Cahiers de Géostatistique

Fascicule 4

PROCESSING DATA WITH A SPATIAL SUPPORT : GEOSTATISTICS AND ITS METHODS

Pierre CHAUVET

Ecole des Mines de Paris
1993

CENTRE DE GÉOSTATISTIQUE,
Ecole Nationale Supérieure des Mines de Paris,
35 rue Saint–Honoré, 77305 FONTAINEBLEAU, France

CAHIERS DE GEOSTATISTIQUE
Fascicule 4
Processing data with a spatial support : Geostatistics and its methods, par Pierre Chauvet, Août 1993, 57 p.

ISSN 1168–2574

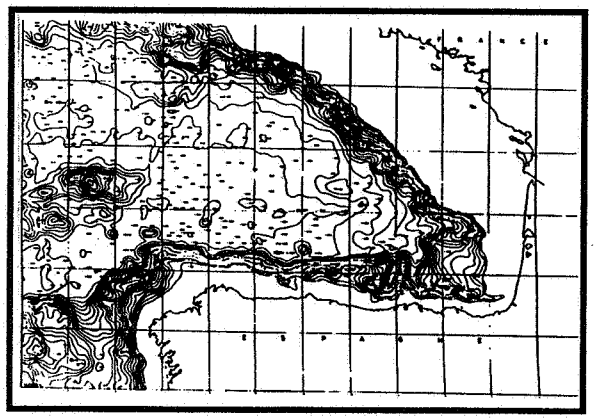
PROCESSING DATA WITH A SPATIAL SUPPORT : GEOSTATISTICS AND ITS METHODS

Pierre CHAUVET

Processing data with a spatial support : Geostatistics and its methods.

Processing data with a spatial support : Geostatistics and its methods

Introduction

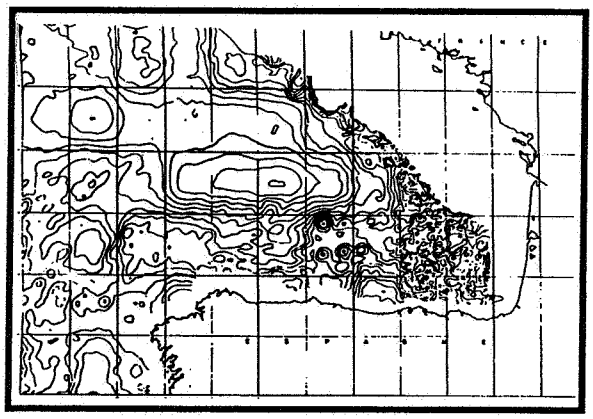


Sea bed depth in the Bay of Biscay

This map answers a simple problem : we wanted to draw the isolines from a data set of sea-bed depths in the Bay of Biscay. In this case, the data were distributed in an irregular way. We used a classical approach of **estimating** at the nodes of a very fine grid, and the isolines were then drawn by a spline type smoothing process.

Although this example is simple, it contains the seeds of many a question.

Through the map we want to show a set of **structural features** of the phenomenon being studied. Indeed, the variable "depth" has characteristics which should be taken into account by a good map: anisotropy (different behaviour according to the directions: the depth does not vary in the same way depending on whether one is parallel or perpendicular to the coastline), heterogeneities (the ocean bed structure is not the same off the Spanish and Landaise coasts), trends (slopes), etc... So, to begin with, these features should be identified and then defined mathematically so as to be able to apply an estimator which takes them into account. Under the name of **variographic analysis** (*), this double work of **interrogating** the data then **modeling** their structural properties makes up the initial and unavoidable step in any solid geostatistical study.



Standard deviation of the previous map

However, one might not wish to stop there. For example, we might ask for an evaluation of the "**quality**" of this first map. And we must begin by defining this notion. For, if a depth is a clearly defined physical unit, which is the same for everyone, the quality of an estimator is an abstract idea which is the result of an arbitrary **choice** — note that the word "arbitrary" is not used in a derogatory sense here.

Therefore, **a quality criterion should be agreed upon** — *a priori* . One possible criterion is the standard deviation of estimation, which is shown on this second map. To put it simply, an estimated value is all the more reliable if its standard deviation of estimation is small.

It can be shown that this standard deviation depends on the structure adapted to the variable being studied, but also on the configuration of the **sampling data** — which is intuitively satisfactory. This

(*) We prefer this expression to "Structural analysis", which is more common but has another meaning in Geology.

criterion is very simple to use and moreover forms the cornerstone of geostatistics. It can also be used to go further and for example ask about the strategy of exploration campaigns, optimal location of further information, or even about the effect of uncertainty concerning the localization of samples.

These questions still remain simple. However, we can be faced with considerably more complicated problems: for example, determine probabilities of submarine rises or **intervals of confidence** on the bathymetric map or even wonder about the length of an underwater cable lying on the seabed between two given points... As well as this, we might want to link the parameters coming from the variographic analysis to geological phenomena and so consider an **interpretive analysis** of the data. The variographic analysis can also help compare sampling methods. In the proposed example, the data came from two sea campaigns which were carried out with two ships and two different types of equipment ; it was up to determine which equipment was the better suited.

All these questions and many more that can be asked about any phenomenon presenting a structure in space or in time are raised in geostatistics. The contribution of geostatistics in the processing of data with a spatial support is always twofold:

- On the one hand it provides data **interrogation tools** which are directed towards the presentation of the intrinsic structure of the variables being studied and of their links with their field of definition;
- On the other hand, it provides a **theoretical framework** which makes it possible to expand algorithms (estimations, numerical simulations, optimizations), answering problems which can be met in the study of spatial processes.

Moreover, the introductory example — though oversimple — which we have just given, could have been taken from many other fields: one of the assets of geostatistics is not to be dependent on any one field of application. That is why, in what follows, little importance should be given to the choice of vocabulary which would seem to favour certain applications. Geostatistics is mainly a set of methods, and the name given to them is immaterial. Should we say "estimation" as in the mining industry, "forecast" as for time series, "interpolation" as in mapping or "assimilation" as in meteorology?... Should we say "white noise" rather than "nugget effect"?... This sort of question will not be considered here.

Background

The origins of geostatistics are to be found exclusively within the mining industry. New estimation methods were developed in the early 50s when "classical" statistics were found to be unsuitable for studying very disseminated deposits. The word "kriging" which was coined in recognition of D.G. Krige and his work on gold mining in South Africa, remains to remind us of the encounter between a mathematical technique of regression and the very real problems to be found in gold mining. Even at the same time, kriging methods were being applied to other minerals such as iron, copper, nickel, uranium...

Two features marked the **early years of geostatistics**. To begin with, on a practical level, computation methods were still very elementary. Therefore publications were packed with approximation formulas, curves, tables, which ended up being a real wealth of information. On a theoretical level, the formalisms being developed were often put within the framework of a given law of distribution. It was not so much a question of the gaussian model (*) — unsuited to disseminated variables — as of the log-normal model which was highly popular in the 50s. So, to sum up, the neologism "geostatistics" was perfectly suited for describing this first stage.

However, in the **second era of geostatistics**, reference to statistical models was dropped. Either we developed models which did not include the distribution laws (linear geostatistics) or we went back to reference models via anamorphosis curves. Parallel to this, we tried to broaden the working hypotheses: this saw the development of non-stationary geostatistics for treating phenomena with a trend, then non-linear geostatistics for solving the problems of exceeding the threshold or change of support — non stationary–non linear geostatistics still remains to be done... New formalisms appeared which went well beyond the classical problems of estimation: simulations (conditional or not), random sets. This methodological abundance could be put into use immediately because of the amazing improvement in the computational methods.

It is not easy to talk about "**third generation geostatistics**" which is currently in full development. As data processing is becoming more and more easy to handle, geostatistics is developing in many different directions. This of course can be seen in the fields of application which are no longer restricted to natural resources (mining, petroleum). But above all, and more fundamental, research is exploring extremely diverse theoretical paths. It is especially interesting to note that we are reconsidering the distribution laws. This in no way means going backwards, but, on the contrary, that we have at our disposal — or that we feel the need for — new theoretical tools which go far beyond the possibilities of structural functions which, up to now, made up the cornerstone of any geostatistical study.

(*) "gaussian" means "normal distribution"

Presentation of geostatistics

Regionalized variable and random function

Any geostatistical study begins with a **set of data distributed in space** (and/or in time). These are **numerical** data. By this we mean that geostatistics is operates on quantities. Qualitative data cannot be used as such unless they can be coded numerically; otherwise they are used mainly to delimit the extension of the range of validity of the geostatistical model being applied.

So mathematically, initially we have at our disposal a certain **function**, usually denoted $z(x)$, defined in a metric space and taking its values in $\mathbf{R}$ (possibly $\mathbf{R}^n$ or $\mathbf{C}$). This function is called the **regionalized variable**. Measures of mining grades, altitude or bathymetry, atmospheric pressure, pollutant content... are all examples of scalar regionalized variables. The measurement of wind is one example of a 2-D vectorial regionalized variable or, in other words, a complex regionalized variable.

It is possible to be satisfied with this level of abstraction and carry out a direct theoretical study on the regionalized variable: this is the purpose of **transitive geostatistics** which only requires very classical mathematical tools and which certainly does not impose any restrictive hypotheses on the variable being studied. The main disadvantage of this approach is that it is largely dependent on the notion of **field**, a **bounded** domain beyond which a regionalized variable is assumed to be zero. It is almost always impossible in a result to take into consideration what is due to the intrinsic structure of the regionalized variable and what is due to the geometry of the field. It is for this reason that it is often essential to take the risk and go beyond just one measure of abstraction.

The step consists in considering the regionalized variable function $z(x)$ as a **realization** of a certain **random function** $Z(x)$. In this case it concerns a freely accepted **methodological choice**, not an attempt to get closer to reality. By this choice, we do not intend to decide on the deterministic or random "nature" of the phenomenon being studied, but we simply choose a range of tools from which we expect reasonable efficacy. And, in actual fact, we have at our disposal all the ammunition coming from probability theory and stochastic procedures.

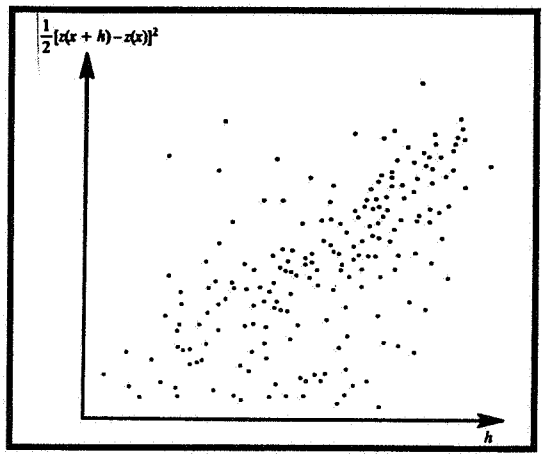
Let us stress the pragmatic character of this choice. The probabilistic model is not an end in itself, but a **tool** that we forge in answer to a problem (estimation, simulation), which, in general, we do not master. Besides, the fact of using probabilities in no way anticipates the very nature of the phenomenon being studied — deterministic or random — but constitutes a vital **choice** which only experience will prove to be appropriate.

Variography

In the probabilistic models that we intend to apply, the simplest tool for measuring an estimator's quality is **variance**, that is to say the quadratic distance in the probabilized space. In this way, the quadratic measure of the distance between two values (random) $Z(x)$ and $Z(y)$ of the random function Z will be given by the function :

$$\gamma(x,y) \quad = \quad \frac{1}{2}\,\mathbf{E}\left[(Z(x)-Z(y))^2\right]$$

(where the symbol $\mathbf{E}$ is the mathematical expectation). The theory enables us to prove that this function γ, called the **theoretical variogram** is the only tool needed to solve "linear" problems (estimation without change of scale, optimization of the sampling network, quality of an estimator).

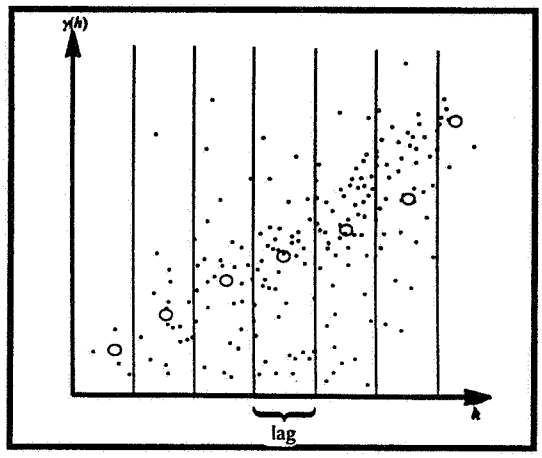


Variogram cloud

As for the regionalized variable, it is therefore natural to examine the mean square deviation between the data set $z(x), z(y)\ldots$ For all pairs of data points, (x, y), the figure shows the scatter diagram between the mean square deviation $(z(x) - z(y))^2$ and the distance $\|x - y\|$.

This scatter diagram, called a **variogram cloud**, is a preliminary tool for examining data, and is comprehensive. It makes it possible to show "abnormal" values, sampling heterogeneities, and possibly the trends. It introduces information which the user has in no way treated. It is an entirely objective picture of the available information.

The variogram cloud is therefore an extremely valuable analytical tool, where "exploratory geostatistics" which is in full development, finds its roots. However, it gives little synthetic structural information and remains difficult to model without strong hypotheses. So we prefer to have a more global structural function, which can express the **evolution** of the mean square deviation between two samples **in terms of the distance** between these two samples. More precisely, we should like to build a function which, given the samples, makes up, in some way, an **experimental version** of the theoretical variogram acquired later on by geostatistical algorithms. Of course, this function exists in the variogram cloud, but in an **implicit** form which makes it unusable.

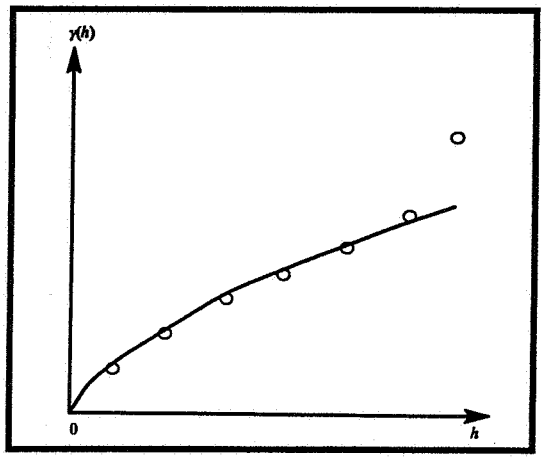


Experimental variogram

In this way, we arrange the data pairs in a network of distance classes. As compared to the variogram cloud, some discretion is introduced, since, apart from the particular case where the data are on a regular network, the choice of the distance classes is partly left up to the user.

The following step consists in calculating, for each distance class, the mean of the corresponding square deviation. For each class, we therefore get a single value, the mean square deviation, and the initial scatter diagram is now summed up by a function defined for a small number of distance values.

This **set of numeric values** is called an **experimental variogram.**



Modeled variogram

Although considerably more synthetic than the variogram cloud and allowing a good structural interpretation of the data, the experimental variogram cannot be used as such in the theoretical formalisms. It has to be expressed as an equation, which makes it possible to give it a value for every possible value of the variable "distance".

The final step in the variographic analysis is therefore to get "the best fit" of a known expression to the experimental variogram.

This **function** is called a **modeled variogram.**

The way to do it is to limit oneself to a reduced set of base models in order to express the modeled variogram. The bibliography contains formulas for the **exponential, spherical** and **monomial** variograms, without forgetting the **nugget effect**. This corresponds to data without spatial structure, hence the geostatistical study goes back to classical statistics.

Let $\{x_i\}$ be the data point set, h any distance and $N(h)$ the number of sample pairs distant from h — eventually with a certain amount of tolerance. So, to sum up, the variography consists in building the experimental variogram

$$\gamma^*(h) \quad = \quad \frac{1}{2\,N(h)} \sum_{i=1}^{N(h)} \left[(z(x_i + h) - z(x_i))^2 \right]$$

to fit a mathematical function to it "as well as possible" and to incorporate the theoretical variogram (required by the theory) to this modeled variogram.

Of course this approach requires taking precautions, firstly mathematical, for all functions cannot represent a variogram, and the models proposed should obey a certain number of constraints (positive or zero values, parity, regularity properties, behaviour at infinity...). However, the range of models proposed and the formalism developed in geostatistics safeguard the user from important mathematical slips.

On the other hand, in what concerns the **physical significance** of the operations, it is absolutely necessary to take care. For example, whatever the data set, it is always possible to calculate an experimental function of the type $\gamma^*(h)$ above. By construction the result will only depend on the distance factor h. But that in no way proves that a probabilistic model can be satisfactorily represented, where the theoretical variogram $\gamma(x, y)$ only depends on the distance $||x - y||$. For example, it can quite easily happen that the experimental variogram cannot be "reasonably" represented by **any** admissible model. Naturally, in this case, it is quite useless to force reality to conform to our models.

The main point of the variographic approach is to ensure the parallel between the numerical operations performed on the regionalized variable and the theoretical developments concerning the random function. We can imagine the most excessive operations on a data set, but they will only make sense within the frame of a consistent model. Conversely, we can imagine purely theoretical developments at the level of the random function, but they will remain abstract mathematics if we do not know how to associate a physical interpretation with them. Now this parallel, which is essential for a **realistic** practical study, is not evident.

The development of the variogram is an excellent illustration of this problem. In order that the spatial mean value which leads to building the experimental variogram $\gamma^*(h)$ can be interpreted as an estimate of the mathematical expectation which appears in the definition of the theoretical variogram, the probabilistic model which is fitted to the data must have the good properties of **stationarity** and **ergodicity**. We shall come back to this later on, but what is important to note here is that this time we are talking about properties **which can be proved false**. Sometimes the numerical manipulations made on the data show up results which are incompatible with these "good" models. From then on it would be useless to pursue mathematical developments, even theoretically correct, which would no longer have a reasonable physical interpretaton.

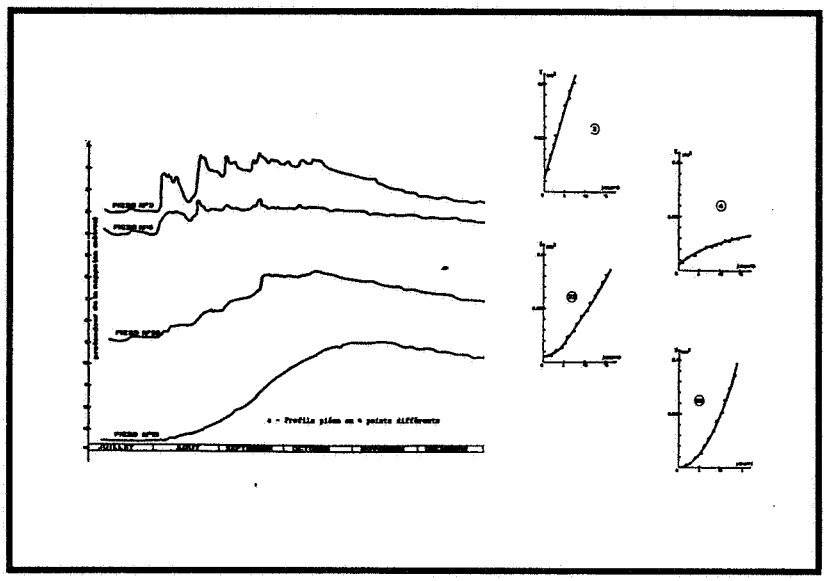
These remarks set out to underline to what extent a geostatistical study depends on the quality of the variography phase. The model, not being an end in itself, but a data tool, it is absolutely necessary to constantly check the suitability of the model to reality. Therefore, it is perfectly normal during a study, because of new information to question the validity of the model, to review the situation. The data should always have the last word in the case of conflict with the model.

A last remark. Even if fitted to the data, the model presents risks. For the domain of validity of a mathematical expression (for example the equation of a theoretical variogram) is unlimited. Now, the fitting of a model always takes place on a bounded domain. In this way nothing can be said about the validity of the model nor the results it leads to, the moment we go beyond the frame in which it was obtained. No crash barrier exists, so it is up to the user not to go beyond reasonable limits, and to be careful not to make the model go further than it can. It is not enough to get the mathematics right. One must also be realistic. So any geostatistical study is associated with the very important notion of **scale** which represents the frame of validity of the methods being considered. All the results in a geostatistical study are subject to its field of validity and work scale.

Structural analysis

Even before going into the equations for the modeled, the experimental variogram enables us to display and analyse some of the data properties. The example below (*) illustrates this.

The data being studied are the piezometric levels taken every hour during the rainy season (July–December) in Korhogo (Ivory Coast). Four piezometers, numbered 3, 4, 33 and 18 respectively, are studied here. Located on a hillside, they correspond to increasingly deep water tables.



Water table levels in Korhogo

The profile of the piezometric levels is easily understood. At piezometer n^0 3, where the water level is close to the surface, the level reacts to each rainfall. Consequently, the piezometric profile is very erratic as it rises every time it rains. At piezometer n^0 4, the thickness of the land acts as a rain absorber and so the profile is already less irregular. With piezometer n^0 33, this effect is increased and the effects of refilling the ground water level abruptly are wiped out by more important longer term behaviour. For piezometer n^0 18, where the underground water level is very deep, the violent rain period (on a daily scale) are barely noticeable: on the other hand, a very significant seasonal trend is observed which is spread over the entire period being studied, and which corresponds to the water accumulation over the entire rainy season.

These results are clearly visible when set down in chronological order and are easy to interpret. From a geostatistical point of view it is important to note that they show up immediately in the time experimental variograms. The variogram of piezometer n^0 3 shows a significant linear increase on a two week scale. Moreover, it reveals at the origin a discontinuity, a **nugget effect** which indicates that there may be big jumps between two adjacent measurements. Piezometer n^0 4 gives the same result although this time the experimental variogram increases more slowly. Compared to piezometer n^0 3, this means that the deviation (mean square) between two separate measures in the same time lapse is less at piezometer n^0 4 than piezometer n^0 3 — or rather that the correlation between the two measures is stronger there. This can also mean that the piezometer n^0 4 data show a more marked structure than those of piezometer n^0 3.

The same effect can be seen with piezometer n^0 33. Even if there is a nugget effect (discontinuity at the origin) we can see this time that the experimental variogram has a parabolic shape. This results in the mean square deviation being weaker at piezometer n^0 33 than at piezometer n^0 4 up to an interval of 6–7 days. However, this order is inversed when the intervals are longer, because, at piezometer n^0 33, the contrast between two measures is mainly caused by the seasonal component which has become predominant. If there is a significant time lapse between two measures, one has to have been taken at the beginning of the rainy season and the other at the end of it. This being the case, there will be a big difference between them, whatever the particular rain-falls throughout the season. In geostatistics,

(*) Example due to J.P. Delhomme, cited in *Journel [1977]*

this situation represents the case where a stationary model is inadequate. Of course this effect is even more exaggerated in the case of piezometer n^0 18 where the nugget effect has completely disappeared: the piezometric profile, like the associated variogram, only represents a significant seasonal tendency.

