RANDOM FUNCTIONS, AND THEIR APPLICATIONS IN GEOLOGY

By Dr. G. MATHERON

(Abstract)

Many regionalized variables occuring in Geology may be interpreted as realizations (sample functions) of random functions (R.F.) and the powerful tools of R.F. theory may be used for describing and/or explaining structural properties of geological phenomena as well as solving some practical problems. Geostatistics stricto sensu is based on the variogram of a stationary (or, more generally, intrinsic) R.F. and implies applications in mining estimation. In the non-stationary case, universal kriging procedures give the best possible estimator of a drift (trend) and may be applied to contouring problems.

In this paper is examined what minimal probabilistic characteristic of a R.F. is really necessary to know in view of solving a given practical problem (global or local linear estimation), and what minimal hypothesis is really required for a possible estimation of this minimal characteristic.

RANDOM FUNCTIONS, AND THEIR APPLICATIONS IN GEOLOGY

INTRODUCTION

The probabilistic theory of Random Function (R.F.) is widely used today in order to describe or to explain some properties of geological phenomena. The reason why is easy to understand. Conventional Geology may explain the main structural features of a given phenomenon, but generally not in a quantitative way. The details of this phenomenon, and its local behaviour, remain impossible to predict with some accuracy, for this phenomenon simultaneously presents a structural side and a random side. Generally, the random part cannot be considered as a simple white noise superimposed to a continuous component - because this "noise" is actually connected with some structural properties of the phenomenon (for instance, its continuity). Thus, there is a great need for a conceptual tool, able to simultaneously take into account both sides of the phenomenon, and to make a synthesis of them. It turns out that the R.F. theory provides us with such a conceptual tool and with the mathematical techniques which are required for applications.

From an epistemological point of view, it is always possible to consider a given phenomenon as a realization of a R.F. But this point of view would remain purely academic, if we were not able to estimate at least partly the probability law of this R.F. Therefore, we must examine the serious problem of statistical inference for R.F.

In order to define a R.F. as a probabilistic object, it is necessary to know at least its space law (i. e., the simultaneous distributions of the random values taken by the R.F. on every finite point set). To solve some important problems, the space law is not even sufficient, but in our context we may neglect these refinements. As a rule, in geology, the available realization of our R.F. is unique, and its numerical values are known only on a finite set of experimental data. On the other hand, the space law depends on an infinite number of unknown parameters, and statistical inference would remain quite impossible if we did not assume some hypotheses (for instance, the stationarity) in order to reduce the number of the parameters we have to estimate.

But such hypotheses (like stationarity) are very strong ones, and very often they cannot be verified, even approximately. Thus, it is of a great methodological importance to accurately answer the two following questions :

1 - What minimal characteristic of a R.F. is really needed in order to solve a given practical problem ?

2 - What minimal hypothesis must we assume to possibly estimate this required characteristic ? (for instance, is it really necessary to assume that our R.F. is a stationary one)

In what follows, I shall only examine some particular cases, connected with the stationarity hypothesis and the problem of linear estimation.

1 - STATIONARY AND INTRINSIC RANDOM FUNCTIONS

In many applications, (particularly in problems such as linear estimation or linear prediction) it is not necessary to know all the space law of a R.F. $Z(x)$, but only its moments of order 1 and 2 (if they do exist), i.e. its expectation $E\,Z(x) = m(x)$ and its covariance :

$$C(x,y) = E\big(Z(x)\ Z(y)\big) - m(x)\ m(y) \qquad (x,y \in R^n)$$

It is often assumed as a widely used but fairly strong hypothesis, that the R.F. is wide-sense stationary , i.e., :

a/ the covariance does exist

b/ the expectation is a constant, and the covariance depends only on the difference $x-y$, but not separately on each of the points x and y.

As a matter of fact, these assumptions are stronger than it is really necessary, and we may at first change them into the following :

a'/ (Intrinsic hypothesis) - The increments $Z(x+h) - Z(x)$ of the R.F. $Z(x)$ are wide-sense stationary (but not necessarily the R.F. itself).

This last hypothesis implies the existence of a linear drift (which may eventually vanish) :

$$E[Z(x+h)-Z(x)] = a\ h = \sum_{i=1}^{n} a_i\ h_i$$

and the existence of an intrinsic variogram :

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

If the variogram remains bounded at the infinity, the intrinsic hypothesis implies the wide sense stationarity. In this case, the following relationship :

$$\gamma(h) = C(o)-C(h)$$

shows that variogram and covariance are perfectly equivalent. But, if the variogram is not bounded, the covariance does not exist at all, and the wide sense stationarity is no longer valid. It is the case, for instance, of the widely used de Wijsian variogram $\gamma(h) = \alpha \log |h|$. Although the covariance does not exist, the variogram enables us to solve in the same way all the linear estimation problems we may encounter.

