RELATIONS BETWEEN THE INDICATORS RELATED TO A REGIONALIZED VARIABLE

Jacques RIVOIRARD
Centre de Géostatistique
Ecole des Mines de Paris
35 rue Saint–Honoré
77300 Fontainebleau – France

0. INTRODUCTION

In the seventies, in order to make predictions for future selective mining, geostatistics had to deal with the problem of estimating reserves above cut–off grades. Since then practitioners have used techniques based on, for example, multigaussian distributions, disjunctive kriging (cokriging of indicators), or more simply kriging of indicators. The indicators above a threshold define the ore at this cut–off or, in a non mining context, define the geometric set of values above the cut–off.

During the past decade the interest in geometric problems has increased, in particular for simulations. So we have seen the development of techniques for simulating random functions $Z(x)$ from indicators $I[Z(x) < z]$ at different thresholds z, or conversely, techniques for simulating sequential geological facies obtained by thresholding a multigaussian random function.

In many such cases several thresholds and indicators corresponding to different random sets related to the variable under study have to be handled simultaneously. These sets depend on each other and their mutual arrangement is an important structural characteristic of the variable. In this paper we propose simple tools to describe this arrangement.

1. SOME THEORY

Thresholding a Random Function (RF) gives Random Sets (RS), the structure of which is related to the structure of the RF. Here we will only characterize structures from pairs of points (x, x+h). The RF Z(x) is supposed to have stationary bivariate distributions (Z(x), Z(x+h)) with covariance $\sigma(h)$. The cut–off z divides the space into the RS of points with value $\geq z$ and its complementary set. The indicator $I[Z(x) \geq z]$ (equal to 1 on the first set, 0 on the other) has the mean $E(I[Z(x) \geq z]) = P[Z(x) \geq z] = T(z)$.

Its covariance

$$\sigma_z(h) = \mathrm{Cov}\{I[Z(x) \geq z], I[Z(x+h) \geq z]\}$$

$$= P[Z(x) \geq z, Z(x+h) \geq z] - T(z)^2$$

and its variogram $\gamma_z(h) = 0.5\ \mathrm{E}\{ I[Z(x) \geq z] - I[Z(x+h) \geq z] \}^2$

$= 0.5\ (P\{ I[Z(x) \geq z] \neq I[Z(x+h) \geq z] \})$

i.e. $0.5\ (P[Z(x) < z, Z(x+h) \geq z] + P[Z(x) \geq z, Z(x+h) < z])$

are functions of the bivariate distribution (Z(x), Z(x+h)). Thus, it is possible to explicitly compute the structure of this indicator under the hypothesis of a given bivariate distribution (such as bigaussian). Note that the indicator $I[Z(x) \geq z]$ and its complement $I[Z(x) < z] = 1 - I[Z(x) \geq z]$ have the same variogram and covariance.

Generally the structure of the indicator $I[Z(x) \geq z]$ changes when the cut–off z varies: apart from the variance which acts as a multiplicative factor, the shape changes as we will see in examples. Recall (e.g. Matheron 1982) that summing the indicator variograms for all cut–offs gives the order 1 variogram of Z(x)

$$\int \gamma_z(h) dz = \frac{1}{2}\ \mathrm{E}\ |Z(x+h) - Z(x)|$$

The cross covariance between indicators at cut–offs z and z' is

$$\sigma_{z\,z'}(h) = \mathrm{Cov}\,(I[Z(x) \geq z],\ I[Z(x+h) \geq z'])$$

$$= P[Z(x) \geq z, Z(x+h) \geq z'] - T(z)\ T(z')$$

It is not necessarily symmetric in h. Knowing these covariances for all cut–offs is equivalent to knowing the bivariate distributions (Z(x), Z(x+h)) for all distances h. Summing these covariances gives the covariance of Z(x)

$$\int\int \sigma_{z\,z'}(h)\ dz\ dz' = \sigma\ (h)$$

Unlike the covariance, the cross variogram mixes the cases +h and –h

$$\gamma_{z\,z'}(h) = 0.5\ E[I(Z(x+h) \geq z) - I(Z(x) \geq z)][I(Z(x+h) \geq z') - I(Z(x) \geq z')]$$

or, if $z \leq z'$: $0.5\ (P[Z(x) < z, Z(x+h) \geq z'] + P[Z(x) \geq z', Z(x+h) < z])$