This is a very basic example. The experimental variograms have simple shapes and the essential differences between the four situations come from their **behaviour at the origin**. This behaviour shows the degree of spatial continuity of the phenomenon or its behaviour on very small scales. But other features of the experimental variogram which are very important for understanding the data, can appear. This is particularly true of the **anisotropy** which can be seen when we work in a space having more than one dimension. It may then happen that the experimental variogram does not have the same behaviour in all the directions, as some of them are favoured for physical reasons (for example outflow phenomena). It would then be fatal to average the valves in every direction. Naturally, if we take into account the anisotropy during the modeling, we increase the number of parameters for the model and give the user more freedom. But at the same time, new constraints of consistency arise and so fitting the theoretical variogram can soon become difficult.

The phenomenon of **nested structures** can also be seen with complex data: the experimental variogram then displays stacked components having different scales which we can try to particularize during the modeling stage. Yet again the number of parameters can be considerably increased. This approach is interesting in that it allows an interpretation of the origins of the data being studied. The analysis of nested structures, especially in its multivariate expansion, resembles data analysis and insofar as the different **scales** are separated, can be compared to a Fourier Analysis. The advantage of the geostatistical method is that the spatial structure is taken into consideration and we are neither constrained by the dimension of the work space nor by a condition of regularity of the data grid. Besides this, if one of the isolated nested phenomena can be associated to noise, a geostatistical method of **filtering** can be proposed and in this case can be compared to signal processing.

Lastly, it should be remembered that the structural analysis can lead to a deadlock, for example that the test results are incompatible with a model $\gamma(x, y)$ depending only on $||x - y||$. This is a typical attack on the hypothesis of stationarity, that an experienced user recognizes immediately. This condition is sufficiently important that special attention has been paid to non-stationary geostatistics for a long time. In this way the intuitive concept of **trend** can be clearly formalized and included in the estimation or simulation computations. However, to avoid giving too many examples here, we shall restrict ourselves to the case of a stationary variogram.

Modeling

Let us go back once more to the modeling stage. It is an essential step as the theoretical formalism of geostatistics "feeds on" mathematical expressions, not a set of test values. It is also a vital step since we abandon the domain of facts (data) for that of speculation. Consequently, this step presents risks, but it also gives the user the opportunity to include non-numeric data, subjective factors and his own experience.

In outline, modeling consists in finding a mathematical expression which best fits the few points of the experimental variogram. So it concerns the classical problem of constructing an interpolation function. We have seen that there are a certain number of theoretical constraints on the final expression and that it is preferable to limit the choice of parameters which represent the model.

In practice, the modeling step is mainly interactive. Using high quality graphical techniques here is particularly advantageous for representing and interrogating the set of (geo)statistical tools — histograms, scatter diagrams, variogram clouds and deferred correlations, variograms etc... — which interpret the structure of the variable being studied. A genuine **dialogue** with the data must be created. And in routine geostatistics the experimental variogram holds the privileged position of representing the structure that we want to interpret mathematically.

Influence of the model

Respect for the structure, for example when fitting the variogram model, constitutes what can be called the **"up stream constraints"** of modeling. However, the choice of a model effects the application of the algorithms: the proposed fit is not unimportant for the rest of the study. So there are always, in variographic analysis, the **down stream constraints**.

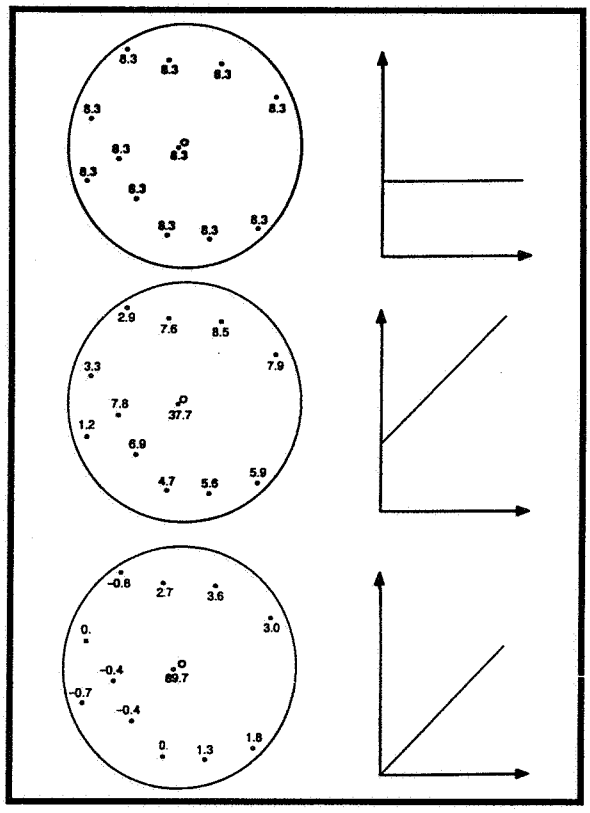
This is an important point. Generally, the experimental variogram leaves a certain amount of freedom when fitting the parameters of the models and, with no additional constraints, there are no irrefutable criteria when settling between several plausible models. Statistical tests could be considered, but such a choice generally means being more specific about the model, which is often impossible in geostatistics because of the amount of data available. In addition, it is **justifiable** to direct the methodological choices in terms of the problem posed. If we go back to the example in the introduction on the bathymetry in the Bay of Biscay, it is evident that we would not want the same features in the final map, depending on whether it was intended for a geography book (maps with general features of the region), for a ship's captain (security map indicating ocean depths), for a submarine captain (peak diagnosis: height, location), for fishermen, etc...

To sum up, there is not only one real model lurking behind the unwavering data, but rather a choice of parameters to be fitted which, of course, honor the data, but are also adapted to the problem in hand. When the result is obtained, the conditions in which the work was done has to be recalled and a **field of validity** associated with this result.

As an example of the "down stream" influence of the choice of the model, we show the optimal weights for a linear estimator at the centre of the circle (*) That is, they are the weights given to each data value when calculating the estimated value at the centre ; these weights sum to 1. Next to each point (and they are irregular in this case) is its weight.

In the case of a **pure nugget effect** model, all the weights are equal and so are worth $\frac{1}{12}$. This is still true of the datum nearest the centre, which is thus no more significant than the other 11. This can be easily explained: the pure nugget effect shows that the mean square deviation between two data is independent of their distance, that the variable being considered has **no spatial structure**. Naturally, this feature is seen at the estimator level. As it is, we find ourselves in the situation of classical not spatial **statistics** .

In the second case, the variogram model still contains a nugget effect component (which could be due to a measuring error), but the variable also shows a linear **structure**. That means that when the distance to the point to be estimated increases, there is a **reduction in the information** coming from a measuring point. Indeed, we observe that the data near the centre to be estimated plays an important role this time (nearly 40% of the total weight of the information). As expected, the other weights tend to decrease when we go further away from the point to be estimated.



Importance of the nugget effect

The third case is that of a linear model. The variable is strongly structured. The absence of the nugget effect indicates the continuity of the variable. As there is no longer the risk of having significant value deviations between close data points, it is clear that an information which is very close to the estimated point will have a decisive role. And indeed, the point near the centre has nearly 90% of the total weight, meaning that the other data are less important. Note that the optimal linear estimator tosses up **negative weights**. Experience shows that this is frequent for strongly structured variables and in theory, nothing forbids this. Moreover, nothing prevents the data from having a weight of more than 100%. This point should not be forgotten, and depending on the study, it can prove either beneficial or

(*) The ideas of "linear estimation" and "optimal weights" will be defined in the next two sections.

disastrous. In topography or bathymetry, that means that it is sometimes possible to obtain an estimator outside the data values — which is not a bad thing: but that also means that it is possible to obtain a negative grade in mining estimation...

Quality criterion

In the preceding example we spoke about the **optimal** linear estimator which means that we have a quality criterion for this type of estimator. Among the geostatistical methods, the **variance** is chosen as the criterion.

The word "variance" is likely to be ambiguous and needs defining. In order to define the quality criterion theoretically, we shall talk about the variance in relation to the **probabilistic model**, that is to say concerning the random function. It therefore concerns a perfectly determined **mathematical concept**. Having said that, note that by choosing the variance as the quality criterion, we are making a **completely free choice**. The advantages are striking. The theoretical calculation of the variance of any linear expression of the random function only requires knowing the variogram. So this criterion is not too demanding in terms of specifying the probabilistic model. However, the choice has its limits, for a variance is quite a poor tool which, in a distribution law, does not "consider" the significant characteristics such as dissymmetries, multimodalities etc... Once more, it is essential to be aware of these limits when choosing the methodology. We should also mention that using more sophisticated tools requires a much more detailed model and doubtless much less realistic than those that satisfy stationary geostatistics.

The actual mathematical computation is not difficult. Let v and w be any two domains in the work space and $|v|$ and $|w|$ their measures (areas in $\mathbb{R}^2$, volumes in $\mathbb{R}^3$, etc...). $\bar{Z}(v)$ is the mean value of the **random function** Z over v :

$$\bar{Z}(v) = \frac{1}{|v|}\int_v Z(x)\,dx$$

and $\bar{Z}(w)$ its mean value over w. These spatial means are **random variables** as is their difference $\bar{Z}(v) - \bar{Z}(w)$. This difference, called the **estimation error**, shows the error made (in the probabilistic model) by estimating $\bar{Z}(v)$ by $\bar{Z}(w)$.

We shall always assume that the estimation of $\bar{Z}(v)$ by $\bar{Z}(w)$ is **unbiased**, that is

$$\mathbf{E}\left[\bar{Z}(v) - \bar{Z}(w)\right] = 0$$

The **extension variance $\sigma_E^2(v,w)$ of the domain v to the domain w** is defined as the estimation error variance and so, taking the unbiasedness hypothesis into consideration, is written

$$\sigma_E^2(v,w) = \mathbf{E}\left[\left(\bar{Z}(v) - \bar{Z}(w)\right)^2\right]$$

Note that this formula is only a general expression, giving no special significance to the v and w domains. In practice, v will indeed be a domain — possibly reduced to only one point — on which we want to estimate the mean value of Z, whereas w will represent the finite set $\{x_\alpha\}$ $(\alpha = 1, \ldots, N)$ of the N data we want to use for this estimation. So we will have:

$$\bar{Z}(w) = \frac{1}{N}\sum_{\alpha=1}^{N} Z(x_\alpha)$$

Be that as it may. If we put

$$\bar{\gamma}(v,w) = \frac{1}{|v||w|}\int_v\int_w \gamma(x-y)\,dxdy$$

this extension variance can be expressed using only the variogram, by the formula :

$$\sigma_E^2(v,w) = 2\bar{\gamma}(v,w) - \bar{\gamma}(v,v) - \bar{\gamma}(w,w)$$

This basic formula calls for several remarks :

- The extension variance $\sigma_E^2(v,w)$ is symmetrical in v and w. Because of this choice of criterion, the quality of the estimation of $\bar{Z}(w)$ by $\bar{Z}(v)$ is systematically the same as that of $\bar{Z}(v)$ by $\bar{Z}(w)$.
- $\sigma_E^2(v,w)$ **is not a conditional variance**, in that it does not depend on the particular values of the data.
- Contrary to this, $\sigma_E^2(v,w)$ strictly depends on the structural function γ, and in particular on its behaviour at the origin which indicates the degree of continuity of Z.
- The extension variance depends on the geometry of the domain to be estimated — through the term $\bar{\gamma}(v,v)$ —, on the geometry of the sampling data — through $\bar{\gamma}(w,w)$ —, and on their relationship — through $\bar{\gamma}(v,w)$ —. For most of the variogram models, $\sigma_E^2(v,w)$ increases when the distance between v and w increases; on the other hand, for a totally unstructured variogram, (pure nugget effect), $\sigma_E^2(v,w)$ no longer depends on this distance.

The formula for the extension variance is all-important because it enables us to calculate the estimation variance of any linear combination of the data: not only mean values, but also integrals, convolutions and even derivatives on condition certain precautions are taken with the limits. There are numerous applications:

- For a given estimation configuration, calculate its variance. This calculation is possible since $\sigma_E^2(v,w)$ does not depend on the particular values taken by the data. Linear geostatistics therefore makes it possible to complete classical linear estimations by evaluating their quality. This type of problem is known as **global estimation.**
- In the same way, the criterion of the extension variance makes it possible to determine an exploratory strategy for a domain, to optimize a sampling data system or to search for an optimal location of additional data while taking into account the spatial structure of the variable under study.
- Likewise, we can use the extension variance for compressing the sampling data: how to remove samples while ensuring a minimum loss of information.
- Alternatively, we can try and find in a family of linear estimators the one that will lead to the minimal extension variance. This technique for calculating the optimal linear estimator, called **local estimation** or more commonly **kriging**, is used very often in routine geostatistics. It will be explained below.

First, beyond the mathematical formulation, we begin by questioning the physical interpretation of the extension variance or, in other words, its transcription in terms of the regionalized variable. In actual fact this transcription can only be established if there are "good" conditions of stationarity and ergodicity (*) which we assume to be satisfied here. **In these conditions** the property of non bias and the expression of the extension variance are explained as follows:

Let v be a domain (reduced to a point if necessary) on which we want to calculate the mean value $\bar{z}(v)$ of the regionalized variable, and let w be a domain in general composed of a finite number of points, on which the regionalized variable is known and whose mean $\bar{z}(w)$ will be taken as the estimator of $\bar{z}(v)$. We assume that the respective geometries and the relative position of v and w are fixed and we shall call the set composed by these two domains v and w the **estimation configuration.**

Then, taking into account the assumptions, if we translated the estimation configuration throughout all the space and if we calculated at each of their locations the experimental estimation error $\bar{z}(v) - \bar{z}(w)$, the non bias property shows that this error would be statistically zero and moreover its mean square value would be equal to the extension variance $\sigma_E^2(v,w)$.

In practice the situation is a little more complicated, firstly because the work domain is always bounded and therefore there is no question of carrying out translations on the entire space, and then because it is quite rare, except in the case of regularly distributed data, to be able to translate exactly the estimation configuration. Therefore the interpretation of the non-bias is more intuitive and less precise, and can be expressed in this way: if we perform an estimation with "approximately" the same estimation

(*) These two vitally important concepts will be described further on, but only briefly, in order to avoid the subtle passage from the model (random function) to the reality (regionalized variable) in this short presentation of geostatistics.

configuration throughout all the regions of the work area, then the errors made will tend to balance out. **There will not be a systematic error at the global results level.** And the experimental variance of the errors made will be of the order of the extension variance.

Kriging

The most elementary problem posed in estimation is to interpolate at a point x where the function z is unknown, given its values are known at N points $z(x_1), \ldots, z(x_N)$. So we try and construct a quantity

$$z^*(x) \quad = \quad \sum_{\alpha=1}^{N} \lambda^\alpha \, z(x_\alpha)$$

where the unknowns of the problem are the weights λ^α(*).

The aim is for this quantity $z^*(x)$ to be as close as possible to the unknown value $z(x)$. If this aim is interpreted on the probabilistic model level, and considering the choice made of a quality criterion, this means that we are trying to reduce the variance of the estimation error $Z^*(x) - Z(x)$. If we assume that the stationarity constraints are satisfied, this variance **is just a special extension variance** which is written:

$$\mathbf{E}\left[(Z^*(x) - Z(x))^2\right] \quad = \quad 2\sum_{\alpha=1}^{N} \lambda^\alpha \, \gamma(x, x_\alpha) \; - \; \sum_{\alpha=1}^{N}\sum_{\beta=1}^{N} \lambda^\alpha \lambda^\beta \gamma(x_\alpha, x_\beta)$$

by simple application of the general formula. It is a quadratic equation, according to the unknown coefficients λ^α which have to be minimized.

We are dealing with a minimization without constraint, provided that the random function Z is assumed to have a zero mathematical expectation. However, this is an exceptional case and in most cases this expectation — assumed constant in the model assuring stationarity — is unknown. Consequently, in order to ensure that the estimation error $Z^*(x) - Z(x)$ will satisfy

$$\mathbf{E}\left[Z^*(x) - Z(x)\right] \quad = \quad 0$$

we are obliged to impose the constraint

$$\sum_{\alpha=1}^{N} \lambda^\alpha \quad = \quad 1$$

This constraint ensures the **unbiased behaviour** of the kriging estimator. Later we shall find ourselves in the case where this constraint must be set up (the estimator used in the previous section to examine the influence of the model was constructed on this assumption).

Consequently, the kriging problem amounts to minimizing a quadratic form under a linear constraint. Ultimately we obtain a system of linear equations

$$\begin{cases} -\sum_{\beta=1}^{N} \lambda^\beta \, \gamma(x_\alpha, x_\beta) \; + \; \mu \; = \; -\gamma(x, x_\alpha) & \forall \alpha \\ \sum_{\beta=1}^{N} \lambda^\beta \; = \; 1 \end{cases}$$

The unknowns of the system are the weights λ^α and the **Lagrangian multiplier** μ. It suffices, after solving the system, to substitute the weights into the kriging estimator $z^*(x)$. As for the **kriging variance**, that is the minimum value σ_K^2 taken by $\mathbf{E}\left[(Z^*(x) - Z(x))^2\right]$, it simplifies to

$$\sigma_K^2 \quad = \quad \sum_{\alpha=1}^{N} \lambda^\alpha \gamma(x, x_\alpha) - \mu$$

where the λ^α and μ are the solutions to the kriging system.

(*) λ^n is the weight associated with the n^{th} point, not the n^{th} power.

Of course we assume that the kriging system has a unique solution which we always try and find. This means that the variogram model should not be too pathologically nasty (the periodic variograms can cause problems...) and that the kriging configuration is not degenerate. Intuitively this last point means that the data used comply with the constraints of quality and location. This being so, the system has a certain number of properties which should be borne in mind:

- Kriging is a multiple linear regression with correlated residuals.
- By construction, kriging is an unbiased estimator, which means that the estimation error has zero expectation in the probabilistic model. When considering the regionalized variable, that means that the average error on point estimations is zero over a large area.
- Kriging is an exact interpolator, which means that when we krige at a sampling point, the kriging system will return the sample value as the estimator (and a kriging variance of zero).
- Finally, by linearity and assuming the data set being used is fixed, kriging a linear combination of point values is identical to the same linear combination applied to the kriging of these point values. If * represents the operator "kriging estimation" and $\mathcal{L}$ is any linear operator,

$$\{\mathcal{L}(Z)\}^* = \mathcal{L}(Z^*)$$

That is why in theory it suffices to present the equations of point kriging. All the other systems (mean values, convolutions, differentials ...) result from it.

Suitability of the model to reality

The formalisms presented up to now make up the backbone of stationary geostatistics. The steps in structural analysis, then estimation (global and especially local) give rise to practical difficulties (fitting the model, choosing a work scale, choosing sampling data with a view to kriging, analysing the results) which call upon the user's experience, but do not usually cause theoretical problems. The methodology of what can be termed "classical geostatistics" has got into its stride.

However, although theoretically rigorous, this methodology is worthless if the mathematical model and the reality being studied bear no resemblance. In this case there is a real risk that the mathematical and computational operations no longer have any physical significance. This risk should never be forgotten throughout a geostatistical study and even more so when the models being used are more complex.

Consequently, from the very beginning, choosing probabilistic methods can pose problems. How can we propose a random function model with all its wealth for phenomena that are in general unique ? In certain cases (time series) we might have several realizations of the process available and therefore a classical inference may be considered: but especially with earth sciences the phenomenon is unique. What meaning can we then give to a probability ? For when the calculations are done, a physical meaning must be given to the results obtained. It is then that the hypotheses of **stationarity** and **ergodicity** intervene.

Stationarity

By this first hypothesis, we assume that the phenomenon being studied presents a certain **structural permanence** in its domain and so the observations made in different parts of the space can be considered as **different realizations of the same process** which we are trying to model. By this hypothesis, we have removed the obstacle of a unique realization. However, on the one hand this assumption is fundamentally **refutable**, for it is quite possible that no stationary model is compatible with the data: then non-stationary geostatistics will have to be considered. On the other hand, from a statistical point of view, we have not yet got out of trouble, since if we now have available several realizations of the process to be modeled, it is clear thar these **realizations are not independent**. It should be noted that in geostatistics this difficulty is usually ignored, insofar as it does not present bias the estimation of the behaviour at the origin of the variogram.

Ergodicity

A model is **ergodic** if the inference of its parameters can be realized from any one of its realizations. This second constraint may not come to mind so clearly when putting it into practice. And yet **we have to assume our models are ergodic**, for otherwise it would no longer be justifiable to compare the

spatial mean values performed on the regionalized variable and the mathematical expectations applied to the random function. Of course, "classical" models include this property, but once again, this assumption can eventually be refuted during a study. In this case, we must have recourse to methods going beyond "classical" geostatistics.

However, it must be noted that the property of ergodicity, like that of stationarity, is essential for applying methods of a probabilistic nature. And finally, the developments of "non-classical" geostatistics, some of which are mentioned below, always try and produce "something" stationary and ergodic.The whole difference between the various methods proposed lies in this "something".

Extrapolating the model

There is one final difficulty which arises, however sophisticated the geostatistical method might be. It is quite normal to remember the limitations of the model for large distances: we know that, even if the domain of validity of the model is infinite, it would be unreasonable to use the values of the variogram for distances greater than those chosen for the model in the computations. Moreover, this is only reasonable, as the model does not always "look after itself" properly: so we can consider cases where an apparently "unrealistic" extrapolation would be associated with a seemingly admissible kriging variance...

It is easily forgotten that this type of problem also arises for short distances. One of the decisive features of the structural function is its mathematical behaviour at the origin (differentiability properties). Now, inevitably, this structural function is fitted from a finite number of data, which therefore does not permit the passage to infinitesimal distances. This does not proscribe the method but means that when we fit a variogram model, **we add information which is not available in the data.** Besides, this situation, which is irreversible, is a good thing: if initiative played no part at certain stages, geostatistics would be reduced to a tautological manipulation of data.

This is an example of a situation where experience is decisive, the geostatistician having at his disposal an unavoidable freedom of choice. As long as we do not have available additional information which could refute the choice, it is justifiable to orient the model fitting, for example, according to the type of problem set. Depending on whether we want an aesthetic map or a risk map, a precise one or one that resembles reality, we will not choose the same variogram. There is nothing dishonest in this, provided that the part left to discretion in the structural analysis is clearly defined in the final comment on the results.

The limits of the linear tool

"Linear geostatistics" is easy to use and calls for few prerequisites. However, because the random function distribution is only taken into account by the variogram, phenomena with very different natures but having the same structural function will be dealt with in the same way, for example by kriging. This confusion would be best avoided but it is the price we have to pay for simplicity.

