

	N	MIN	MAX	MEAN	MEDIAN	VARIANCE	STD DEV	SKEWNESS	KURTOSIS
FE	99	4540	68519	33368	35550	2.4e08	15493	0.003	-1.01

Table 2.4 – Basic statistics for Iron.
Tabella 2.4 – Statistiche di base per il Ferro.

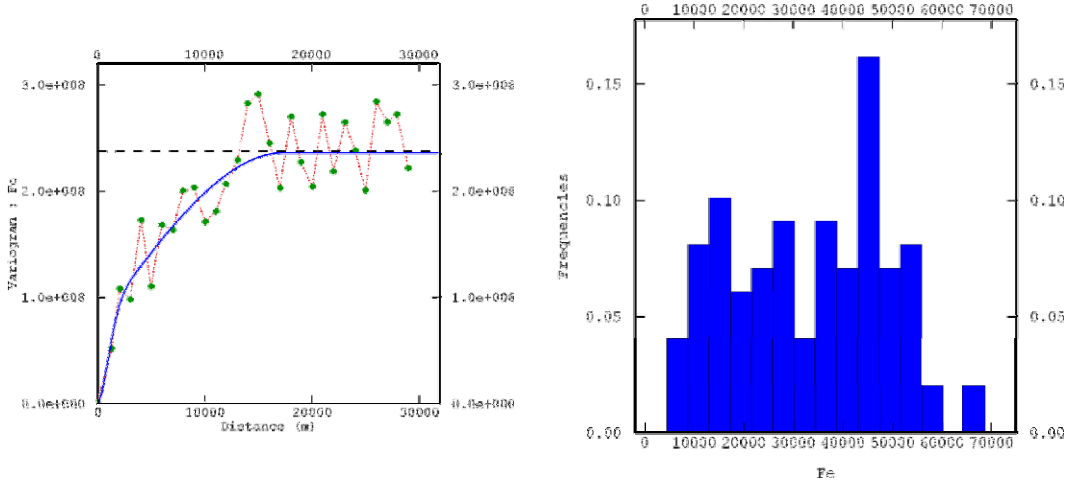


Figure 2.18 – Variogram and histogram for Iron.
Figura 2.18 – Variogramma e istogramma per il Ferro.

With the theoretical nested variogram model:

$$\gamma(h) = 7.40e007 * gaus^{(2.49e003)} + 1.62e008 * sph^{(17.37e003)}.$$

Chromium

Chromium variogram does not show any clear nested structure. The rapid increase of variability at small scales, maybe due to the abundance of very small

values populating the first class of the histogram, does not allow to catch the multiple structure. Small values, especially if related to geochemical variables, are often supposed to be affected by analytic errors, and such uncertainty is confirmed by small scale variogram. Even if it is reliable to have the same nested structure of the other variables, we cannot hazard to model the short range component with only one lag.

	N	MIN	MAX	MEAN	MEDIAN	VARIANCE	STD DEV	SKEWNESS	KURTOSIS
CR	104	4.46	101.76	41.16	42.37	592	24.32	0.061	-1.05

Table 2.5 – Basic statistics for Chromium.
Tabella 2.5– Statistiche di base per il Cromo.

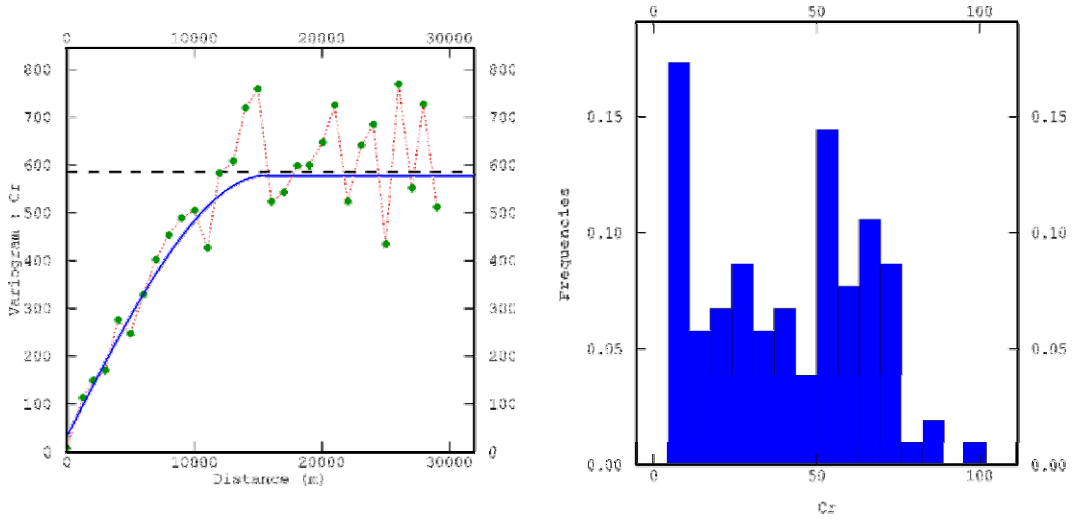


Figure 2.19 – Variogram and histogram for Chromium.
Figura 2.19 – Variogramma e istogramma per il Cromo.

With the theoretical variogram model:

$$\gamma(h) = 30.52 + 547.78 * sph^{(15.67e003)}.$$

Copper

In agreement with Chromium, also Copper does not show a clear nested structure, with some degree of nugget effect and only one spherical model. Such as in Chromium case, Copper histogram presents a large accumulation of very small values.

	N	MIN	MAX	MEAN	MEDIAN	VARIANCE	STD DEV	SKEWNESS	KURTOSIS
CU	104	1.15	46.44	19.60	18.81	148	12.18	0.136	-1.24

Table 2.6 – Basic statistics for Copper.
Tabella 2.6– Statistiche di base per il Rame.

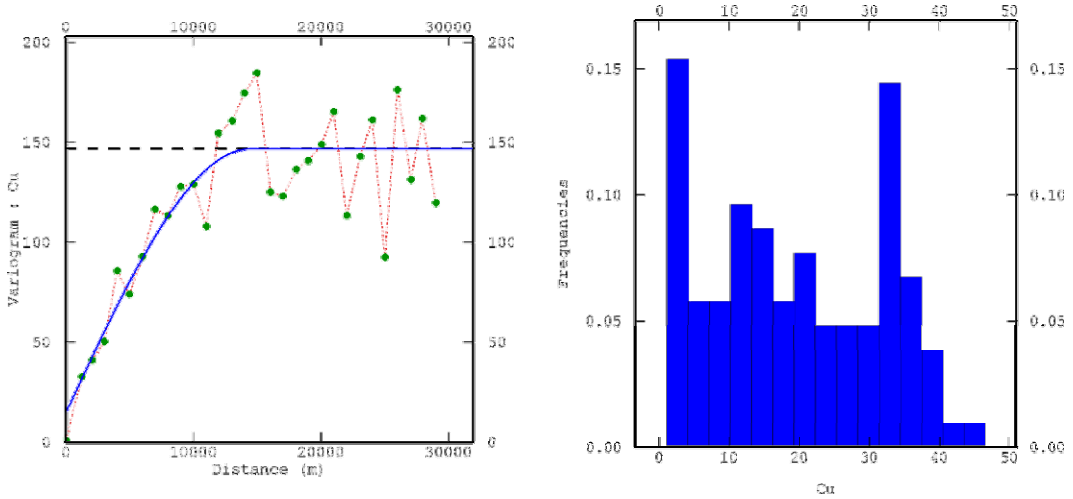


Figure 2.20 – Variogram and histogram for Copper.
Figura 2.20 – Variogramma e istogramma per il Rame.

With the theoretical variogram model:

$$\gamma(h) = 14.38 + 132.21 * sph^{(14.46e003)}.$$

Nickel

As the two previous variables, also Nickel shows some nugget effect with no clear nested structures. Once again, its histogram presents several units in the lowest classes.

	N	MIN	MAX	MEAN	MEDIAN	VARIANCE	STD DEV	SKEWNESS	KURTOSIS
NI	104	2.49	45.82	24.05	25.36	166	12.87	-0.133	-1.19

Table 2.7 – Basic statistics for Nickel.
Tabella 2.7– Statistiche di base per il Nichel.

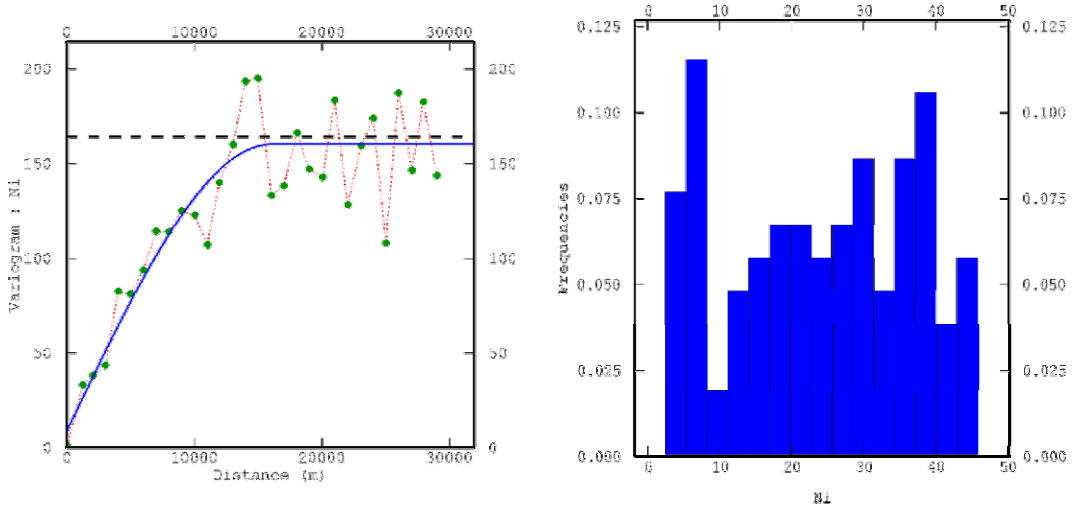


Figure 2.21 – Variogram and histogram for Nickel.
Figura 2.21 – Variogramma e istogramma per il Nichel.

With the theoretical variogram model:

$$\gamma(h) = 8.55 + 151.63 * sph^{(16.05e003)}.$$

Vanadium

Variogram of vanadium shows a clear nested structure with two spherical models. Its histogram is closer to a normal-like one than the previous variables.

	N	MIN	MAX	MEAN	MEDIAN	VARIANCE	STD DEV	SKEWNESS	KURTOSIS
V	104	12.04	151.45	65.24	68.31	900	30.00	0.128	-0.65

Table 2.8 – Basic statistics for Vanadium.
Tabella 2.8– Statistiche di base per il Vanadio.

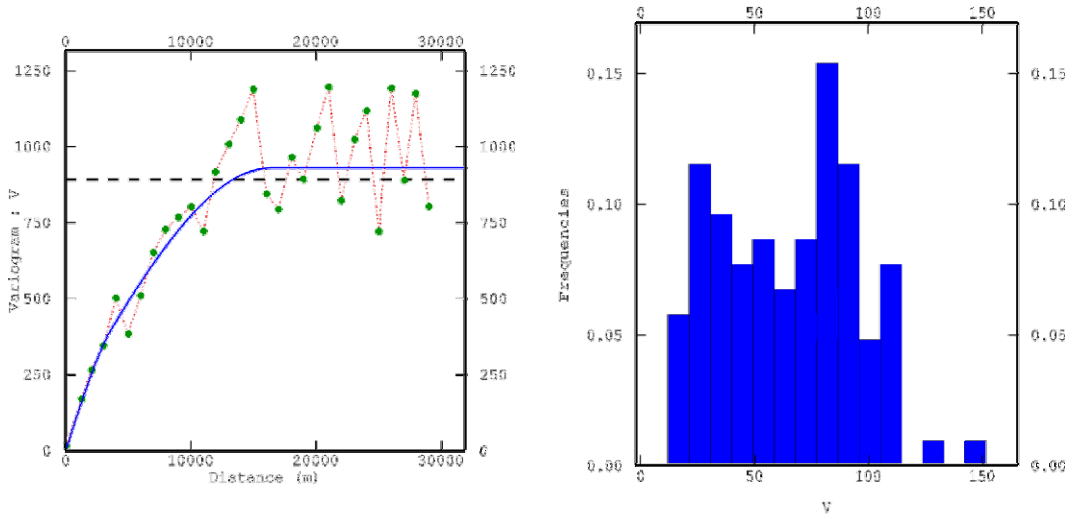


Figure 2.22 – Variogram and histogram for Vanadium.
Figura 2.22 – Variogramma e istogramma per il Vanadio.

With the theoretical nested variogram model:

$$\gamma(h) = 140.66 * sph^{(3.76e003)} + 788.98 * sph^{(16.40e003)}.$$

Zinc

Once again, Zinc variogram shows a well defined nested structure, with two spherical models. The histogram is somewhat normal-like.

	N	MIN	MAX	MEAN	MEDIAN	VARIANCE	STD DEV	SKEWNESS	KURTOSIS
ZN	104	11.99	124.86	69.81	73.43	935	30.57	-0.221	-0.93

Table 2.9 – Basic statistics for Zinc.
Tabella 2.9– Statistiche di base per lo Zinco.

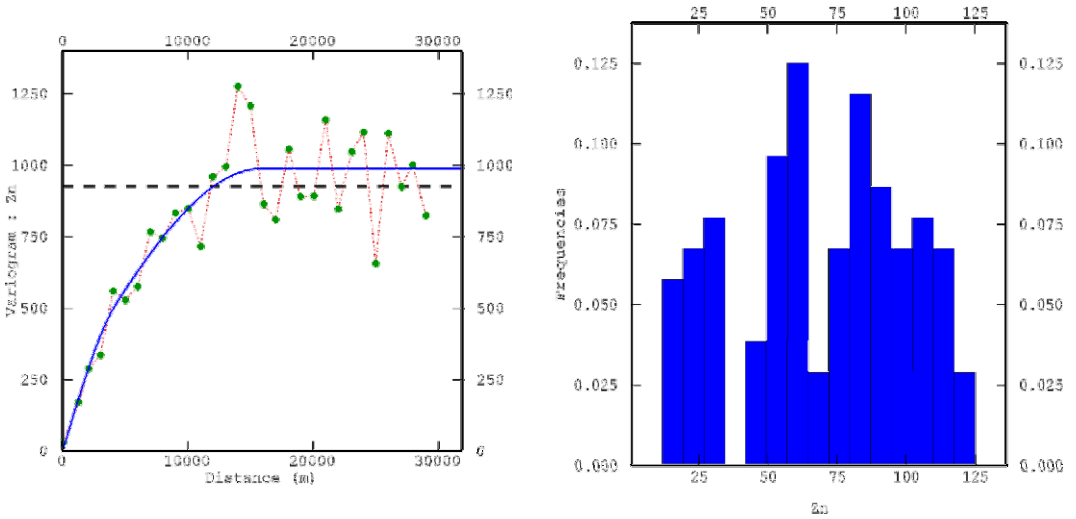


Figure 2.23 – Variogram and histogram for Zinc.
Figura 2.23 Variogramma e istogramma per lo Zinco.

With the theoretical nested variogram model:

$$\gamma(h) = 208.60 * sph^{(4.33e003)} + 779.56 * sph^{(15.86e003)}.$$

Lead

Lead histogram reveals a certain unbalance of the distribution with a positive skewness. Its variogram and generally the application of geostatistical technique could be more correctly implemented if the variable were logarithm transformed (the natural logarithm *ln*). Nevertheless, if we observe the histograms of the original and the transformed variable, we can note that such improvement has not happened.

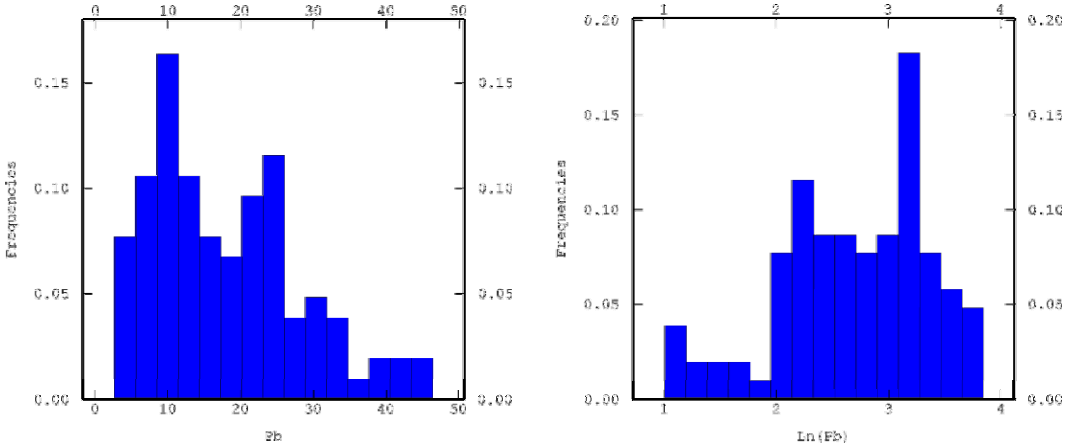


Figure 2.24 – Histograms of Lead (left) and natural logarithm of Lead (right).
Figura 2.24 – Istogramma del Piombo (sinistra) e del suo logaritmo naturale (destra).

Basic statistics are:

	N	MIN	MAX	MEAN	MEDIAN	VARIANCE	STD DEV	SKEWNESS	KURTOSIS
PB	104	2.75	46.29	18.00	16.44	110	10.47	0.678	-0.13
Ln(PB)	104	1.01	3.83	2.69	2.80	0.44	0.67	-0.59	2.85

Table 2.10 – Basic statistics for Lead and natural logarithm of Lead.
Tabella 2.10 – Statistiche di base per il Piombo e per il suo logaritmo naturale.

where it is clear that the variable has not improved the normal-like behaviour (compare the skewness coefficients). In agreement with statistical results, also variogram of transformed variable is not clearer than the original one.

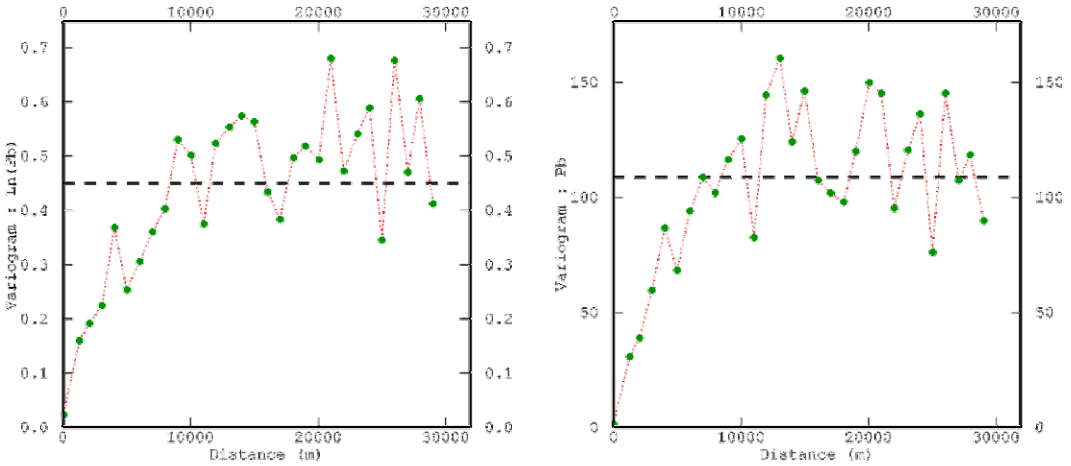


Figure 2.25 – Histograms of Lead (left) and natural logarithm of Lead (right).
Figura 2.25 – Istogramma del Piombo (sinistra) e del suo logaritmo naturale (destra).

In order to avoid the distortions due to computation of variogram of a transformed variable, and considering that the gain from such transformation is about null, we'll continue processing the untransformed variable.

Lead variogram shows a very clear nested structure, with the highest short range component of the variables.

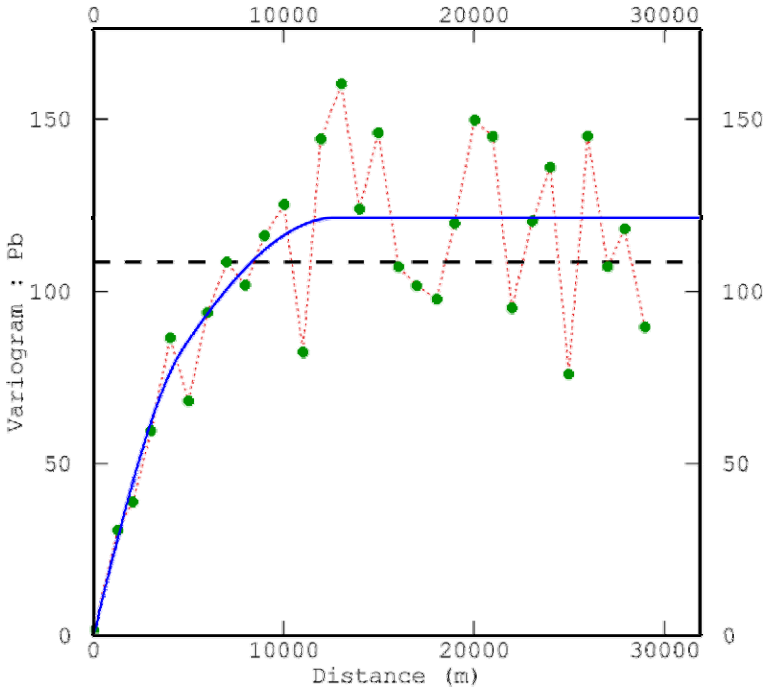


Figure 2.26 – Nested variogram of Lead.
Figura 2.26 – Variogramma nidificato del Piombo.

With the theoretical nested variogram model:

$$\gamma(h) = 40.42 * sph^{(4.72e003)} + 80.95 * sph^{(12.68e003)}.$$

Mercury

Variogram of mercury shows a complete lack of spatial structure. Data seem to be totally uncorrelated and the resulting model is a pure nugget one, with sill close to the total sample variance. Skewness value and the highly skewed histogram reveal a strong unbalance toward lower classes.

	N	MIN	MAX	MEAN	MEDIAN	VARIANCE	STD DEV	SKEWNESS	KURTOSIS
HG	102	0.0002	0.35	0.06	0.04	4.0e-03	0.06	2.222	5.67

Table 2.11 – Basic statistics for Mercury.
Tabells 2.11 – Statistiche di base del Mercurio.

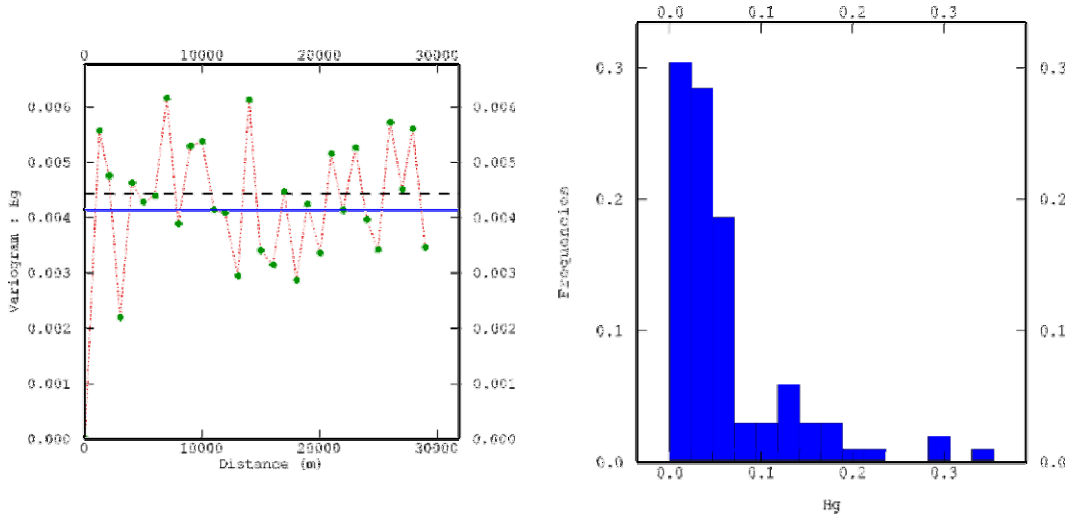


Figure 2.27 – Variogram and histogram for Mercury.
Figura 2.27 – Variogramma e istogramma del Mercurio.