We will see (next section) that the way the RS are arranged for two cut–offs $z \leq z'$ can be conveniently described using the conditional probabilities

$P\ [Z(x+h) \geq z'\ |\ Z(x+h) \geq z,\ Z(x) < z]$ denoted $P_h\ [\rightarrow \geq z'\ |\ \rightarrow \geq z]$

$P\ [Z(x+h) < z\ |\ Z(x+h) < z',\ Z(x) \geq z']$ denoted $P_h\ [\rightarrow < z\ |\ \rightarrow < z']$

The first one is the probability that, going from a point x with value $< z$ to a point $x+h$ with value $\geq z$, the second value is $\geq z'$. In other words, it is the probability with which, getting into the domain of values $\geq z$, one meets a value $\geq z'$. The second one is the probability with which, getting into the values $< z'$, one meets a value $< z$. The first probability describes the "border effects" within the domain $\geq z$, the second one the border effects within the domain $< z$.

In practice, such probabilities can be computed from pairs $(Z(x), Z(x+h))$ of points separated by h. They are expected to be equal for $+h$ and $-h$ only if the distributions of pairs are symmetric in h. This will be supposed here from now on. Then the cross variogram can be written $(z \leq z')$

$$\gamma_{zz'}(h) = P[Z(x) \geq z', Z(x+h) < z]$$

whereas the simple one becomes

$$\gamma_z(h) = P[Z(x) \geq z, Z(x+h) < z]$$

We then get $(z \leq z')$ $\quad P_h[\rightarrow \geq z' \mid \rightarrow \geq z] = \gamma_{zz'}(h) / \gamma_z(h)$

$$P_h[\rightarrow < z \mid \rightarrow < z'] = \gamma_{zz'}(h) / \gamma_{z'}(h)$$

So comparing the cross variogram between two indicators to their simple variograms gives our conditional probabilities directly. All these results are not original, but seem to have been insufficiently exploited for structural analysis as well as for the choice of estimation or simulation techniques. We will now see what these probabilities physically mean, then how they can be used.

2. SOME TYPES OF ARRANGEMENTS

Some schematic figures will help in understanding. In each case the RF Z(x) takes values 1, 2, 3, 4 with given probabilities (the names : mosaic, residual, and diffusion will be justified in section 4.).

In the mosaic model (Figure 1), the space is divided into compartments. Each compartment is given one of the four values, according to their probabilities, and independently from the other compartments. So, when going from a point x with value 1 to a point $x+h$ with value $\neq 1$, this point takes any of the other values 2, 3, 4 without preference. Going into the domain C2 of points ≥ 2, the probability to reach a value ≥ 3 does not depend on h either. In the same way, going into the domain < 3, the probability to reach a value < 2 does not depend on h. To sum up, probabilities such as $P_h[\rightarrow \geq 3 \mid \rightarrow \geq 2]$ and $P_h[\rightarrow < 2 \mid \rightarrow < 3]$ are constant in h.

In a diffusion-type model (Figure 2), when leaving the domain of points 1, we meet first values 2. When leaving the 2's, there is a transition to 1 or 3, depending on whether we are going upwards or downwards. Probabilities like $P_h[\rightarrow \geq 3 \mid \rightarrow \geq 2]$ or

P_h [→ < 2 | → < 3] are not constant. They start at 0 (no contact between 1 and 3) and then increase with h.