The geostatistician is responsible for evaluating the need for more highly developed methods. So let us consider kriging as an example, which is the best (meaning minimum variance) linear estimator. But is the range of linear estimators the "best" choice? And is the criterion of the minimum variance sufficient? For a Gaussian random function we know that the regressions are linear ; in such a case we can easily show that the kriging estimator is associated with the conditional expectation : consequently, we are assured that kriging is the best possible measurable estimator. But what happens when our distributions are very far from the Gaussian type? Experience shows that for asymmetrical distributions, linear estimators are no longer suitable and that even for solving simple point estimation problems, non linear geostatistical methods should be used.

Likewise we cannot be certain that the variance criterion is always good enough. For example it does not always "recognize" asymmetries, multimodalities, etc... The symmetry along v and w of the extension variance $\sigma_E^2(v, w)$ can justifiably be considered as unsatisfactory. Moreover the variance which we minimize in kriging is not conditional, so that the weights will be the same whether they are in a region of weak or strong data values.

These disadvantages exist and should never be lost from view. However, they are probably of little importance compared with the temptation of applying unrealistic models which have nothing in common with the data. For example, it would be tempting to use the medians rather than the expectations, for reasons of robustness. Unfortunately, the slightest calculation involving information from different supports immediately becomes intricate if we use the medians, and we are led either to make assumptions

on the model which cannot be proved, or propose *ad hoc* approximations which falsify the work. The advantages of the operation have completely evaporated.

In summing up this introduction to linear geostatistics, the geostatistician's sense of analysis and initiative are of the utmost importance if he is to bring his study to a satisfactory conclusion. There is nothing more foreign to the geostatistical spirit than a program of the "black box" type.

Expansions

The geostatistics mentioned up to now is elementary and its necessary developments are numerous. An incomplete list is presented below.

These extensions of "basic" geostatistics are of two kinds, depending on whether we are modeling phenomena which cannot be treated with the "basic" tools (non-stationary geostatistics, multivariate geostatistics) or whether we are replying to new methodological problems (non-linear geostatistics, random sets, conditional simulations etc.).

Non-stationary geostatistics

It often happens that no stationary variogram model is compatible with the data being studied. This can be due to the presence of a trend, in the intuitive sense, having a regular behaviour and at a scale which is similar to the size of the domain being studied (low frequency phenomena). In that case, the best course of action is to use classical methods to show up "something" stationary in this more complex phenomenon.

A first approach — a very intuitive one — consists in proposing a dichotomy. We try and describe the global variable as the sum of two components, one that could be treated using linear geostatistics methods and the other one considered as a drift (or trend). This approach is interesting as it is generally accepted by naturalists. Besides, it is the approach used when examining time series. This doubtless explains why this approach was formalized first under the name of universal kriging.

However, this mistakenly simple approach should be handled with care. For there is no reason why the demands of the naturalist, who requires a physical explanation to the dichotomy, should meet those of the geostatistician, who wants a stationary component to be present. Moreover, using the variographic analysis with a view to universal kriging is quite difficult. That is why an alternative approach was proposed : the theory of intrinsic random functions of order k (IRF-k). This time the approach is to propose a **transformation** of the data which goes back to "something" stationary. This transformation is similar to a differentiation of order k and so results in **filtering** out the most regular components of the phenomena.

The first advantage of this second approach is to permit us to use a much wider range of structural functions than the family of variograms. These are the generalized covariances. In this way, we are able to model variables which are much more complex than those in stationary geostatistics. If K denotes the generalized covariance and f^l the family of the basis functions describing the drift we filter, the kriging in IRF-k — or **intrinsic kriging** — is written

$$z^*(x) = \sum_{\alpha=1}^{N} \lambda^\alpha z(x_\alpha)$$

where the λ^α are the solutions to the equations

$$\begin{cases} \sum_{\beta=1}^{N} \lambda^\beta K(x_\alpha, x_\beta) + \sum_l \mu_l f^l(x_\alpha) = K(x, x_\alpha) & \forall \alpha \\ \sum_{\beta=1}^{N} \lambda^\beta f^l(x_\beta) = f^l(x) & \forall l \end{cases}$$

In this formula, the unknowns μ_l are Lagrangian parameters. There are as many as there are basis functions f^l to filter. The intrinsic kriging formula is the same as that obtained in kriging with a variogram, except that the function $-\gamma$ must be replaced by K. Consequently, the characteristics of intrinsic kriging are like those of kriging with a variogram : linearity, exact interpolation, non-bias, etc...

The study of non-stationary phenomena from the IRF-k viewpoint presents numerous advantages :

- The automatization of the variographic analysis is much easier than for universal kriging ; in particular it avoids the biases which are innate in the treatment by dichotomy. However, the former comments are still true : a blind variographic study is inadvisable.

- We can easily deduce a **dual presentation** from the intrinsic kriging system :

$$z^*(x) = \sum_{\alpha=1}^{N} b^\alpha K(x, x_\alpha) + \sum_l c_l f^l(x)$$

where the coefficients b^α and c_l, which do not depend on the point to be estimated x, are the solutions of a linear system having the same structure as the intrinsic kriging system. In this way, by storing these coefficients, it is easy to estimate at as many points as desired extremely rapidly.

- The formalism of the IRF-k makes it possible to prove the **formal equivalence between splines and kriging.**

Multivariate geostatistics

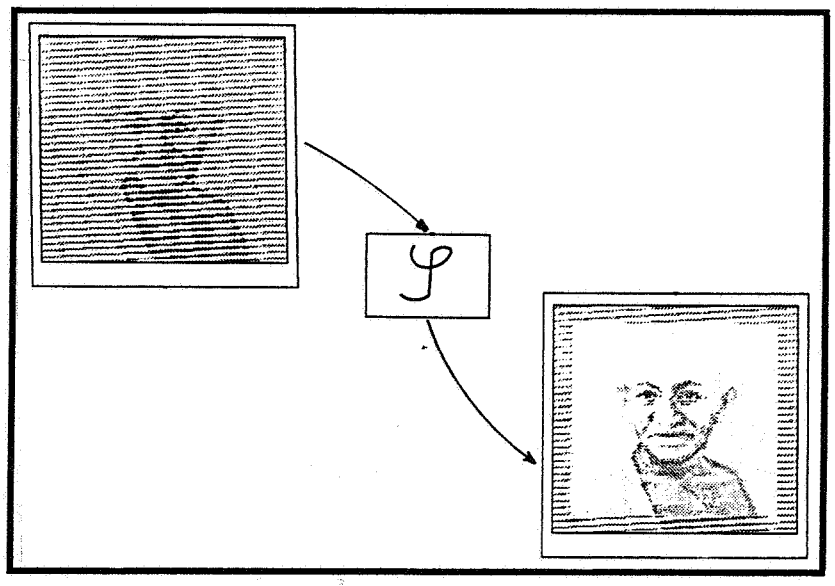
In practice, being able to treat several variables simultaneously is important. In linear geostatistics we know how to treat pairs of variables whose joint structure is defined by **cross variograms**

$$\gamma_{YZ}(x, y) = \frac{1}{2} \mathbf{E}[(Y(x) - Y(y))(Z(x) - Z(y))]$$

where Y and Z denote the two random functions to be studied. When doing the variographic analysis, we must make sure of the theoretical regularity of the set made up of all the cross variograms and the univariate variograms. The variography can then prove difficult...

Multivariate geostatistics can be applied in many ways.

- **Cokriging.** This involves a linear estimation of a variable with data provided on other variables if necessary. Common sense shows that cokriging is of little interest for non-correlated variables (neither one provides information on the other), or, quite the contrary, if the linear correlation is too strong, the data is superfluous. That said, there are many applications of cokriging. For example, we may want to cokrige a signal using noisy data (**filtering**). Here cokriging seems a generalization of Signal Processing, but it has the property of not being constrained either by the spatial dimension or by the lay-out of the data. Note that nothing prevents the noise from being structured or from having a non zero mean, or even from being correlated to the signal. The figure below gives an example of filtering by means of cokriging, applied to an image.



Filtering a noise by cokriging

Processing data that is full of **errors of location** is another example of using cokriging. This problem is specially met when treating the results from maritime campaigns. In general, the locations of data from the same navigation profile are well known relative to each other, but there may well be an

important imprecision on the relative position of two distinct profiles. Besides, the uncertainty of the location has an important effect on the variographic analysis. In this way, when two profiles cross, it may happen that two measurements which are considered to be located at the same point have different values, which gives an **"apparent" nugget effect** to the experimental variogram. Modeling a regionalized variable where the location is uncertain becomes very difficult in the non-stationary framework.

- **Variables linked by linear equations** (especially equations with partial derivatives). Examples of these problems abound : uranium deposits where the grades and radioactivity (which is approximately a convolution product of the grade) are treated simultaneously, meteorology where wind and pressure are linked by equations, environmental flow configurations etc... The equations should be incorporated firstly on the basis of the global structural model. But the approach of using the data (cokriging neighborhood) is also strewn with problems because of the links between the variables.
- **Kriging analysis**. The idea is to reveal basic components with different scales (different frequencies) in the global structural model, and to consider the variables under study being made up of superpimposing these components. The next step consists in studying these basic components separately and hoping to to be able to give them a physical significance. This is a sort of fusion between data analysis and a Fourier analysis. We are trying to give a precise meaning to the notions of components and anomalies. However we are asking a lot of the model and consequently the dialogue with the naturalists and the demands of realism are vital.
- **External drift**. Here we use two variables having different properties. The variable of interest is known from a very small number of reliable data, whereas the data on the second variable are much more numerous and are supposed to give information on the general structure of the variable of interest. As an example, petroleum can be mentioned, where the few well data are completed with very rich seismic information. The heterogeneity of the variables and the small number of data on the variable which interests us make it impossible to cokrige in good conditions. Therefore, we use the second variable as a guideline which sets out the main shape (**drift** role) of the variable in question. This technique will no doubt develop with the multiplication of satellite images which add to the ground data (in soil science, agronomy, bathymetry etc.)

Simulations

In the introduction, we mentioned the calculation of the length of an underwater cable on the seabed as a possible utilization of a bathymetric map. Kriging is unsuitable for this sort of problem, because an estimated map does not give a true picture of the structure of the variable. The estimation procedure (here simple interpolation), results in a **smoothing** of the shapes. By minimizing the variance, we get rid of the high frequency points and fluctuations. In this way, in the example of the underwater cable, using a kriged map would lead to **underestimating** the necessary length — probably most considerably. Let us just mention here that this problem is basically non-linear and so it is not at all surprising to see that kriging is unsuitable.

So we aim to construct an acceptable image of reality, that is a **numeric model** giving the structural characteristics of the variable being simulated. What is more, nothing stops us from making a large number of these models. They are used to visualize structures, but also to make it possible to solve problems (for example non-linear) which cannot be solved theoretically. We can also make simulations of industrial processes (for instance mining) on these numeric models. However, we must nevertheless note that this technique does not replace an estimation. It can be shown that a point simulation makes a poor estimator (as compared with the usual criterion) because the variance is double of that of kriging.

An **unconditional simulation** is a numeric model giving the distribution and the structure (variogram) of the variable. There are many techniques, one of which is the **turning bands** method. The aim is first to produce one dimensional simulations, using techniques like those used in time series (autoregressive processes, moving averages), then to spread the results obtained in $\mathbf{R}^n$ by integration along all the directions in space. Indeed, geostatistics provides the formulas for passing from $\mathbf{R}^1$ to $\mathbf{R}^n$, which are especially simple for $n = 3$. Note that the turning band method is not basically different from some reconstruction techniques used in tomography.

There is no reason for the values of an unconditional simulation to be close to those of the data. There is only a similarity at a statistical level and not on a level with particular values. With **conditional simulations**, we ask the numeric models to respect the data values, as well as having the statistical

likeness. The simulation is thus "held in place" by the constraints which will be stronger, the more there are conditioning data. This time we have a model at our disposal which not only resembles statistically reality but in addition is close to this reality at the neighboring data points.

The interest of conditional simulations is clear for solving the questions we do not know how to attack on a theoretical level. For example, by using a large number of conditional simulations we can evaluate numerically the confidence intervals of the estimator — something we do not know how to do theoretically just with linear geostatistics tools. However, we must take care. In conditional simulation the structural model is very much in demand and care should be taken to see that the conclusions proposed are significant and not only the result of artefacts (problems of discretization, truncation, etc...). Besides, a considerable number of values always have to be simulated. At this stage finding rapid algorithms is crucial. The domain of conditional simulations is a particularly busy field of research.

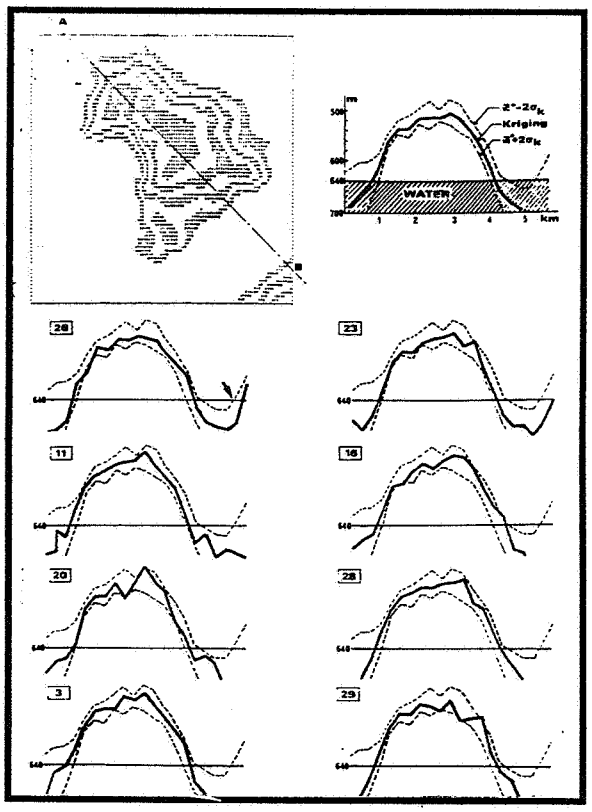
As an example of conditional simulation we propose a vertical section of an oil dome. The map shows depth contours of the roof of the structure obtained by kriging, and the first section represents the cross section along A-B of this estimated roof. The interval $Z^* \pm 2\sigma_K$, can also be seen, where Z^* is the kriged value and σ_K the kriging standard deviation.

This is a typically non-linear problem. We must estimate the trapped oil volume, that is the volume between the roof of the structure and the water level which we assume is known. A major difficulty and theoretically unsolvable one is determining the limits of this volume, or the intersection of the roof of the structure and the water level.

Let us first examine the shape which, in general, is much more irregular in the conditional simulations than in kriging. In order to visualize this effect, the simulated values are represented by reference to the interval $Z^* \pm 2\sigma_K$.

We note that on simulation 20 the simulated values can leave this interval. This confirms that, at least vis-a-vis the minimum variance criterion, the conditional simulation can prove to be a very poor estimator.

Incidentally, the example of simulation 26 draws attention to the risks we run if we limit ourselves to only one simulation. Indeed, we see the simulated roof goes up on the right side of the cross section and goes beyond the water level. In this simulation, the structure which is likely to contain oil is therefore in two parts, and, if proved right, would be of prime importance for estimating the volume. In actual fact, if we observe a large number of simulations, we realize that this is rare and consequently it would be unwise to base an entire study on the very special properties of simulation 26, even more so as it concerns a zone simulated in extrapolation and associated with a very strong theoretical variance.



Simulation of an oil reservoir

Methodological analysis

The discussion on estimation/simulation enables us to end the analysis on the role and importance of the model in the geostatistical approach. By simulating a given structure conditionally, we can show the impact of the choice of the structural model, or visualize the effect of a change of parameter (range, nugget effect, etc...). At the same time, by kriging from conditioning data, we show the difference of behaviour between estimation and conditional simulation.

This comparison is not entirely theoretical. It also gives information on the way models react to certain situations (missing data...) and on the soundness of the results. In this way, it is possible to

develop the geostatistician's experience by non-quantitative data which are very important when making choices : for example, when deciding between two equally plausible models for the variographic analysis or when fixing a research program of neighborhood for the kriging analysis.

To illustrate this point, we provide five pages of figures at the end of this document which illustrate some important notions :

- Importance of the **range**. In a stationary model, the range is the distance beyond which two point data no longer have a significant correlation. This parameter corresponds to the intuitive idea of a zone of influence. All things being equal in other respects, the larger the range, the more the phenomenon will be structured on a large scale. Contrary to this, if the range tends towards 0, there is a transition towards the nugget effect and total absence of spatial structure.
- Figure 1 shows four non-conditional simulations, constructed with a spherical variogram model, for four increasing values of the range. As expected for ranges 50 and 75 we can see the outline of important structures on the scale of the domain, and which were completely absent, especially for range 10. Moreover, although the model used for these simulations is stationary, if we performed a variographic analysis on the simulations of ranges 50 and 75, we would surely come to the conclusion of a non-stationary phenomenon — **on the scale of the domain.**
- Moreover, it is clear that if these simulations were data from which we wish to perform a kriging, the procedure for locating the neighborhood should not be the same for ranges 10 and 25 on the one hand, and ranges 50 and 75 on the other. Indeed, in the first case, the absence of any prominent global structure makes it impossible to commit a serious error of estimation, if the estimator is close to global mean. Therefore, it is important to use a maximum amount of information when the mean value will be a suitable estimator: Contrary to this, for the long ranges, the phenomenon shows a strongly structured global geometry and it is therefore of the utmost importance, in order to make an estimation, to have well located information, and above all, close to the point to be estimated. This example illustrates an empirical rule (it is not a theorem and exceptions can be found!). When a phenomenon has a weak structure, it is usually better to favour the **quantity** of data (classical statistics) . When the phenomenon is strongly structured, it is the **quality** of the data which prevails.
- Importance of the **type of model**. Figure 2 shows four simulations of models having equivalent practical ranges. The difference between these models is primarily due to the behaviour at the origin. The spherical and exponential models have a **linear** behaviour at the origin which corresponds to a **continuous** random function. The "cubic" model is twice differentiable at the origin, which corresponds to a random function that is once differentiable. Finally, a random function having a gaussian variogram is indefinitely differentiable.

  As the range is quite big in relation to the simulated field, we can see structures take shape on these four simulations. But for models having a linear behaviour, these structures have very rugged boundaries, especially the exponential model. Already the cubic model has much sharper boundaries and the zones which take shape have strong structures. In the gaussian case, the isolines are very smooth and it is likely that a variographic analysis on the scale of half the field would indicate the presence of a drift which is absent in the underlying model.
- Importance of the **nugget effect** — figure 3. For a phenomenon with fixed global variance, we increase the part due to the nugget effect in the model. The effect of destructuration goes without saying. For a 100% nugget effect, we are dealing with a totally unstructured white noise.
- Figures 4 and 5 have a common presentation. A first figure proposes "reality", in fact a non-conditional simulation made with models having equivalent practical ranges. The difference between the spherical and the gaussian structures — already mentioned — can be seen again. Then on these simulations, considered as the reality, we carry out a sampling (the same for the two models). From this sampling, we perform the kriging in unique neighborhood — using all the data. This result is completed by the standard deviation map.
- We have already shown that kriging smooths and that a kriged map is more regular than reality. Naturally this effect is more noticeable for the spherical model as the initial shape was much more erratic. Contrary to this, the effect on the gaussian model is less noticeable and the kriged map is very close to reality. This means that in the present circumstances, it would be unimportant to mistake reality for the kriging estimator. However, it should be noted that there is a significant number of data points (100) and that the gaussian model might give rise to artefacts if the kriging was constrained by too small a number of data.

Non-linear geostatistics

Simulations are one way of solving non-linear problems. However they run the risk of asking too much from the model, and making the simulated numerical values say more than is realistic. For this reason when possible, it is advisable to look for a theoretical approach to non-linear problems. Three types of questions are treated in non-linear geostatistics.

- **Support effect** : The basic formula of the extension variance makes it possible to calculate the variogram of a regularized random function. But this formula is inadequate if we want to construct a **change of support model**, a model for the bivariate distribution $\left(Z(x), \bar{Z}(v)\right)$. This is a problem particularly in mining geostatistics, where the samples and selection units have very different volumes. It is important to know the distribution of the selection units conditionally to the sample distribution.
- Still taking mining geostatistics as an example, we often make choices (cutoffs, selections) on the basis of Z^* estimators, not of real values Z. Kriging yields a smoothed version of reality and so clearly it does not have the same distribution. This effect is called **information effect**. So we must be able to model the bivariate distribution $(Z(x), Z^*(x))$.
- Finally, we are often led to estimate not the values (point or regularized) but the **probabilities of exceeding the threshold**. In mining, the notion of **cutoff** is very important, as for constructing grade/tonnage curves. This is fundamentally a non-linear notion.

It is to answer these questions of non-linear geostatistics that disjunctive kriging was developed. It is an intermediary tool between linear regression (kriging) and regression (conditional expectation). Stronger than linear geostatistics, disjunctive kriging needs a modeling of the bivariate distribution and is no longer satisfied with moments of order 2 (variogram). So we see the notion of distribution reappear, which was absent from linear geostatistics. And, as in general we do not know how to handle arbitrary distribution theoretically, an important tool in non-linear geostatistics is the **anamorphosis**, a transformation which aims to convert the initial random function to a random function with a known distribution (generally gaussian).

Finally, let us note that disjunctive kriging demands strict stationarity in order to infer the bivariate distribution. In this way, even if the range of questions we can tackle theoretically is considerably enlarged compared to linear geostatistics, nonetheless the range of phenomena we can handle is reduced. The development of non-linear, non-stationary geostatistics remains an open question.

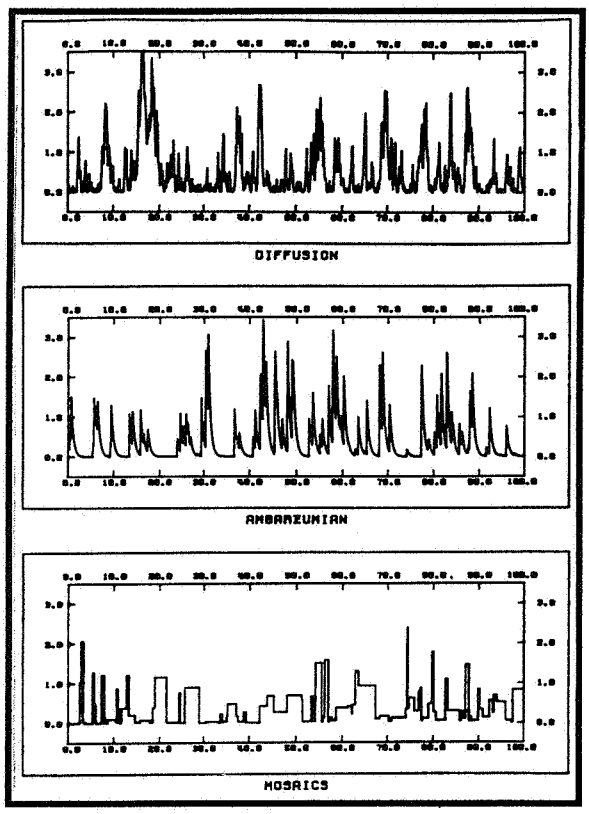
Perspectives

Geostatistics can choose from a wide range of tools which has grown over the years. However we should not forget that these tools may prove to be inadequate for modeling the fundamental characteristics of certain complex phenomena, and that as a result it will be necessary to look for new tools.