With the theoretical pure nugget variogram model:

$$\gamma(h) = 4.13e-003.$$

If we consider the analytic mean error equal to the 10 percent of the median value of the whole dataset, we can fix the median analytic variance to 1.6e-005 ppm^2, that is much lower than nugget variance value. It means the small scale variability is intrinsic in the dataset and does not depend from the analytic uncertainness. By transforming the variable into its natural logarithm, we can improve the meaning of variogram.

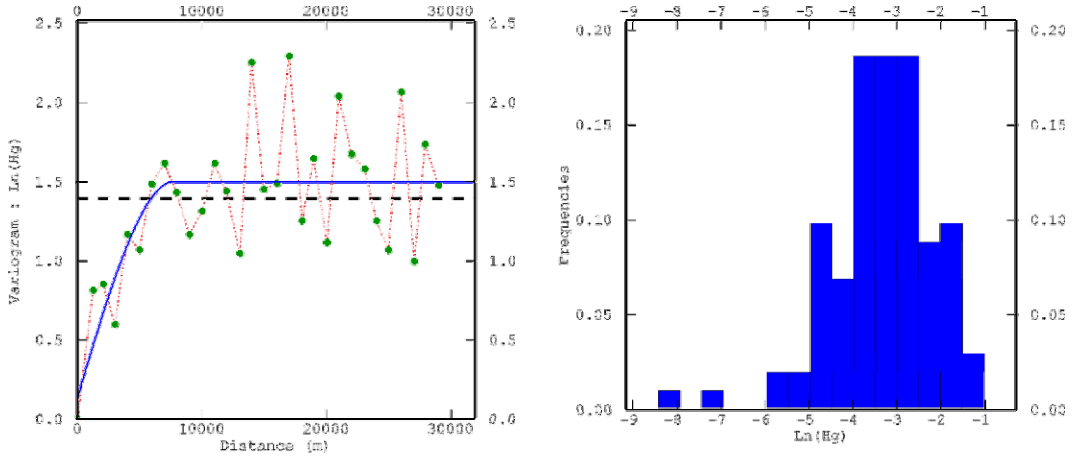


Figure 2.28 – Variogram and histogram for natural logarithm of Mercury.

Figura 2.28 – Variogramma e istogramma per il logaritmo naturale del Mercurio.

With the theoretical variogram model:

$$\gamma(h) = 0.13 + 1.36 * sph^{(7.76e003)}.$$

Cadmium

Experimental variogram of Cadmium does not reveal any clear structure, and observing the histogram we can propose to check the role of the highest class values in variogram definition.

	N	MIN	MAX	MEAN	MEDIAN	VARIANCE	STD DEV	SKEWNESS	KURTOSIS
CD	104	0.02	2.00	0.15	0.10	4.8e-02	0.22	6.243	50.58

Table 2.12 – Basic statistics for Cadmium.
Tabella 2.12 – Statistiche di base per il Cadmio.

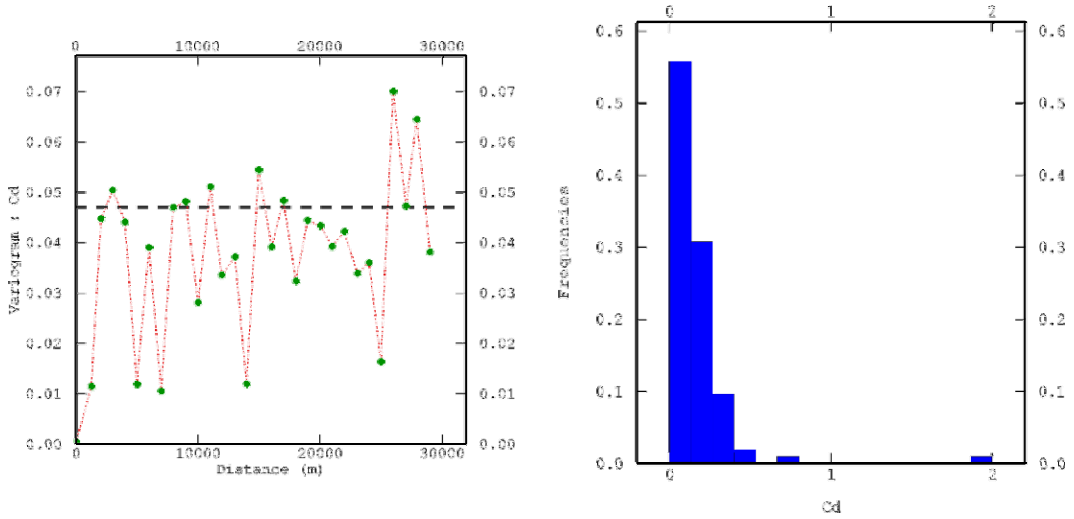


Figure 2.29 – Variogram and histogram for Cadmium.
Figura 2.29 – Variogramma e istogramma per il Cadmio.

If we filter out such class (red highlighted), the variogram is somewhat clearer and it can be conveniently modelled.

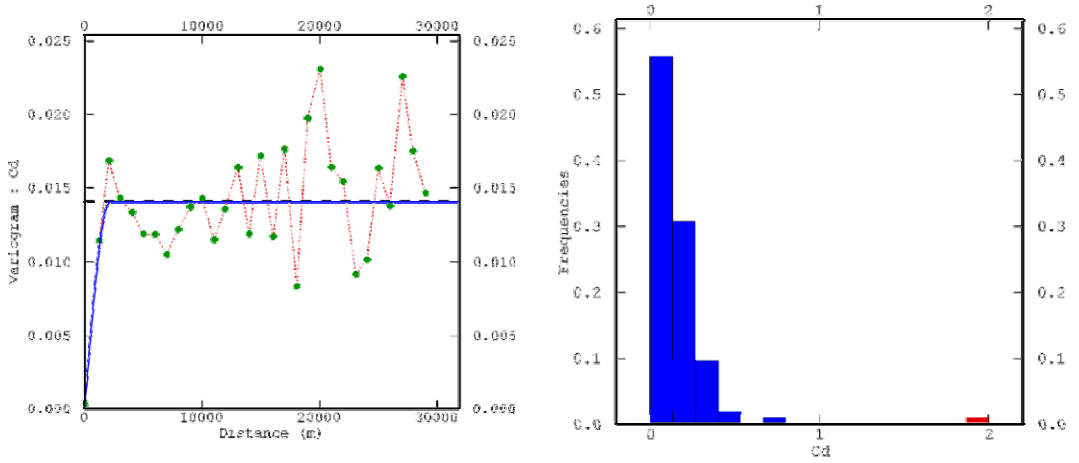


Figure 2.30 – Variogram and histogram for filtered Cadmium. Outlier class highlighted in red.
Figura 2.30 – Variogramma e istogramma del Cadmio filtrato. Le classi estreme sono evidenziate in rosso.

With the theoretical variogram model:

$$\gamma(h) = 0.014 * sph^{(2.13e003)}.$$

In this case the logarithmic transformation of the variable, even improving the shape of histogram making it more normal-like, does not increase the conformity of the variogram.

Arsenic

Experimental variogram of Arsenic reveals a somewhat high degree of small scale variability, with a important role of the nugget variance. Nevertheless, by computing the variance contribution from analytic measurements as 10% of the median value, about 16 ppm^2, we must rescale the nugget effect, such that only the 50% is effectively due to intrinsic small scale variability.

	N	MIN	MAX	MEAN	MEDIAN	VARIANCE	STD DEV	SKEWNESS	KURTOSIS
AS	104	6.47	50.77	20.85	20.18	63	7.96	0.859	1.26

Table 2.13 – Basic statistics for Arsenic.
Tabella 2.13 – Statistiche di base per l'Arsenico.

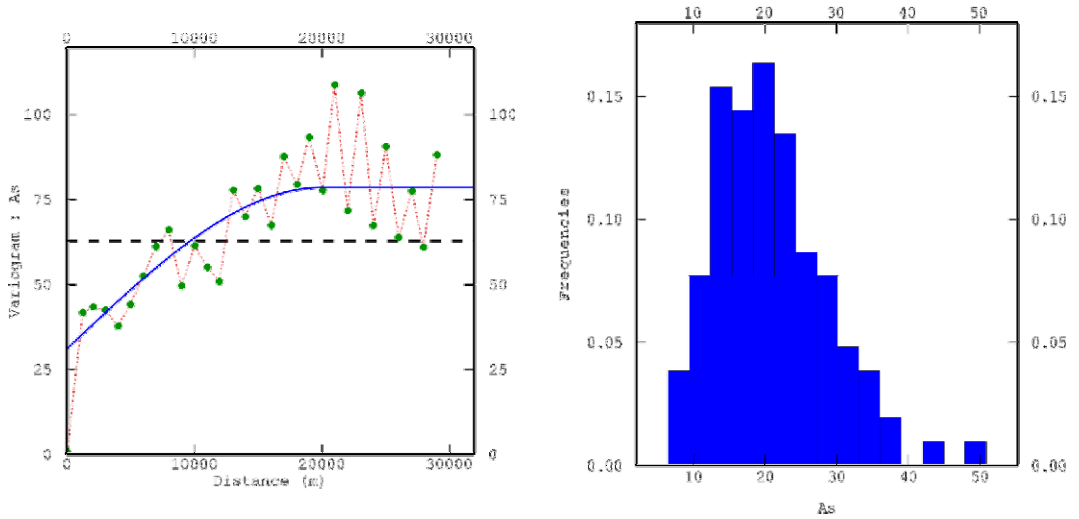


Figure 2.31 – Variogram and histogram for Arsenic.
Figura 2.31 – Variogramma e istogramma per l'Arsenico.

With the theoretical variogram model:

$$\gamma(h) = 30.90 + 47.61 * sph^{(20.00e003)}$$

4. The estimation

Once defined the variograms and modelled them in the most appropriate way, we can implement the final target, that is the spatial estimation, focused on the complete knowledge of the spatial distribution of the different trace metals, in the investigated area.

By analyzing results of structural analyses, we can gather the variables into two main group:

- Nested (Aluminium, Fe, V, Zn, Pb);
- Non nested (Chromium, Cu, Ni, As).

Eventually, mercury has been *ln*-transformed, in order to obtain a clearer variogram structure, and Cadmium has been filtered of highest values. Some example for each group will be shown here, in order to realize how the variogram influences the estimation process.

4.1 Nested structures

The nested variogram model of Zinc is well defined and it will be used to implement kriging of components, in order to decompose the two spatial components contributing to the distribution of variable. Kriging of the global structure gives origin to the map in figure 2.32, where the constant global mean is 72.67 *ppm*. A unique neighbourhood has been chosen because variogram range represents the driving indication for the kriging search radius definition.

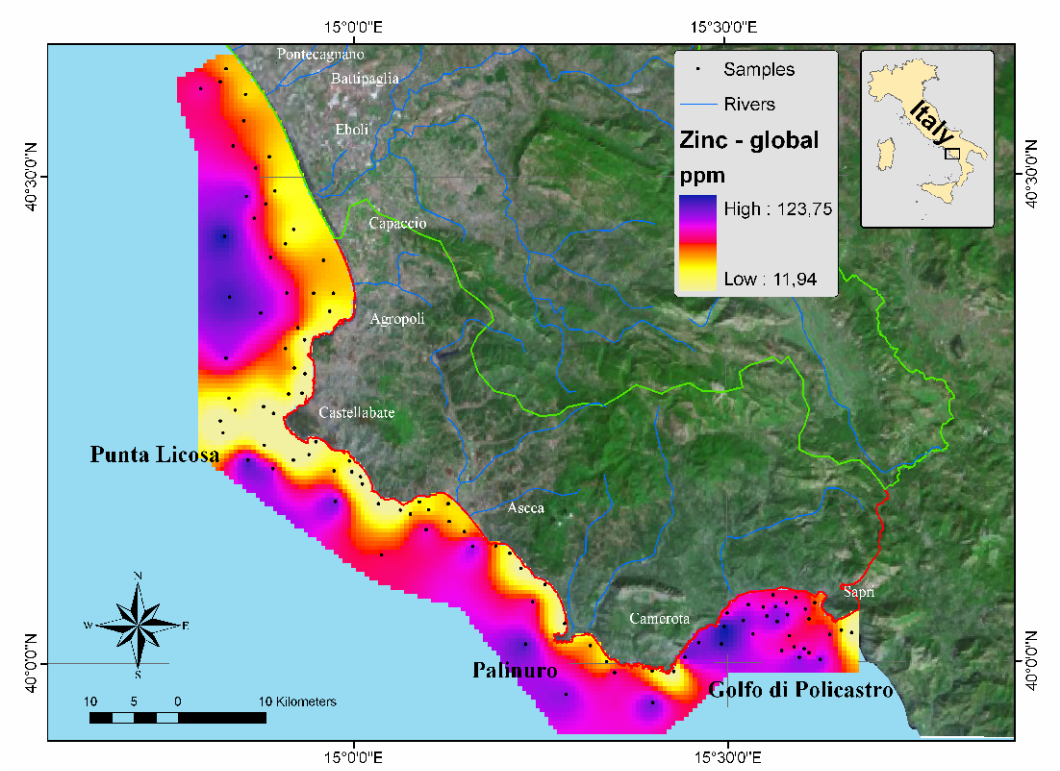


Figure 2.32 – Kriged map of Zinc, based on the nested variogram.
Figura 2.32 – Stima con Kriging dello Zinco, basata sul modello nidificato.

The spatial distribution reveals a complex structure, in which a global smoothed variation is overlapped to a more local one, with some spots widely distributed in the domain (Durrani and Badr 1995). Lowest concentration values are located mainly along the shoreline at the shallowest depth as in Punta Licosa, while high values are concentrated in deepest areas as Golfo di Policastro. The greatest bull's eyes effect is concentrated around those high value samples that are isolated from the others out of Palinuro and Punta Licosa shores. Kriging variance map is

minimized at the close neighbourhoods of samples and increase slowly with going further.

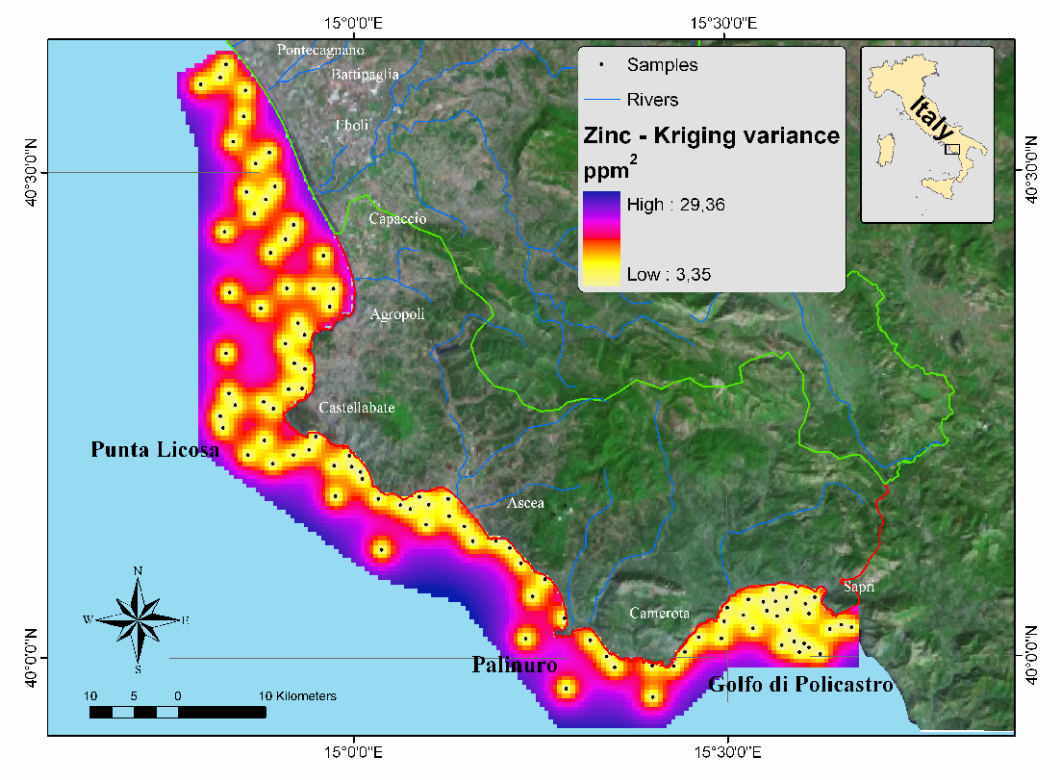


Figure 2.33 – Kriging variance map of the estimated map based on nested variogram of Zinc.
Figura 2.33 – Mappa della varianza di kriging relativa al modello nidificato dello Zinco.

Cross validation confirms the good quality of the estimation, with the scatter plot of real and estimated values revealing a correct balance between under- and over-estimation, and the histogram of the standardized errors showing their normal distribution. As always, estimated values are slightly less spread than real ones. The bands are related to ± 2.5 standardized error.

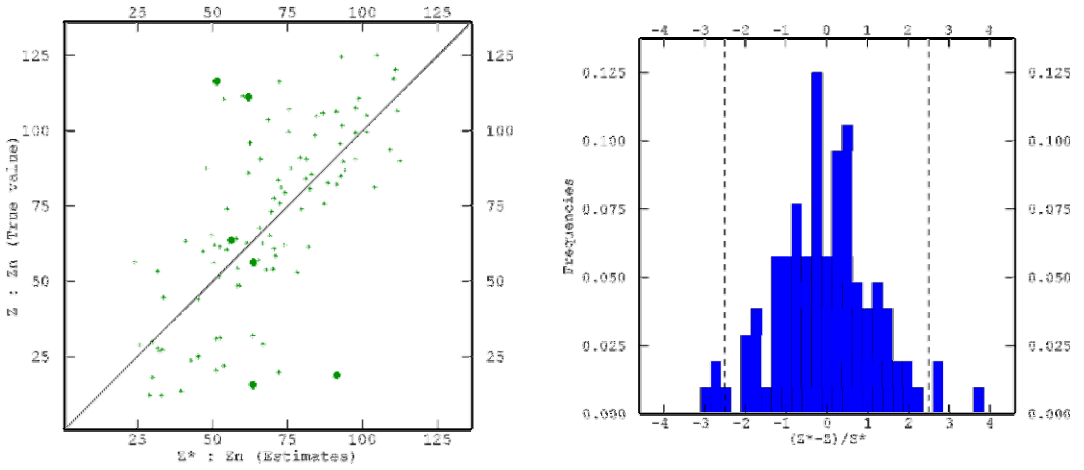


Figure 2.34 – Left: scatter plot of known and estimated values. Right: histogram of relative errors.
Figura 2.34 – A sinistra: scatter plot dei valori noti e stimati. A destra: istogramma degli errori relativi.

	SCATTER PLOT	HISTOGRAM			
	$\rho(Z,\tilde{Z})$	MIN	MAX	MEAN	STD DEV
ZN	0.69	-3.02	3.83	-0.03	1.20

Table 2.14 – Basic statistics for estimates for Zinc.
Tabella 2.14 – Statistiche di base per le stime dello Zinco.

By using kriging to filter the spatial components, we can decompose the spatial structure. In the figure, the long range map looks like the global one except for the absence of the punctual spots of high and low values.

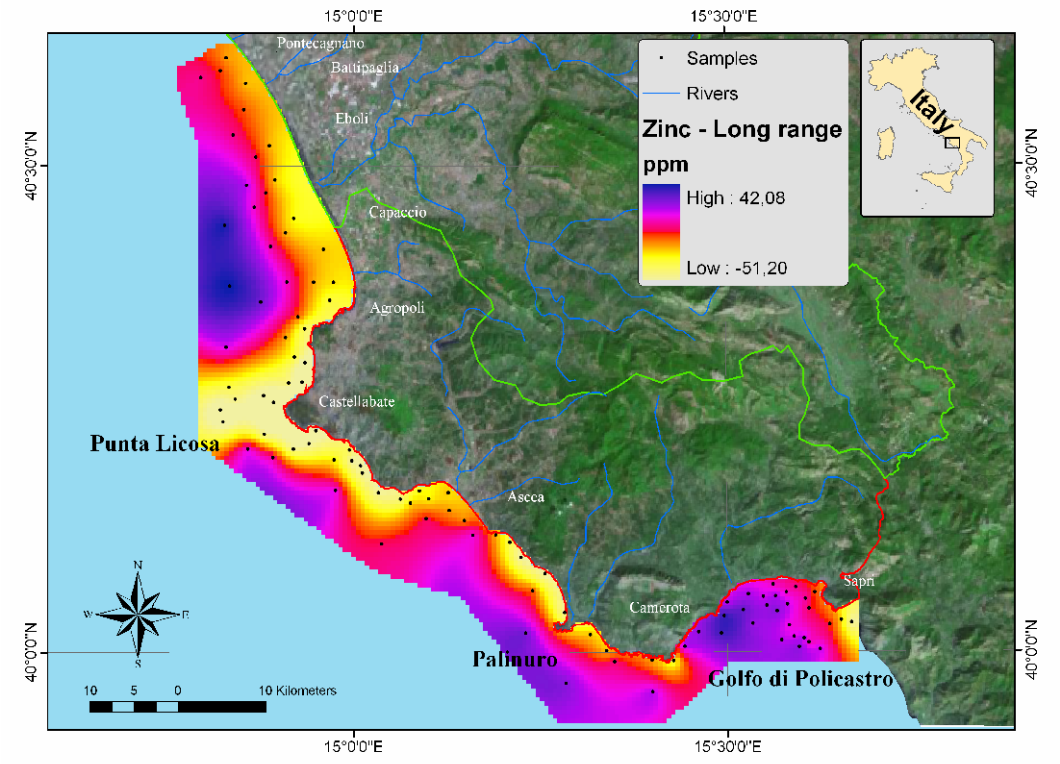


Figure 2.35 – Kriged map of the long range component of Zinc (subtracted of the global mean value).
Figura 2.35 – Stima con kriging della componente regionale dello Zinco (sottratta del valore medio globale).

It represents the low frequency variation of the field, with slow and smoothed variability spread over the whole domain. As in the global map, high values are concentrated in deep water and in Golfo di Policastro, while the lowest values are mainly located along the shoreline and out of Punta Licosa. The estimated values are subtracted of the mean (72.67 *ppm*) and does not represent a quantitative estimation of real values.

The short range map reveals a series of rapid oscillations around the mean filtered value.

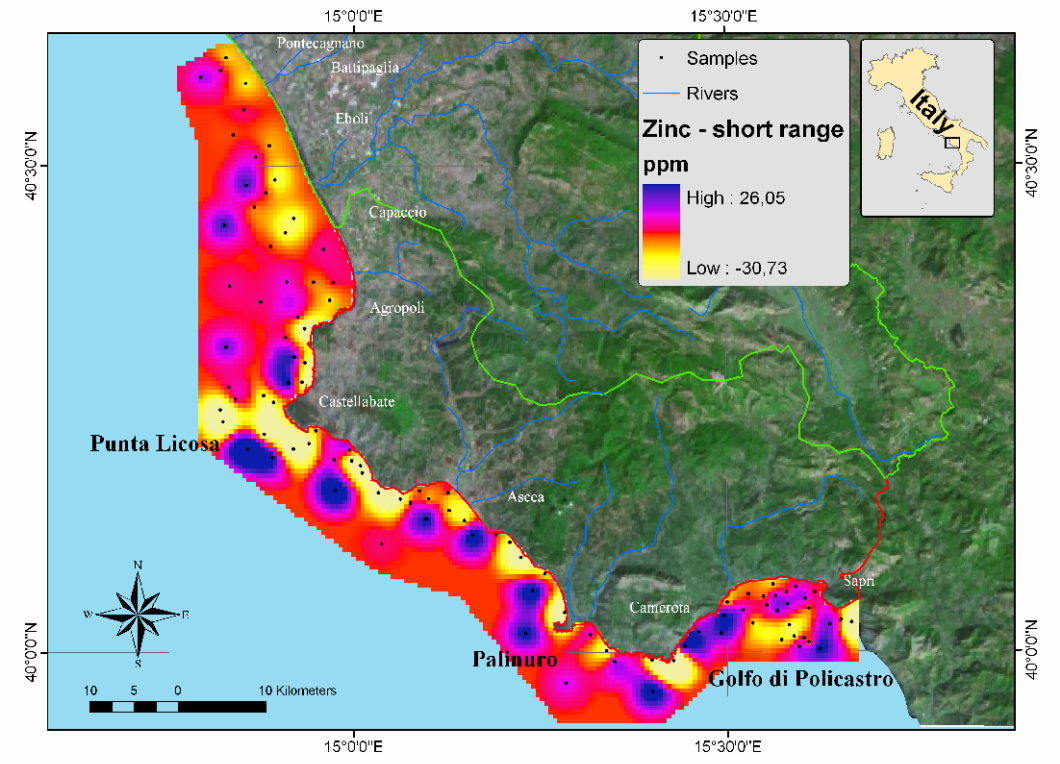


Figure 2.36 – Kriged map of the short range component of Zinc (subtracted of the global mean value).
Figura 2.36 – Stima con kriging della componente locale dello Zinco (sottratta del valore medio globale).