2 - AN EXAMPLE : THE BROWNIAN MOTION

Let us now examine a simple example of a process satisfying the intrinsic hypothesis a'/, but not the wide sense stationarity, and show how misleading may sometimes be the usual procedures of statistical inference.

Let us denote by $Z(t)$ a brownian motion on the straight line $-\infty < t < \infty$, a realization of which is known on an interval $0 \leq t \leq L$. $Z(t)$ is a process with stationary independent increments, and is well characterized by its linear variogram $\gamma(h) = |h|$. Note that - this variogram not being bounded - there exists no stationary covariance.

But nevertheless, the usual procedure of statistical inference will give an estimation of it - which as a matter of fact will be a pure artefact.

First, to estimate the expectation $m = E(Z(t))$ (which really does not exist), we shall compute the experimental mean :

$$\overline{Z} = \frac{1}{L} \int_0^L Z(x)\, dx$$

then, putting

$$C^*(x,y) = (Z(x)-\overline{Z})(Z(y)-\overline{Z})$$

for x,y belonging to [0,L], we shall compute the sum :

$$(2\text{-}1) \qquad C^*(h) = \frac{1}{L-h} \int_0^{L-h} C^*(x+h,\, x)dx$$

and consider its numerical value as an estimator of our (non-existing) covariance.

Let us now compute the expectation of $C^*(h)$. With the variogram $\gamma(h) = |h|$, as it can easily be shown, ([3]) we get :

$$E(C^*(x+h),\, x)) = \frac{2}{3} L + \frac{x^2 + (x+h)^2}{L} - 2x - 2h$$

Substituting this result to (2-1), we get :

$$(2\text{-}2) \qquad E(C^*(h)) = \frac{1}{3} L - \frac{4}{3} h + \frac{2}{3} \frac{h^2}{L}$$

For the variance (the real value of which is infinite), we get the estimator $C^*(o)$, the expectation of which

$$E\left(C^*(o)\right) = \frac{1}{3} L$$

depends only on the length of the interval (O,L) we have chosen. Clearly, $C^*(o)$ and $C^*(h)$ are pure artefacts. Even the slope at the origin has been altered ($\frac{4}{3}$ instead of 1 for the true variogram). The bias introduced by statistical inference are so strong that we always (but only apparently) get a confirmation of our (wrong) starting hypothesis concerning the existence of covariance.

We may add, in this case, that this alarming result would have been avoided, had we used the unbiased expression

$$\gamma^*(h) = \frac{1}{2(L-h)} \int_o^{L-h} \left(Z(x+h)-Z(x)\right)^2 dx$$

as an estimator of the variogram.

3 - THE GLOBAL ESTIMATION PROBLEM

Let us assume now that we have to estimate the mean value :

$$m(V) = \frac{i}{V} \int_V Z(x)\, dx$$

of the R.F. Z(x) in a given volume V, knowing only the numerical values $Z(x_i)$ (sample values) taken by Z(x) on a finite set of points x_i, regularly distributed in volume V. To solve this global estimation problem, we may use the sample mean :

$$m^* = \frac{1}{n} \sum_{i=1}^{n} Z(x_i)$$

In this case, the estimation variance [3], [4], is given by :

$$D^2(m(V)-m^*) = \frac{1}{V^2} \int_V \int_V \gamma(x-y)dx\,dy -$$

$$- \frac{2}{n\,V} \sum_i \int_V \gamma(x-x_i) + \frac{1}{n^2} \sum_i \sum_j \gamma(x_i-x_j)$$

This well known formula depends only on the variogram (and on the geometry of our sampling). A further examination shows that our estimation variance chiefly depends on the values taken by $\gamma(h)$ in the neighbourhood of the origin (for $|h| \leq a$, where a denotes the sample spacing). It is a very pleasant circumstance, because it also turns out that statistical inference for the variogram $\gamma(h)$ itself is, in general, reasonably possible only for the first experimental points [3]. In other words, we can get a good knowledge of our variogram only in the neighbourhood of the origin, but nothing else is really required in order to compute the estimation variance.

The preceding result apparently depends on the intrinsic hypothesis. Actually, it is possible to get free from this hypothesis, by just assuming the existence of a (non intrinsic) variogram $\gamma(x,y)$, which separately depends on points x and y (and no longer only on their difference), provided that for each fixed h the function $\gamma(x,x+h)$ does not vary too quickly with x. For, by putting

$$\overline{\gamma}(h) = \frac{1}{V(h)} \int_{V(h)} \gamma(x,x+h)dx \qquad (3\text{-}2)$$

(V(h) denoting the set of points x such that $x \in V$ and $x+h \in V$) it can be shown that the estimation variance depends only on the beha-

viour of $\bar{\gamma}(h)$ in the neighbourhood of 0, exactly as in the intrinsic case ([4]), and again the estimation of this part of $\bar{\gamma}(h)$ is reasonably possible.