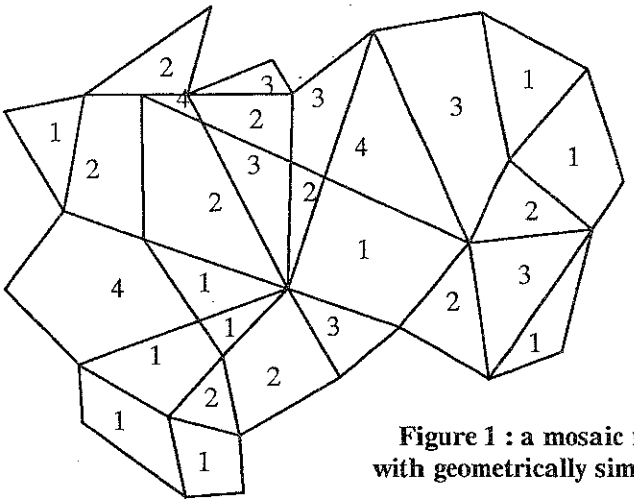

Figure 1 : a mosaic model with geometrically simple tiles

It is possible to imagine less strict transitions, Figure 3. Imagine that the values ≥ 3 are anywhere within the domain C2 of points ≥ 2, but are rarer within the borders of this domain. Leaving the 1's, we will meet, not always but still preferentially, values 2. The probability P_h [→ ≥ 3 | → ≥ 2] increases with the first distances h, but starts from a positive value, expressing a possible contact between values 1 and values ≥ 3.

Imagine now – **model with indicator residuals**, Figure 4 – that within the domain C2 of values ≥ 2, the values ≥ 3 are randomly distributed no matter the proximity of the frontier of C2. There are no border effects within C2: leaving the 1's, we meet values 2 or ≥ 3 without any preference: P_h [→ ≥ 3 | → ≥ 2] is independent of h. There is no longer any tendency to transit at value 2 when going from the 1's domain to the ≥ 2 values. Similarly, the values 4 being distributed randomly and without border effects within the domain ≥ 3, there is no tendency to transit through 3 when entering the ≥ 2. However there is a hierarchy in this model, and what happens when entering the domain of values < a cut–off is not the same as when entering the domain ≥ a cut–off. Here, leaving C3, the probability P_h [→ < 2 | → < 3] to reach a value 1 depends on h : it is small for small h (because the frontiers of C3 are more in contact with values 2), and increases afterwards.

Many other arrangements may be thought of, e.g. that the values 4 are distributed randomly and without any border effects within the domain C3, but that these values ≥ 3 themselves avoid the borders of C2. Or that C3 is preferentially located on the right parts of C2, etc. Here we will only consider the simple types.