Such high frequency variability produces several hotspots presenting high and low values (anomalies). In general, positive anomalies are widely spread in deep water, while lowest one are close to the shore, but in some zones of Golfo di Policastro, such tendency is inverted, with a negative anomaly in the deep centre of the gulf.

An interesting aspect is the variation range of both the maps. The long range map presents a range of about 93 ppm, while the short one only 57 ppm. It confirms the greatest variation of the low frequency component that is the driving factor in spatial distribution of the variable. Conversely, the short range component is just like an adding signal that leads positive and negative anomalies to a more stable track. If we move the interpretation to the temporal framework, we can assume that the long range component is a more stable natural and old factor defining the spatial structure, while the short one is a more recent one that can reasonably be lead to some external interference, like anthropic presence.

4.2 Non nested structures

Chromium will be used as example for non nested structures. Its processing is a classical example of Ordinary Kriging application. The estimated map shows a spatial distribution very similar to the Zinc one.

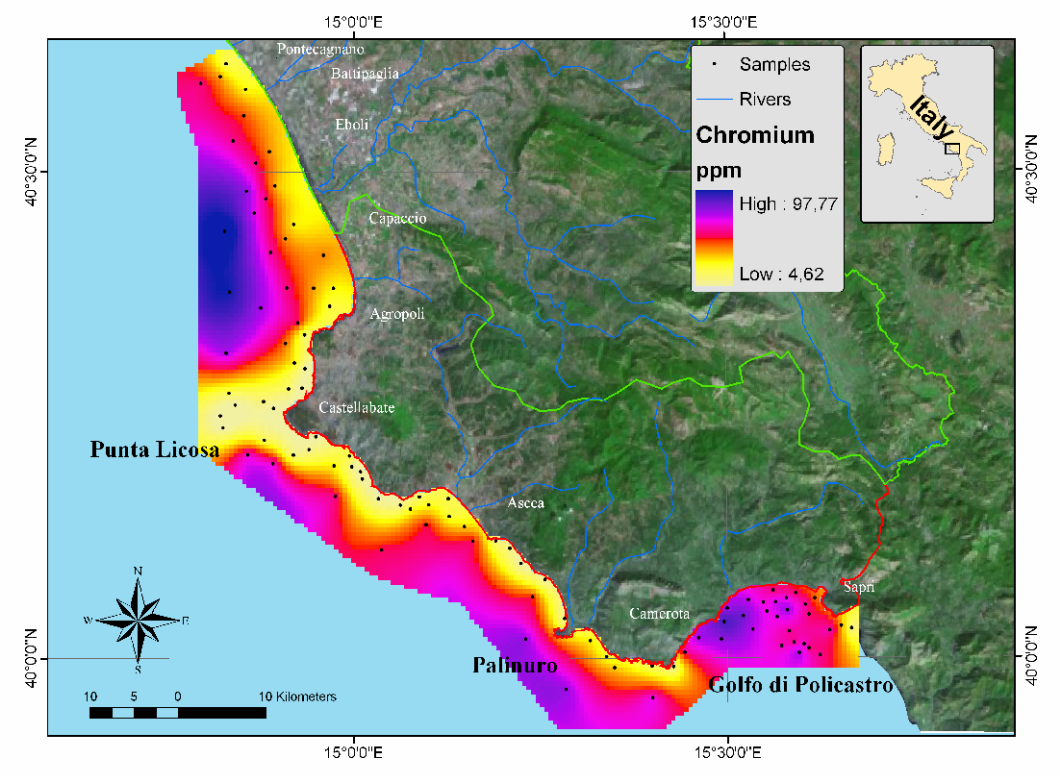


Figure 2.37 – Kriged map of Chromium.
Figura 2.37 – Stima con kriging del Cromo.

Again, concentration values are mainly driven by grain size, with highest values concentrated in Golfo di Policastro and in deep water and lowest ones spread along the shoreline and out of Punta Licosa. Cross validation and kriging variance map confirm the affordability of the map.

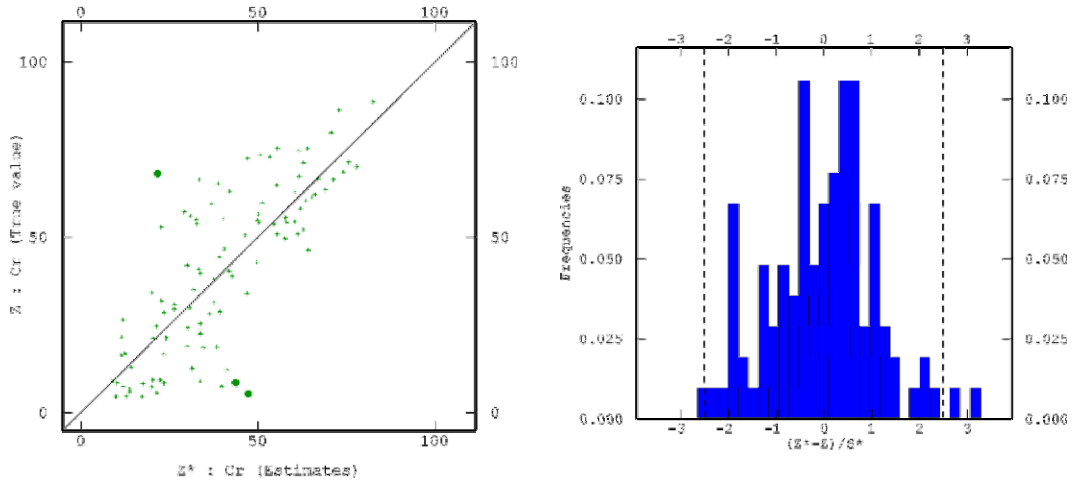


Figure 2.38 – Left: scatter plot of known and estimated values. Right: histogram of relative errors.
Figura 2.38 – A sinistra: scatter plot dei valori noti e stimati. A destra: istogramma degli errori relativi.

	SCATTER PLOT	HISTOGRAM			
	$\rho(Z,\tilde{Z})$	MIN	MAX	MEAN	STD DEV
CR	0.79	-2.63	3.27	-0.03	1.11

Table 2.15 – Basic statistics for estimates for Chromium.
Tabella 2.25 – Statistiche di base per la stima del Cromo.

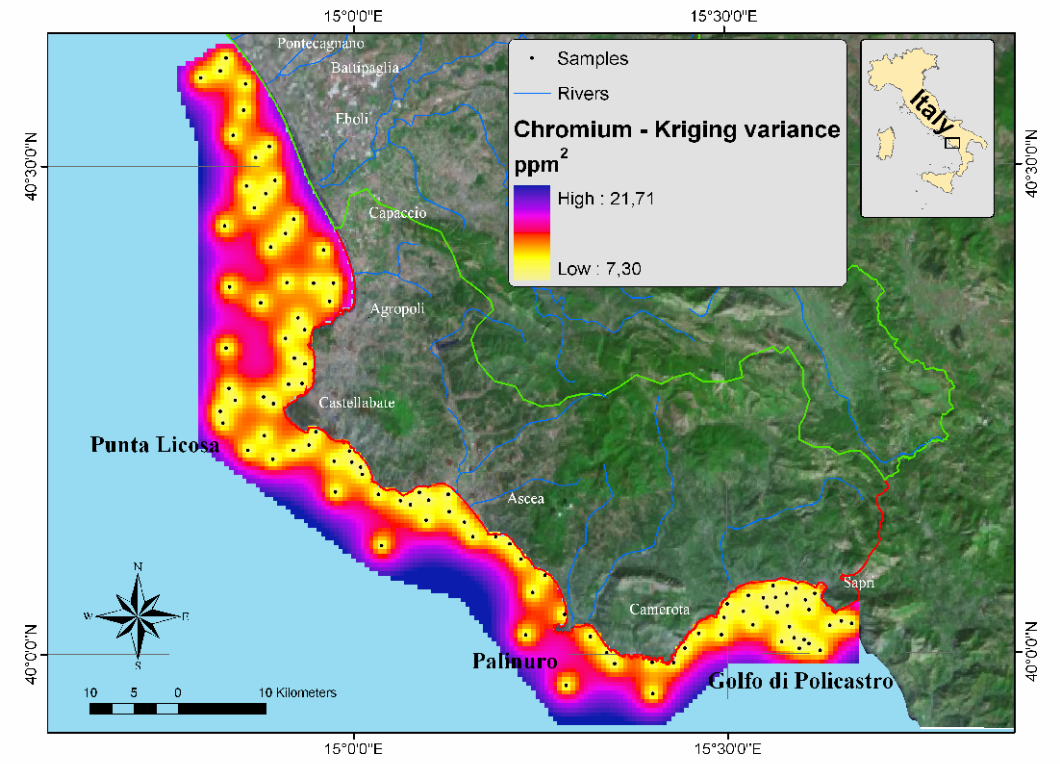


Figure 2.39 – Kriging variance map of the estimated map of Chromium.
Figura 2.39 – Mappa della varianza di kriging per il Cromo.

Minimum value of kriging variance are quite high, because of the non zero value of nugget variance.

4.3 Log-transformed variables

As described in the previous chapter, log-normal kriging is based on the application of the method to the logarithmic transformed variable, in the case it does not present a convenient normal distribution. As noted in structural analysis, mercury shows such highly skewed histogram and it has been *ln*-transformed. In this case the variable shows all positive values and a shift is not needed.

Log-normal kriged map shows a somewhat irregular spatial distribution of mercury concentration values, with several spots and high frequency variations.

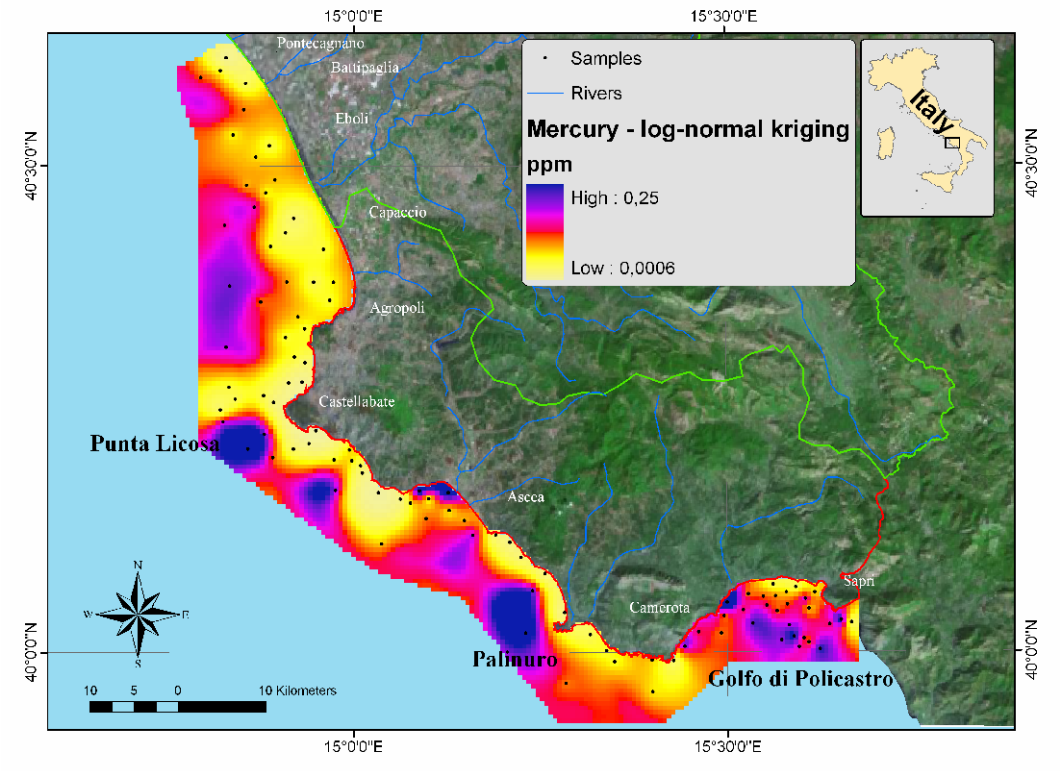


Figure 2.40 – Estimated map of Mercury computed with Log-normal kriging technique.

Figura 2.40 – Stima con Log-normal kriging del Mercurio.

Nevertheless, the general course of the field is kept and the usual spatial correlation with grain size is somewhat clear. If compared with results from IDW (inverse distance weighted) and SPLINE methods, the kriged map is much clearer and consistent with original data (Lancaster and Salkauskas 1986).

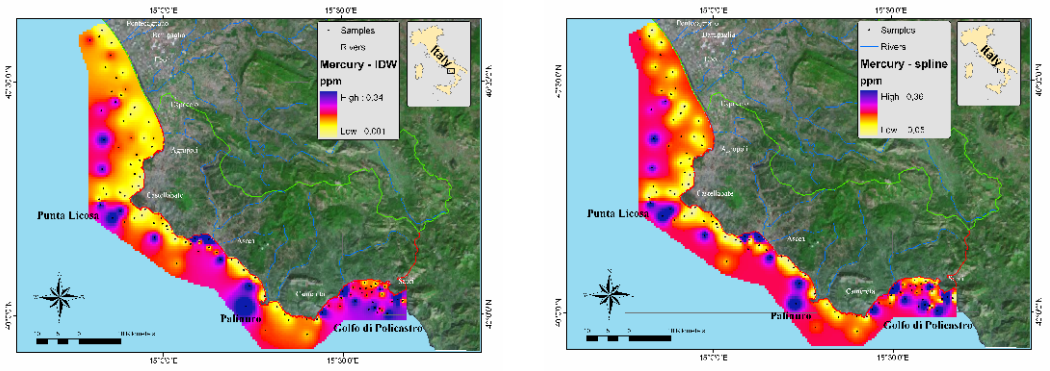


Figure 2.41 – Estimated map of Mercury computed with IDW and spline techniques, respectively on the left and on the right.

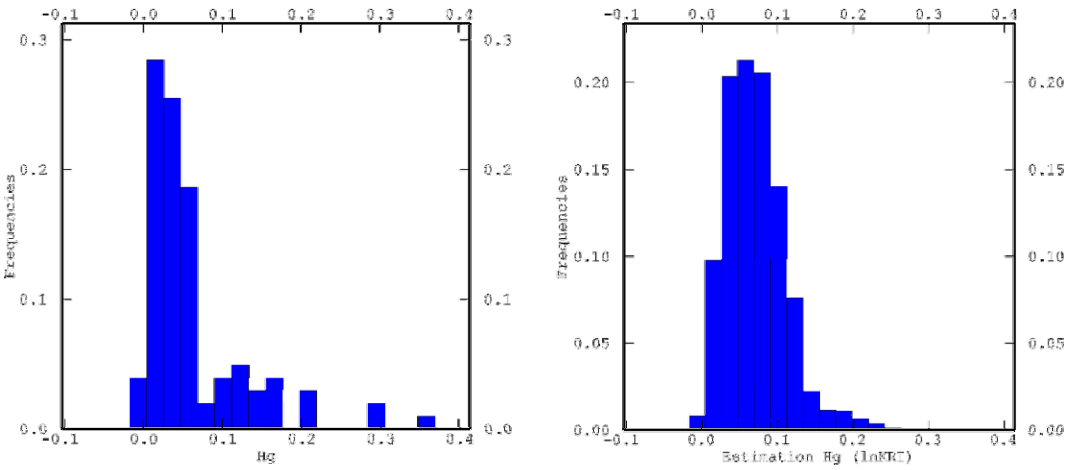
Figura 2.41 – Stima con IDW e spline del Mercurio, rispettivamente a sinistra e a destra.

Both the methods presents the classical bull's eyes effect, more evident in spline map, while the last one, because of the smoothing behaviour of the 3rd degree polynomials, produces also negative values (Wingle 1992; Collins and Bolstad el. source). By observing the basic statistics we can note how the kriged map is the less spread, because of the features of kriging algorithm described before.

	N	MIN	MAX	MEAN	MEDIAN	STD DEV	SKEWNESS	KURTOSIS
Hg raw data	102	0.0002	0.35	0.06	0.04	0.06	2.22	5.67
Est. Hg log-kri	5150	0.0006	0.25	0.07	0.06	0.04	0.98	4.66
Est. Hg IDW	5150	0.0014	0.34	0.06	0.05	0.03	1.95	11.25
Est. Hg spline	5150	-0.05	0.36	0.06	0.06	0.03	2.95	22.87

Table 2.16 – Basic statistics comparing the results of different estimation techniques.
Tabella 2.16 – Statistiche di base per il confronto dei risultati delle diverse tecniche di stima

The best reliability of kriged map is more evident from the observation of the histograms.



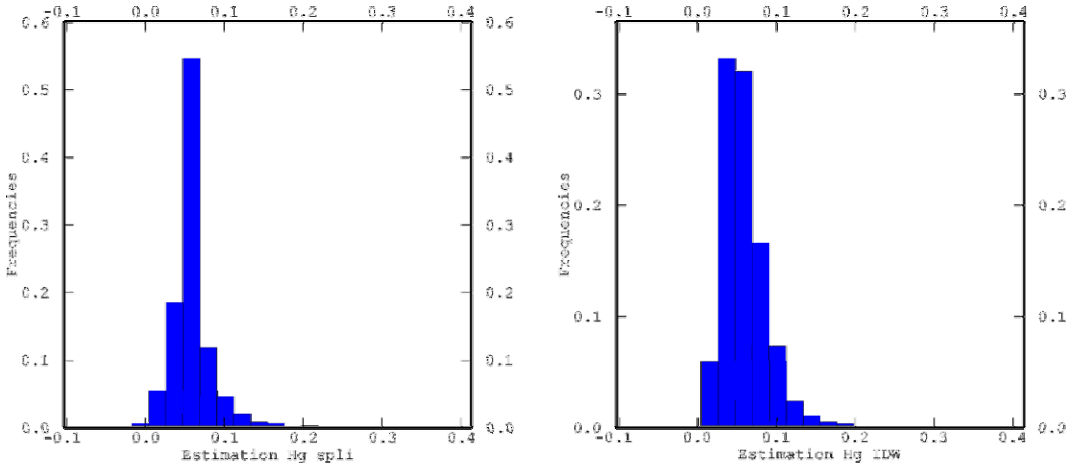


Figure 2.42 – Histograms of original data and estimated values with lognormal kriging, spline and IDW, respectively from the top and from the right.
Figura 2.42 – Dall'alto a sinistra: istogramma dei dati originali e delle stime con log-normal kriging, spline e IDW.

Where the skewed shape of original data is respected better in kriged map.

5. Discussion

We presented some example of geochemical variables processing approach, based on stationary geostatistical technique. A preliminary statistical analysis allowed to investigate on the correlations among the variables and to divide them in two main groups. Aluminium, Fe, Cr, Cu, Ni, V, Zn, Pb result strictly correlated, while arsenic, mercury and cadmium are far from such cluster. A detailed structural analysis and an accurate variography complicate such gathering defining two more classes. Aluminium, Fe, V, Zn present a nested variogram with two or more spatial components, while for Chromium, Cu, Ni, As the spatial variability can be conveniently described by only one model. Finally, lead, mercury and cadmium are treated separately, analyzing each case in details.

Different forms of kriging are applied to the different groups of variables, illustrating the use of spatial component filtration and the influence of skewed distributions or outliers in datasets. Each map is checked by cross validation technique or by observing the basic statistics of results.

The main evidence is the almost complete dependence of nearly all the variables from grain size distribution. Finest sediment, concentrated in deeper water and in Golfo di Policastro area, seems to cause an accumulation of each trace metals, while

the zone out of Punta Licosa, with its shallower bottoms and its more active gravel sand, does not allow a high degree of sedimentation.

For what concerns the variables that present a nested structure of variability, it is very interesting to propose a geochemical interpretation of such situation. Large scale component, in fact, can be regarded as a globally varying background value that represents the natural oscillating level of each of these chemical species (Cheng, Agteberg et al. 1994; Salminen and Tarvainen 1997; Vivo, Boni et al. 1997; Yunker, Macdonald et al. 1999; Singh, Müller et al. 2003). Background values are assumed to be not a fixed threshold that is constant for the whole spatial domain, but, more reasonably, a slowly varying surface (range of values) that had been evidently shaped by natural constraints. Conversely the high frequency component can be regarded as an adding signal with more rapid and small scale oscillation. Thus, such component can be reasonably regarded as the result of a more local control of grain size anomalies, hydrothermal activities, that are widely spread in the area, or a somewhat anthropic input.

Such interpretation is obviously less affordable than the one proposed for the large scale component (Sprovieri et al. 2005). It is much easier to suppose a natural background interpretation for the low frequency slowly varying component, rather than for the more irregular one. Thus, anthropic incidence can be at least one of the factor responsible for such structure.

CHAPTER 3
OPTIMIZATION OF SAMPLING DESIGNS

The optimization of sampling strategies is a well known argument that has been exhaustively discussed in literature (Lloyd and Atkinson 2001; Lloyd and Atkinson el. source). The quality of data and the reliability of the measure is strictly dependent rather than on the analytic errors, on the choice regarding the location of sampling sites. The key feature of geostatistical applications is the strict dependence of such methodologies on the results of variographic analyses. Such analyses are based on statistical processing of hard data (the samples) that are deeply influenced by their spatial distribution.

In fact geostatistics is exactly somewhat statistics (or better stochastics) in spatial context and it is consequently strictly dependent on spatial features of hard data. The main of such spatial features is just the sampling scheme. The sampling structure, the sampling resolution and the total amount of samples (informations) are essential parameters for variogram computation and just such instrument will be employed in this work, in order to optimize sampling strategy (Smith and Williams 1996).

The problem will be divided in two main branches:

- Optimizing a new sampling plan;
- Improve the efficiency of an existing sampling plan.

The first concern will be based on the definition of the technique finalized to optimize the faculty of a sampling plan, to catch the whole variability of the system. Variograms will be used to analyze the differences between two classical sampling techniques: i) the systematic one and ii) the stratified random one.

The second point will focus on the enhancement of effectiveness of an existing sampling plan for which the minimum number of additional samples will be outlined (Goovaerts 2001; Kanevski, Parkin et al. 2004; Goovaerts, Jacqueza et al. 2005; Goovaerts el. source; Goovaerts el. source). Geostatistical simulations will be used to quantify the spatial uncertainness of the field, while the variogram will be involved in the process of minimization of the small scale variance (Castrignanò and Buttafuoco 2004).

1.Optimization of a new sampling plan

Sampling strategy is a very important aspect of spatial analysis. As said before the information deriving from hard data are deeply influenced by samples location and their distribution in spatial domain. In particular, with fixed economic availability and with the consequent fixed number of potential samples, the quality of representation of variability structure of the variable is strictly dependent on the way the available information sources are located in the spatial domain.

That is the reason for which sampling strategy is so important in spatial analysis and its optimization is a key concept in such kind of applications. The author wants to present an application of variographic instruments as control of the influence of the sampling plan on the variability structure and thus propose the best strategy.

Many sampling approaches have been defined in literature (Burrough and McDonnell 1997) and each of them is reasonably applicable in different conditions. Nevertheless, here a generalization of an optimal choice is proposed, focusing on the capacity of applying it conveniently in any situation.