In the classical domain which we have considered in this presentation up to now, let us remember some of the developments in view.

- For instance it is clear that the cross variogram, which is a pair function by construction, does not "see" the systematic shifting between variables (which are present in flow phenomena) : meteorology, pollution, alluvial deposits etc... In this case, there is an answer, at least in the stationary frame, that is to work in cross covariance rather than in variogram. In this case, the theoretical basis exists. The practice has to be developed.
- Another field of research : the **change of scale**. More and more frequently, the use of satellite images as a means of giving more information at ground level (agriculture, soil science, oceanography etc.) forces us to work with heterogeneous data representing very different sized domains. This poses problems of scale conversion, in addition to variables which are often not additive (for example data of a geometrical kind).
- Another very important domain : modeling **spatial/temporal phenomena**. We can no longer merely "photograph" a regionalized variable, but must describe its evolution in time. Adding another dimension does not solve the problem, as the evolution obeys other laws or at least is limited by inequality constraints. Multivariate geostatistics has to develop for the study of variables linked by equations.

- Another subject for reflexion : what information does the simultaneous understanding of the law of distribution and covariance bring on the structure of a variable? (the figure below shows three examples of realizations of processes having the same covariance (in this case exponential) and histogram.



Three simulations with the same distribution and the same covariance

The first simulation is a diffusion model.

The second one is an Ambarzumian process (exponential decrease on a Poisson model). It is especially interesting, because it is asymmetrical. It is evident that the distribution of $Z(x+h)|Z(x)$ — the distribution of $Z(x+h)$ **conditionally** to $Z(x)$ — is not identical to the distribution of $Z(x)|Z(x+h)$. This property is not "seen" by the variogram...

Lastly, the third example is a mosaic model. The process is piecewise constant.

A tool that cannot distinguish three such different phenomena is unsuitable. That is why we must choose stronger methods than geostatistics of order 2, which confines itself to using **moments**. A promising analytical tool is the **"deferred" scatter diagram**, a scatter diagram between the random variable $Z(x)$ and $Z(x+h)$. But the application of these tools which are considerably more refined than those of linear geostatistics requires a much better quality data, and especially very strong conditions of stationarity.

The list is not yet closed as geostatistics is developing by leaps and bounds. We have not mentioned here **random sets** which is a vast and promising domain in full expansion. It concerns a section of geostatistics which bridges the gap with mathematical morphology, by granting more importance to the **geometrical** aspect of the data. The mathematical tools applied are **set oriented** rather than being numerical. The many different applications not only give images which look more like reality (with, for example, in stratigraphy the constraints of inclusion), but also approach typically non-linear problems, as, for example, questions of connexity which are important for flow phenomena (oil wells, propagation of pollutants). The field of random sets, which will be applied more and more in the future, is also a more fundamental field of research in what concerns the problem of the inference of models and the conditioning of simulations.

Link with other methods

Its very nature makes geostatistics a subject where mathematical methods and fields of application encounter one another. Some of these methods have led to theoretical innovations (for instance random sets), but the majority are quite classical and because of this, are used for problems which are quite different from those encountered in geostatistics. Besides, practitioners of natural sciences have mathematical tools adapted to their field of study for their particular problems. Consequently, geostatistics is certainly not an isolated subject. Therefore it is essential to place it in relation to other similar methods, and eventually to review its fields of application.

On the methodological level, geostatistics proposes a basic choice :

- the analysis of the **spatial structure** of the data

and its application are characterized by three factors :

- we work with **sampling data** which provide us with incomplete knowledge of the field being studied.
- we work in a metric space, but, *a priori*, of any dimension. In particular, we are not obliged to go back to $\mathbf{R}(1\ D)$.
- Finally, we do not have to go back to a regular grid. The methods proposed are applicable whatever the distribution of the data.

These four features of geostatistics will help compare similar methods.

Probabilistic methods

Apart from the transitive methods, geostatistics operates on **probabilistic models**. As this formalism is based on a criterion of variance, it is mainly on the **first two moments** that the work is carried out. Consequently, in stationary hypothesis, the suitable tool is the covariance function whose mathematical characteristics are defined by **Bochner's theorem**. This theorem, which links the structural function and its Fourier transform makes it possible to compare geostatistics with the **spectral methods**. When the hypotheses of stationarity must be extended, we must resort to a generalization of Bochner's theorem, but the technical complications involved do not change this parallel. Consequently, when teaching the theory of geostatistics, it is also interesting to present the random functions from the point of view of **harmonic analysis**, so as to allow for the comparison with the standard methods, for example in **signal processing**. In this way, it can be seen that the cokriging of errors is nothing more than a **filtering**, and that kriging analysis resembles a **frequential analysis**. The practical difference is based on the fact that geostatistics works well in more than one dimension and, above all, it is totally unnecessary to have regularly spaced data.

Whatever the hypothesis of stationarity, we always go back to the handling of random quantities having a variance. Consequently, the mathematical frame of work is a $\mathbf{L}^2$ space of random variables and so the minimization of a variance can be presented as a **projection** of random functions on a Hilbert space. In this way we have a geometrical presentation of kriging. The characteristics of the kriging are then an immediate result of the projection theorem : for example the **theorem of additivity**, proved in the universal kriging framework is just the standard theorem of the three perpendiculars.

We can finally consider the tie between geostatistics and **stochastic processes** — or random function methods. As geostatistics is generally used in spaces having several dimensions, it makes little use of formalisms such as **Markov chains**, **waiting times**, **queues**, etc... because it does not benefit from an ordered structure in its domain of work. More general, geostatistics may thus appear less fitted to the special processing of **time series**, especially because transferring to $\mathbf{R}^n$ of tools which are specifically unidimensional seems to be extremely difficult. However, developments in the geostatistical study of $\mathbf{R}^1$ could throw new light on this field where, in truth, other methods which are perfectly suitable, already

exist. In any case, methods such as **diffusion processes** are already applied in geostatistics, not in the "geographical" work space, but in the "state space" taken by the variable (transition of stratigraphic features for example). Moreover, the **turning band** simulation method which starts by simulating unidimensional processes, sometimes makes use of tools from the time series : **moving averages**, autoregressive processes, or, more generally, **ARIMA** processes.

Statistics

If only because of the name geostatistics, we should consider the link between geostatistics and classical statistics. In fact, this is rather complex.

The first point to be observed is that, historically, geostatistics broke away quite early from the questions concerning the **distribution** of the regionalized variable in order to concentrate on the "spatial structure" aspects. In fact, this break is rather deceptive as we know that the choices made — that of the linear estimator and the variance as a quality criterion — are optimal in the case of a gaussian random function and become less suitable as we "move away" from the gaussian case. In other words, the underlying distribution is important even if it is not explicitly considered in the formalism of linear geostatistics. Moreover, by the use of anamorphoses, non-linear geostatistics takes the underlying distribution into account indirectly. Indeed, current theoretical developments are moving towards an analysis and modeling of bivariate distributions.

The real problem between statistics and geostatistics probably comes from the role allocated to the model. For the geostatistician, the probabilistic model is essentially a **tool** which is forged for the specific needs of the problem. A "real" model is not hiding behind the sampling data, any more than there is a "real live" parameter set to be found. Applied geostatistics should be considered as more than a mathematical method but as a **physical science**, where experience counts for more than the model. That is why **statistical inference**, or the fitting of the model, is simply taken as an **approximation** which we are obliged to do unless we know the regionalized variable exhaustively, and any additional information which crops up during the study **must be used** to verify the model. The purely operative role given to the model for the geostatistician explains that it is not normal to make statistical tests in order to evaluate the quality of the variographic analysis, the best test being the comparison with reality. Consequently, a parameter of the model which has no physical significance is at best a step in the calculations and at worst a parasite presenting artefacts.

This point of view shows up the difficulties in communicating with the world of statistics. The best-known illustration of the differing points of view is the fitting of the variogram model and more especially of its behaviour at the origin. It is quite possible to devise statistical tests which measure the distance of a model from the experimental variogram. The problem is that, quite rightly, several tests can be considered, they in turn demanding hypotheses on the model and the conclusions can be radically different. Let us just mention here that this approach is proposed in the Bluepack software program and indeed the answers are sometimes ambiguous. Now, we have already seen how the model can affect the results (simulation or estimation). So unless the geostatistician has solid criteria, he will prefer to use his own experience in making his choices — while free to do so — in terms of the problem in hand. It will be up to him to correct the model when additional information makes it possible (and *ipso facto* requires this...).

A truly statistical approach is only possible when there is a underlying model, that is when we work on a numerical model. Contrary to this, when we work on samples of real variables, the philosophy of a geostatistical study could, if pushed to the extreme be expressed in the following way : the moment the algorithms are mathematically correct, there are no right or wrong results, but only suitable or unsuitable results, a result being meaningless unless accompanied by a list of operating instructions (**laboratory conditions** in physics).

Other methods

Geostatistics can be used to study geometric forms, especially in its transitive presentation (non-probabilistic). What makes it different from methods like **image analysis** and **mathematical morphology** is the use of fragmentary, sampled data. Consequently, it presents an **interpolative** aspect which is not present when working on a totally known image. However, this difference is diminishing, now that satellite images are being used more and more as a complement to ground data. Besides, geostatistics used to be set apart from mathematical morphology because of the structural tools employed. Mathematical morphology uses continuous objects whereas geostatistics only uses finite objects (a point

pair for variograms, several points for generalized covariances). But here too the difference is disappearing and random set geostatistics closely resembles mathematical morphology.

Kriging without a change of support **is** an interpolator. Consequently, we can consider its ties with the methods of **numeric interpolation**. Contrary to methods such as "distance inverse", "inverse of the square root of distances" etc..., contrary to the polynomial fitting by least squares, the role of geostatistics is to take into account the intrinsic structure of the variable to be interpolated and so does not propose an "all-round" estimator. For the quasi-totality of natural variables, there is not any "down stream" constraints, for example on the analytical expression of the estimator. We do not have the same viewpoint when interpolating a car's coachwork or the wing of a plane on the one hand, or a pollutant grade or an atmospheric pressure on the other hand. Geostatistics is bound by "up stream" constraints. Therefore, the theorem which expresses the formal equivalence between interpolation using **splines**, and kriging (note that "**splines d'ajustement**" are like kriging with a nugget effect), is an interesting subject to examine. Although there are very differing points of view in their utilization , splines and kriging are one and the same thing. This does not mean however that setting up a kriging given in terms of splines, or the reverse, is easy!

Kriging analysis resembles a frequential analysis, insofar as it separates a global phenomenon into components with different scales. However, in its multivariate version, it also resembles **data analysis**, because it aims to show the similarities and the differences of the variables being examined simultaneously. But in kriging analysis it concerns a structural similarity , shown at the cross variogram level, rather than a statistical similarity. It must be noted that the characterization of a global multivariate model with a kriging analysis in view, relies on an analysis in "**principal components**" performed not on the regionalized variables but on the parameters of the structural functions.

Finally, we should consider the link between geostatistics and **fractals**. The fractal or non-fractal characteristic of modeled random functions in geostatistics (or their realizations) depends on the behaviour of the variogram at the origin. However, this is not characteristic. Therefore a gaussian random function on **R**, having a stationary exponential covariance, will be fractal with a 1,5 dimension (*), whereas a Markov chain having two states with the same covariance will not (1 dimension). We should note that here we are talking about a **fractal dimension** and not an **self-similar property**. A stationary random function (non constant...) cannot be self-similar. On the other hand, when the hypothesis of strict stationarity is removed, geostatistics using IRF-k makes it possible to model the fractal phenomena just as well as the self-similar phenomena. This said, these properties should be taken into account with care, exactly like the behaviour of the variogram at the origin, because they cannot be analysed strictly from a finite number of samples and so proceed from an extrapolation towards infinitely small scales. For the same reason, we cannot be sure that the fractal characteristic of a random function is a decisive property, at least as far as simulations go. For, in the end, it is always a finite number of values that we construct...

(*) Hausdorff–Besicovitch dimension

Fields of application

Geostatistics is mainly a collection of **methods** and so in theory is not reserved for any particular domain. In one way this is an advantage insofar as algorithms and practical aspects can be tested on the most widely diverse variables. However, it is also a handicap : as each subject has its own formalisms, habits and vocabulary, it is sometimes difficult to promote a course of action considered too vague. That is why, in the past, geostatistics favoured domains which were receptive to its new methods.

Original fields of application

Geostatistics was first developed in **mining**. The vocabulary ("nugget effect" referred to as "white noise" in other domains) shows this clearly. The circumstances were favourable as it was the mining world itself (or at least some of its representatives) who were looking for new methods. In the 1950s, the problems were of a statistical kind (bias correction) ; however, it is interesting to note that once contact was made, geostatistical research finished by covering all the steps in a mining project : from **prospection** to **open pit optimization**, not forgetting **estimation**. So the development of non-linear geostatistics at the beginning of the 70s was warranted at the outset solely because of problems in mining. Non-specialists certainly should not perceive mining as a monolithic entity and even a general subject such as geostatistics should consider different approaches for gold, nickel, uranium or coal mining. In this way, because of the variety of studies undertaken, geostatistics has been able to develop its resources considerably, both in its applications as well as in its theoretical development. In spite of the more difficult international situation today, geostatistics is still thriving.

However, despite the seemingly good conditions, a lot of time was necessary before geostatistics was accepted as a standard tool. Almost twenty years passed by between the first treatise on geostatistics and the first utilization in routine of geostatistical estimation software programs. Barely less time was needed before geostatistics was accepted in the **oil** industry, but now it is a privileged domain of application. Yet again, communication was one of the reasons for this lapse of time. The oil industry has its own way of tackling problems and its own vocabulary which are far removed from those of mining geostatistics. Another difficulty comes from the fact that oil companies have highly developed, consistent data processing environments to which the "newcomer" — geostatistics — must fit. But once these difficulties were overcome, the oil industry proved challenging profitable by bringing up a new range of methodological problems for geostatistics to solve. The first particularity of oil wells comes from the non-stationary nature of the geological structure under study. The second one is the considerable importance of the geometric aspect : discontinuities (faults, contacts), heterogeneities of the land, sequential order of the facies, problems of connectivity. Moreover, quality data are generally very costly and thus rare, therefore it is important to be able to work on reliable numeric models. Consequently, on the one hand we had to develop the **non-stationary aspect of geostatistics**, especially where software programs of **variographic analysis** and **estimation** are concerned, and on the other hand we had to develop set theory methods to model the geometries. That is why the application of the theory of **random sets**, although almost twenty years old, is currently regaining ground. Finally, a study in the oil industry nearly always has to process dissimilar data : few data from the wells and a great number of geophysical data. That is why we must also develop the tools of multivariate geostatistics, especially the **external drift** which was first used for the integration of seismic data to well mapping.

Other applications

A list of examples will illustrate the wide range of geostatistical case studies.

- **Meteorology** : analysis and mapping of the geopotential (*). Incidentally, this problem was one of the very first applications of **cokriging** in routine. We should note that the problem of mapping

(*) isovalue surface of the atmospheric pressure

equipotentials is complicated by the constraint of having consistent maps for the different values of the atmospheric pressure. Besides, these meteorological data had the particularity of being very heterogeneous, as oceans are obviously sampled much less than the continents. The **neighborhood search** was thus a crucial problem.

Meteorology is known to be one of the privileged fields in processing data with a spatial support. Indeed, under the name of **objective analysis**, **L.S. Gandin** (U.R.S.S.) proposed an interpolation method equivalent to universal kriging in 1965.

- **Forestry estimation** : This field is the second example for which an equivalent estimation method was proposed independently (**B. Matérn**, Sweden, 1959). But because of their discrete properties, the tree counts can be better defined by more suitable models, as, for example, the **point processes**.
- **Agriculture, soil science** : like forest estimation, these domains currently benefit from aerial data, especially satellite images. Geostatistics is thus confronted with the problem of **heterogeneous data** which is complicated by the quantity of information, and with the problem of **change of scale**, the different variables having supports of several orders of different magnitudes.
- **Material sciences** : geostatistics is used for example in **fracture statistics** and relies on **random set** models. The geostatistical study of thin sections resembles the techniques of mathematical morphology and image analysis.
- **Geochemistry** was the privileged domain of application of multivariate kriging analysis, in order to give a specific sense to the notion of **anomaly** and **regional**.
- **Geophysics** uses geostatistical methods to solve difficult problems, as for example the **deconvolution** of gravimetric values, or the **filtering** of diurnal anomalies of magnetism. Geostatistics is an original way of approaching the famous "inverse problem".
- **Cartography** (mapping) and **bathymetry** (mapping seabeds) are privileged domains for using geostatistics : **interpolation**, **compacting data**, using chararcteristic data (summets, watersheds, thalwegs...), accounting for measurement errors or **uncertainties of localization**.
- **Climatology** can take advantage of the multivariate geostatistics approach. In this way **rainfall data** can be linked to topography taken as the **external drift**, or to more geometrical factors such as the orientation of the dominant wind. These phenomena are naturally asymmetrical and so require finer tools than simple variograms (**bivariate distributions**).
- **Fisheries** might not jump to mind as an example : **global estimation** methods are used for the evaluation of the fish population which is difficult because of the special sampling conditions and the fact that the population being studied is a moving one with spatial characteristics varying throughout the day and the seasons... Let us just mention that a ruling of ICES (*) has suggested using geostatistics for the estimation of fish populations.

Future prospects

- We have brought up the subject of using satellite images or thin section data. Generally speaking, **image processing** can be treated by using geostatistical tools : **filtering** a noise, **compacting** data, **frequential analysis**. It is true that today this type of problem is usually studied in signal processing. Besides, the current geostatistical software may not be suited to the volume of data encountered in image processing. Contrary to this, the geostatistical tools of **fine variographic analysis** should bring more precision in the image representation. The **change of support models** should enable the use of images jointly with samples having other origins, and **kriging analysis** seems to be a suitable tool to help understand the underlying physical phenomena which are responsible for the general aspect of an image.
- At present the geostatistical study of **numeric models** is rarely employed, except in academic studies *a posteriori* of simulations (conditional or not). It is probable that computational problems would crop up — possibly in the sense of simplifying the algorithms — tied to the special distribution of the data in the space. However, the most important will certainly be to take into account the equations governing these data. A recent study was made in this direction and led to a new development : a geostatistical study on a **complex variable**..
- If the model is not constructed on an Euclidean space, a new **distance measure** will have to be defined, then the notions of **stationarity** and **ergodicity** will have to be defined for the probabilistic

(*) International Council for Sea Exploration

model. This is not a trivial question and is already being encountered in the problems concerning the earth's surface (meteorology on the scale of hemispheres).

- **Civil engineering** is another domain where development is possible, more precisely in the study of the **stability** or the **deformation** of the soil beneath constructions. However we must consider an important theoretical difficulty for this type of problem, the modeling of the **change of support**, given that the variables being studied (for example compression, penetration resistance, etc...) are mostly non-additive and consequently do not lend themselves to a linear approach.
- The most encouraging prospects are to be found in the **environment** domain. This is a subject which is not only of great interest today but which brings up important geostatistical questions. The common feature of environment data whether, they have natural or human origin (pollutants), is that their spatial distribution is ruled by physical equations. Flow phenomena obey the partial differential equations, which simply must be taken into account by the structural model, otherwise the estimations or simulations will not be credible. Therefore the processing of **variables linked by equations** is at the heart of environmental geostatistics. Work carried out up to now shows to what extent the way of approaching the problem depends on the particular form of the equations ruling the phenomena.

  The equations occur as a consistency constraint on the results but they are also present at the data level. In this way as initial information we can have both a variable and its gradient (or its Laplace operator) which must be integrated in a **cokriging** estimator. Often the gradient data intervene as **boundary conditions**, which increases the importance of the **geometry** of the problem. Finally, if the boundary conditions concern the **behaviour at infinity**, the problem becomes more difficult with questions of approximation, convergence, truncation...

  The preceding questions lead us to model the evolution of spatial structures in time. **Spatio–temporal models** are not limited to the addition of a dimension. We must consider the form of the equations of evolution, and doubtless also the **inequality constraints** which restrict the speed of evolution. This is a domain of research to be explored because of its potential wealth.

  A last remark : the term "environment" should be taken in a wide sense and in particular, meteorology should once again lead in the methodological developments in geostatistics.

Three examples

The aim of this brief presentation is to set out some aspects of a geostatistical study, using two real cases. The third case is an example of interpreting structural functions.

Bathymetric survey on the site of the "Titanic" (*)

This study was carried out in order to map the seabed in the immediate proximity of the wreck of the Titanic. The figures proposed show the steps in a real variographic analysis, and the importance of the choice of the kriging neighborhood.

As is usually the case for marine data, the available information is distributed along the **profiles** (Figure 6). The data is very close on the profiles (every 50 metres) whereas the profiles are much more spaced out . We find ourselves directly presented with an **anisotropy**, not of the regionalized variable — at present we do not have any information on this subject — but of the information. Consequently at each stage, we shall have to make sure that this situation does not cause artefacts. In particular, when checking the data, we must ask the ratio of the sampling grid : why precisely this orientation, why not have made a square grid ? One of the essential conditions for carrying out a good variographic analysis is indeed to have **non–preferential information**. The data should have an unbiased statistical significance. This does not mean that geostatistics can only work on random data, but that qualitative data *a priori* should be carefully considered **before** any statistical processing, either by delimiting the homogeneous sub-zones or by explicitly incorporating these data in the model.

As it is, the choice of the campaign complied with the constraints unknown to the bathymetric variable studied. Consequently, no bias was to be expected, *a priori*. However, when checking the data, we see that they prefer one direction. Indeed, at the work scale requested, (200 or 300 metres) , we cannot have information on the bathymetric structure in the direction perpendicular to the profiles. Therefore we are obliged to restrict ourselves to the variograms in the direction of the profiles and without additional information we **must** find an isotropic model. Indeed, we have no elements available for considering any modeling of an eventual anisotropy.

This situation is typical in geostatistics. Lacking decisive arguments, we fall back on the "minimum model". Any other choice would be arbitrary and the freedom left to choose the parameters would be misleading. The "minimum model" is also arbitrary but we run less risks with it and it will be the easiest to correct in case it is refuted later on . For reasons of prudence and also realism, at each modeling stage it is wise to adopt a **principle of economy** : avoid adding hypotheses and parameters which cannot be verified.