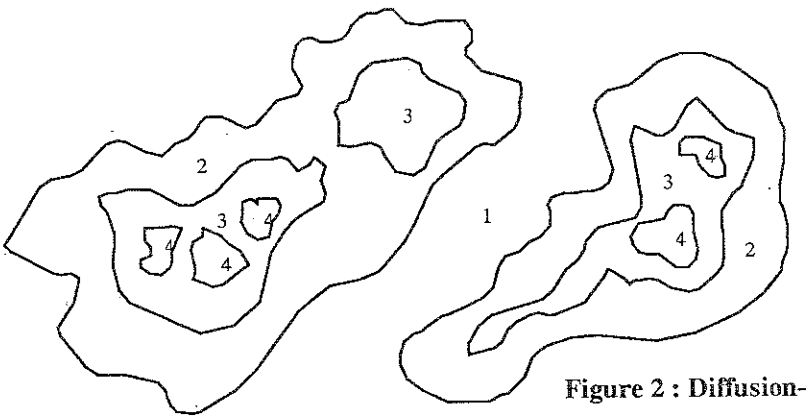

Figure 2 : Diffusion–type model

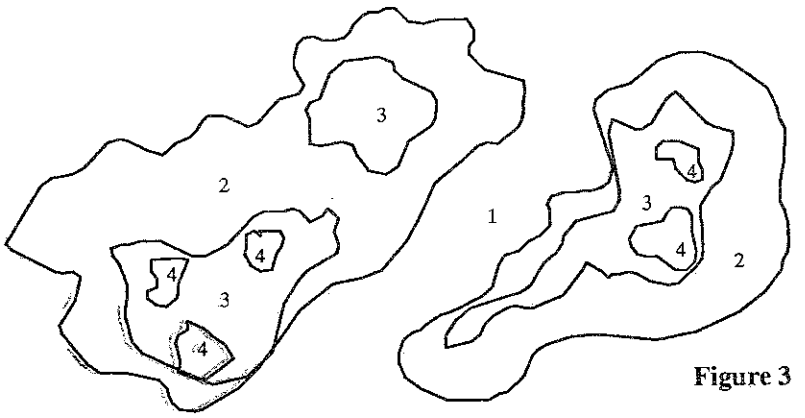

Figure 3

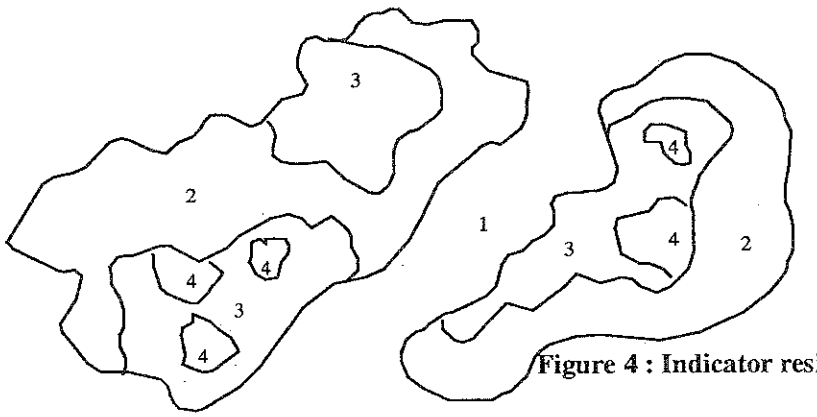

Figure 4 : Indicator residual model

In short, and in term of indicator variograms :

If the cross variograms of indicators $\gamma_{zz'}(h)$ are proportional to the variograms of the lower threshold indicator $\gamma_z(h)$, it is the indicator residual model. The sets of points $\geq$ the different thresholds are nesting without border effects one within the other.

If the cross variograms are proportional to the variograms of the larger cut–off, it is also an indicator residual model, but in the other direction ("decreasing" instead of "increasing"). Now it is the sets with values lower than the different cut–offs which are nesting one within the other without border effects.

If the ratios $\gamma_{zz'}(h)/\gamma_z(h)$ are constant in h whatever the thresholds z and z', we have the mosaic model. There is no border effect whether increasing or decreasing.

Lastly, if the ratios $\gamma_{zz'}(h)/\gamma_z(h)$ increase with h, there are border effects, increasing and decreasing, like in the diffusion–type models.

However, when the cut–offs z and z' correspond to close probabilities, their cross and simple variograms are of course close to each other, even in a diffusion–type model. Moreover for quite distinct cut–offs, there are in practice always border effects, more or less well marked. Taking them into account is a matter of appreciation and the above indications have to be moderated. The best is to look at real examples.

3. REAL EXAMPLES

Example 1

These data are herring acoustic densities, which were provided by K. Foote and I. Rottingen (Laboratory of Marine Research, Bergen, Norway). Petitgas (1991) has made a geostatistical study of them. Each value is in fact the regularized density over a 1 nautical mile segment along parallel transects. In the units that were used, 37% of the 986 values exceed $10^2 = 100$, 12% $10^3 = 1000$, and 3% $10^4 = 10000$.

The fluctuations that can be seen on the variograms of the corresponding indicators I2, I3, I4 show that the structures are not well known: Figure 5, distances in nautical miles. However there is a clear destructuration of the simple variograms when the threshold increases. The cross variogram between two indicators, e.g. I2 and I3, looks like I2, although it is not similar. The ratio between these two (the probability to exceed 1000 when entering the > 100 domain) shows a rapid increase, which indicates small border effects within this domain. Petitgas chose to neglect these effects and used an indicator residual model to obtain a disjunctive kriging map (see next section).