Thus, we have found out a first answer to our basic question. As far as a global problem is concerned, neither stationary nor intrinsic hypotheses are really necessary. We only have to estimate the beginning of the graph of the function $\bar{\gamma}(h)$ appearing in (3-2) - and this is generally possible. We may also notice, from an experimental point of view, that the mean value $\bar{\gamma}(h)$ of a non-intrinsic $\gamma(x,y)$ in a given volume V must be estimated exactly by the same procedure that an intrinsic $\gamma(h)$ in the same volume V. Thus, in a sense, (but only for the global problem) it is quite legitimate to treat a non intrinsic R.F. in the same way that an intrinsic one, provided that the available data form a regular covering of the volume V we have to estimate.

4 - THE LOCAL ESTIMATION PROBLEM

Let us now examine the local estimation problem. Knowing the numerical values taken by the realization of a R.F. Z(x) on a given set S of experimental points x_i, we now have to estimate the true (unknown) value $Z(x_o)$ at a given point $x_o \in S$, or, more generally, the value of a weighted average $\int \mu(dx)\, Z(x)$, with a given measure μ, the support of which does not intersect the set S.

In the wide sense stationary case, this problem can easily be solved by the techniques of the linear prediction [1], provided that

the expectation $m = E(Z(x))$ is known. If the expectation is not known, or if the R.F. is intrinsic but not stationary, a slight modification of this technique leads to the kriging procedure, which is well known in Geostatistics. ([3], [4]) In order to apply this procedure, it is necessary to know fairly well the variogram $\gamma(h)$. Actually, the intrinsic hypothesis is not necessary in itself, but is only required in order to provide a good statistical inference for the variogram $\gamma(h)$.

To what extent is it again possible to get free from any intrinsic hypothesis? The answer here is only partly positive ([2], [5]). Let us denote by $Z(x)$ a non stationary R.F., by :

$$m(x) = E(Z(x))$$

its expectation (which is called <u>the drift</u>), and assume that in some neighbourhood V of each space point x_o the drift is well approximated by the following expression :

$$(4\text{-}1) \qquad m(x) = \sum_{\ell=0}^{k} a_\ell f_\ell(x) \qquad (x \in V)$$

in which the $f_\ell(x)$ are known (a priori chosen) functions, for instance polynomials, and the a_ℓ are unknown numerical coefficients (to be estimated). Let us also assume we know the covariance $C(x,y)$, or the variogram $\gamma(x,y)$ of the residuals $Z(x) - m(x)$. Then, if experimental data are available on a space point set S, <u>universal kriging</u> (U.K.) procedure gives the optimal solution for the three fundamentally different following problems :

a/ Estimating the drift itself (the well known problem of "trend surface analysis" will here, perhaps, encounter its happy end) -

for instance, in geophysics, estimating a regional anomaly - Note that in the finite Gaussian case, the U.K. estimator of the drift is identical to the maximum likelihood estimator.

b/ Estimating the real (unknown) value of $Z(x)$ in points $x \notin S$, with obvious applications to contouring problems. Note that for a point $x_i \in S$ on which the experimental $Z(x_i)$ is known, the U.K. estimator is identical to $Z(x_i)$ itself : the U.K. is an exactly fitting interpolation procedure.

c/ At last, estimating a moving average on a set $S' \neq S$, for instance, in mining problems, estimating the grade of a given panel.

In each of these problems, we can easily get the corresponding (optimal) estimation variance. For instance, in contouring problems, the map itself is completed by an isovariance map indicating the precision with which each point is known.

The main problem which arises in the applications consists in identifying the real (unknown) variogram of the residuals (the underlying variogram). Although fairly advanced, this problem is not entirely solved now. It appears that a sort of "quasi-stationarity" condition will be required for the residuals : for instance, a condition expressing that the variogram $\gamma(x,y)$ of the residuals may be approximated by the relation :

$$\gamma(x,y) = \varpi\ \gamma(x-y)$$

where $\gamma(h)$ is an intrinsic variogram (to be estimated) and ϖ a slowly varying factor we can consider as a constant on the neighbourhood V in which (4-1) remains valid. Thus we are not entirely unable to treat the non-stationary case - and this will be my general conclusion.

BIBLIOGRAPHY

[1] - CRAMER, H. and LEADBETTER, M.R. - Stationary and related stochastic processes - J. Wiley and Sons, New York, 1968.

[2] - HUIJBREGTS, Ch. and MATHERON, G. - Universal kriging (an optimal method for estimating and contouring in trend surface analysis) - CIMM, IX International Symposium, Montreal, June 1970.

[3] - MATHERON, G. - Les Variables régionalisées et leur estimation - Masson, Paris, 1965.

[4] - MATHERON, G. - Osnovy prikladnoï geostatistiki - Mir, Moscow, 1968.

[5] - MATHERON, G. - Le krigeage universel - Les Cahiers du Centre de Morphologie Mathématique, Fontainebleau, 1969.