As said before, each sampling strategy is opportunely applied in any well defined situation, but all the methods are grouped into two main divisions:

- systematic strategies;
- random strategies.

As easy realizing, the first group concern all the methods based on regular division of the spatial context and the systematic allocation of samples, while the second one regards a random distribution of information sources.

1.1 Systematic strategies

Systematic sampling strategies are based on regular distribution of sites of investigation, based on a simple division of the spatial domain in a regular grid, or by following some specific demand, as isovalues lines or transects. The main typologies are (Burrough and McDonnell 1997):

- *pure systematic sampling*, that is based on the simple division of the spatial domain in a regular grid and on the sample allocation in each cell of such grid;

- *transect sampling*, that is based on the distribution of each sample along defined transects;
- *isovalues sampling*, that is based on the distribution of samples along isovalues lines of the investigated variable (whereas they are known).

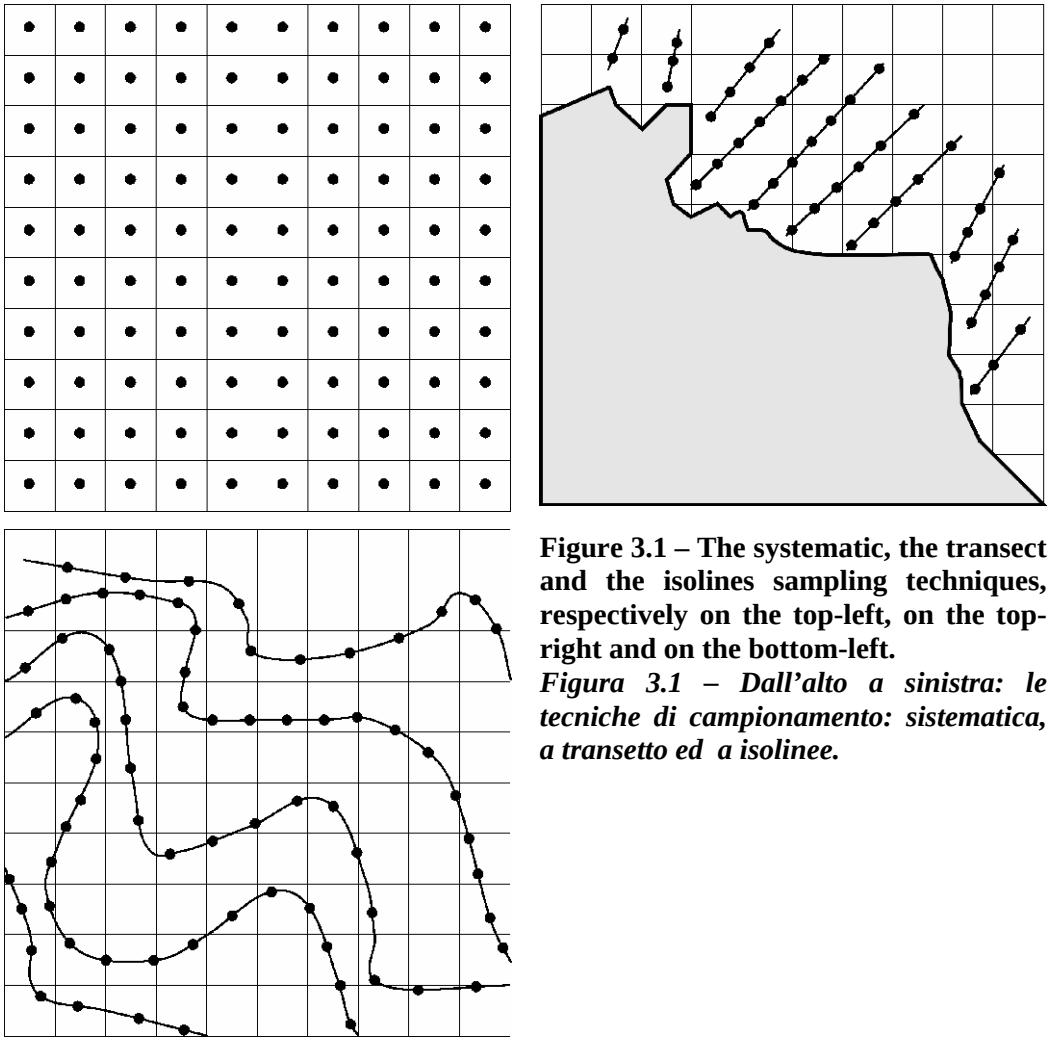


Figure 3.1 – The systematic, the transect and the isolines sampling techniques, respectively on the top-left, on the top-right and on the bottom-left.
Figura 3.1 – Dall'alto a sinistra: le tecniche di campionamento: sistematica, a transetto ed a isolinee.

The main feature of such kind of methods is the capacity to cover uniformly the whole domain of investigation, and to explore with details the elements of the system that are partially known. The transect sampling strategy is focused on the idea to investigate some variable that is expected to vary principally along a certain direction, while the isovalues sampling method is usually used for the vectorization of raster map, for which the aim is to create a punctual set of isovalues for a certain variable.

The common feature of such methodologies is the lack of relation between the variability of the variable and the choice on samples allocation. In pure systematic sampling only the resolution of the grid is defined before (usually as function of economic availability), with no consideration on possible over- or under- estimation of the field, while the other two methods make some assumption on the behaviour of the variable, but with no precise knowledge.

1.2 Non systematic strategies

As affirmed in previous chapter, environmental variables are assumed to be better interpreted by stochastic modelling, more than by deterministic ones. It has been explained how the main geostatistical principle is to consider the variable as a random regionalized function and to regard the samples as realizations of such function.

Following such principle, we can assume that the best way to create an unbiased sampling plan is to define it in a random framework (Burrough and McDonnell 1997; Burrough el. source). Nevertheless, a pure random sampling strategy can lead to an excessive economic waste, because a complete coverage of the domain is assured only by a very large amount of samples. In order to join the unbiasedness of random methods and the complete investigation of the site, the stratified random strategy has been defined.

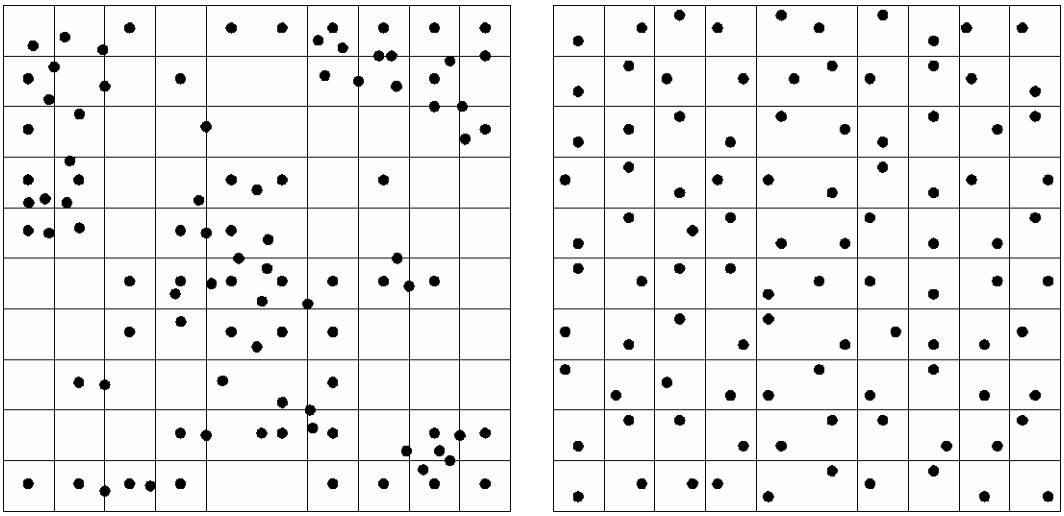


Figure 3.2 – The pure random and the stratified random sampling strategy, respectively on the left and on the right.
Figura 3.2 – Le strategie di campionamento casuale pura (a sinistra) e casuale

stratificata (a destra).

Such methodology is based on the random location of each sample in each cell of the regular grid previously created over the domain. In this way the complete coverage of the area is guaranteed by the regular grid and the unbiasedness of the investigation is assured by the randomness of samples location within the grid.

The main problem regarding the bias introduced by sampling strategy regards the relationship between the scale of variability of the field and the scale of the exploration grid. When the frequency of the sampling method is so precise, the information at that scale are redundant, while the ones at other scales are scarce or totally missing. This is the key concept will be discussed in this section. In particular, variogram will be used to test how the stratified random sampling can be moulded as the optimizing method for generic spatial analysis.

1.3 The effect of sampling strategy

For what concern the influence of sampling approach on the experimental variogram, two main features will be analyzed in details: i) the small scale lags variability and ii) the pairs abundance.

Looking at the two variograms computed respectively on a systematic and on a stratified random sampled variable, we can note how the main differences regard the smallest scale investigable and the distribution of pairs abundance among the different spatial scales.

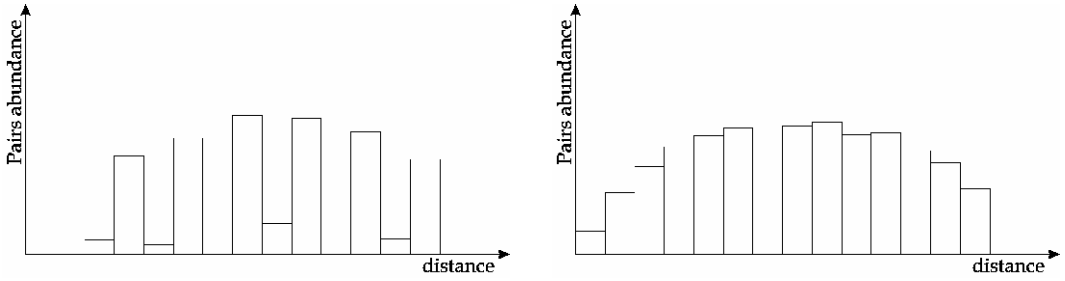


Figure 3.3 – Effect of the different sampling strategies on the pairs abundance.
Figura 3.3 – Effetto delle diverse strategie di campionamento sulla disponibilità delle coppie.

In particular, it is clear how, in the case of systematic sampling the smallest lag is far to be the smallest investigable spatial scale and the pairs abundance is biased

through the different spatial scales. There are many pairs at the multiples of the grid resolution value and very few at the intermediate scales.

Such relevant features are very important for the computation of experimental variogram, that is deeply influenced by the pairs distribution over the lags and that can be consequently interpreted in a more or less correct way.

1.4 The simulation test

In order to test the differences between the systematic and random sampling strategies and to analyze their influence on the variogram computation, a simulation test is proposed. It consists on the simulation of several iterative sub-samplings of a known artificial surface with different grid resolutions, and the computation of variographic analysis on them.

1.4.1 The artificial surface

An artificially surface has been built with Matlab®, simulating a stationary random function on a 350 x 350 mt. grid with 1 mt. resolution (Hu el. source).

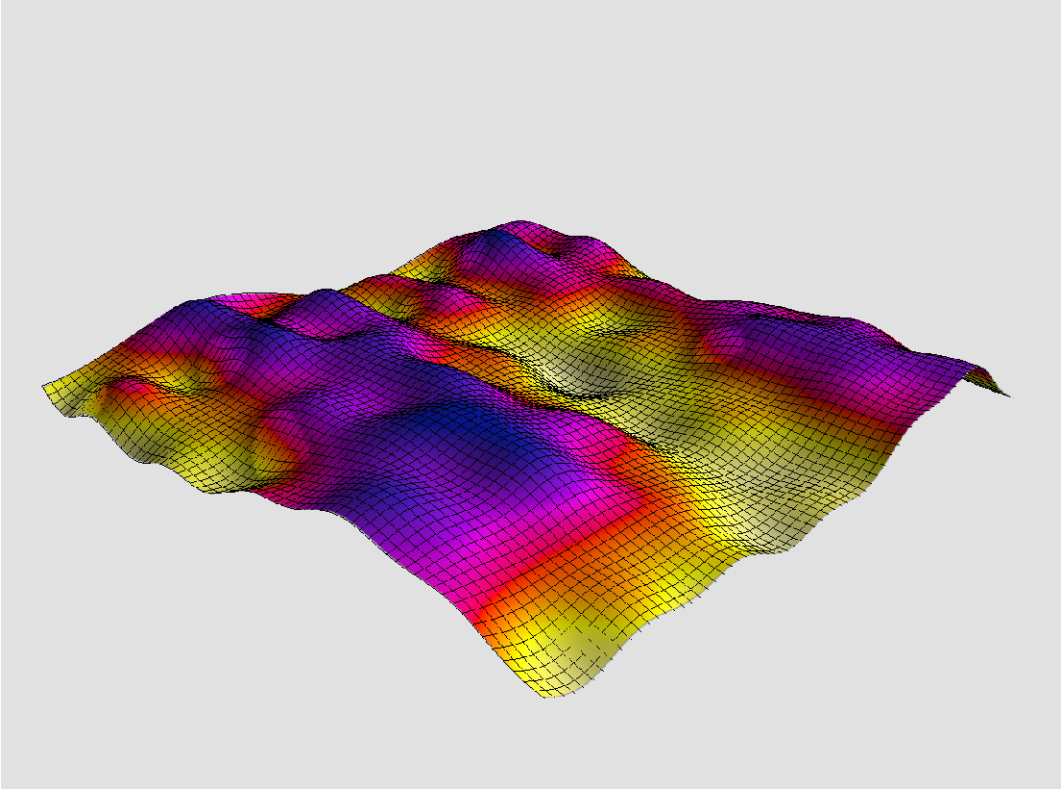


Figure 3.4 – The artificial surface.

Figura 3.4 – La superficie artificiale.

The surface presents several small oscillations, with an homogeneous small scale variation of the variable. It is stationary and its two-dimensional polynomial global trend is an horizontal plane. Histogram of the variable shows a normal like behaviour, with mean and median values coinciding and a skewness of -0.04.

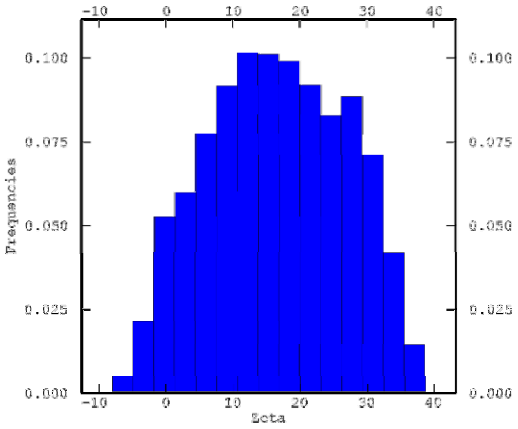


Figure 3.5 – Histogram of test variable 'Zeta'.
Figura 3.5 – Istogramma della variabile test 'Zeta'.

BASICS STATISTICS FOR 'Zeta' variable	
Min	-7.93
Max	38.57
Mean	16.60
Median	16.55
Variance	104.58
Standard deviation	10.23
Skewness	-0.04
Kurtosis	2.11

Table 3.1 – Basic statistics for test variable 'Zeta'.
Tabella 3.1 – Statistiche di base per la variabile test 'Zeta'.

Experimental variogram computed for East-West and North-South directions (lag = 10 mt.; # of lags = 35), reveals a slight geometric anisotropy, with the northward range larger than the southward one.

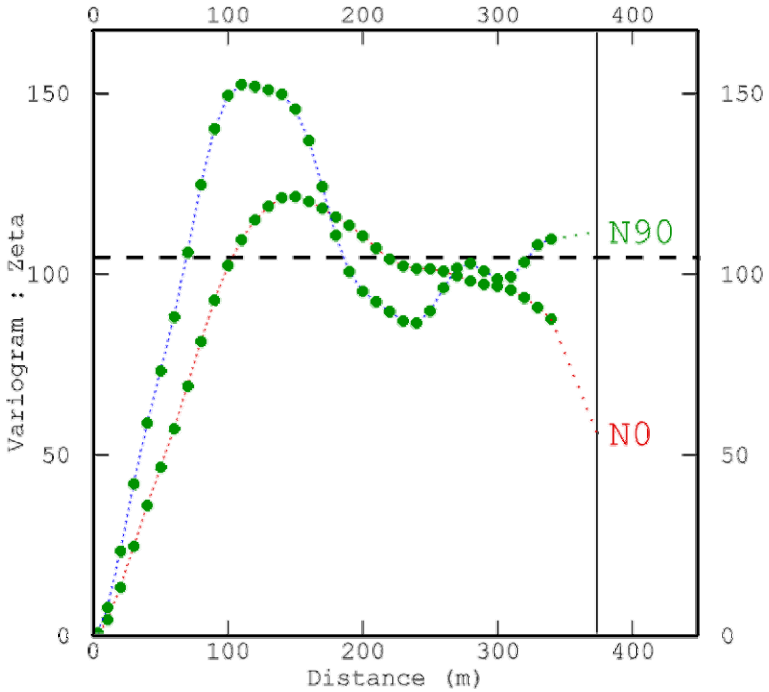


Figure 3.6 – Experimental directional variogram of test variable 'Zeta'. Lag value = 10 mt.; # of lags = 35.
Figura 3.6 – Variogramma direzionale sperimentale della variabile test 'Zeta'. Lag = 10 m. # di lags = 35.

Each structure shows an evident periodicity, reflecting the periodic variation of the surface with larger oscillations in eastward direction. As confirmed by the observation of the shape of the surface, the variograms reveal a very homogeneous variability at small scale lags, with a parabolic behaviour within the first 20-30 mt. In general, the experimental variograms are very regular, because of the continuity and large abundance of information.

One interesting aspect is the frequency of the oscillations of the variograms. In both the directions the semivariance values complete a whole oscillation within the 250 mt. lag.

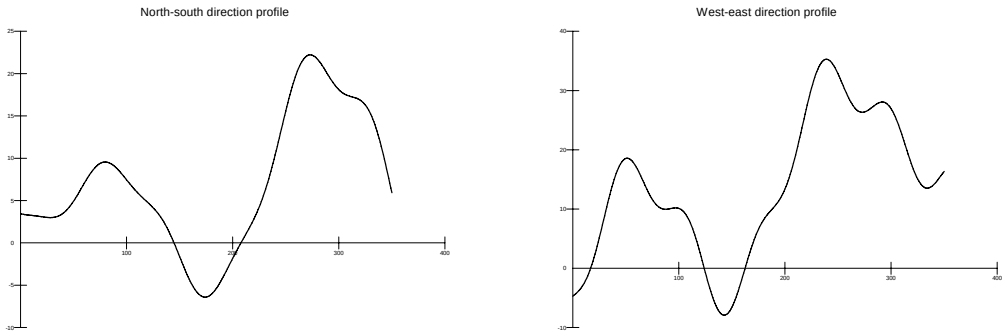


Figure 3.7 – Orthogonal profiles for the test surface.
Figura 3.7 – Profili ortogonali della superficie di test.

If we observe the profiles for the two directions, we can note just an oscillation period of about 200-250 mt. This is the confirmation of the importance of variogram as instrument for structural analysis of spatial variables.

1.4.2 The subsampling procedures

As mentioned before, in order to examine the differences between the two kind of sampling procedures (the systematic one and the stratified random one), an iterative subsampling of the original dataset has been implemented with different spatial resolutions. The ISATIS package from geovariances allows to compute two kinds of sampling selections on points: i) a regular one in which, once defined the resolution, samples are selected at the centre of each cell, and ii) a stratified random one in which, once defined the grid, the samples are chosen randomly within each cell (Malyuk el. source).

Subsampling has been implemented for resolutions 2, 5, 10, 20, 30, 40, 50, 60, 70 mt., respectively for regular and random method, and experimental variogram has been computed for each subset.

In the figure 3.8, an example of regular and random subsampling for 20 mt. resolution is shown.

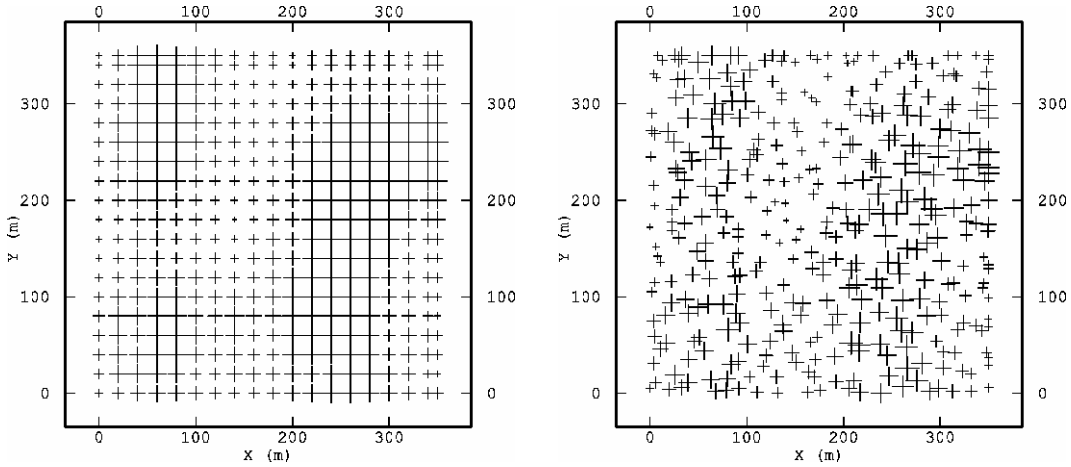


Figure 3.8 – Two examples of systematic and stratified random sampling techniques, respectively on the left and on the right.
Figura 3.8 – Due esempi di tecniche di campionamento sistematica (sinistra) e casuale stratificata (destra).

It is clear how the total amount of spatial scales covered by random sampling is much more than the one of regular sampling, while complete coverage of the spatial domain is generally guaranteed by both the methods.

Structural analysis has been computed on both the series of sub sampled datasets and a general decrease of reliability of variograms has been observed. The main concern is just the way and the rapidity with which such variograms lose meaning.

Omnidirectional variograms have been computed and compared with the original one.

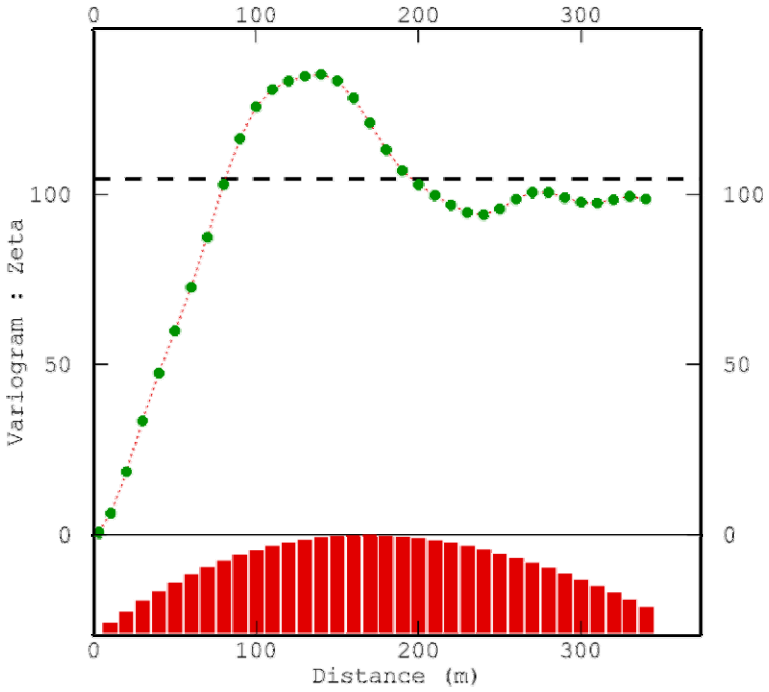


Figure 3.9 – Omnidirectional experimental variogram of test variable 'Zeta'. Lag = 10 mt.; # of lags = 35.
Figura 3.9 – Variogramma omnidirezionale sperimentale per la variabile test 'Zeta'. Lag = 10 m. # di lags = 35.

Regularly and randomly sampled subsets are respectively on the left and on the right side of the following figures.