In this way, the variograms are computed in one dimension, in the direction of the profiles. Figure 7 shows these variograms at steps 50, 100 and 500 metres. The last figure regroups these three results and confirms their consistency. The conclusion drawn is that a **stationary model is clearly unacceptable**. Indeed, on the first two figures, we do not see a sill, which seems to indicate that there is no stationarity until at least 5000 metres. At the 500 metre step, we just might admit a sill phenomenon at about twenty kilometres, but this sill would be three times the variance of the data which is incompatible with a stationary model. Naturally, this effect of the variance could be due to a strong anisotropy (and then the sill in other directions would be very much below the variance), but there is no way of checking this model. The principle of economy would then suggest finding a non-stationary model, while hoping that if indeed there is an anisotropy, it will be taken into account by the drift. The variographic analysis is thus undertaken using IRF-k with the BLUEPACK program.

A first "tentative" approach using default options results with IRF-k, having a linear and spline generalized covariance. The resulting map, obtained by kriging with neighborhoods of 12 points (default

(*) This example is provided with an authorisation from TAURUS INTERNATIONAL / TITANIC VENTURES. To preserve confidentiality, the coordinates and depths are not specified, nor are the details of the sampling campaign.

values) is inadmissible — see figure 8. The reason for this is that nothing in this estimation obliges the program to use data from different profiles. In order to estimate the values next to a given profile, the kriging neighborhood only uses data from this profile. So on the one hand we work in IRF-1 — with the data almost aligned — which leads to a **quasi-singularity of the kriging system.** On the other hand, seen from the point to be estimated, the data are only distributed at best in a 180° angle — which means that we are extrapolating. Paradoxically, it is only when we are quite far between two profiles that we find less inadmissible conditions : the neighborhood contains information on each of the two profiles, the system is fixed size and the isolines become plausible.

It is possible to constrain the program to use the data belonging to at least n distinct profiles, for the variographic analysis as well as the kriging. By taking the minimum value $n = 3$, the program examines an IRF-2 this time, having a linear generalized covariance. The kriging, carried out with a default neighborhood of 16 points, is shown in figure 9. We note a definite improvement, especially in the areas where the profiles are regularly spaced and informed. However, there are anomalies as soon as the program has trouble finding the data which are well distributed and come from three different profiles.

We shall increase the minimum number of profiles to be used. Trying $n = 5$, and neighborhoods having 24 points, the result is a disaster (figure 10). We observe, *a posteriori*, that this is due to a generalized covariance model which is too regular (cubic).Therefore, when kriging, the estimation value is extremely sensitive to the closest data, and so magnifies the slightest fluctuations in the data. In return, when we are far enough from any datum, the estimator is aligned more to the drift.

From figure 8, it was obvious, *a priori*, that we would obtain anomalies — and this estimation was only carried out as an academic example. On the other hand, in figure 10, **the disastrous result teaches us something** : it shows that the IRF-2 model is numerically unstable, on account of the neighborhoods chosen and the intrinsic structure of the data. In such a case the variographic analysis protects itself by adopting a too regular generalized covariance which, linked to a second degree drift leads to completely erratic estimators.

We must learn from this experience. To impose 3 as the minimum of the number of profiles is insufficient, according to figure 9. Moreover, we have just seen that a "quadratic drift" destabilizes the estimation. Having seen figure 10 we shall impose a linear drift on the program, everything else being equal. In these conditions the program examines a linear and cubic model (figure 11). This model is quite close to that of figure 8 as regards regularity and degree of drift ; however it is identical to that of figure 10 in the neighborhood structures.

This time, the result can be considered as satisfactory. Doubtless we should smooth the isolines a little in order to "rub out" the few irregularities caused locally by the slightly defective neighborhoods. However the transversal structure which in figures 8 and 9 could have been attributed to artefacts and which was totally hidden by interference in figure 10 can be clearly seen here. This cañon and its structure are already known.

In these circumstances **it would be advisable to make another fine variographic analysis,** which would determine the structure of the cañon and the rest of the domain. The available data did not make this possible. At the very most we could have removed the cañon data from the variographic analysis and propose a structural model for the rest of the domain. The bathymetry would have been slightly better on most of the map and probably a little less good around the cañon.

The last run on this case study will be used to try to economize computational time. We are looking for the impact of the number of points in the neighborhood, all the other conditions being identical to those of the preceding estimation. By taking the default value of 8 points, we obtain figure 12, which, in comparison with the preceding estimation, is only slightly less good. Now, in the present case, the kriging system is only 11×11, whereas it was 27×27 in the preceding computation. For work in routine — which was not the case here — the time saving would probably justify the minor degradation noticed.

Figure 13 proposes the map of the kriging standard deviations for this last estimation. The reconnaissance profiles can be seen but not the cañon, which was predictable in theory. This is normal as we considered a global structural model so the kriging standard deviation only depends on the geometry of the information and not on the data values. A more precise study would not be satisfied with this result...

The lesson to be learnt from this example is the difficulty in developing a purely automatic procedure of variographic analysis and estimation. An essential element in a real study is the dialogue between the geostatistician and the data.

Moreover, going back to purely statistical quality criteria is likely to shatter many an illusion. Consequently, in this study, the program would be best suited to figure 10. This is where there was the most freedom to fit the parameters to the best of its statistical tests. And indeed, representing the values in the immediate proximity of the data is excellent. Unfortunately, what interests the user is precisely not what is taking place near the data but elsewhere. So, in order to lead to a good result, the model must contain more information than what is in the data. This additional information (here the choice of the neighborhood size and the imposed degree of the drift) shows to what extent the geostatistician must take **responsibility** in conducting a practical study.

Simulations of heterogeneous reservoirs

When working with subterranean reservoirs of water, oil or storage, we must have at our disposal a 3-D mapping of the hydrodynamic properties of the reservoir (porosity, permeability, ...).

These variables can be simulated by using classical geostatistical tools. But there is a problem when their distribution is no longer homogeneous in the space studied, especially when these variables differ from one area to another having different types of deposition. This is what we call heterogeneous reservoirs. The simulation procedure we propose consists first in simulating the geology of the reservoir in the form of its lithofacies.

The HERESIM software package developed by the IFP(*) and the Centre de Góstatistique mainly concerns fluvio–deltaic sedimentary environments. The methodology is based on a geological analysis of the reservoir which leads to the vertical partitioning of the reservoir into **homogeneous stratigraphic units** from the point of view of sedimentology (type and proportion of the facies) and genetics (form and dimensions of sandstone elements). Each unit is characterized by its deposition, particularly by a paleo-horizontal surface (surface taken as horizontal at the time of deposition) which will be used as a **reference level** and by its proximity to the neighboring units.

The data for these simulations come from wells through the different units making up the reservoir, and for which the lithofacies series is known, either by core drilling (sampling and analysis of rock samples), or after examining the different drill logs available (continuous recordings of geophysical variables along the boreholes). The first stage consists in delimiting the passages of the different units by the boreholes and to extend these limits to the entire reservoir, then to divide the information of each well to the step of the simulation grid.

Inside each unit, the lithofacies are sequenced according to the deposition (for example from the most shale to the most sandstone), and the variables studied are indicators of the lithofacies. Once the unit is horizontal (compared to its reference level) the well data provide different information :

- vertical proportion curves which represent the proportions of the lithofacies in a horizontal slice. These vertical proportion curves sum up the geological sequences of the unit. Figure 14 (1) shows a vertical proportion curve representing a unit going from mainly sandstone facies at the base to shale at the top. This vertical distribution is usually non-stationary.
- horizontal proportion curves which represent the proportions of the different facies along any horizontal section of the reservoir. Horizontally, the stationarity hypothesis is tolerated in most cases such as the one presented. However, depending on the work scale, and often for deltaic deposits, it is more appropriate to consider a horizontal non-stationarity.
- Experimental simple and cross variograms of the indicators of the different facies, calculated either along the boreholes or parallel to the reference level.

Proportion curves are not modeled. Conversely, experimental variograms should be fitted using a consistent model. Indeed, a single gaussian random variable is examined, each lithofacies fitting to a value interval of this variable. The thresholds of these intervals are limited by the experimental values of the proportions of the different lithofacies and consequently expand for each horizontal slice of the unit, as shown on the vertical proportion curve. They can also differ horizontally in the case of horizontal non-stationarity.

The spatial structure is fitted for each stratigraphic unit, using a single model with a few parameters (one for each direction in space) : an example for fitting is shown in figure 14 (2). In this case the method involves simulating the underlying gaussian variable conditioned by the wells and then converting the numeric result into a type of facies with the proportion curves.

(*) Institut Français du Pétrole

Figure 15 shows a simulation of a vertical cross section, where we can clearly see that each unit presents proportions and distributions having different facies : in particular the second unit (from the top) has hardly any shale. The horizontal section at this altitude cuts across the four units.

The petrophysical variables (porosity, tensorial components of the permeability, ...) are then also simulated or simply allocated to each cell making up the reservoir conditionally to the lithofacies type. They can be used later as input data for a fluid flow simulation program.

This method gives highly satisfactory results. Moreover, it has the advantage of needing few parameters and of being flexible enough to take into account additional information such as the distribution of lithofacies using proportion curves.

On the structural analysis of radioactivity–grade

The data studied here come from a borehole campaign made on a uranium deposit (*). Consequently, on each borehole we have simultaneously at our disposal a chemical analysis (carried out on the cores) and a continuous recording of the radioactivity, performed by lowering a probe into the borehole.

The cores provide us with vital information on the deposit, by giving us the grade value (the variable which interests us) as well as enabling the geologist to get a qualitative picture of the mineralization. However, they have two disadvantages :

- They are costly. Consequently cores cannot be taken from all the boreholes at the final sampling stage of the closely drilled deposit.
- Their vertical location is imprecise because of the core recovery rate which is never 100%. The sample comprises fragments which could have shifted within the core drill, even causing a global displacement of the sound pieces of the core.

Contrary to this, the radiometric data are economical, numerous and localized with great precision. However, as sampling the deposit progresses, the percentage of radiometric data becomes large compared to the chemical analyses. But the radiometry has a major disadvantage: it is an **indirect measurement**. Indeed nobody exploits a deposit of radioactivity ... To simplify the matter, we can say that the radioactivity variable is like a **convolution product** of the grade, the parameters of the convolution depending mainly on the drilling conditions (diameter of the hole, presence of mud, casing) and the geometry of the mineralization. This shows that it is unrealistic to propose a deterministic model for this convolution.

That is why an essential preliminary stage of investigating uranium deposits — one which was tried out in mining companies long before geostatistics — lies in fitting a radiometry/grade regression model which will make it possible, when we have only the radioactivity data, to convert these data into the variable of interest. Compared to these statistical methods, geostatistics considers the spatial structure, both of the grade and of the radioactivity. We shall confine ourselves to explaining the effect of convolution geostatistically and only in the vertical direction, but it is understood that the coregionalization model makes it possible later to apply the methods of multivariate geostatistics and especially cokriging.

Figure 16(1) shows the normed covariance (the **auto-correlation** function) along the boreholes of the grade data. This experimental covariance is very admissible: it was obtained in favorable conditions (a multitude of regularly spaced data), and the result shows a consistent structure : a stationary model $C(h)$ is admissible with a linear behaviour at the origin and a range in the order of 1,20 metres. To be more precise, we note that the variogram which is associated to it by the classical formula $\gamma(h) = C(0) - C(h)$, **does not reach** the value $C(0)$ for $h = 1,2m$. This effect is due to an anisotropy, the grades being more erratic horizontally than vertically — but this discussion concerns tridimensional modeling which we shall not consider in this presentation.

Figure 16 (2) shows the auto-correlation function for the radioactivity. In order to make a comparison, the previous curve has been represented by a dotted line. Once again, we see a pronounced vertical structure, but there are significant differences between the two curves :

- Although a stationary model is still probably admissible this time, we see that the range has still not been reached at 3 metres. On the scale of several metres, the radioactivity thus presents a more

(*) For confidentiality the units are not cited. We worked on standardized covariances, that is autocorrelation functions.

definite structure than the grade. The "range of influence" is greater for the radioactivity than for the grade.

- The curve shows a concavity at the origin which could be modeled by a parabolic behaviour whereas the auto-correlation of the grades had a linear behaviour. This is an experimental example of the relation of **regularization** which exists between the two variables. The radioactivity presents greater regularity than the grade at short distances : needless to specify that this was known to uranium practitioners from experience...
- Between 0 and 2,5 metres, the auto-correlation of the radioactivity is always stronger than that of the grade. The effect of greater regularity of the radioactivity is seen at all these distances.

Figure 16 (3) shows the (vertical) experimental cross variogram between radioactivity and grade. We might think, *a priori*, that this tool is hardly suited to structural sampling. Indeed, the cross variogram is symmetrical by construction, that is to say that, in a given direction it does not take into account the order of the variables. Now, in uranium deposits, we can have leaching phenomena (*), which means a systematic displacement of the uranium $U(x)$ compared to its radioactive tracer $Ra(x)$. In this way, the distribution $U(x)$ knowing $Ra(x+h)$ — usually denoted $U(x) \mid Ra(x+h)$ — is likely to be appreciably different from that of $U(x+h) \mid Ra(x)$, which the cross variogram will not "see". It is for this reason that it would be better to calculate the experimental **cross covariance** C_{URa}, whose theoretical model is defined as

$$C_{URa}(x,y) = \mathbf{E}[U(x).Ra(y)]$$

or in a stationary case such as here :

$$C_{URa}(h) = \mathbf{E}[U(x).Ra(x+h)]$$

This function is not necessarily symmetrical along h.

Let us note here the formula linking the cross variogram and the cross covariance :

$$\gamma_{URa}(h) = C_{URa}(0) - \frac{C_{URa}(+h) + C_{URa}(-h)}{2}$$

This expression proves that knowing the cross covariance implies knowing the cross variogram, whereas the converse is false : this confirms that the cross covariance is a more precise analytical tool than the cross variogram.

In the circumstances, a preliminary investigation is essential. The cross covariance is very close to being symmetrical around the axis $h = 0$. This means that on the scale of the deposit there is no significant systematic shift between the grade and radioactivity measurements, and this whatever caused the shift. There is no **systematic** measuring error of the position of the cores in comparison with the radioactivity recordings and neither has there been a vertical migration of one of the variables compared to the other one, since the deposit was formed. Consequently, by a purely numeric verification, we are now in a position to state or confirm a geological conclusion and so reassure the drillers on the quality of their work.

In the case of the deposit being studied, this first analysis of the cross covariance was no surprise to the geologists. However, this same curve presents unexpected properties. We can see that the cross covariance does not have its maximum along $h = 0$. In fact the correlation between two measurements, one of radioactivity and the other of grade, is somewhat better when these measurements are at a distance of $\pm 10cm$ vertically, than when they are considered to be located at the same place. This means an uncertainty of relative positioning of the grade and radiometric measurements of precisely $\pm 10cm$, and not systematic. In other words, when two radioactivity peaks are only several centimetres apart, it is misleading to want to allot to one rather than the other a peak of single grades which is opposite them. And likewise it is misleading to want to "correlate horizontally" between boreholes at a distance of some 10 metres apart, mineralized parts located on the cores, on account of the inaccuracy on the depth of these cores. This time, the geologists did not predict these results.

Further work on the deposit proved that these remarks and others have clearly indicated to what extent it would be mistaken to consider a geostatistical study as merely an aesthetic pastime, and the importance of communication between the different specialists involved.

(*) dilution of the ore, followed by an outflow of the solution.

Spreading the word

It is not easy to describe development in geostatistical methods. This is partly due to the basic, pluridisciplinary character of geostatistics which can be used in the most unexpected situations. And then there is the risk of being dogmatic when defining the limits of "orthodox" geostatistics, handing out good and bad marks, granting or not the trademark "geostatistics" to a method.

There is neither a French nor an international directory of geostatisticians or research centres at present and it is not our intention to create one here, but rather to give a few examples to illustrate the state of the art.

Location of geostatistical activities

The origins of geostatistics are to be found in mining where the problems posed by gold deposits in South Africa encountered the probabilistic formalism developed by French and Soviet schools. At the beginning of the 70s, an American school, helped by unrivalled computational facilities, developed in its turn.

Today, we have little information on the progress of geostatistics in the ex-USSR, which seems to be characterized by a definite division between university and industrial methods. In South Africa, developments in geostatistics are of an industrial nature, and are carried out by mining companies. In what concerns teaching and research, two main axes can be determined, European and North American. However, it should be noted that the majority of those teaching geostatistics in the U.S.A. received their initial training in Europe.

The birthplace of geostatistics is to be found in Europe:

- The Centre de Géostatistique at the ENSMP has been training specialists for more than twenty five years. The majority of the senior teaching staff throughout the world have spent time — sometimes a long time — here. As well as its research activities, which were at the origin of non-linear geostatistics, non-stationary geostatistics and random sets among others, the Centre gives post-graduate courses mainly for foreign students, and specialized courses for industrialists. Moreover, research workers from the Centre are frequently sent abroad to companies or universities. Lastly, the Centre runs summer schools of introductory courses on geostatistics, as well as training days, which are more centred on new developments in geostatistics.

These different activities can be found in all the centres offering geostatistics in Europe, such as:

- L'INPL, National Polytechnical Institute of Lorraine,
- Polytechnical Institute of Zürich,
- L'ETSECCP, of Barcelona (Catalonia University)
- Lausanne, Rome, Dublin, Lisbonne, Leeds University ...

As well as this, geostatistics is also taught within the EEC's ERASMUS program. Besides this, certain EEC programs may request a specific geostatistics program, for example in the fishing domain where a European ruling explicitly advocates estimation using geostatistical methods. We can also mention the many research organisms or company laboratories working with or on geostatistics in France:

- BRGM Orléans,
- Institut Français du Pétrole,
- INRA Avignon,
- COGEMA,
- IRSID,
- IFREMER,
- Oil companies (CFP, SNEA),
- ORSTOM ...

Without forgetting the universities of Nancy, Grenoble, Paris VI...

In industry, the systematic use of geostatistical methods was to be found only in mining companies: Penaroya, COGEMA (the word "geostatistics" was coined at the CEA). With the development of non-stationary geostatistics and random sets, the oil companies incorporated geostatistical algorithms into their standard software. In other branches of industry routine use of geostatistics is often done bit by bit. In the field of environment, we look forward to a development such as that known in the mining and oil fields. Let us finish by mentioning GEOVARIANCES, a company which is completely committed to providing geostatistical services.

The situation is somewhat similar in the USA, in what concerns the industrial development of geostatistics: mining and especially oil companies. In some cases, this comes from Europe. The Centre de Géostatistique BLUEPACK software program is very successful in the oil domain. Among the many universities teaching geostatistics and offering research programs, let us mention:

- Stanford university, California
- Tucson university, Arizona
- Lawrence university, Kansas
- Iowa State University ...

and, in Canada

- Polytechnical School of Montreal.

In the rest of the world, countries with a strong mining tradition make use of geostatistics: South Africa (gold, precious stones, platinum), Australia (gold, base metal), French speaking Africa (uranium, phosphate, manganese... but also oil and forests), South America: Peru and Chile (copper), Brazil (semi-precious stones, asbestos, oil), Venezuela (oil). This industrial approach sometimes includes teaching programs and research methods :

- GEOVAL in Australia (Sydney, Perth),
- Santiago University in Chili,
- School of Mining of Ouro Preto in Brazil, ...

We also know that geostatistics is taught in China, that India is developing a project for a research centre, etc...

Communication

The Centre de Geostatistique took the initiative in promoting the **First International Geostatistical Congress** near Rome in 1975. Several years later there was sufficient interest to make this sort of communication systematic and at the second congress (Lake Tahoe, California, 1984) it was decided to hold a congress every four years. At the congress in Avignon in 1988, **IGeostA** an international geostatistical association was set up to ensure communication within the geostatistical community. A newsletter, *De Geostatisticis* is sent out regularly to members. There is also a similar review, *Geostatistics*, in the USA.

At the moment there is no geostatistical journal. Papers on earth sciences are mostly published in *Mathematical Geology*, which is edited by the International Association of Mathematical Geology. But as the fields of application of geostatistics is becoming more varied, papers are more and more dispersed in publications : reviews on statistics and probabilities for theoretical research, on stereology for some works on random sets, on geophysics for petroleum geostatistics, etc...

Available software

As with the methods, it is difficult to list software reserved for geostatistics or using it only partly. Therefore this list only sets out to give partial information, even on the French level.

The Centre de Geostatistique at the ENSMP pioneered the use of geostatistics in an industrial context, and as from the middle of the 70s, developed an exclusively geostatistical software library which is currently set forth into two program collections:

- GEOSMINE, adapted to mining and non-linear geostatistics: deposit simulation, disjunctive kriging, selectivity curves.
- BLUEPACK, specialized in non-stationary geostatistics : variographic analysis in IRF-k, mapping, simulations. As an example, surveys of oil structures use BLUEPACK.

These programs, also developed by the Centre de Geostatistique, depend on a data base structure. They are commercialized in collaboration with GEOVARIANCES which itself also commercializes

- OPMINE, software for simulating open pit mining.

Besides this, the Ecole des Mines, in collaboration with IFP (French Petroleum Institute) has created and distributed the program

- HERESIM, software for the conditional simulation of heterogeneous reservoirs, setting out to model complex stratigraphic structures encountered in petroleum exploration.

The BRGM (Bureau of Geological and Mining Research) also offers a series of programs,

- GDM, more interested in mining geostatistics but also including functions from the geostatistics domain. Like GEOSMINE, this program also depends on a data base structure.

The COGEMA has created

- SERMINES, a mining estimation program system.

The Centre de Geostatistique and the SNEA(P) have collaborated in the construction and commercialization of the program

- KRIGEPACK, directed towards the automatic variographic analysis and mapping with the kriging system of non-stationary regionalized variables.

Two programs from abroad are worth mentioning :

- GEO-EAS, an interactive program of variographic analysis and kriging, proposed by the EPA (Environment Protection Association) in the USA. This software is in the public domain.
- REGARDS, distributed by Trinity College, DUBLIN. This is a program of "exploratory geostatistics", which is entirely directed towards interactivity and permitting highly precise structural analyses.

It is much more difficult to list software in which geostatistics plays but a secondary role. The programs of the National Meteorology in France use algorithms based on cokriging which were first set up with the collaboration of the Centre de Geostatistique. Moreover, we know that certain universities in the USA (Arizona) produce and distribute programs containing geostatistics. We also come across geostatistical software in the most unexpected places, for example in provincial universities in the ex-USSR. Lastly, the production of software using geostatistics can be made by industrial companies (for example, FLUOR in the USA).

The following software including geostatistics can be quoted as an example :

- GEOCAD : modeling program mainly using discrete splines, developed by INPG in Nancy together with STANFORD University.

- CARTOLAB, automatic mapping program distributed by INPG and GEOVARIANCES.
- The GEOSTOKOS company in the U.K. distributes a family of graphic software which uses geostatistics as well.
- In Denmark, two graphic software programs, UNIRAS and IRAP contain a kriging estimation.
- Likewise, the American software SAS explicitly proposes kriging.