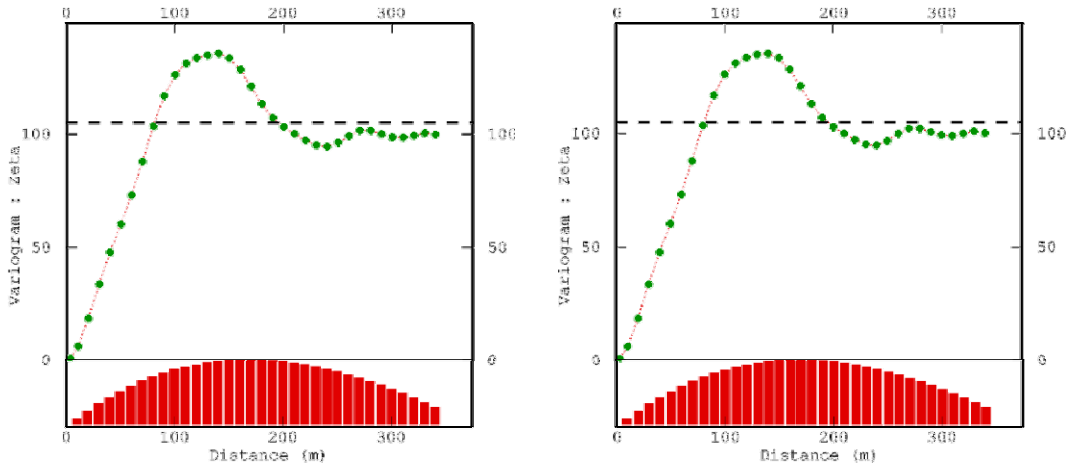


Figure 3.10 – Omnidirectional variograms of sub-sampled test variable 'Zeta'. Systematic and stratified random sampling respectively on the left and on the right. Sub-sampling resolution = 2 mt.
Figura 3.10 – Variogrammi omnidirezionali della variabile test 'Zeta', subcampionata secondo la tecnica sistematica (sinistra) e quella casuale stratificata (destra). Risoluzione = 2 m.

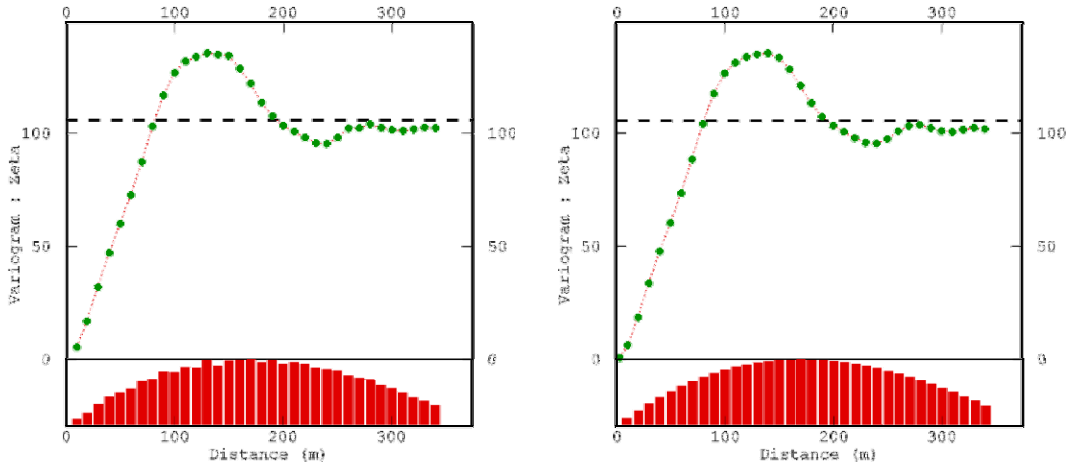


Figure 3.11 – Omnidirectional variograms of sub-sampled test variable 'Zeta'. Regular and stratified random sampling respectively on the left and on the right. Sub-sampling resolution = 5 mt.

Figura 3.11 – Variogrammi omnidirezionali della variabile test 'Zeta', subcampionata secondo la tecnica sistematica (sinistra) e quella casuale stratificata (destra). Risoluzione = 5 m.

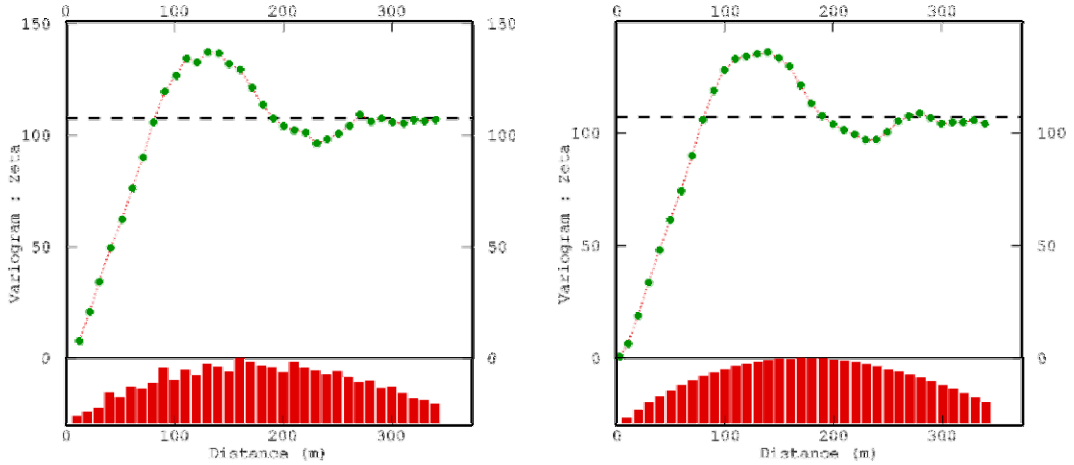


Figure 3.12 – Omnidirectional variograms of sub-sampled test variable 'Zeta'. Regular and stratified random sampling respectively on the left and on the right. Sub-sampling resolution = 10 mt.

Figura 3.12 – Variogrammi omnidirezionali della variabile test 'Zeta', subcampionata secondo la tecnica sistematica (sinistra) e quella casuale stratificata (destra). Risoluzione = 10 m.

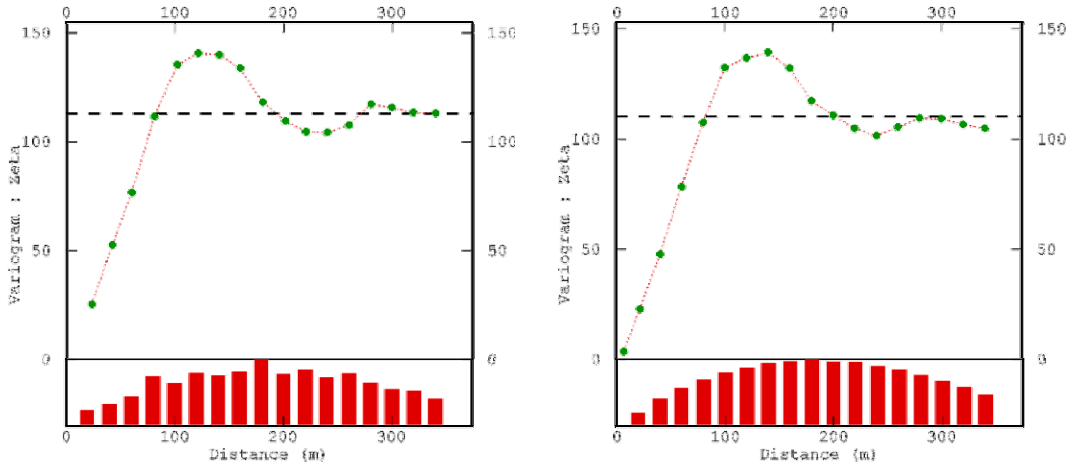


Figure 3.13 – Omnidirectional variograms of sub-sampled test variable 'Zeta'. Regular and stratified random sampling respectively on the left and on the right. Sub-sampling resolution = 20 mt.

Figura 3.13 – Variogrammi omnidirezionali della variabile test 'Zeta', subcampionata secondo la tecnica sistematica (sinistra) e quella casuale stratificata (destra). Risoluzione = 20 m.

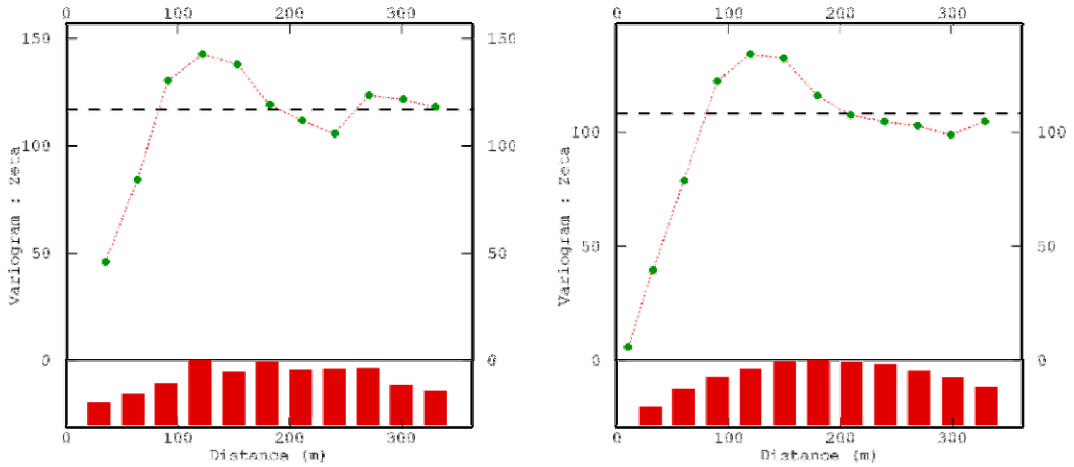


Figure 3.14 – Omnidirectional variograms of sub-sampled test variable 'Zeta'. Regular and stratified random sampling respectively on the left and on the right. Sub-sampling resolution = 30 mt.

Figura 3.14 – Variogrammi omnidirezionali della variabile test 'Zeta', subcampionata secondo la tecnica sistematica (sinistra) e quella casuale stratificata (destra). Risoluzione = 30 m.

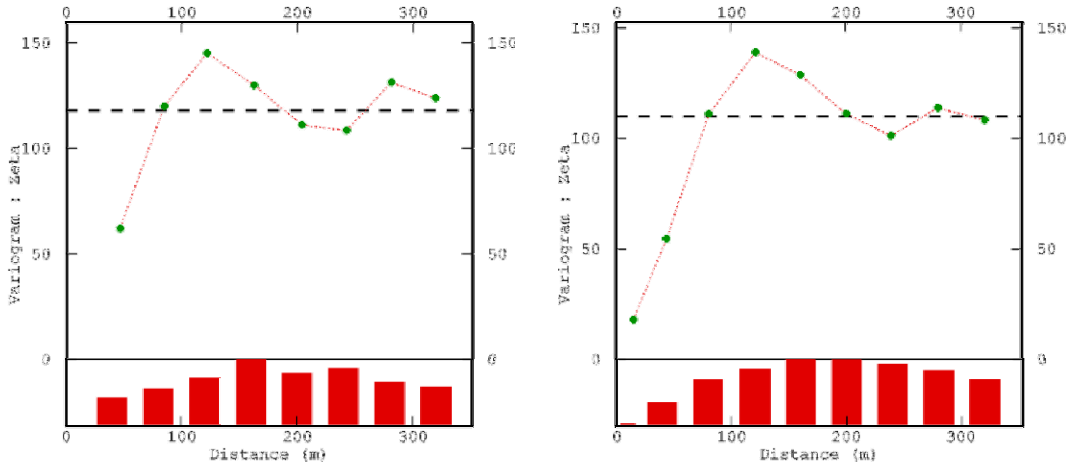


Figure 3.15 – Omnidirectional variograms of sub-sampled test variable 'Zeta'. Regular and stratified random sampling respectively on the left and on the right. Sub-sampling resolution = 40 mt.

Figura 3.15 – Variogrammi omnidirezionali della variabile test 'Zeta', subcampionata secondo la tecnica sistematica (sinistra) e quella casuale stratificata (destra). Risoluzione = 40 m.

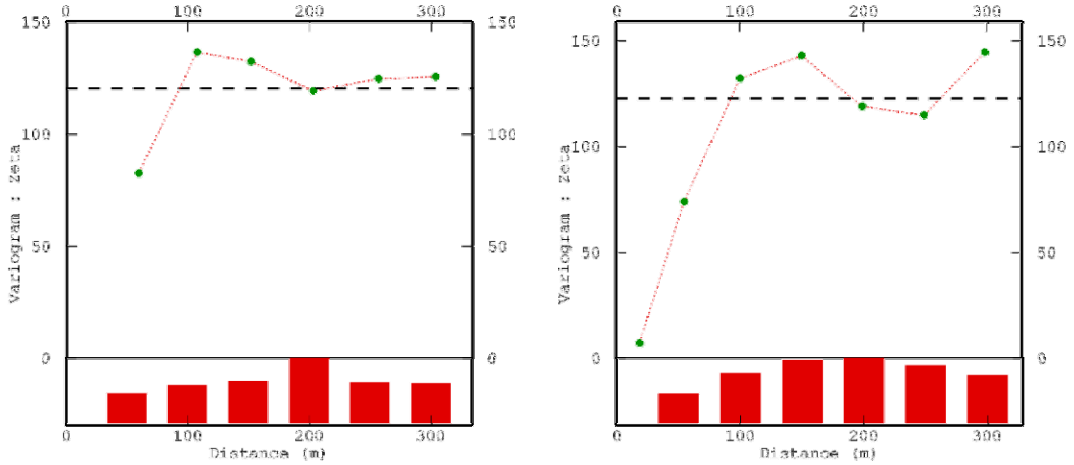


Figure 3.16 – Omnidirectional variograms of sub-sampled test variable 'Zeta'. Regular and stratified random sampling respectively on the left and on the right. Sub-sampling resolution = 50 mt.

Figura 3.16 – Variogrammi omnidirezionali della variabile test 'Zeta', subcampionata secondo la tecnica sistematica (sinistra) e quella casuale stratificata (destra). Risoluzione = 50 m.

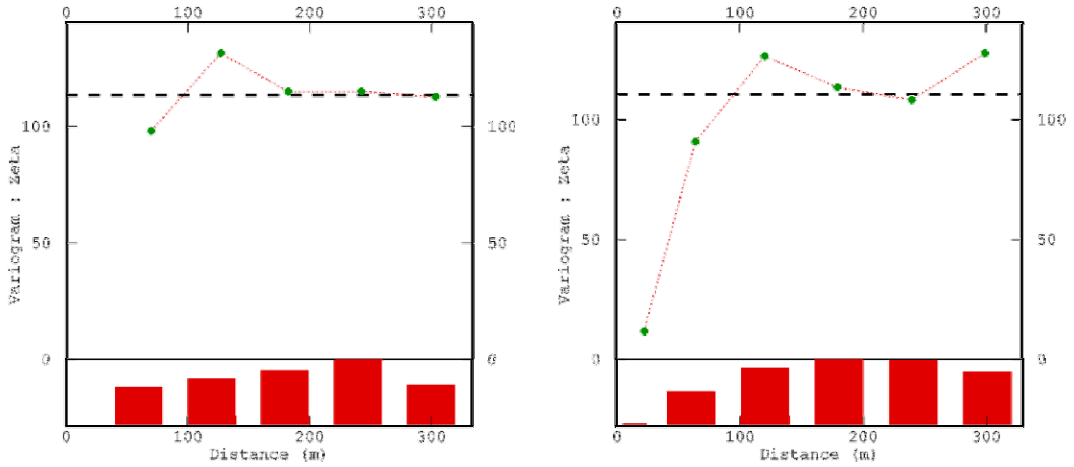


Figure 3.17 – Omnidirectional variograms of sub-sampled test variable 'Zeta'. Regular and stratified random sampling respectively on the left and on the right. Sub-sampling resolution = 60 mt.
Figura 3.17 – Variogrammi omnidirezionali della variabile test 'Zeta', subcampionata secondo la tecnica sistematica (sinistra) e quella casuale stratificata (destra). Risoluzione = 60 m.

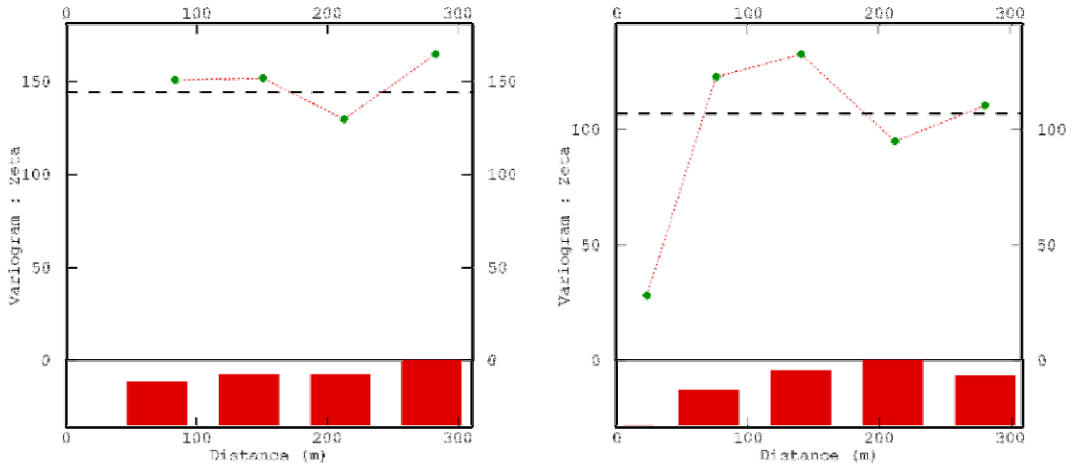


Figure 3.18 – Omnidirectional variograms of sub-sampled test variable 'Zeta'. Regular and stratified random sampling respectively on the left and on the right. Sub-sampling resolution = 70 mt.
Figura 3.18 – Variogrammi omnidirezionali della variabile test 'Zeta', subcampionata secondo la tecnica sistematica (sinistra) e quella casuale stratificata (destra). Risoluzione = 70 m.

The results of variographic analysis reveal the presence of many differences in the behaviour of the two groups of subsets. First of all the change in the histograms of the pairs abundance. As the subsampling resolution decreases, the number of pairs available for semivariance computation decreases uniformly in random approach and irregularly in systematic one. In particular, an overabundance of pairs, for lags coinciding with sampling scale or multiple of it, is relevant in regular

approach. The uniformity of informations is crucial for an unbiased computation of semivariance values, that, in this case, results poorly reliable. The irregular oscillation of variogram values, as in 10 mt. scale in regular sampling, can be related just to such unbalanced abundance of pairs. Conversely, in random sampling, the pairs abundance decreases in regular way, so that, even the variogram being less reasonable, the semivariance values contributing to its shape are wholly consistent and affordable.

Another very important aspect is the characterization of the small spatial scale variability, that is completely lost in regular sampling, after the first sampling scales. The systematic sampling strategy assumes to have one information on each cell of the grid and consequently to fix the smallest lag exactly to the value of the sampling grid resolution. It leads to the complete lack of information regarding the micro structural variability and to the quantification of high nugget effects, often not reflecting the reality.

Looking at the experimental variograms of figure 3.10 - 3.12 (from 2 mt. to 10 mt. scales) we can note how the small scale variability in regular sampling is lost immediately and some nugget effect is modelled. Conversely, random sampling represents correctly the uniform variation of the first lags, that reveals the homogeneity of the variable (compared with variogram of raw data).

Going on with subsampling scale, we can note how the nugget effect modelled by systematic sampling variograms is even more evident, while random subsets show only a bit increase of it. Eventually, at the scale 70 mt. the systematic sampling variogram is somewhat a pure nugget model, while the random one still preserves the correct shape.

To better understand the differences between the two approaches, we can observe the directional variograms for the 50 mt. sampling scale. In the figure, the systematic and the stratified random sampling strategy, respectively on the left and on the right, are represented. Lag = 25 mt. # of lags = 14.

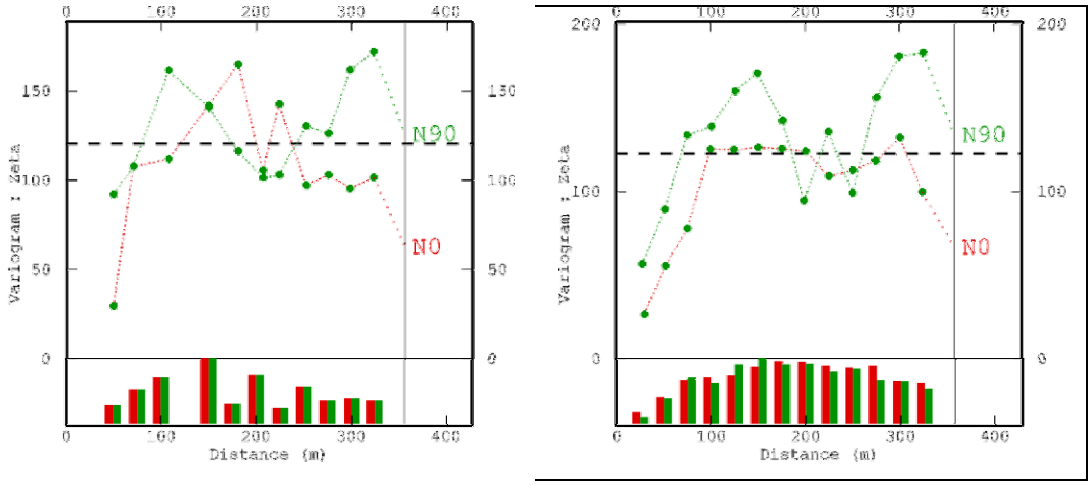


Figure 3.19 – A zoomed image of the directional experimental variograms for the sub-sampled test variable 'Zeta', at the resolution value of 50 mt. Lag value = 25 mt.; # of lags = 14.
Figura 3.19 – Un'immagine ingrandita del variogramma sperimentale direzionale per la variabile test 'Zeta' subcampionata con risoluzione 50m. Lag = 25 m. # di lags = 14.

It is clear how, in the case of directional variograms (with zero angular tolerance) the differences between the two methods are even more evident. First of all, the pairs abundance are hardly biased and completely absent for lags 50 and 125 mt. in systematic approach and then the shape of the variograms that are much less meaningful and irregular. In stratified random strategy, the structure is hold and the original anisotropy is honoured too.

1.4.3 Discussion

An iterative subsampling of a known surface with 1 mt. resolution, for 9 different grid resolution, form 2 to 70 mt., using a systematic regular sampling strategy and a stratified random one has been simulated. Structural analysis on the subsets has been implemented, computing omnidirectional variograms for each scale and comparing results with the one computed on raw data. Since the first subsets, for which the grid resolution is reasonably high, the differences among the two approaches appear clear.

The systematic sampling approach is not able to keep the information of the small scale variability and loses it, assuming a increasing nugget effect (that is not real). Meanwhile it presents several unreliable oscillations, due to the biased amount of pairs, that are denser at the scale equal to the sampling one or at the multiples of it, and fewer in the others. Conversely, variograms computed on

randomly sampled subsets, preserve the original shape of the raw data variogram and, even decreasing in details with the increase of sampling scale, they represent conveniently the real variability structure of the variable.

Such crucial differences reveal the importance of the choice of sampling strategy on environmental analysis planning. Especially for what concerns the spatial estimation of environmental variables, the correct and reasonable reproduction of the real structure of the field is a crucial aspect that is mostly reflected on the interpretation practice. As seen on results of simulation, an unreliable image of the variable represents the base for a incorrect understanding of the real natural processes.

If we consider the economic aspect, we can observe how, with the same economic waste, we can obtain more detailed informations on the subject, simply choosing to locate the samples in a random way and not over a systematic grid. Moreover, with stratified random sampling, we can reduce the number of samples, and obtain the same degree of quality of knowledge on the variable that we could obtain only by increasing consistently the economic waste.

One concrete application is the geochemical datasets of Port of Naples, coming from the I.A.M.C. C.N.R. (The Coastal Environment Institute of the National Research Council), sampled following the national directions of I.C.R.A.M. (The Italian Institute for Marine Research). Most of the docks of the inner part of the port have been examined implementing a detailed program for monitoring the distribution of pollutants in marine sediments coming from the commercial divisions of the harbour. Following economic accessibilities of the Institute and the national standard for sampling strategies, a systematic grid with 50 mt. resolution has been applied on the site.

In the “Darsena di Levante” dock, with its squared shape 350 mt. wide, such sampling plan has led to a total amount of about 50 samples, regularly distributed trough the spatial domain. The number of samples could seem enough, in such a small area, but the resulting experimental variogram does not reflect this assumption (in the figure the experimental variogram for vanadium computed with lag = 50 mt. and # of lags = 6) .

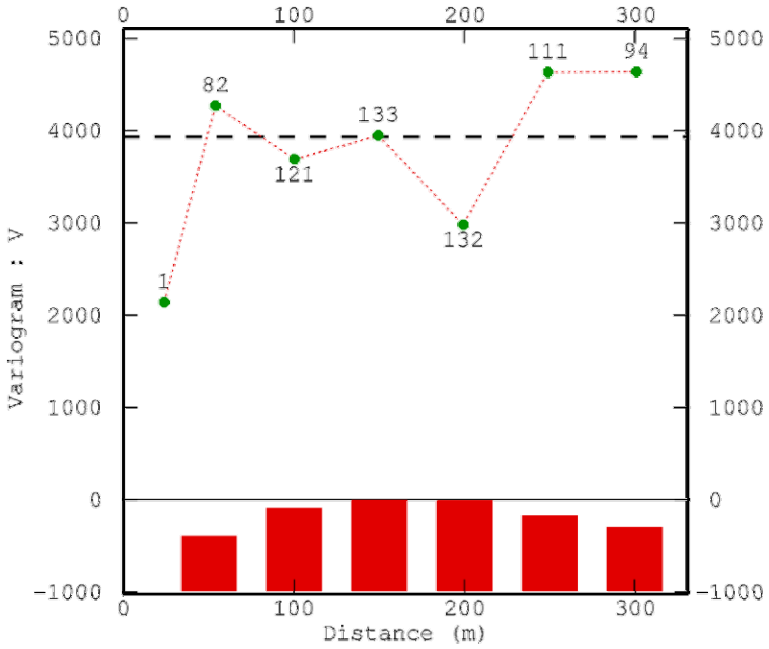


Figure 3.20 – Experimental omnidirectional variogram for vanadium in the 'Nuova Darsena' dock in port of Naples. Lag value = 50 mt.; # of lags = 6.
Figura 3.20 – Variogramma sperimentale omnidirezionale per il Vanadio nella 'Nuova Darsena', nel porto di Napoli. Lag = 50 m. # di lags = 6.

The shape of the variogram is not so far from a pure nugget model and its use for spatial modelling is very hard. The semivariance value at the first lag should not deceive because the amount of pairs contributing to such calculus is only one. If we observe the comparison between regular and random sampling in figure 3.16, for the simulation test at 50 mt. sampling scale, we can note how the results are comparable with the one obtained in a real application. Almost certainly, if they had chosen a stratified random sampling strategy for such dock, they could have obtained a more useful and reasonable description of the real spatial structure of the variable, investing the same resources.

2. Improving the performances of an existing sampling plan

Sometimes happens that the spatial analysis of sampled data reveals some lack in coverage of the whole spatial structure of the variable. Such lack is often related to the sampling strategy that is not able to catch entirely the behaviour of the field. In this case, an infilling procedure is rather needed, in order to cover the existing holes. Assumed to have the chance to take into account some additional samples, the problem is where to locate them (Pilger, Costa, Koppe).

Many authors have discussed on this argument, presenting several different hypotheses. The one presented here is a somewhat discussion of the work of Pilger, Costa, Koppe, to which the comparison between estimation variance and nugget variance is added.

As mentioned before, geostatistical simulations have been used to locate the area with maximum dispersion in kriging estimations, in order to suggest the sites where locating additional samples. An iterative check of the consistence of variogram has been computed in order to minimize such increase of sample amount.

The main concern is related to the definition of the nugget variance of experimental variogram. As defined in previous chapter, it is regarded as the sum of the intrinsic small scale variability of the variable and the analytic uncertainness of laboratory analyses. Just the relationship between these two factors is a trivial concept in geostatistics, whereas the quantification of the nugget effect is regarded as an instrument to evaluate the contribute of the analytic uncertainness to the irregularity of the small spatial scales in map estimation.

In particular, the zero lag semivariance value, assumed to be always zero, can be computed as a positive value only if we consider two different results of laboratory analysis. There will never exist an analytic procedure that will guarantee zero errors for any measure; it is intrinsic in any measurement process.

Thus, the minimum value of nugget effect must be fixed to exactly the analytic error variance:

$\lim_{h\to 0} \mathrm{Var}(h) = Var_S$. Nugget effects less than such quantity are not reasonable.

When experimental variograms show very high values of nugget effect (if compared with analytic variance), the aim is to propose an infilling procedure that minimizes the numbers of additional samples, reducing the nugget variance until it reaches the estimated value of analytic variance.

3. The case of 'Nuova Darsena' dock

The Nuova Darsena is one of the dock of the harbour of Naples, where the I.A.M.C. (Coastal Environment Institute of the Italian National Research Council)

planned an investigation program to asses the quality of the waters and levels of pollutants accumulation in marine sediments. The dock is a square tub 350 mt. wide, located in the inner part of the port, in front of the internal edge of the offshore wharf.

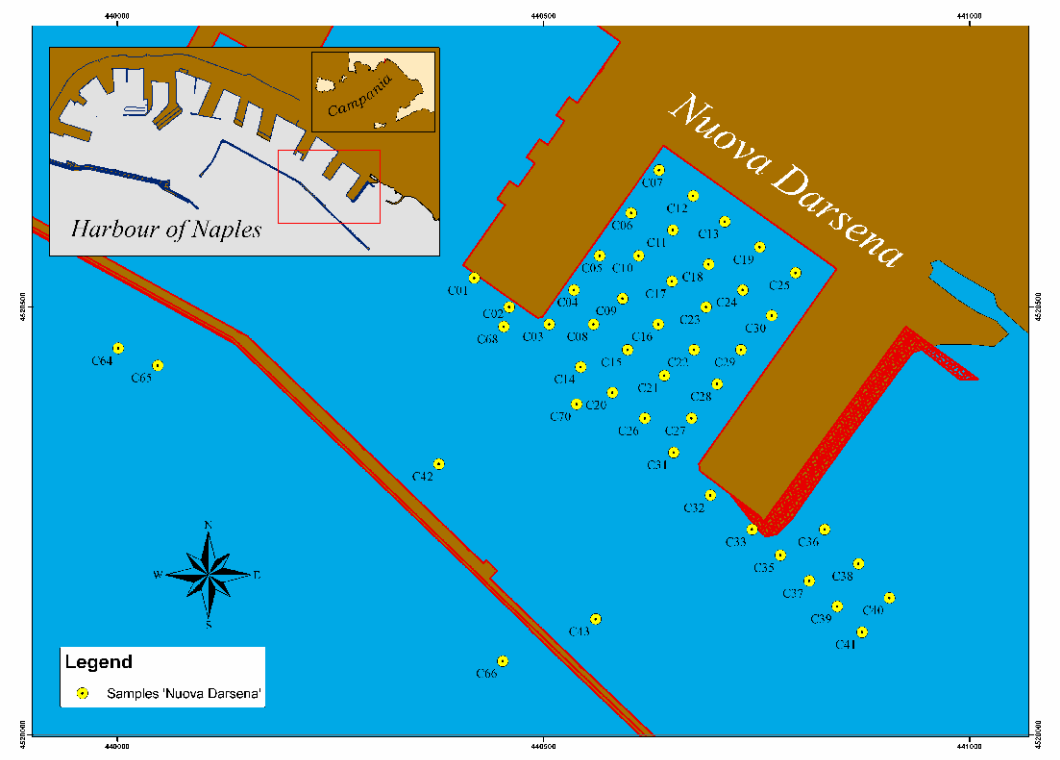


Figure 3.21 – Location map of the samples in 'Nuova Darsena' dock, in the port of Naples
Figura 3.21 – Mappa di ubicazione dei campioni nella 'Nuova Darsena', nel porto di Napoli.

The sampling strategy chosen is a systematic one with 50 mt. grid resolution, with a total amount of 30 samples inside the dock and 17 out of the basin. Each sample consists on a corer 3,00 mt. long from which 5 different levels have been identified. Experimental variogram has been computed for iron, with lag = 50 mt. and # lags = 7 for the first level (0 - 20 cm.).

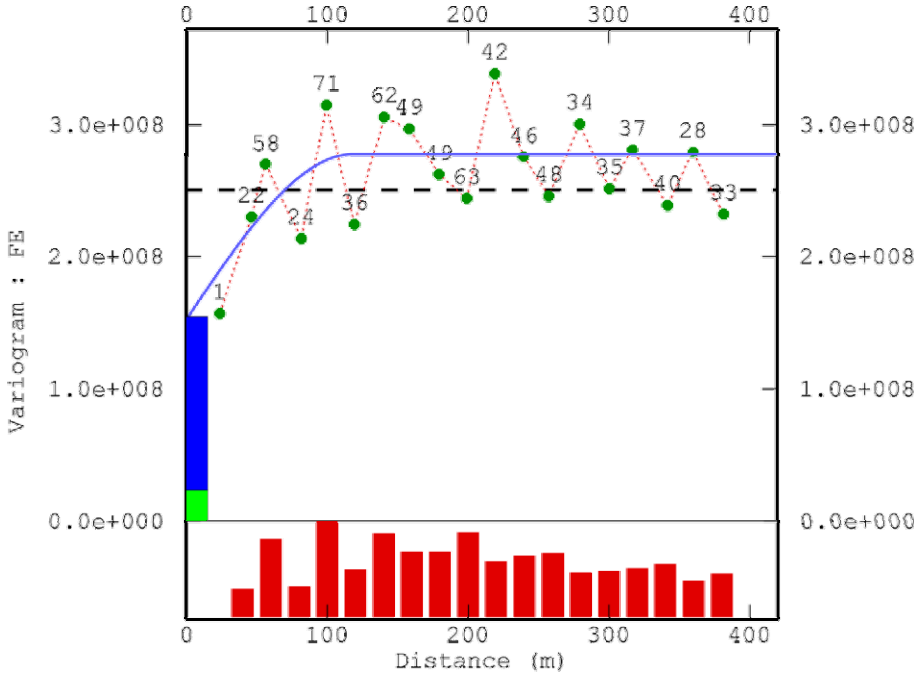


Figure 3.22 – Experimental omnidirectional variogram for Iron. Lag = 50 mt.; # of lags = 7.
Figura 3.22 – Variogramma sperimentale omnidirezionale per il Ferro. Lag = 50 m. # di lags = 7.

As clear from the figure, the nugget effect (~1.5e008) is regarded as the sum of analytic error variance (computed as 10% of the median value of the raw data), in green and the intrinsic small scale variability of the variable, in blue. We can note how the bars of pairs abundance are biased with unbalance among lags coinciding to the sample scale and its multiples and the others lags. Such bias could be even more evident by computing directional variograms on a rotated reference axis, but it is not the target of the work.

Especially when the target is an environmental variable it is assumed that high variations at very small scales are unreliable. Thus, the non zero value of the nugget effect, rather than be constituted by analytic variance, can be explained only by the sampling strategy that has been not able to model correctly the regularity of the variable.

3.1 The cyclic approach

The aim is to reduce the nugget variance until it reaches the analytic variance. Geostatistical simulations can come in our aid.

The approach is iterative and assumes i) a cyclic computation of Sequential Gaussian simulations (SGS), with the location of additional sample located in the cell that shows averagely the highest variability, ii) the selection of such sample from the kriged map and iii) its addition to the original dataset and eventually iv) the computation of experimental variogram of the updated dataset (Pilger, Costa et al. el. source). At each step, the varying nugget variance is compared with the analytic variance and the process is stopped when the two values converge.

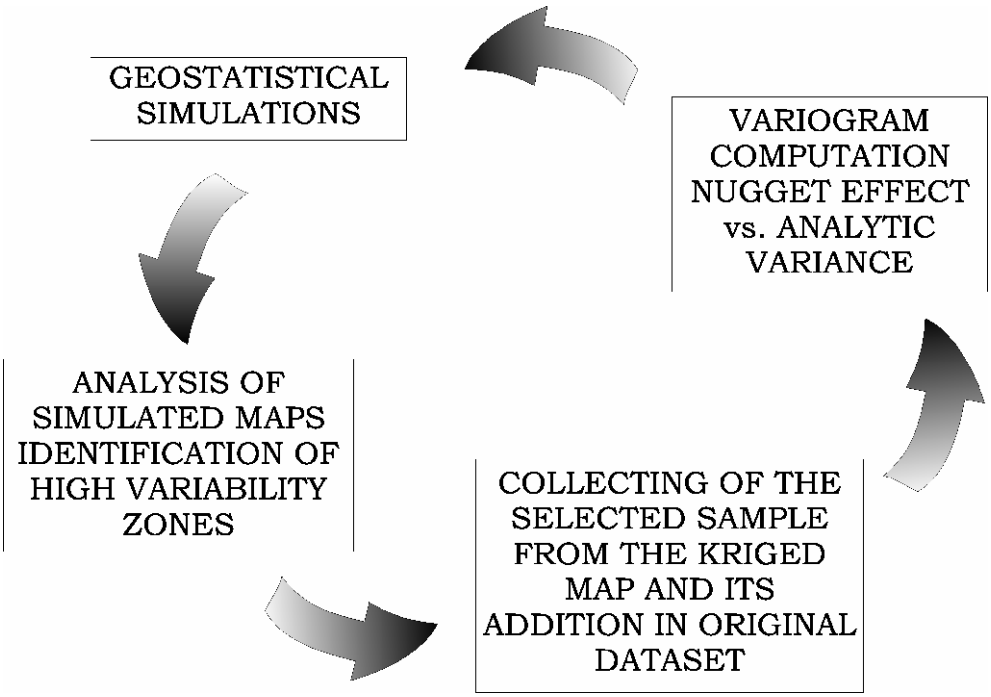


Figure 3.23 – Conceptual scheme of the procedure used to locate the additional points.
Figura 3.23 – Schema concettuale della procedura utilizzata per ubicare i campioni addizionali.

When the aim is to infill the original dataset with real information, the new sample can be collected directly on the site, and not extracted by kriged map. In this way the sampling strategy can be defined as dynamic, composed by a preliminary exploration step and a further infilling with additional samples. Obviously the synopticity of the sampling process becomes a driving factor.

In the case of Nuova Darsena, 100 Sequential Gaussian Simulations have been computed on Iron concentration values, for the superficial level, and the area (the

grid cell) with maximum standard deviation of the 100 simulated maps has been highlighted.

In the figure the first, the 10th, the 50th and the 80th simulation are represented as examples.

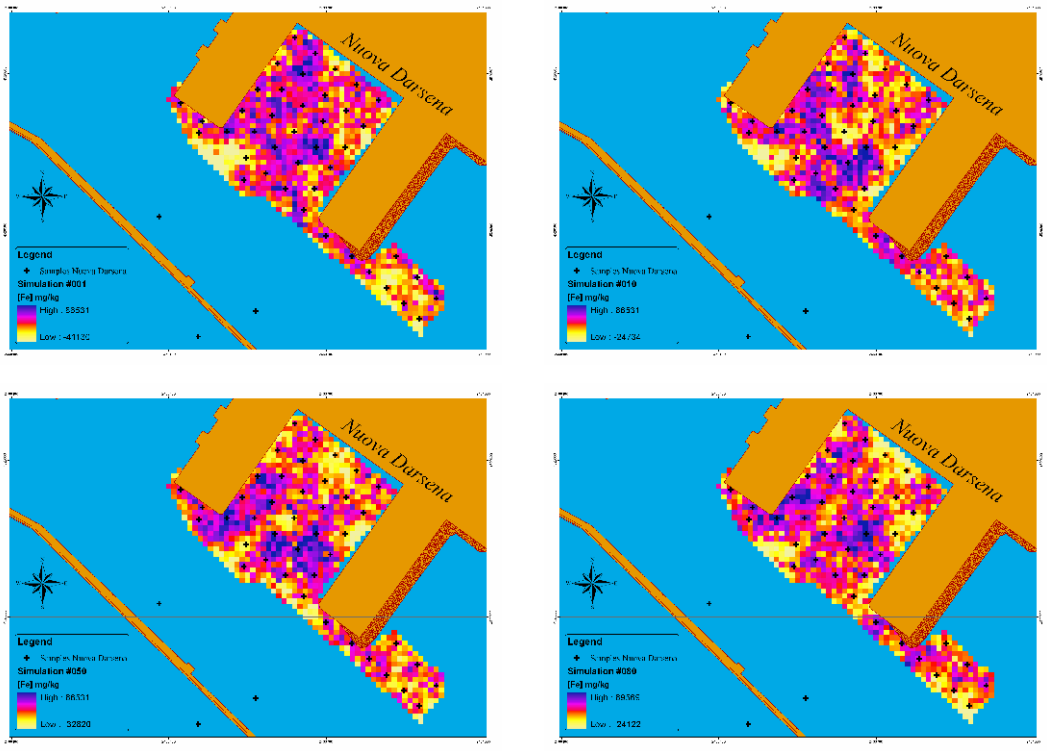


Figure 3.24 – Four examples of simulated maps. The first, the 10th, the 50th and 80th simulated map, respectively from the top-left to the down-right.
Figura 3.24 – Quattro esempi di mappe simulate. Dall'alto a sinistra: la prima, la decima, la cinquantesima e l'ottantesima mappa simulata.

Negative values are explained by the negative weights of the simple kriging system, that cannot take into account the natural meaning of the variable. However it does not matter, because our use of simulations is focused on the quantification of the absolute value of spatial variability.

Summarizing the first 100 simulations, we can compute any statistic measure of dispersion. ISATIS allows to calculate the mean and the standard deviation value of the whole set of simulated map. The map of mean and standard deviation of realizations is represented in figure 3.25.

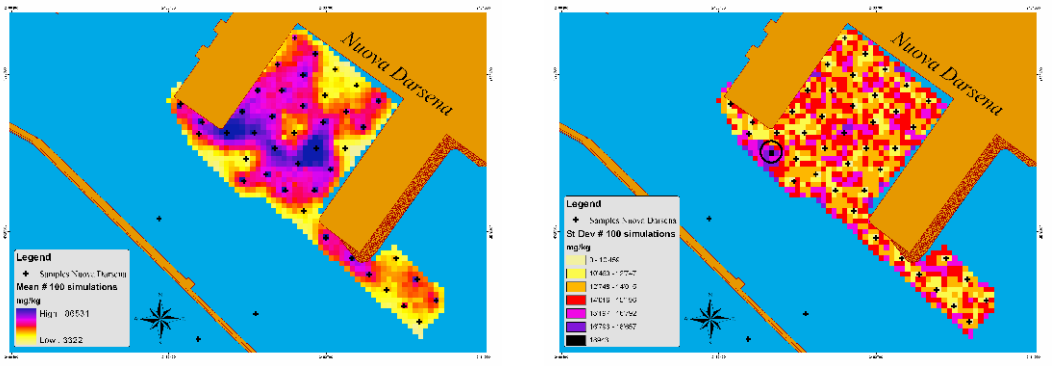


Figure 3.25 – The maps of mean and standard deviation of the 100 simulated maps, respectively on the left and on the right. In the black circle the cell with the maximum value of standard deviation.
Figura 3.25 – Le mappe della media (sinistra) e della deviazione standard (destra) delle 100 mappe simulate. Il cerchio nero indica la cella che presenta il picco di deviazione standard.

The highest score for standard deviation (18943 mg/kg) of the 100 simulations is located in the cell highlighted by the black circle. Thus, we will extract that cell value from the kriged map.

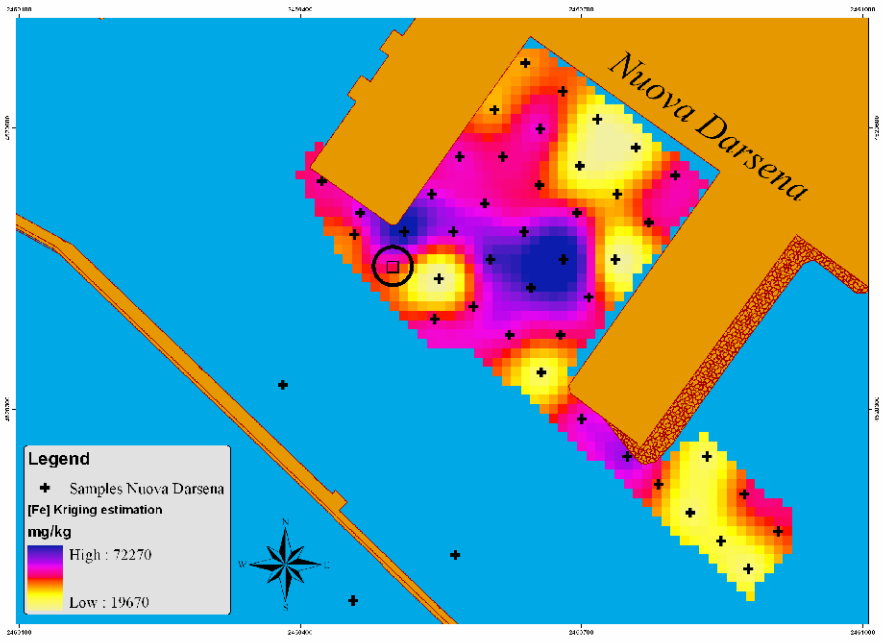


Figure 3.26 – The kriged map from which the additional samples has been extracted.
Figura 3.26 – La stima con kriging da cui il campione addizionale è stato estratto.

In that point, the iron estimation is 41694 mg/kg. We can add such value to the original dataset and compute the resulting variogram.

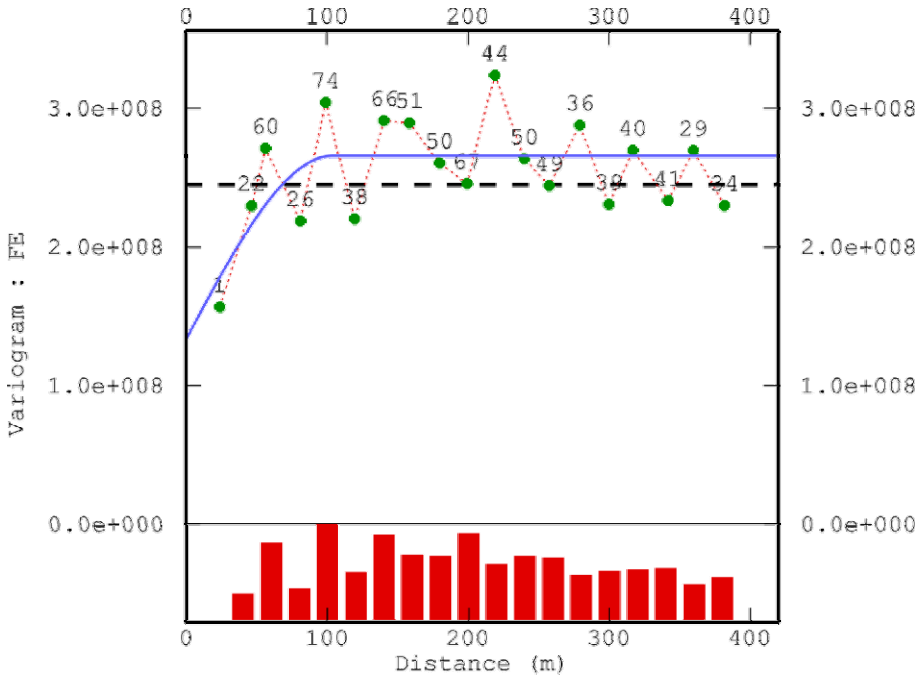


Figure 3.27 – The variogram of the dataset added of one point.
Figura 3.27 – Il variogramma del dataset arricchito di un punto.

The resulting nugget effect has decreased from 1.5e008 to 1.33e008. Kriging can be applied again to this new dataset and the cycle can be carried on checking the variation of the nugget effect. Following the same process, six samples have been added, obtaining a decrease of nugget effect until setting it to zero. In the figure the six added samples taken by the recalculated kriged map.