Needless to say, this list is far from being exhaustive. Stories get back to us of geostatistical software of mining estimation being written by South African companies (precious metal or stones). As well as this, we might say that geostatistics is used by many without their knowing it, since most of the statistical software include methods of local regression based on cubic splines, which, mathematically, is the same as kriging in IRF-k — minus the essential step of the variographic analysis...

Bibliography

This list contains a limited number of references mainly in French, which have been chosen because they treat the subject as a whole. The Ecole des Mines de Paris (ENSMP) publications are available from the library at the Centre de Góstatistique.

Background references

Matérn B. [1960] : Spatial Variation, in *Almaenna Fœrlaget* — Stockolm. Reedited in [1986] by Springer Verlag, Berlin

Gandin L.S. [1965] : Objective analysis of meteorogical fields, *Israel program for scientific translations* — Jerusalem
(Two textbooks which introduce the processing of regionalized variables both in the forestry domain and in meteorology. The objective analysis leads to a formalism identical to universal kriging.)

Matheron G. [1963] : Traité de Géostatistique Appliquée — Editions du B.R.G.M., Paris
(Linear geostatistics directed towards mining applications.)

Matheron G. [1965] : Les Variables Régionalisées et leur estimation — Masson, Paris
(Mathematical bases for linear geostatistics within the stationary or intrinsic frame. A large part is given to transitive formalism.)

Matheron G. [1975] : Random sets and integral geometry — J. Wiley & Sons, New York
(The boundary between mathematical morphology and geostatistics. The creation of the theory of random sets.)

General works

David M. [1977] : Geostatistical ore reserve estimation — Elsevier Scientific Publishing Co., Amsterdam, Oxford, New York

Rendu J.M. [1978] : An Introduction to Geostatistical Methods of Mineral Evaluation, *South African Institute of Mining and Metallurgy Monograph Series* — Johannesburg
(Two books oriented towards mining geostatistics.)

Cressie N. [1991] : Statistics for spatial data — J. Wiley & Sons, New York
(The state of the art at the beginning of the 90s, presented from a statistician's point of view. The book also talks about the domain of mathematical morphology, markovian fields, etc...)

Linear geostatistics

Matheron G. [1970] : La théorie des Variables Régionalisées et ses applications — ENSMP, Paris
(Also in English. Basic course on linear geostatistics, up to universal kriging included.)

Chauvet P. [1991] : Aide-mémoire de Géostatistique Linéaire — ENSMP, Paris
(Including a presentation of IRF-k.)

Variography

Journel A. [1977] : Géostatistique Minière *(Thèse)* — ENSMP, Paris

Journel A. & Huijbregts Ch. [1978] : Mining Geostatistics — Academic Press, London
(The thesis translated into English. General presentation of geostatistics, mainly in the mining field. It gives numerous practical examples, particularly concerning the variographic analysis.)

Extension variance

Formery Ph. & Matheron G. [1962] : Recherche d'optimum dans la reconnaissance et la mise en exploitation des gisements miniers, in *Annales des Mines, Novembre 1962* — Paris

Krigeage

Rivoirard J. [1984] : Le comportement des poids de Krigeage *(Thèse)* — ENSMP, Paris
(Precise analysis of the weightings of the kriging in stationary or intrinsic hypothesis.)

Chauvet P. [1988] : Réflexions sur les pondérateurs négatifs du Krigeage, in *Sciences de la Terre* — Nancy
(Comments on the link between the probabilistic model and the resulting maps from a problem which often bothers users : negative kriging weights.)

Non-stationary geostatistics

Delfiner P. & Matheron G. [1980] : Les fonctions aléatoires intrinsèques d'ordre k — ENSMP, Paris
(An introduction to the IRF-k formalism directed towards the application in software of automatic variographic analysis.)

Chilès J.P. [1977] : Géostatistique des phénomènes non stationnaires (dans le plan) *(Thèse)* — ENSMP, Paris
(Synthèse du traitement dans le plan de données non stationnaires, complétée par un examen des méthodes de simulation.)

Multivariate

(For cokriging see Linear geostatistics 1970.)

Delfiner P., Delhomme J.P. & Pélissier-Combescure J. [1983] : Application of geostatistical analysis to the evaluation of petroleum reservoir with well logs, *Annual Logging Symposium of the SPWLA* — Calgary
(Introduction to the external drift method.)

Wackernagel H. [1985] : L'inférence d'un modèle linéaire en Géostatistique Multivariable *(Thèse)* — ENSMP, Paris
(Introduction to the kriging analysis.)

Dong A. [1990] : Estimation Géostatistique des phénomènes régis par des équations aux dérivées partielles *(Thèse)* — ENSMP, Paris

Daly C. [1991] : Applications de la Géostatistique à quelques problèmes de filtrage *(Thèse)* — ENSMP, Paris

Simulations

Freulon X. [1992] : Conditionnement du modèle gaussien par des inégalités ou des randomisées *(Thèse)* — ENSMP, Paris
(Comments recalling the turning band methods, and developments on the simulation of geometrical structures and simulations under constraints.)

Non-linear geostatistics

Matheron G. [1976] : A simple substitute for conditional expectation : the disjunctive kriging, in *Proceedings NATO ASI "Advanced Geostatistics in the Mining Industry", Octobre 1975* — Reidel, Dordrecht
(The initial paper on disjunctive kriging.)

Rivoirard J. [1991] : Introduction au Krigeage Disjonctif et à la Géostatistique Non Linéaire — ENSMP, Paris
(A course on non-linear geostatistics.)

Various methods

Séguret S. [1991] : Géostatistique des phénomènes à tendance périodique (dans l'espace-temps) *(Thèse)* — ENSMP, Paris
(Comparison of signal processing and filtering techniques applied to marine geology.)

Boulanger F. [1990] : Modélisation et Simulation des Variables Régionalisées par des Fonctions Aléatoires Stables *(Thèse)* — ENSMP, Paris
(Generalization of simulation models to a non-gaussian frame, using the ARMA methods of the time series.)

Dubrule O. [1981] : Krigeage et Splines en cartographie *(Thèse)* — ENSMP, Paris
(Kriging considered as an interpolator, and the spline-kriging equivalent.)

Langlais V. [1990] : Estimation sous contraintes d'inégalités *(Thèse)* — ENSMP, Paris
(See also "Positive Kriging", R.J. Barnes & T.B. Johnson : in records from the third International geostatistical congress.)

Jeulin D. [1991] : Modèles morphologiques de structures aléatoires et de changement d'échelle *(Thèse de Docteur ès Sciences Physiques)* — Université de Caen
(A synthetic approach of random set models.)

International geostatistics congress proceedings

Guarascio M., David M. & Huijbregts Ch. ed. [1976] : Advanced Geostatistics in the Mining Industry — Reidel, Dordrecht

Verly G., David M, Journel A. & Maréchal A. ed. [1984] : Geostatistics for Natural Resources Characterization — Reidel, Dordrecht

Armstrong M. ed. [1989] : Geostatistics — Kluwer Academic Publishers, Dordrecht

The records from the first three International geostatistical congresses contain theoretical developments as well as examples of application. We can also quote

Armstrong M. & Matheron G. [1987] : Geostatistical case studies — Reidel, Dordrecht
(A particularly representative group of examples of applications.)

In the records of the " Application of computers and operation research in the mining industry" symposium (APCOM), there are papers on geostatistics.

Taking stock of geostatistics...

The word "geostatistics" has been in the Petit Larousse dictionary for the last ten years. However, the article in the Encyclopedia Universalis which puts the emphasis on non-linear geostatistics is more instructive. To sum up, the following works should be cited:

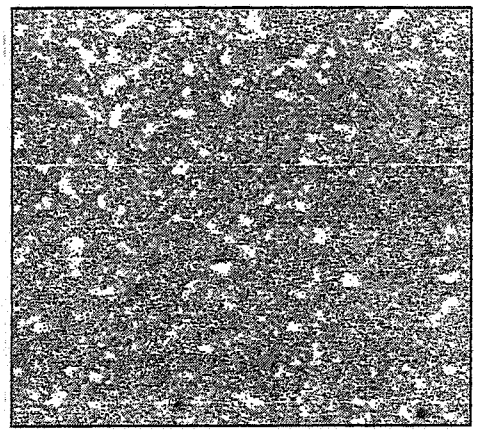
Matheron G. [1978] : Estimer et Choisir — ENSMP, Paris

in English :

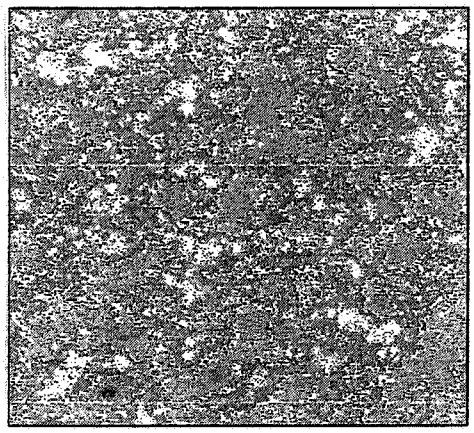
Matheron G. [1989] : Estimating and Choosing — Springer Verlag, Berlin

This textbook proposes a synthetic look at geostatistics and at the significance of classical operations in routine during a study. The all-important notions of stationarity and ergodicity are minutely analysed, as well as the fundamental problems of the adequation of a model to reality and of the real place for probability theory in geostatistics.

IMPORTANCE OF THE RANGE

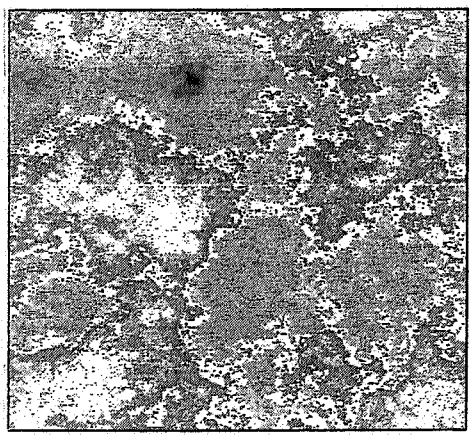


Range 10



Range 25

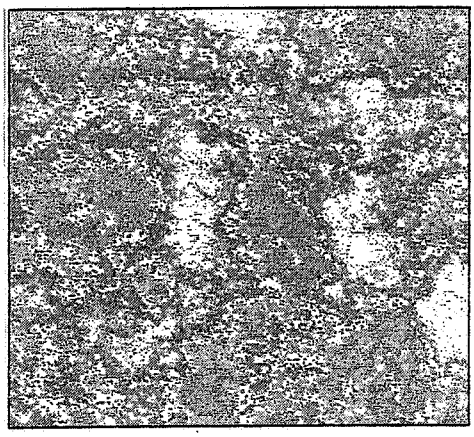
Range 50



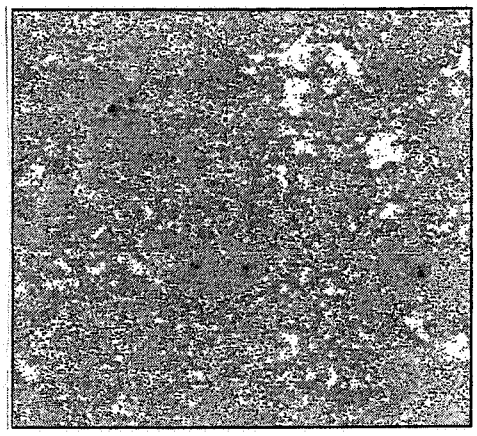
Range 75

Figure 1

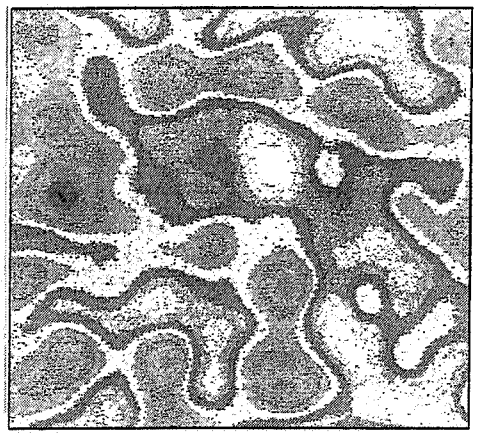
IMPORTANCE OF THE MODEL



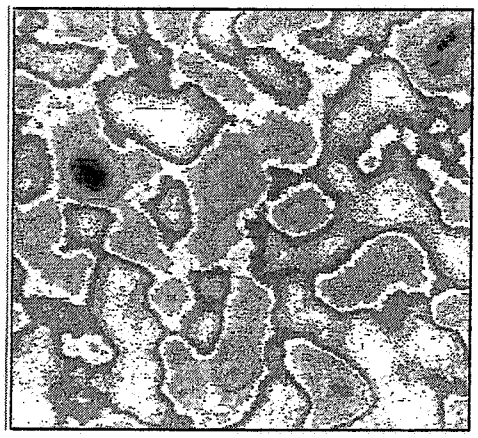
Spherical (range 40)



Exponential (range 13.33)



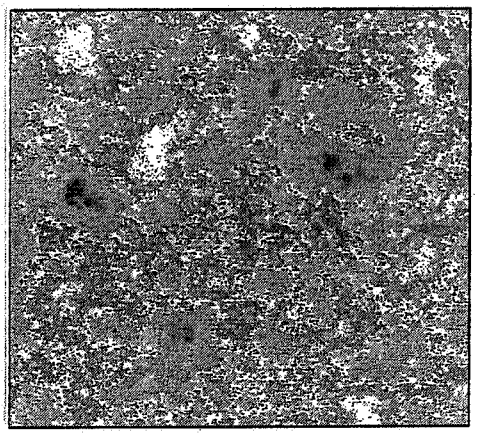
Gaussian (range 23.12)



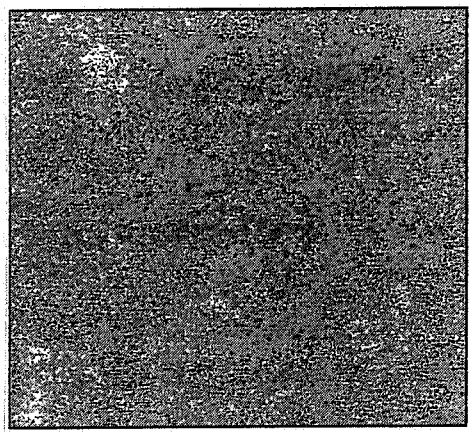
Cubic (range 40)

Figure 2

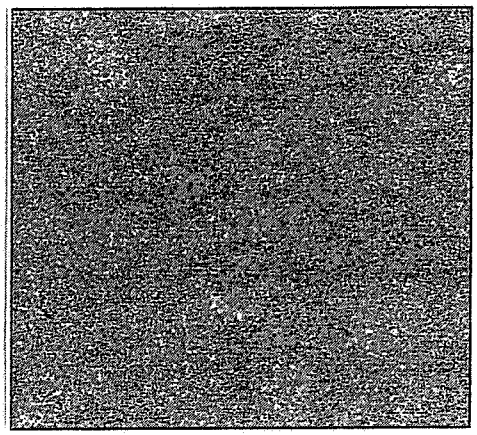
IMPORTANCE OF THE NUGGET EFFECT



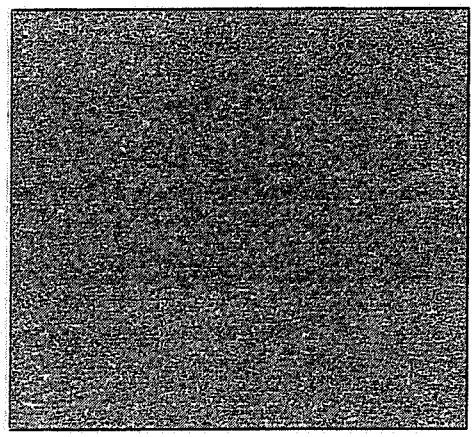
100% Spherical



2/3 Spherical – 1/3 Nugget



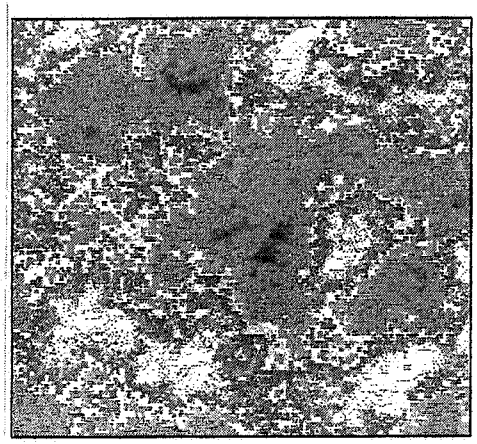
1/3 Spherical – 2/3 Nugget



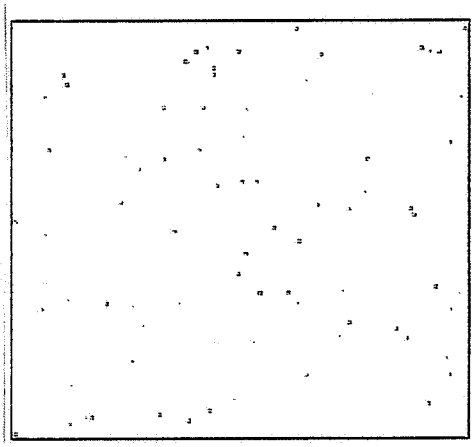
100% Nugget

Figure 3

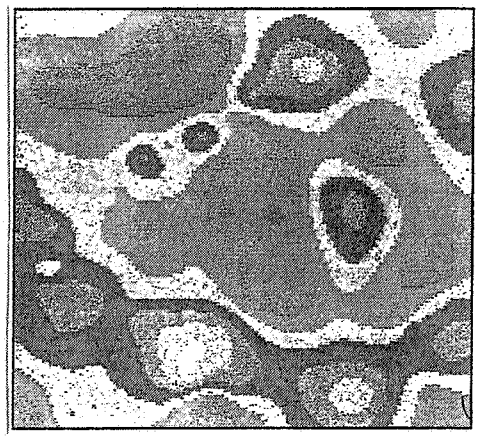
COMPARISON REALITY–KRIGING : SPHERICAL MODEL



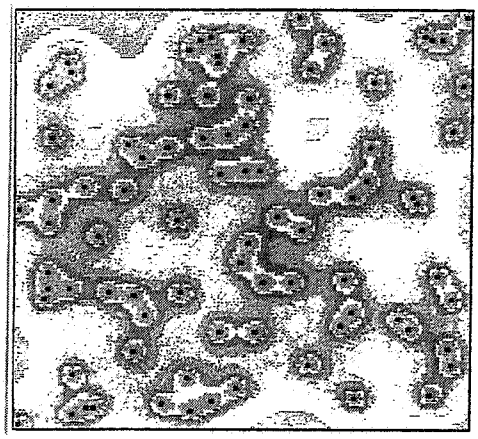
Reality



Samples



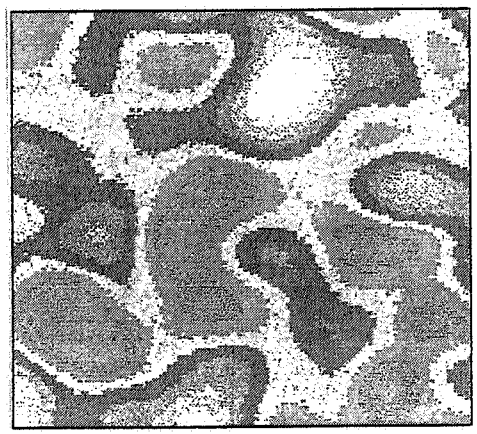
Kriging



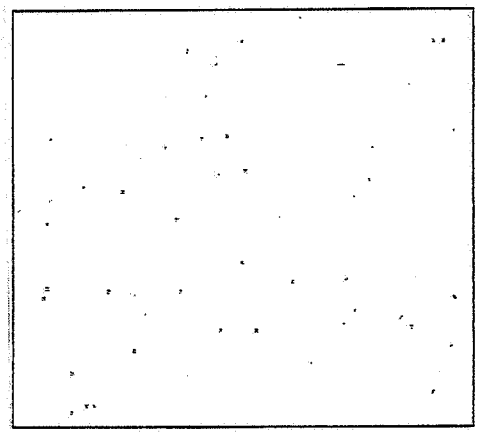
Kriging standard deviation

Figure 4

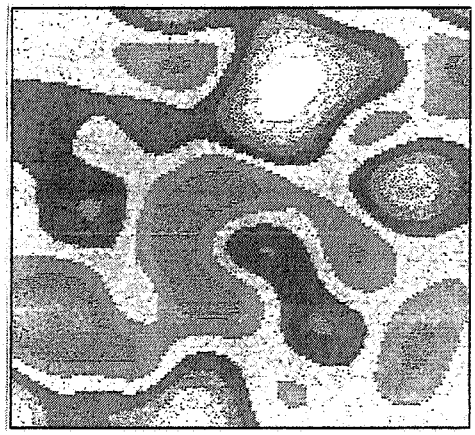
COMPARISON REALITY–KRIGING : GAUSSIAN MODEL