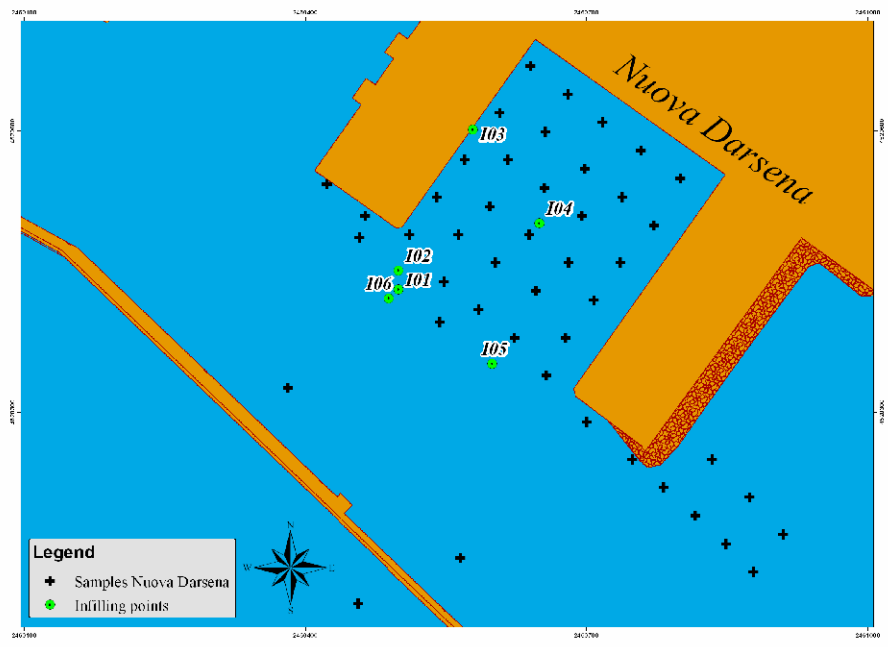
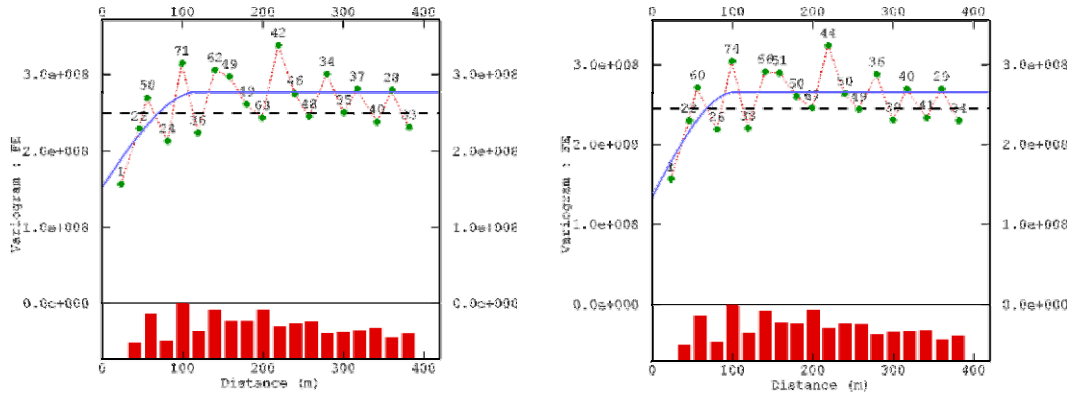


Figure 3.28 – The location map of all the six additional points.
Figura 3.28 – La mappa di ubicazione dei 6 campioni aggiuntivi.

3.2 The behaviour of nugget effect

Even being characterized by only few samples, the small scale variability decreases after the first additions and goes to zero with the sixth sample. In the figure the variograms computed after each addition.



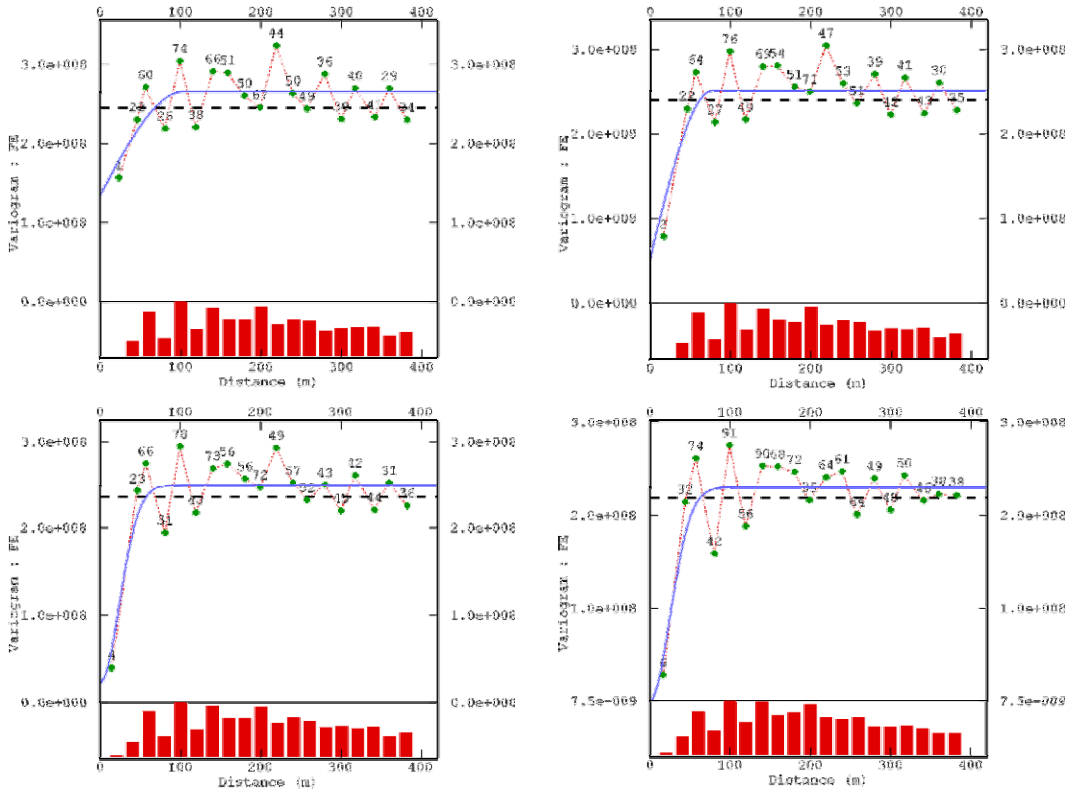


Figure 3.29 – The six variograms after the six addition, respectively from the top-left to the down-rigth.
Figura 3.29 – Dall'alto a sinistra: i 6 variogrammi relative alle sei aggiunte.

The plot of decreasing nugget variance reveals the relationship with analytic uncertainness variance (in red).

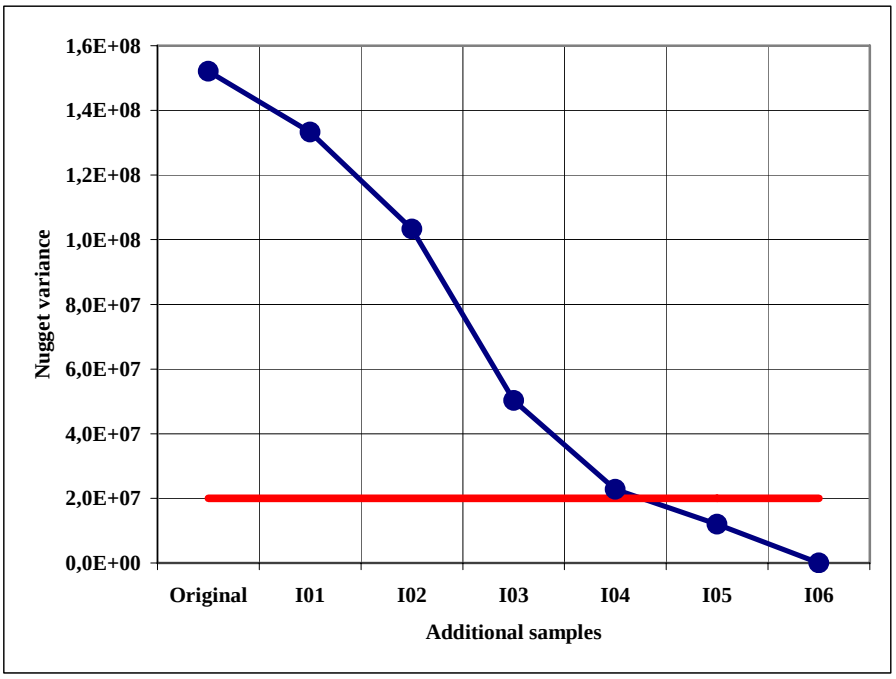


Figure 3.30 – Plot of the decreasing value of nugget variance with the increase of additional points. The threshold value is reached at the fourth point.
Figura 3.30 – Grafico del valore decrescente della nugget variance. Il valore soglia viene raggiunto al quarto punto.

Looking at the plot, we can note how at the fourth addiction, the nugget variance reaches the analytic variance and beyond such value, it decreases further. Following the principles defined before, we can stop the iterative process at the fourth sample and obtain a somewhat more reasonable variogram. In fact it would be useless to invest in further samples when analytic uncertainness introduces a contribute to the global variance that is larger than the one obtained just from the additional samples.

3.3 Discussion

The approach described for the infilling of an existing sampling plan wants to represent an indication on how to rebuild the sampling strategy for successive surveys. The main concern is the target of decreasing the nugget variance of modelled variogram, in order to reduce it to being comparable to the analytic variance. Such aim is obtained by infilling the sampling grid with some additional points that are located by defining the risk zones (Koppe) as the ones presenting the highest variability (with the aid of geostatistical simulations). It is clear that the

small scale variability computed with only few samples is poorly representative, but the resulting variogram is anyhow indicative.

CHAPTER 4
FILTERING SIDE SCAN SONAR MOSAICS

In most of branches of Science, one of the most popular problem is the image treatment. Very often the base of interpretation and scientific discussions is simply a confuse image. Microscopes and telescopes return images of very different scale phenomena and their interpretation is often based on somewhat confuse slide or film. It is clear how one of the central need is, thus, the practice of image processing, finalized to improve the reliability of data and to extract the needed information.

In this bound, mathematics and statistics have contributed with several methods, that in many cases returned optimal results and allowed to generate correct interpretation of different phenomena. One of the most common need is the removal of the noise. Salt and pepper noise is often present on images coming from scientific instruments, mainly due to the sensibility of film grain (ISO value) that must be set to high values in order to catch the details of the phenomenon. Band-pass filter, in deterministic approach, or k-order moment filter in statistics one, are the most commonly used approaches.

In this work, kriging of spatial component will be applied to an acoustic seabottom mosaic (Side Scan Sonar mosaic image), and used as filtering technique, in order to remove noises due to acquisition errors. Such kind of methodology, although often used to treat image, for which the salt and pepper noise is usual, must be carefully applied in these cases, in order to avoid to remove also the signals related to real structures.

1. Filtering with kriging

Kriging of spatial components is applied to a portion of a side scan sonar mosaic coming from a survey made in gulf of Augusta (Siracuse – Sicily) by the Coastal Environment Institute of the Italian National Research Council (I.A.M.C. C.N.R.).

Variographic analysis reveals a marked nugget effect in the variability structure of data, related to noises and spikes, widely spread through the domain. Kriging of spatial component, used as filtering technique, can rebuild the surface, filtering out

the small scale variance and obtaining a satisfying results (Bourgault 1994; Wen and Sinding-Larsen 1997). Most attention must be reserved to the use of such technique, in order to confuse real signals with noise and filter them out.

1.2 The side scan sonar dataset

Side scan sonar images are digital elaborations of the back scattering signal of the seabottom recorded by the instrument. A hundreds KHz order signal is pulsed downward by the transmitter and the back scattering wave is recorded by the receiver. The result is a digitally converted 8-bit (grey scale) map that represents the acoustic image of the seabottom. Often used to locate and to classify marine features, like sediment typology or different species of marine plants, side scan sonar is a even more diffuse technique for seabottom investigation. The lattice chosen for this work is 2.3 x 1.9 km wide, with 1 meter resolution, showing several sedimentological structures as erosive traces and pull-up depositional formations. Different grain size facies are present, going from coarse sand to pelitic clay.

The parcel chosen is a portion of a raw mosaic, made by joining numerous survey routelines.

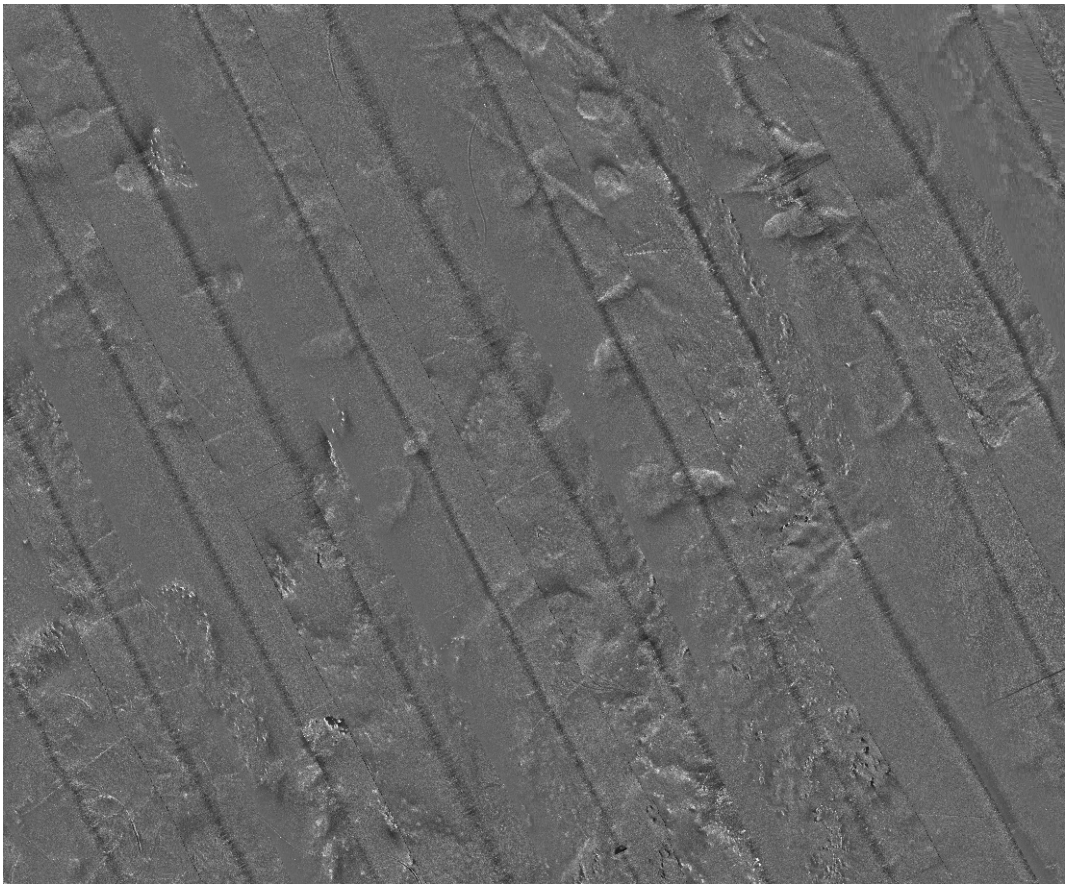


Figure 4.1 – The portion of Side Scan Sonar mosaic used as test data.
Figura 4.1 – Il taglio del mosaico Side Scan Sonar utilizzato come test.

The image reveals several sedimentary accumulation structures that are sufficiently clear, except for the presence of oblique lines revealing an imperfect removing of slant range signal (the line joining the two channel signals in side scan sonar traces).

Nevertheless, if we observe the image with a larger zoom, we can note the presence of a diffuse noise.

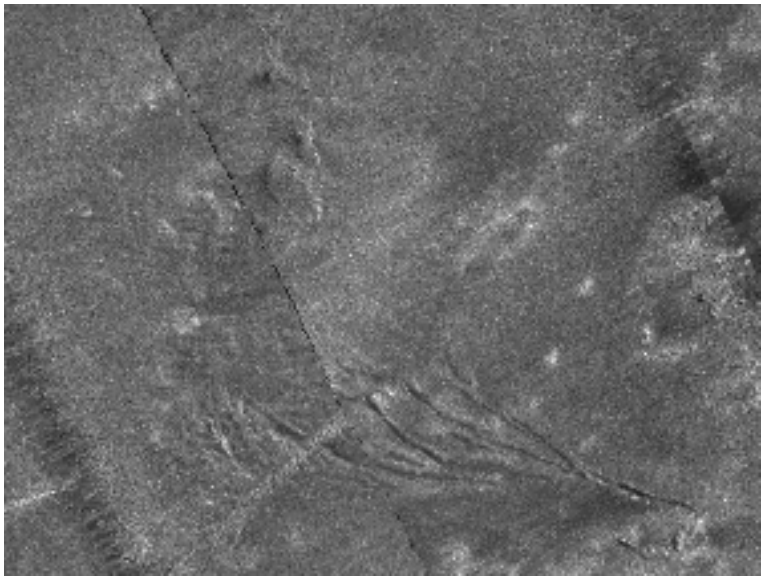


Figure 4.2 – A zoomed detail of the test mosaic, where the *salt and pepper* noise is evident.
Figura 4.2 – Un dettaglio ingrandito del mosaico test, dove il rumore sale e pepe è netto.

Such kind of noise is the typical *salt and pepper* noise, that is a generally a high frequency noise, with a defined variance, that is uniformly added to the original signal.

Analyzed from a geostatistical point of view, the grey scale value for each pixel can be regarded as a random regionalized function with the form:

$Z(x) = \mathrm{Y}(x) + \varepsilon$; where the variable $Z(x)$ is composed by a real random function $\mathrm{Y}(x)$, that is assumed to have a certain variability structure, and an uncorrelated noise ε with a defined variance.

We can observe the histograms and the qq-plot (the scatter plot of the quantiles of the two datasets).

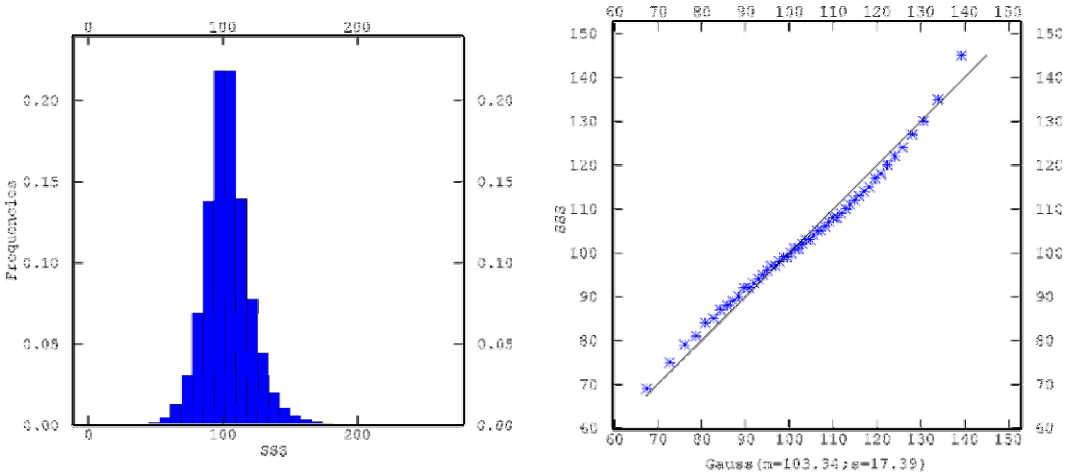


Figure 4.3 – The histogram and the *qq*-plot of the test mosaic.
Figura 4.3 – Istogramma e qq-plot del mosaico test.

The variable is almost normal-like, with some bias trough the highest classes, related to positive spikes. Its qq-plot, computed comparing it with a gaussian variable, with a mean of 103.34 and standard deviation of 17.39, confirms the normal-like behaviour. Basic statistics are represented in the table 4.1:

	SSS original (grey scale pixel value)
MIN	13
MAX	254
MEAN	103.34
MEDIAN	102
VARIANCE	302.57
STD DEV	17.39
SKEWNESS	0.73
KURTOSIS	6.52

Table 4.1 – Basic statistics for test mosaic.
Tabella 4.1 – Statistiche di base per il mosaico test.

1.3 Structural analysis

Due to the absence of correlation, we can regard the resulting variogram as the sum of different spatial component and treat them singularly with kriging (Bleines, Deraisme et al. 2004).

The resulting variogram, computed orthogonally in the two directions, with lag = 1 mt. and # of lags = 50, is represented in the figure.

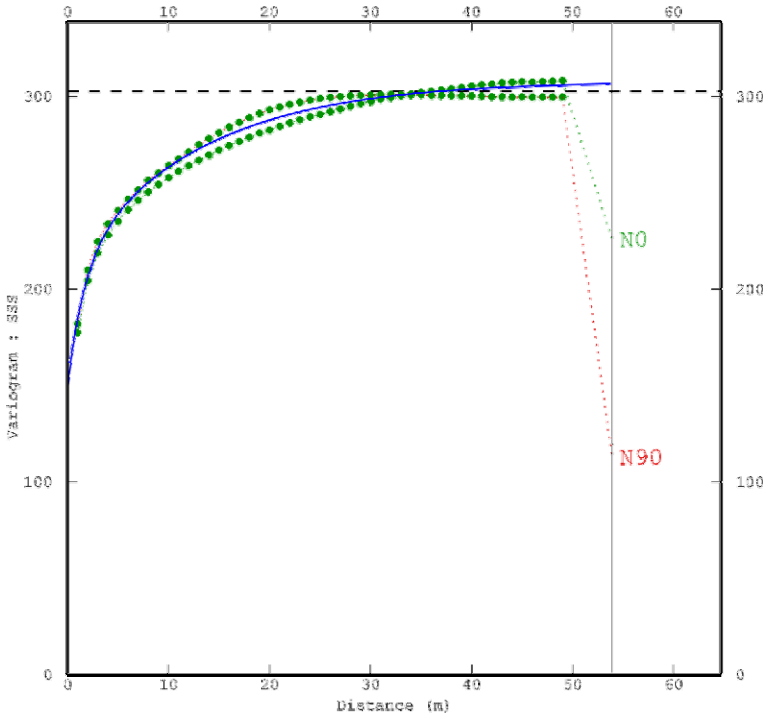


Figure 4.4 – Directional (orthogonal) variograms of test mosaic.
Figura 4.4 – Variogrammi direzionali (ortogonali) del mosaico test.

The coinciding shape of the two directional experimental variograms reveals a marked isotropy of the field. The experimental values can be modelled by two isotropic nested exponential models and a marked nugget effect. The model is $\gamma(h) = 150 + 60 * exp^{(5)} + 98 * exp^{(38)}$. Such so high nugget effect is due mainly to the noise and in part to the small scale signals related to some real microstructures that are hard to be discriminate by filtering techniques.

1.4 Kriging of spatial components

Kriging of spatial components can be used to filter out such nugget effect and rebuild the grid with no small scale variability (Bourgault 1994; Wen and Sinding-Larsen 1997).

The resulting image is the one represented in figure 4.5.