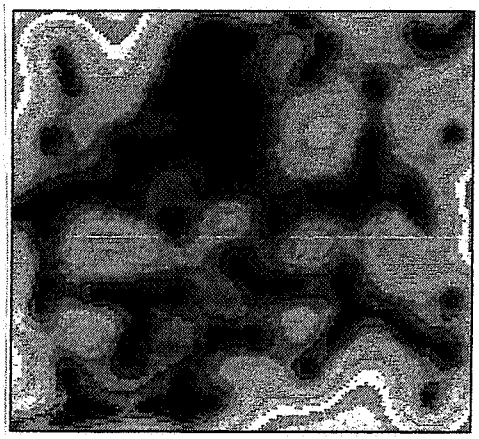
Reality



Samples



Kriging



Kriging standard deviation

Figure 5

"TITANIC" : AVAILABLE INFORMATION

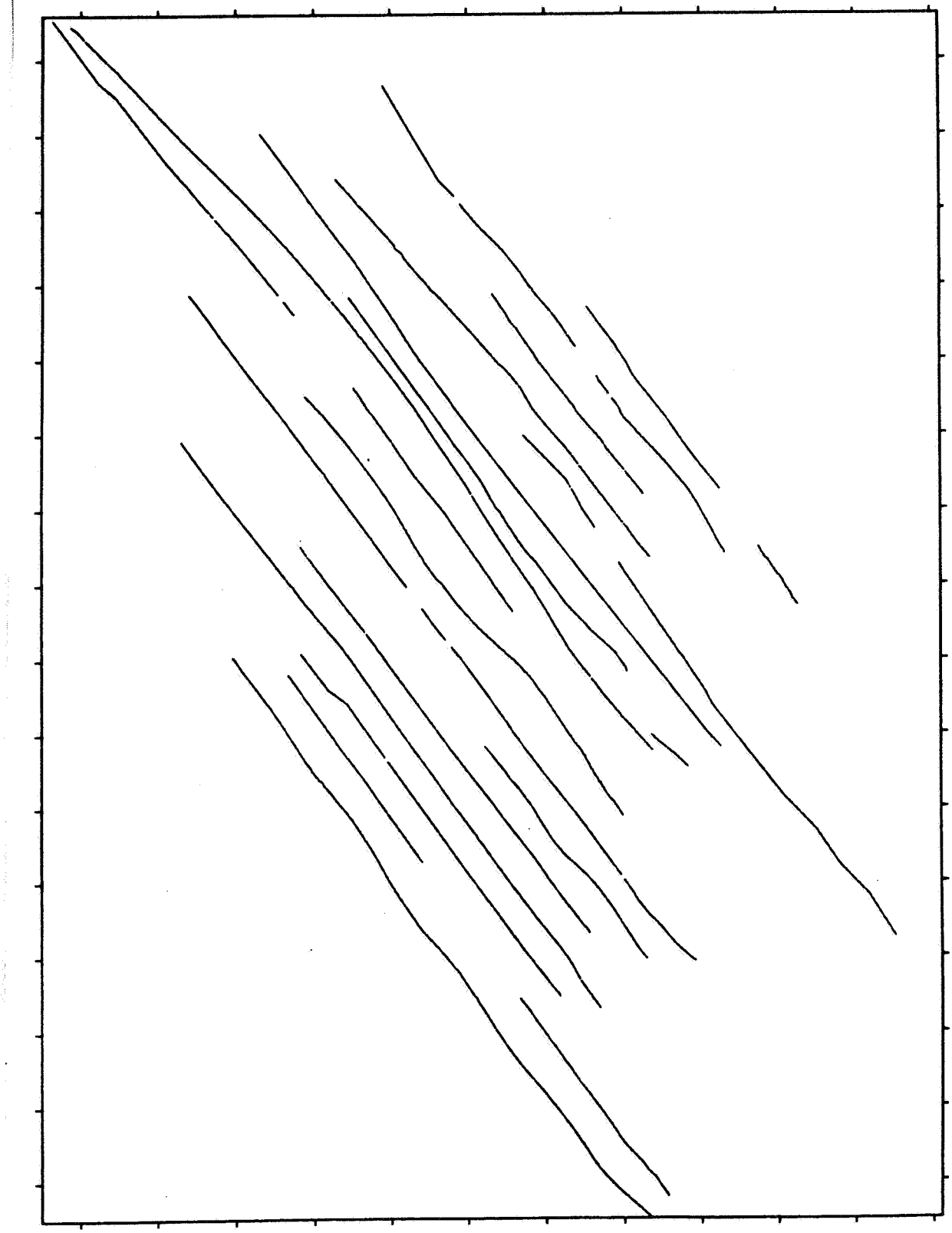
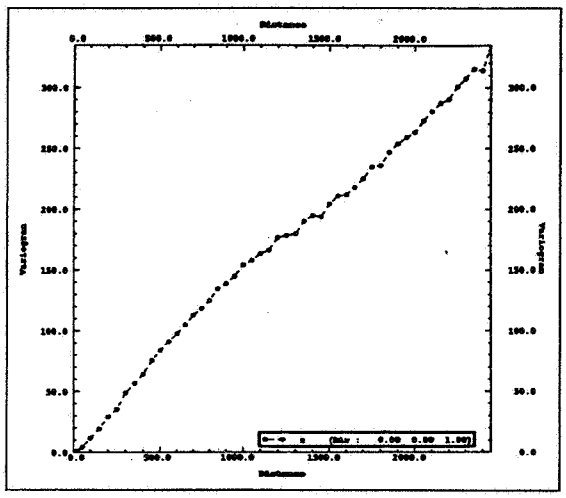
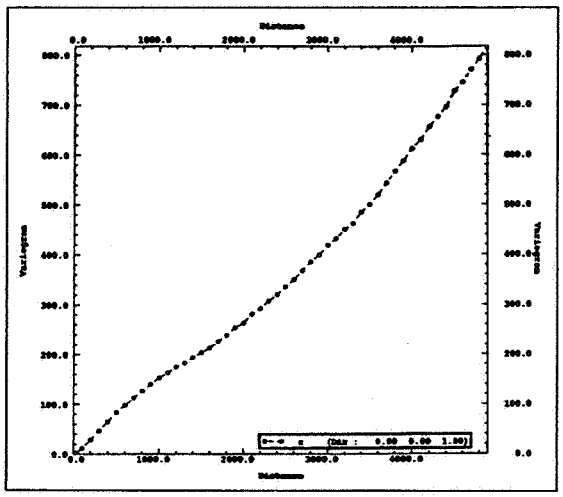


Figure 6

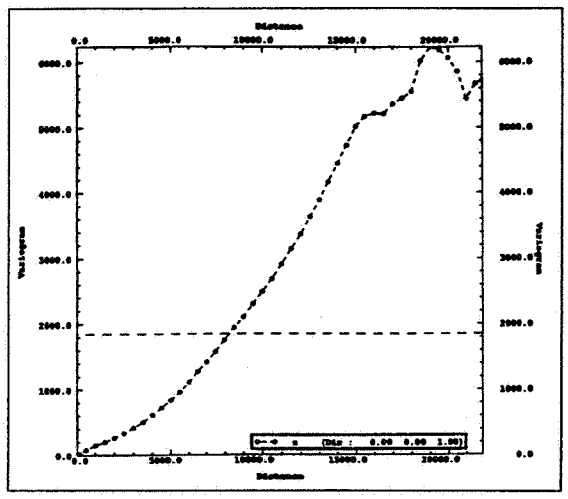
"TITANIC" : VARIOGRAMS ON PROFILES



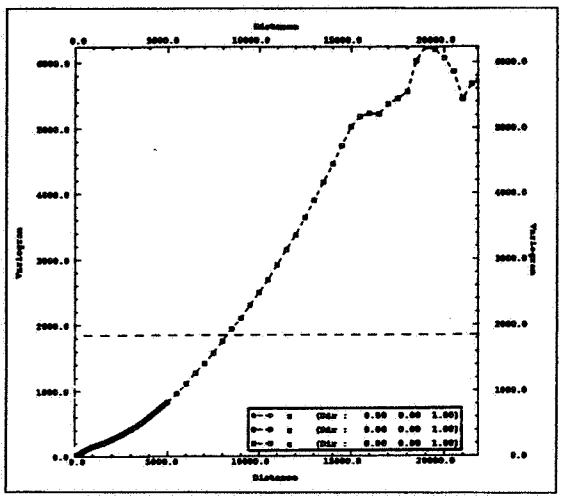
Step : 50 meters



Step : 100 meters



Step : 500 meters



Superimposing of variograms

Figure 7

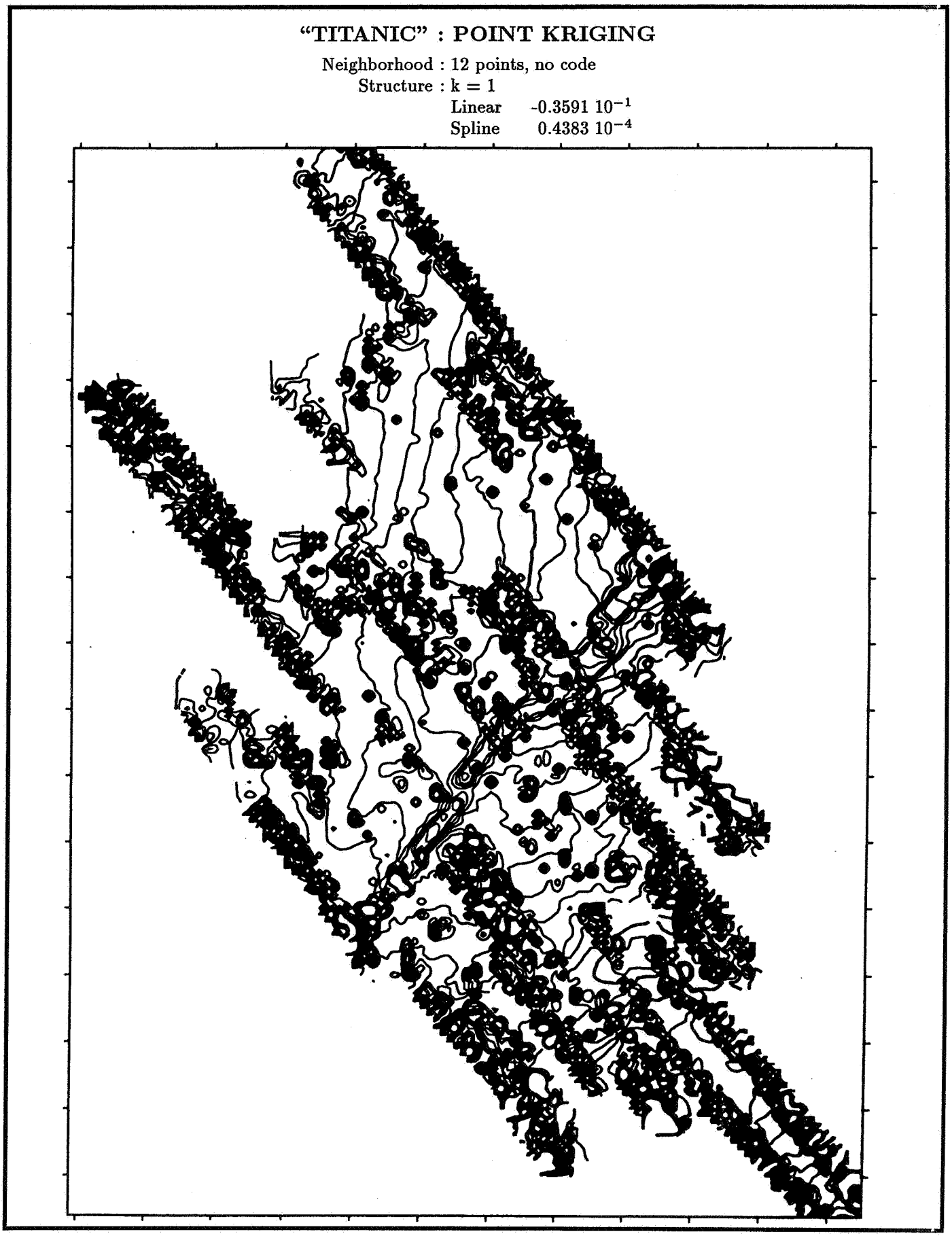

"TITANIC" : POINT KRIGING
Neighborhood : 12 points, no code
Structure : k = 1
Linear -0.3591 10^-1
Spline 0.4383 10^-4

Figure 8

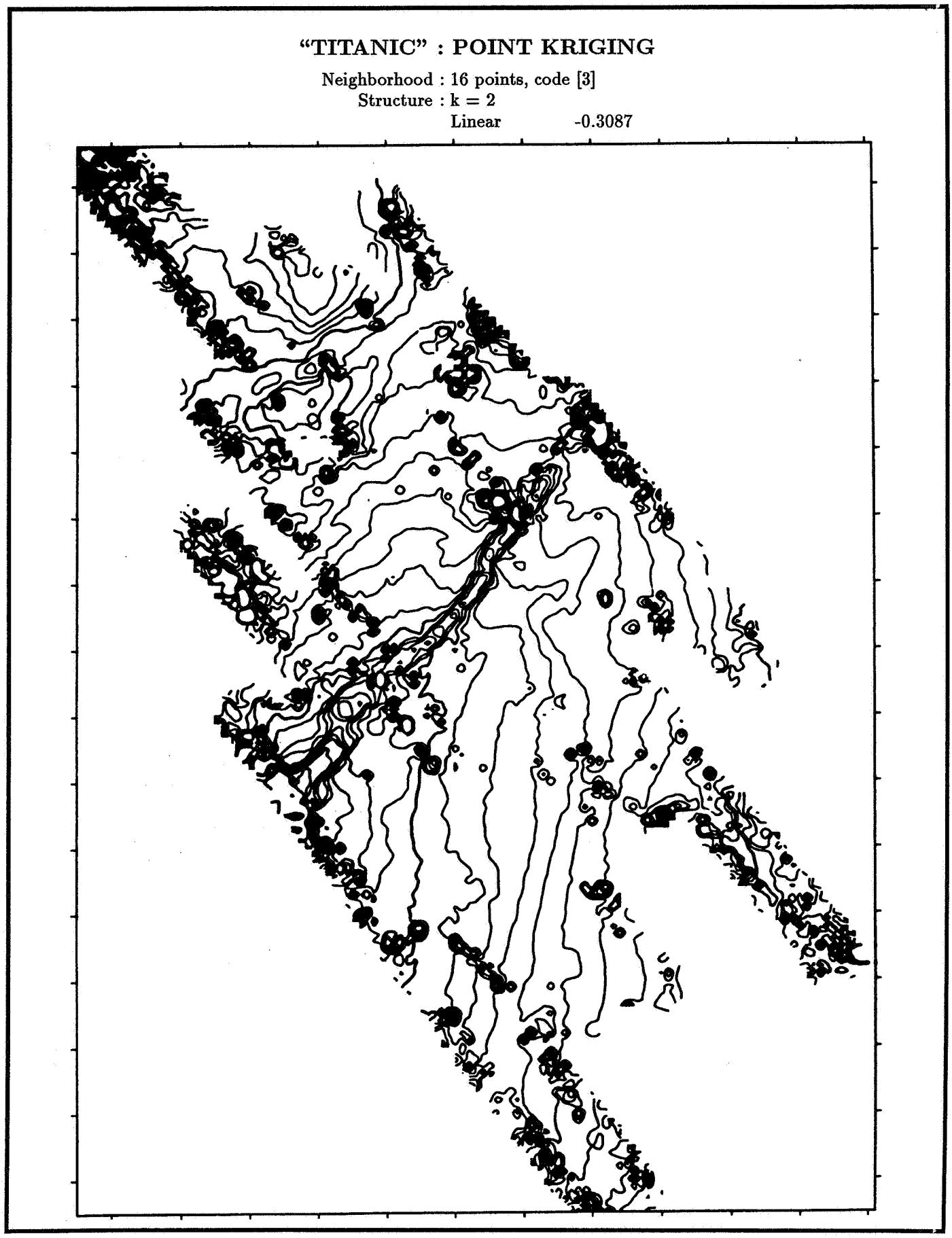

"TITANIC" : POINT KRIGING
Neighborhood : 16 points, code [3]
Structure : k = 2
Linear -0.3087

Figure 9

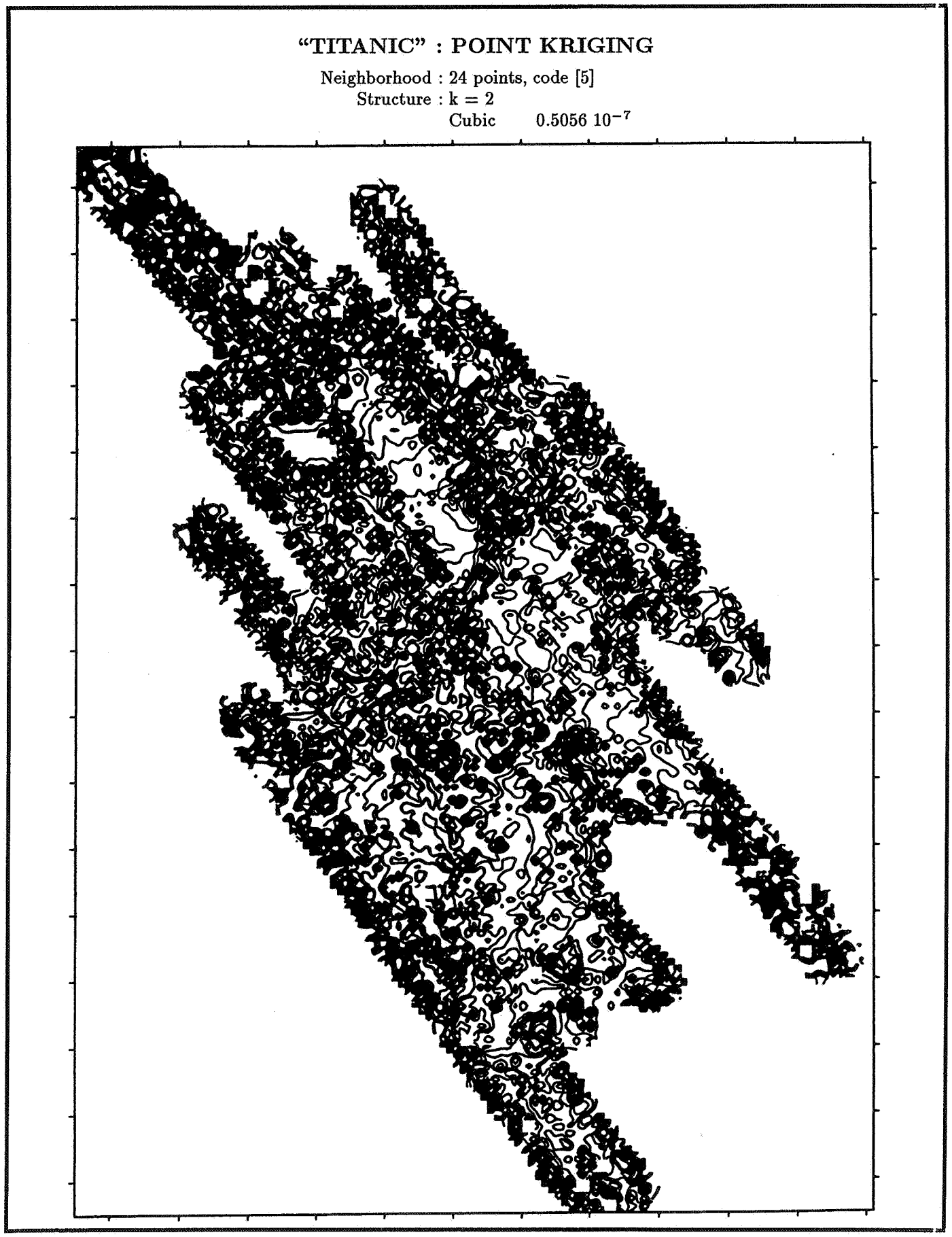

"TITANIC" : POINT KRIGING
Neighborhood : 24 points, code [5]
Structure : k = 2
Cubic 0.5056 10⁻⁷

Figure 10

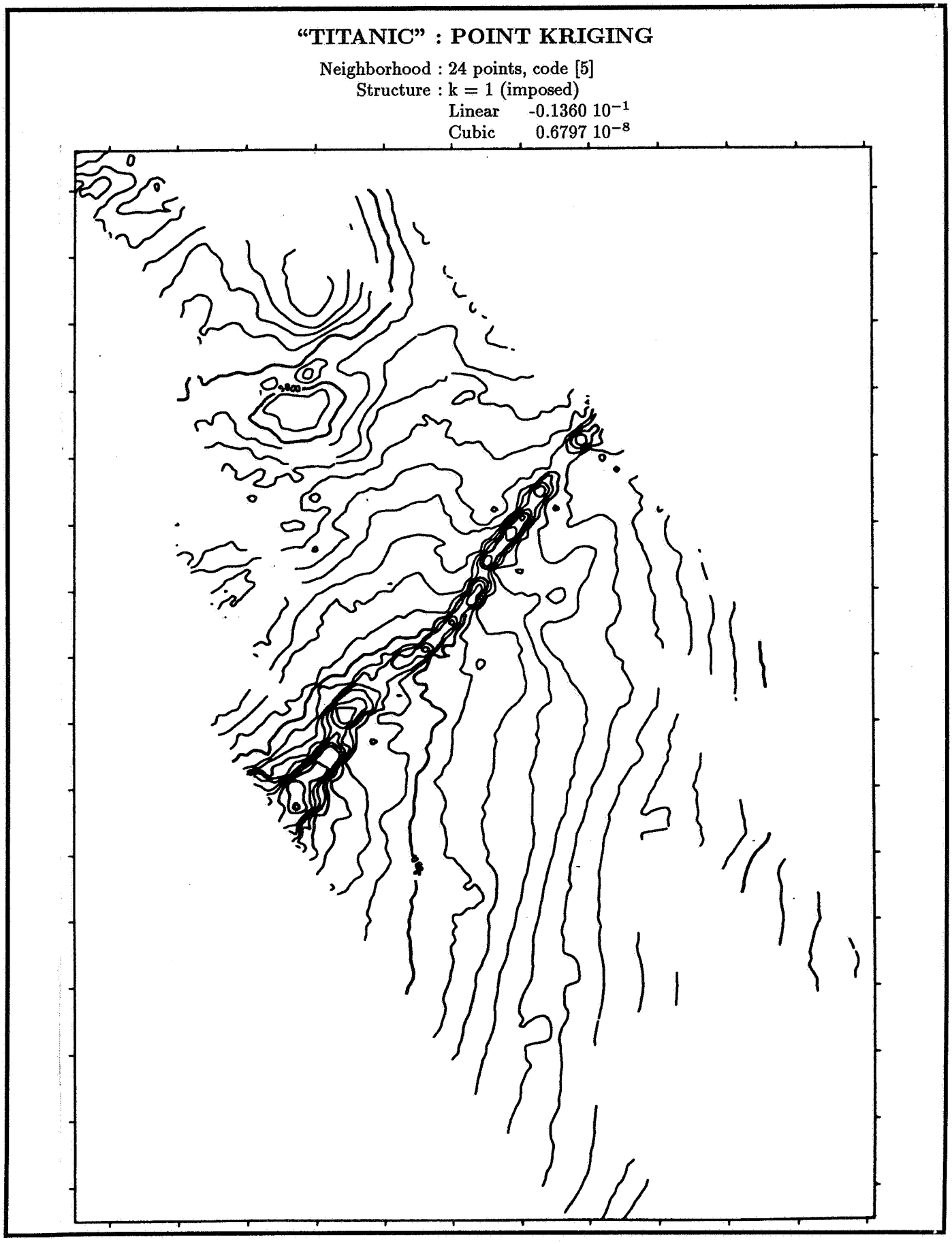

"TITANIC" : POINT KRIGING
Neighborhood : 24 points, code [5]
Structure : k = 1 (imposed)
Linear -0.1360 10^-1
Cubic 0.6797 10^-8
0
-800

Figure 11

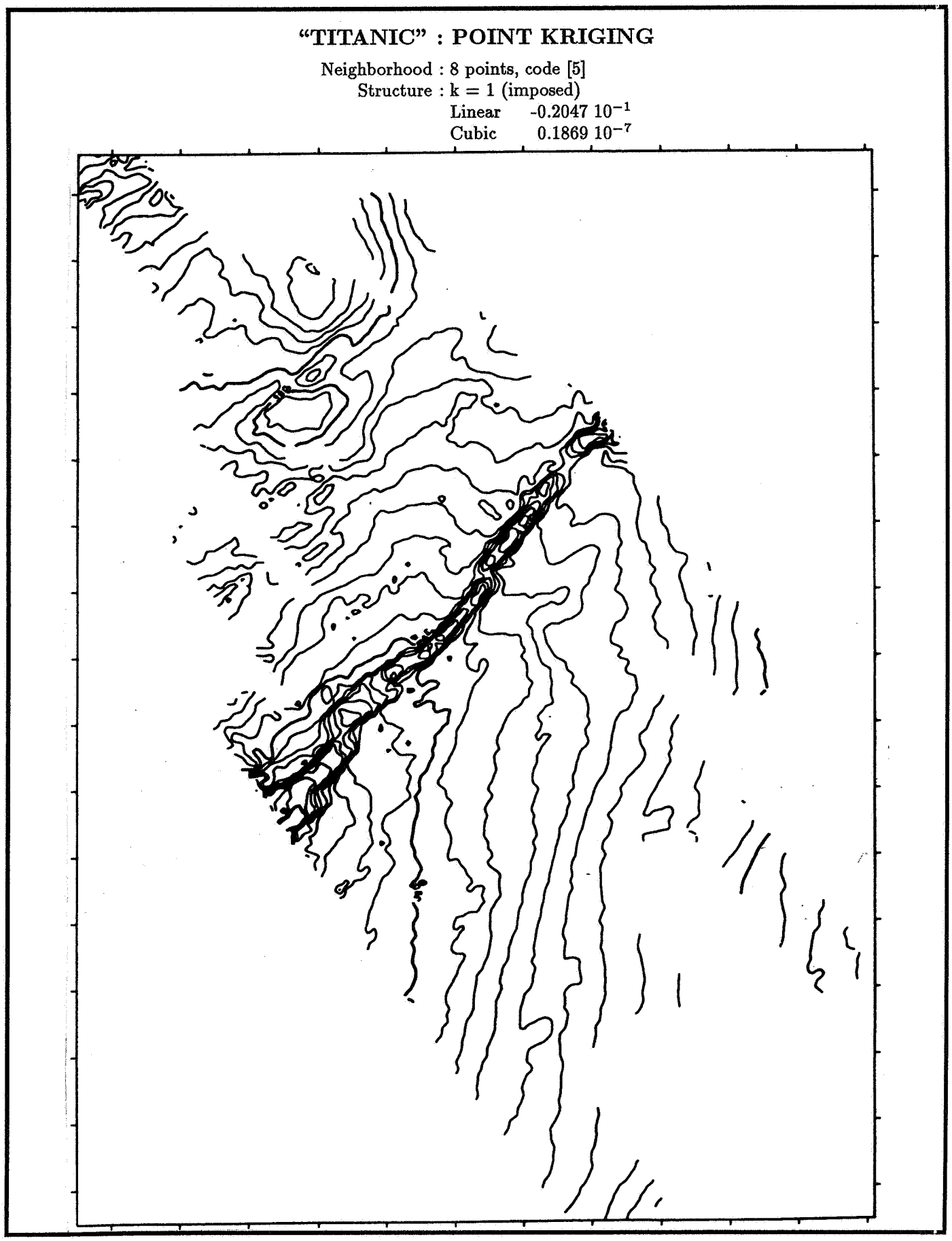

"TITANIC" : POINT KRIGING
Neighborhood : 8 points, code [5]
Structure : k = 1 (imposed)
Linear -0.2047 10^-1
Cubic 0.1869 10^-7

Figure 12

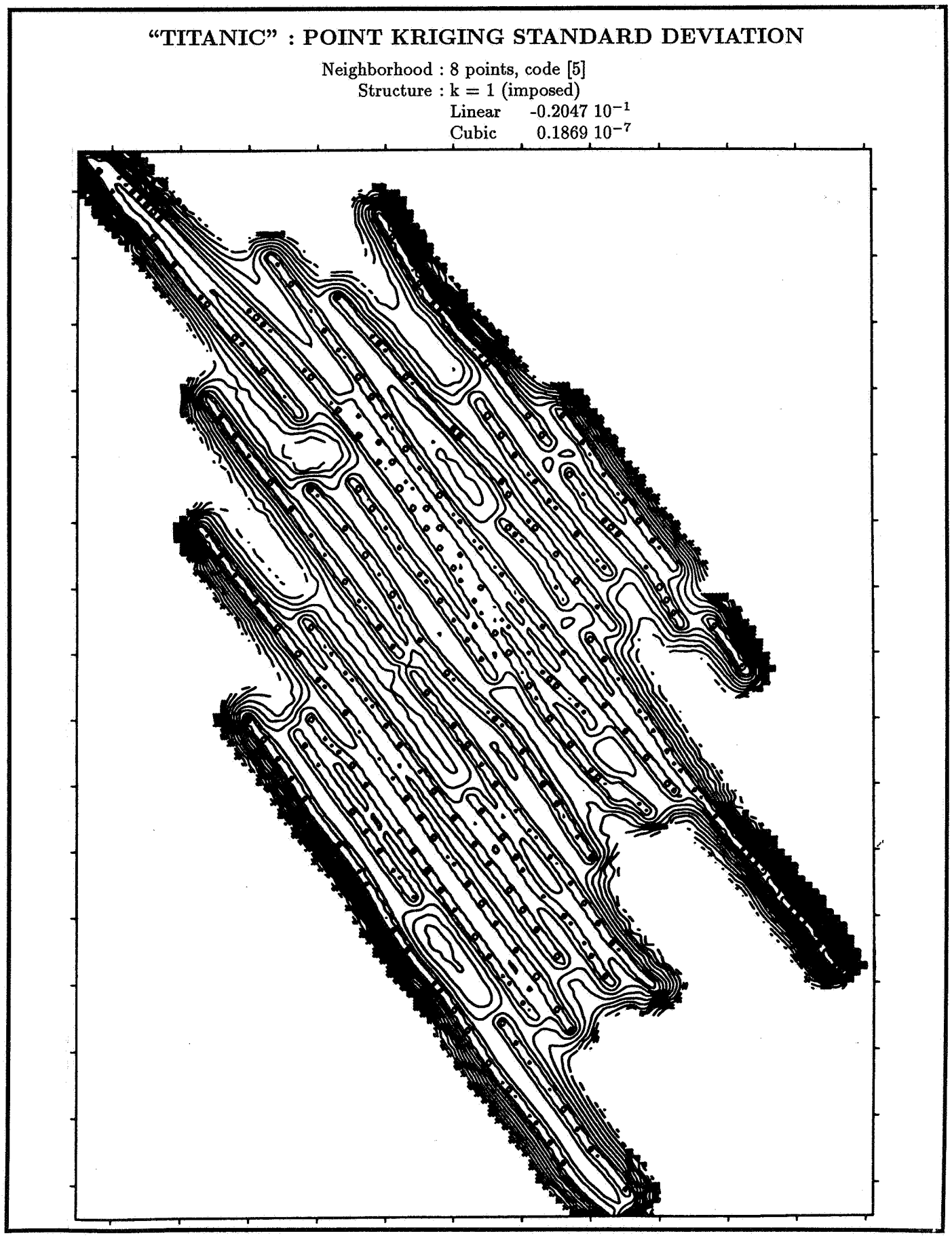
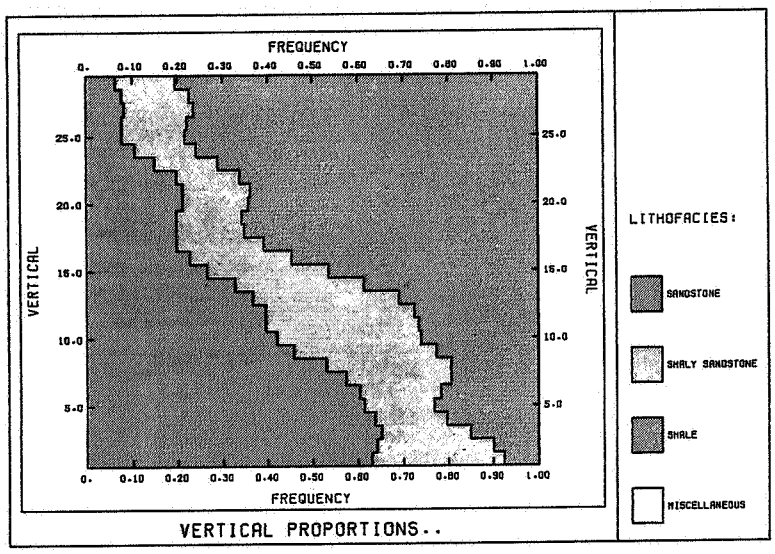
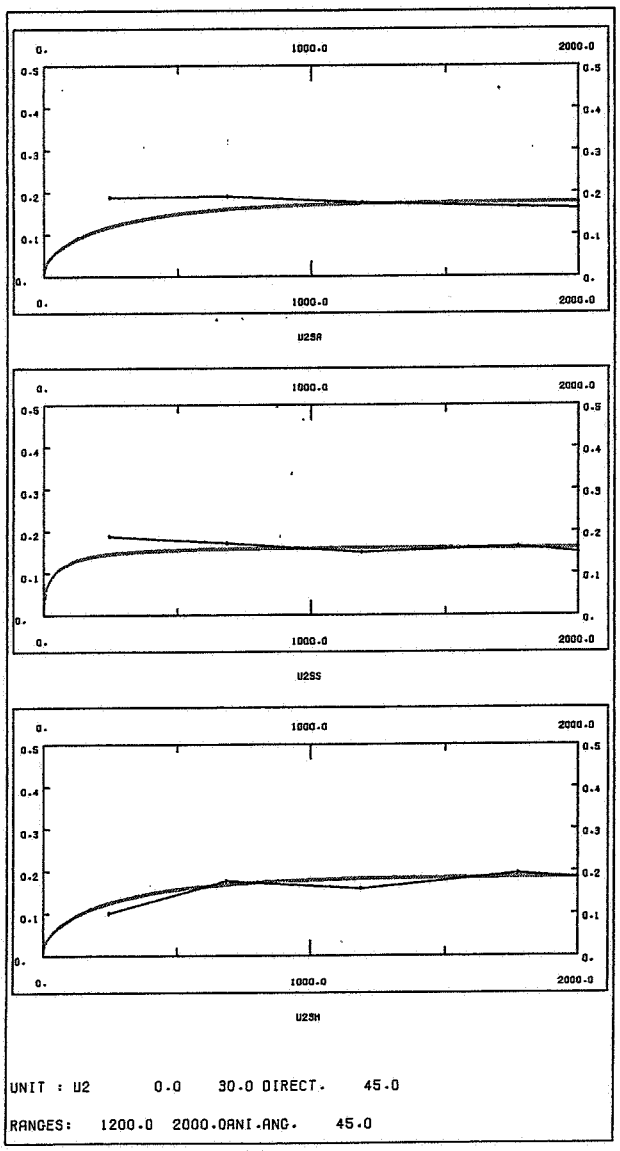

"TITANIC" : POINT KRIGING STANDARD DEVIATION
Neighborhood : 8 points, code [5]
Structure : k = 1 (imposed)
Linear -0.2047 10^-1
Cubic 0.1869 10^-7

Figure 13

HETEROGENEOUS RESERVOIR : STRUCTURAL ANALYSIS





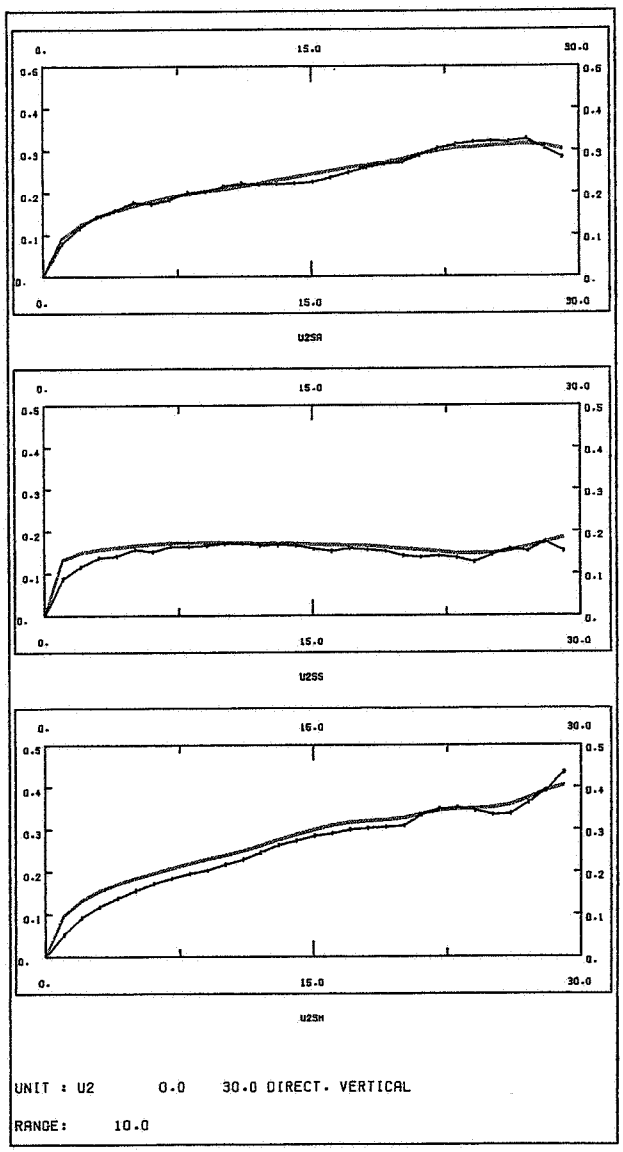


Figure 14

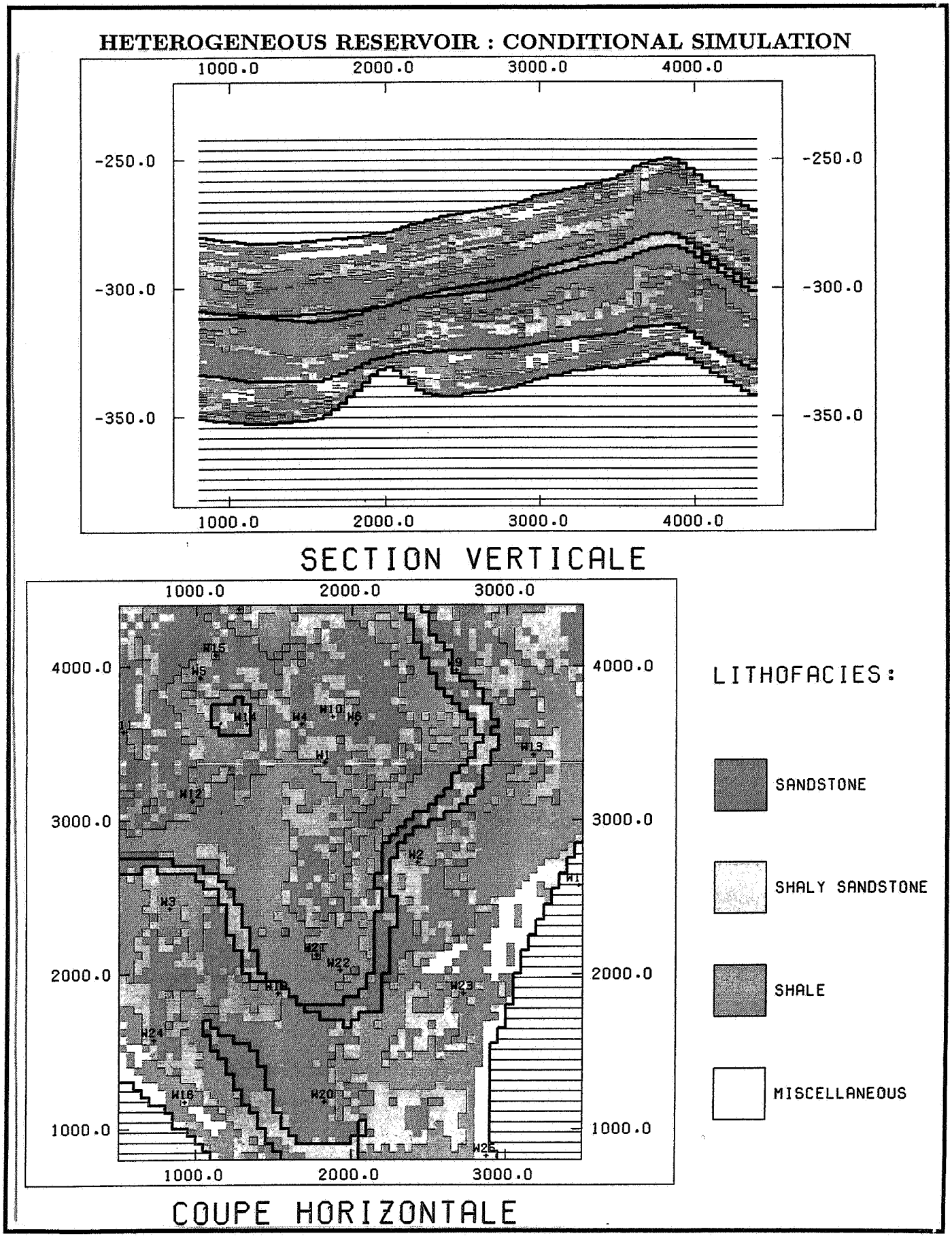
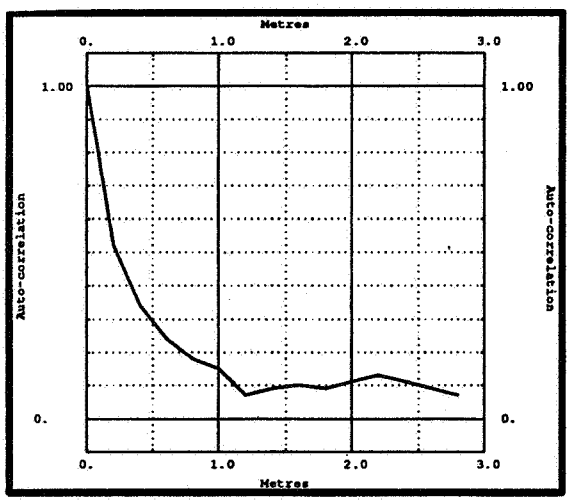
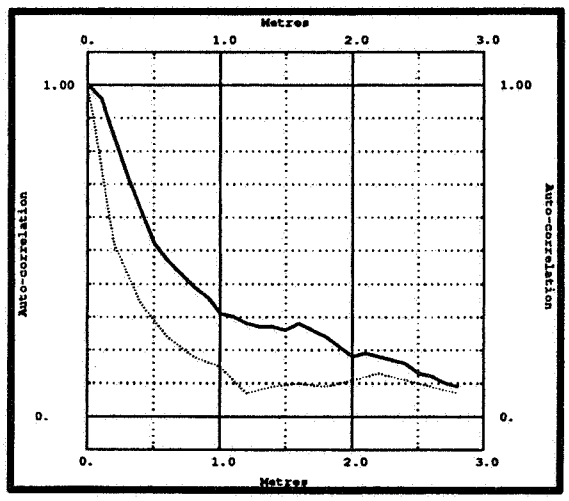

HETEROGENEOUS RESERVOIR : CONDITIONAL SIMULATION
1000.0
2000.0
3000.0
4000.0
-250.0
-300.0
-350.0
SECTION VERTICALE
LITHOFACIES:
SANDSTONE
SHALY SANDSTONE
SHALE
MISCELLANEOUS
COUPE HORIZONTALE

Figure 15

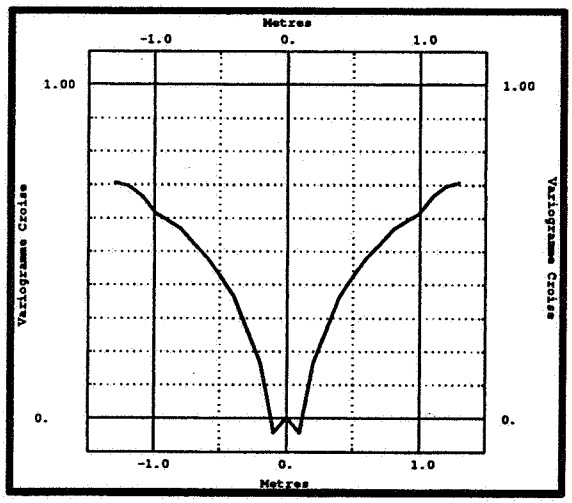
RADIOACTIVITY–GRADE VARIOGRAPHY



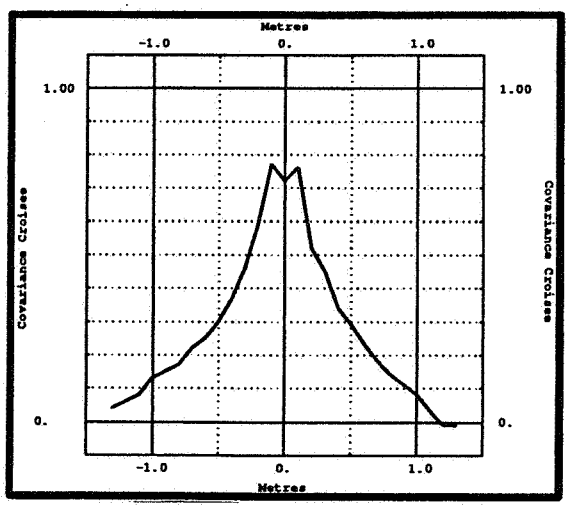
Grade covariance



Radioactivity covariance



Cross variogram



Cross covariance

Figure 16

Les Cahiers du Centre de Morphologie Mathématique

Fasc. 1 : *Le krigeage universel*, par G. Matheron. Paris : ENSMP, 1969. 82 p.

Fasc. 2 : . *Cours de géostatistique*, par G. Matheron. Paris : ENSMP, 1969. 82 p.

Fasc. 4 : . *Théorie des ensembles aléatoires*, par G. Matheron. Paris : ENSMP, 1969. 54 p.

Fasc. 5 : . *La théorie des variables régionalisées et ses applications*, par G. Matheron. Paris : ENSMP, 1971. 212 p.

Fasc. 7 : . *Estimer et choisir*, par G. Matheron. Paris : ENSMP, 1978.175 p.

Cahiers de Géostatistique

Fasc. 1 : *Compte rendu des journées de géostatistique 6–7 Juin 1991 Fontainebleau*, éd. par C. de Fouquet. Paris : ENSMP, 1991. 261 p.

Fasc. 2 : *Aide–mémoire de géostatistique linéaire*, par P. Chauvet. 4e éd. Paris : ENSMP, 1994. 210 p.

Fasc. 3 : *Compte rendu des journées de géostatistique 25–26 Mai 1993 Fontainebleau*, éd. par C. de Fouquet. Paris : ENSMP, 1993. 205 p.

Fasc. 4 : *Processing data with a spatial support : geostatistics and its methods*, par P. Chauvet. Paris : ENSMP, 1993. 57 p.

Fasc. 5 : *Compte rendu des journées de géostatistique 15–16 Juin 1995 Fontainebleau*, éd. par C. de Fouquet. Paris : ENSMP, 1995. 248 p.

Imprimé en France

ISSN 1168–2574