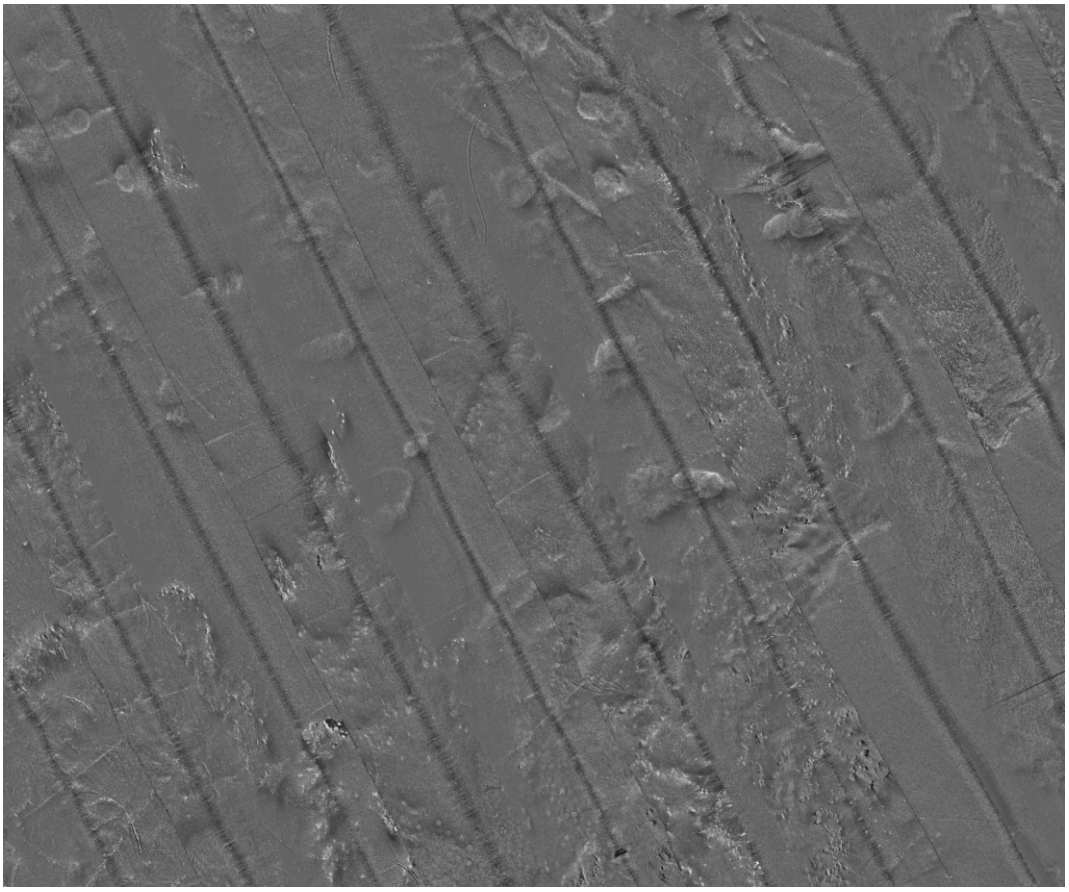


Figure 4.5 – The resulting image filtered with kriging.
Figura 4.5 – L'immagine risultante filtrate con il kriging.

In the filtered image the structures are more evident and most of the salt and pepper noise has been removed. We can observe the resulting variogram of such grid.

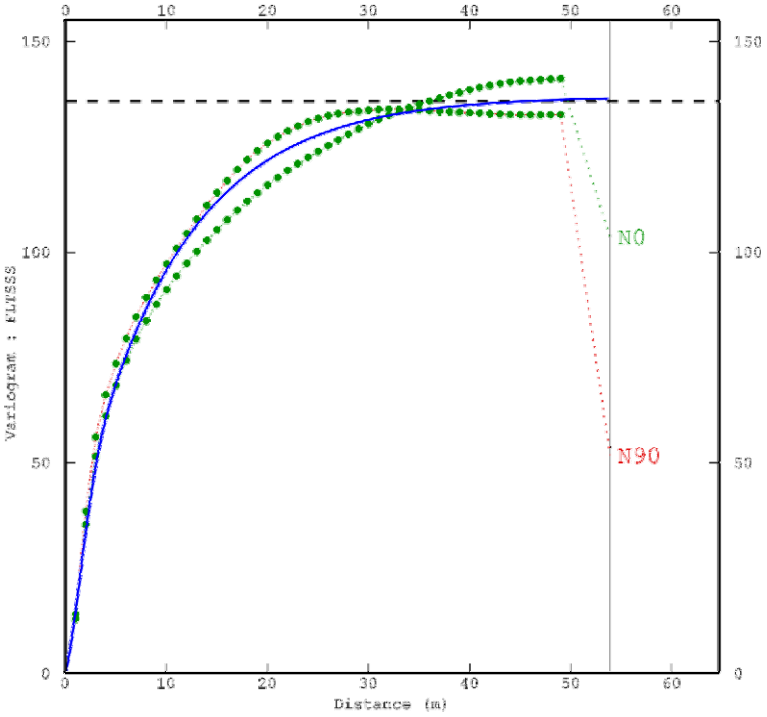


Figure 4.6 – Directional (orthogonal) variograms of filtered variable.
Figura 4.6 – Variogrammi direzionali (ortogonali) della variabile filtrata

With theoretical nested model: $\gamma(h) = 25 * gaus^{(4)} + 112 * exp^{(30)}$. Such model reveals a large uniformity at small spatial scale confirmed by the absence of nugget model. Moreover, the type of short range model, the gaussian one, typical of very uniform functions, validates such regularity. Basic statistics confirm the increase of normality of the variable, with a lower value of skewness parameter.

	SSS filtered (grey scale pixel value)
MIN	14
MAX	253
MEAN	103.34
MEDIAN	103.11
VARIANCE	135.81
STD DEV	11.65
SKEWNESS	0.53
KURTOSIS	7.82

Table 4.2 – Basic statistics for filtered mosaic.
Tabella 4.2 – Statistiche di base del mosaico filtrato.

As any kriged variable, the filtered one is less disperse than the original one, with lower standard deviation (Wackernagel 2003).

We can observe two zoomed examples to better appreciate the results of filtering. The original and the filtered set respectively on the left and on the right of the following figure.

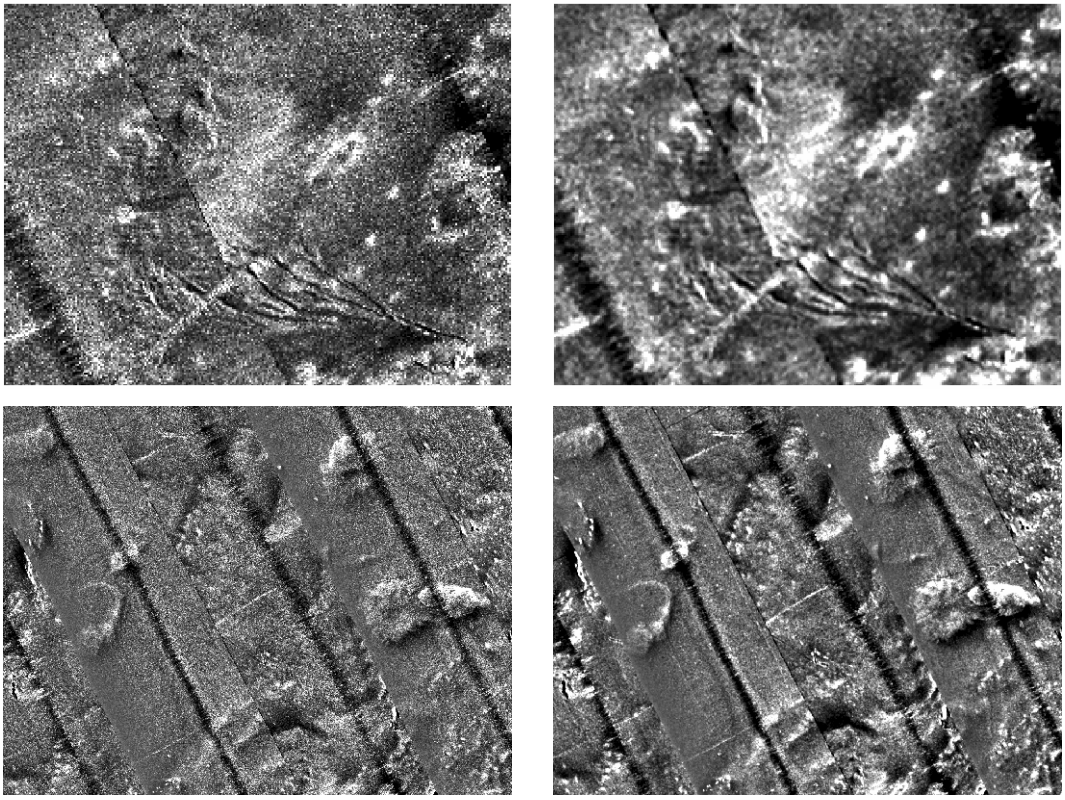


Figure 4.7 – Two zoomed details of original and filtered data, respectively on the left and on the right.
Figura 4.7 – Due dettagli ingranditi del dato originale (sinistra) e di quello filtrato (destra).

The salt and pepper noise has disappeared and the image is clearer. Such improvement is particularly effective for the final aim of the processing practice of side scan sonar data, that is the target recognition. Such procedure is based on the vectorization of all the structures and elements that can be regarded as known marine constructions and be used as indicator factors for the characterization of seabottom.

A further analysis of results can be made by Fourier transform, that allows to examine the frequencies of the signal and to quantify the contribute of each

frequency to the whole structure. In the figure the plot with the original and the filtered signal analyzed by Fourier transform (Lancaster and Salkauskas 1986).

It is evident how, in the filtered signal, all the highest frequencies are filtered out. Their amplitude, in fact, is highly reduced and, most of times, hold to zero. Such result confirms the efficiency of the filtering method that is able to clean the signal from the highest frequencies, responsible for the underlying noise.

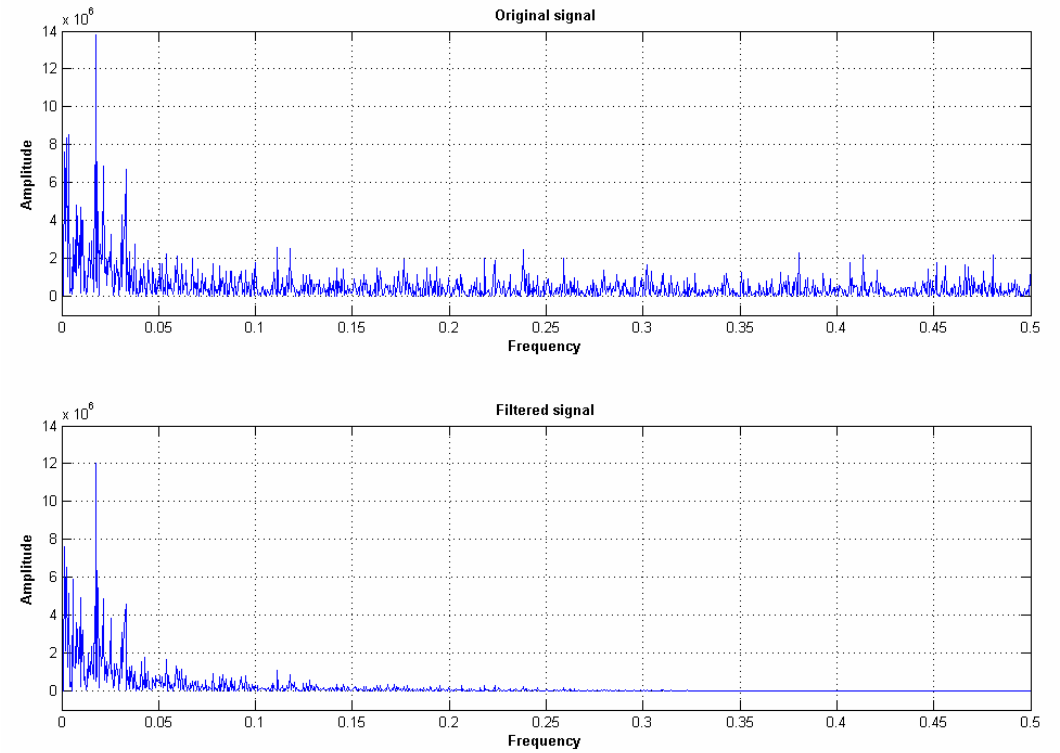


Figure 4.8 – Results of frequency analysis of original and filtered signal, respectively at the top and at the bottom.
Figura 4.8 – Risultati dell'analisi di frequenza del segnale originale (in alto) e di quello filtrato (in basso).

2. Discussion

Kriging of spatial components has been used to treat side scan sonar images, in order to filter out the fixed variance noise that is the cause of a poor interpretation of the information. Variography revealed the presence of a marked nugget effect in experimental variograms and kriging has been used to filter such component out.

The resulting filtered image is clearer and most of the salt and pepper noise has disappeared. Maybe some real microstructures have been partially hidden by the procedure, but the filtered mosaic is surely more adapt for seabottom classification.

Frequency analysis, computed by Fourier transform confirmed the efficiency of the filtering method.

CONCLUSIONS

The study of environmental science and its several applications are deeply influenced by the quality of data and its reliability. In general each data should represent faithfully the reality of the field and be regarded as a affordable information on which the interpretation of natural phenomena can be based. Unfortunately, such informations are not continue in the space and they are measured only in few points. Fitting techniques are thus called in aid, to reproduce an estimation of the reality that will be the stand for the understanding of the occurrences. Here the importance of the spatial analysis.

Most of times the observed phenomenon is the result of a complex mixture of factors influencing its behaviour, such that a simple model is not sufficient to completely represent it. Stochastic approach is based on the application of the theory of the regionalized random function (Armstrong 1998) to natural phenomena. The principle is that such natural complexity can be poorly represented by deterministic models but can be more reasonably reproduced in a stochastic framework.

The measured values in any point are regarded as the realizations of a random function that is defined on the basis of the measured variable. Thus the variable is substitute by its stochastic model and the interpretative process continues on such model.

Geostatistics stands for the application of such theory to spatial environmental variables. The modelling of the natural variable is carried on by one principal step (the variography) and some others following that are defined depending on the scientist demands (spatial estimations or simulations). Variography is the core of spatial analysis: the study of the spatial structure of the variable is functional to any estimation, simulation or other spatial computations can be made. The definition of experimental variogram, and its fit with continue deterministic functions, is the principal step that requires the maximum attention and experience by the user. Once modelled the variability spatial structure of the function, typically the kriging estimate is computed, in order to model the spatial distribution of the original variable. Kriging guarantees the best estimation with the smallest unbiased

estimation error and joins the estimation with the measure of its statistical consistency (the variance of estimation).

For such reasons, kriging, and more generally geostatistics, is more and more often used in environmental spatial analysis and largely diffused in the most common softwares.

In this PHD work the ISATIS geostatistics package from the french Geovariances (Bleines, Deraisme et al. 2004) has been widely used to present some applications to some different environmental datasets coming from the Neapolitan center of the Coastal Environment Institute of the National Research Council (I.A.M.C. C.N.R.).

Here the applications presented in the work.

- The geochemical dataset coming from the analyses computed on the marine sediments of the southern Campanian continental shelf has been processed with classical variography and *kriging of spatial components* has been implemented in order to define the geochemical backgrounds and their manifestation at different spatial scales. In particular the regional component has been interpreted as the real background value, globally varying in the domain, while the local component has been attributed to some non natural (anthropogenic impacts) or high frequency natural component added to the intrinsic small scale variability of the field. A further detailed variographic analysis has been implemented for all the eleven trace metals analyzed and, for each of them, the correct spatial models have been defined.
- One of the most important factors determining the quality of information and the reliability of the interpretation is the sampling strategy. Here variography has been applied to the optimization of a new sampling plan and to the minimization of wastes for the improvement of an existing one. In the first case the stratified random sampling strategy has been compared with systematic one from a geostatistics point of view. The subsampling of a simulated surface has been implemented with several decreasing resolution and experimental variograms have been computed on the resulting data vectors. The influence of the different strategies on the nugget effect has been examined in details and the optimization of the procedure has been defined. In the second case geostatistical

simulations have been implemented to model the spatial uncertainness and to locate the high risk points in a model area examined with an existing sampling plan. The additional samples have been located just in this areas and the process is iteratively checked by variographic analysis, in order to minimize the economical wastes. In particular, the ratio between the varying nugget effect and the analytic variance introduced by laboratory measure errors has been analyzed in details, with optimizing procedures. A practical application on a real dataset coming from survey made in the Port of Naples has been carried on.

- One of the most used geophysical method of investigation of seabottom is the *side scan sonar*. One of the step of the processing procedure of such datasets is the filtering of the mosaic image, in order to remove the noise coming from additional sources during acquisition. *Salt and pepper* noise is one of the most common. Variographic analysis has been used in this case to define the spatial component related to such noise and kriging of spatial components has been implemented to filter it out. Visual comparative tests, together with Fourier frequency analysis, have been presented to demonstrate the validity of the method.

The aim of this work wants to be the definition of the correct use of Geostatistics. As affirmed before, this discipline is more and more diffused in environmental science and many and many software host different kriging and variography applications. The negative aspect of such trend is the meaning of geostatistics that most scientists take into account. Geostatistics is based on a modelling approach (the modelling of the environmental variable by a random regionalized function) and as such is more sensible to subjective interpretation of the different tools used. In particular, the variogram modelling and the use of kriging cannot be considered as a automatic process and cannot be used as a black box practice. Results from fitting procedures are the basis elements for the interpretation of environmental phenomena and the attention required by the processing practice cannot be submitted to automatic methods.

REFERENCES

Armstrong, M. (1998). Basic Linear Geostatistics, Springer.

Bleines, C., J. Deraisme, et al. (2004). Isatis Software Manual, 5ft edition, Geovariances & Ecole des Mines de Paris.

Bourennane, H., S. Salvador-Blanes, et al. (2003). "Scale of spatial dependence between chemical properties of topsoil and subsoil over a geologically contrasted area (Massif central, France)." Geoderma **112**: 235-251.

Bourgault, G. (1994). "Robustness of noise filtering by kriging analysis." Mathematical Geology **26** (733-752).

Burrough, P. A. "Beyond GIS: the development of spatial analysis tools for modelling the physical environment."

Burrough, P. A. and R. A. McDonnell (1997). Principles of Geographical Information Systems, Oxford University Press.

Buttafuoco, G. and A. Castrignanò (2005). "Study of the spatio-temporal variation of soil moisture under forest using intrinsic random functions of order *k*." Geoderma **128**: 208-220.

Carlon, C., A. Critto, et al. (2001). "Risk based characterisation of contaminated industrial site using multivariate and geostatistical tools." Environmental Pollution **111**: 417-427.

Castrignanò, A. and G. Buttafuoco (2004). "Geostatistical Stochastic Simulation of Soil Water Content in a Forested Area of South Italy." Biosystems Engineering **87**: 257-266.

Cheng, Q., F. Agteberg, et al. (1994). "The separation of geochemical anomalies from background by fractal method." Journal of Geochemical Exploration **51**: 109-130.

Clark, I. (2000). Practical Geostatistics 2000.

Deutsch, C. V. (2002). Geostatistics**:** 697-707.

Durrani, S. A. and I. Badr (1995). "Geostatistically controlled field study of radon levels and the analysis of their spatial variation." Radiation measurements **25**: 565-572.

Dutter, R. (2003). Geostatistics, Vienna University of Technology.

Facchinelli, A., E. Sacchi, et al. (2001). "Multivariate statistical and GIS-based approach to identify heavy metal sources in soils." Environmental Pollution **114**: 313-324.

Ferraro, L., T. Pescatore, et al. (1997). "Studi di geologia marina del margine tirrenico: la piattaforma continentale tra Punta Licosa e Capo Palinuro (Tirreno Meridionale)." Bollettino Società Geologica Italiana **116**: 473-485.

Franklin, R. B. and A. L. Mills (2003). "Multi-scale variation in spatial heterogeneity for microbial community structure in an eastern Virginia

agricultural field." FEMS Microbiology Ecology **44**: 335-346.

Goovaerts, P. Combining Minimum Error Variance and Spatial Variability in the Mapping of Environmental Variables.

Goovaerts, P. "Kriging vs Stochastic Simulation for Risk Analysis in Soil Contamination."

Goovaerts, P. (2001). "Geostatistical modelling of uncertainty in soil science." Geoderma **102**: 3-26.

Goovaerts, P., G. M. Jacqueza, et al. (2005). "Geostatistical and local cluster analysis of high resolution hyperspectral imagery for detection of anomalies." Remote sensing of environment **95**: 351-367.

Hu, J. J. "Methods of Generating Surfaces in Environmental GIS Applications."

Isaaks, E. H. and R. M. Srivastava (1989). An Introduction to Applied Geostatistics, Oxford University Press.

Journel, A. and C. J. Huijbregts (2004). Mining Geostatistics, The Blackburn Press.

Jr., F. C. C. and P. V. Bolstad. "A Comparison of Spatial Interpolation Techiques in Temperature Estimation."

Kanevski, M., R. Parkin, et al. (2004). "Enviromental data mining and modelling based on machine learning algorithms and geostatistics." Environmental

Modeling & Software **19**: 845-855.

Lancaster, P. and K. Salkauskas (1986). Curve and Surface Fitting An Introduction, Academic Press.

Lee, Y. "An Introduction to Spatial Statistics."

Lloyd, C. D. and P. M. Atkinson. "Scale and the spatial structure of landform: optimising sampling strategies with geostatistics."

Lloyd, D. and P. Atkinson. "Designing optimal sampling configurations with ordinary and indicator kriging."

Malyuk, B. I. "Surozh Deposit: Brief Comparative Geostatistics Using Different Software."

Morris, S. J. (1999). "Spatial distribution of fungal and bacterial biomass in southern Ohio hardwood forest soils: fine scale variability and microscale patterns." Soil Biology and Biochemistry **31**: 1375-1386.

Myers, D. E. (1989). "To Be or Not to Be Stationary? That Is the Question." Mathematical Geology **21**: 347-362.

Petitgas, P. (1997). "Sole egg distributions in space and time characterised by a geostatistical model and its estimation variance." Journal of Marine Science **54**: 213-225.

Pilger, G. G., J. F. C. L. Costa, et al. "Additional samples: Where they should be located."

Raspa, G. (2000). Il ruolo della Geostatistica della modellistica ambientale. A. Aria**:** 89-99.

Raspa, G. (2004). Dispense di Geostatistica applicata.

Raty, L. and M. Gilbert (1998). "Large-scale Versus Small-scale Variation Decomposition, Followed by kriging Based on a Relative Variogram, in Presence of a Non-stationary Residual Variance." Journal of Geographic Information and Decision Analysis **2**: 102-125.

Reisa, A. P., A. J. Sousab, et al. (2004). "Combining multiple correspondence analysis with factorial kriging analysis for geochemical mapping of the gold–silver deposit at Marrancos (Portugal)." Applied Geochemistry **19**: 623-631.

Rosenbaum, M. and M. Soderstrom. (1996). "Geostatistics as an Aid to Mapping." from http://gis.esri.com/library/userconf/europroc96/PAPERS/PN11/PN11F.HTM

Russo, F. (1990). I sedimenti quaternari della Piana del Sele, studio geologico e geomorfologico, Università degli studi di Napoli "Federico II"**:** 168.

Salminen, R. and T. Tarvainen (1997). "The problem of defining geochemical baselines. A case study of selected elements and geological materials in Finland." Journal of Geochemical Exploration **60**: 91-98.

Singh, M., G. Müller, et al. (2003). "Geogenic distribution and baseline concentration of heavy metals in sediments of the Ganges River, India." Journal of Geochemical Exploration **80**: 1-17.

Smith, M. L. and R. E. Williams (1996). "Examination of methods for evaluating remining a mine waste site. Part I. Geostatistical characterization methodology." Engineering Geology **43**: 11-21.

Soliani, L. (2005). "Statistica univariata e bivariata parametrica e non-parametrica per le discipline ambientali e biologiche."

Sprovieri, M., et al. (2005). "Heavy metals in top core sedimtns from the Southern Campania Shelf (Italy): hints to define large scale geochemical backgrounds." In press

SurfStat Australia.

Trevisan, S. (2004). Geostatistica nel contesto idrogeologico ed ambientale. Dipartimento di Geologia, Paleontologia e Geofisica. Padova, Università degli studi di Padova**:** 183.

Vivo, B. D., M. Boni, et al. (1997). "Baseline geochemical mapping of Sardinia (Italy)." Journal of Geochemical Exploration **60**: 77-90.

Wackernagel, H. (2003). Multivariate Geostatistics An Introduction with Application (3rd ed.), Springer.

Wen, R. and R. Sinding-Larsen (1997). "Image filtering by factorial kriging.Sensitivity analysis and application to Gloria side-scan sonar images." Mathematical Geology **29**: 433-468.

Wingle, W. L. (1992). Examining Common Problems Associated with Various Contouring methods, Particularly Inverse-Distance Methods, Using Shaded Relief Surfaces. Geotech, Colorado.

Yunker, M. B., R. W. Macdonald, et al. (1999). "Natural and anthropogenic inputs of hydrocarbons to the Strait of Georgia." The Science of Total Environment **225**: 181-209.