In comparison, the ratio of the cross variogram between I2 and I3 on the variogram of I3 (the probability to be lower than 100 when entering the < 1000 domain) shows much more important border effects, which conforms to the increasing model with indicator residuals.

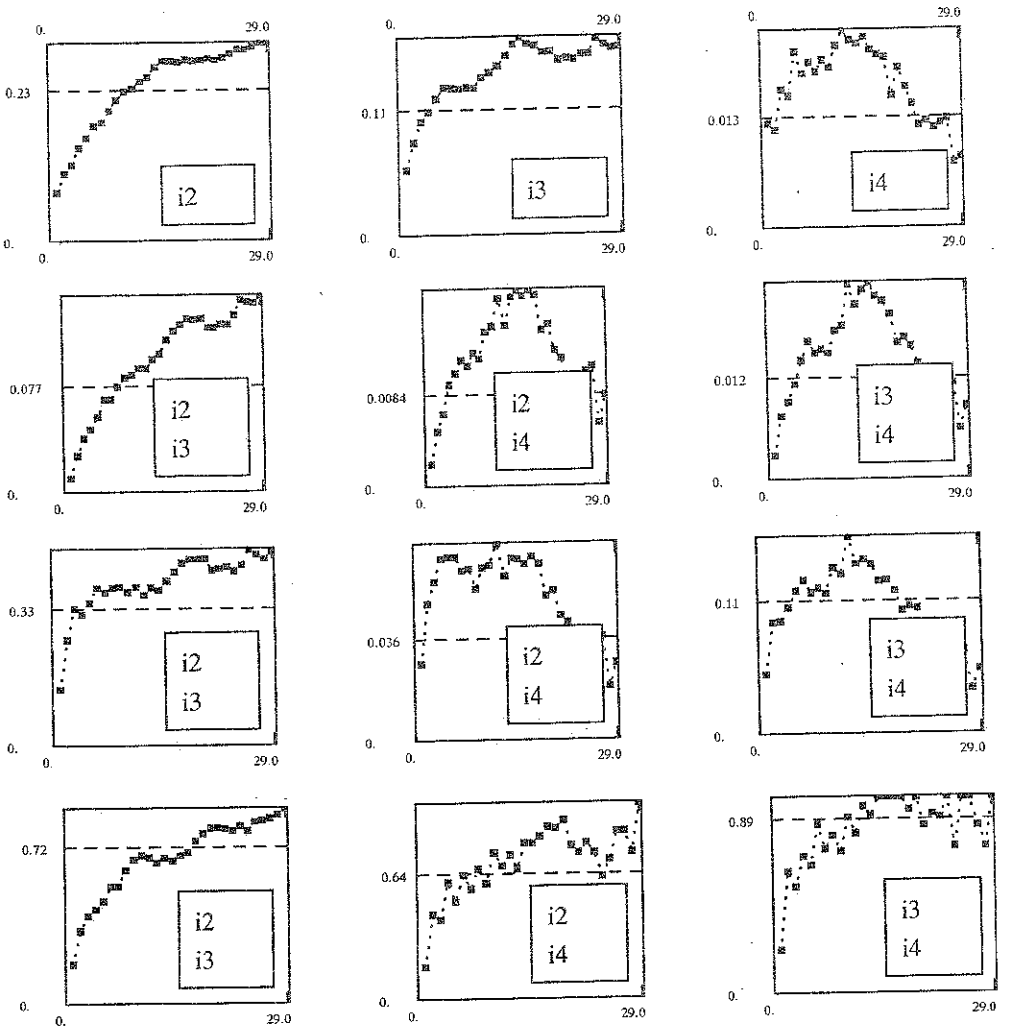

Figure 5 : thresholds 10^2, 10^3, 10^4 on fish densities.

From top to bottom:

- variograms of indicators;
- cross variograms of indicators;
- conditional probabilities upwards: $P_h\ [\rightarrow \geq 10^3 \mid \rightarrow \geq 10^2]$, ...;
- conditional probabilities downwards: $P_h\ [\rightarrow < 10^2 \mid \rightarrow < 10^3]$, ...

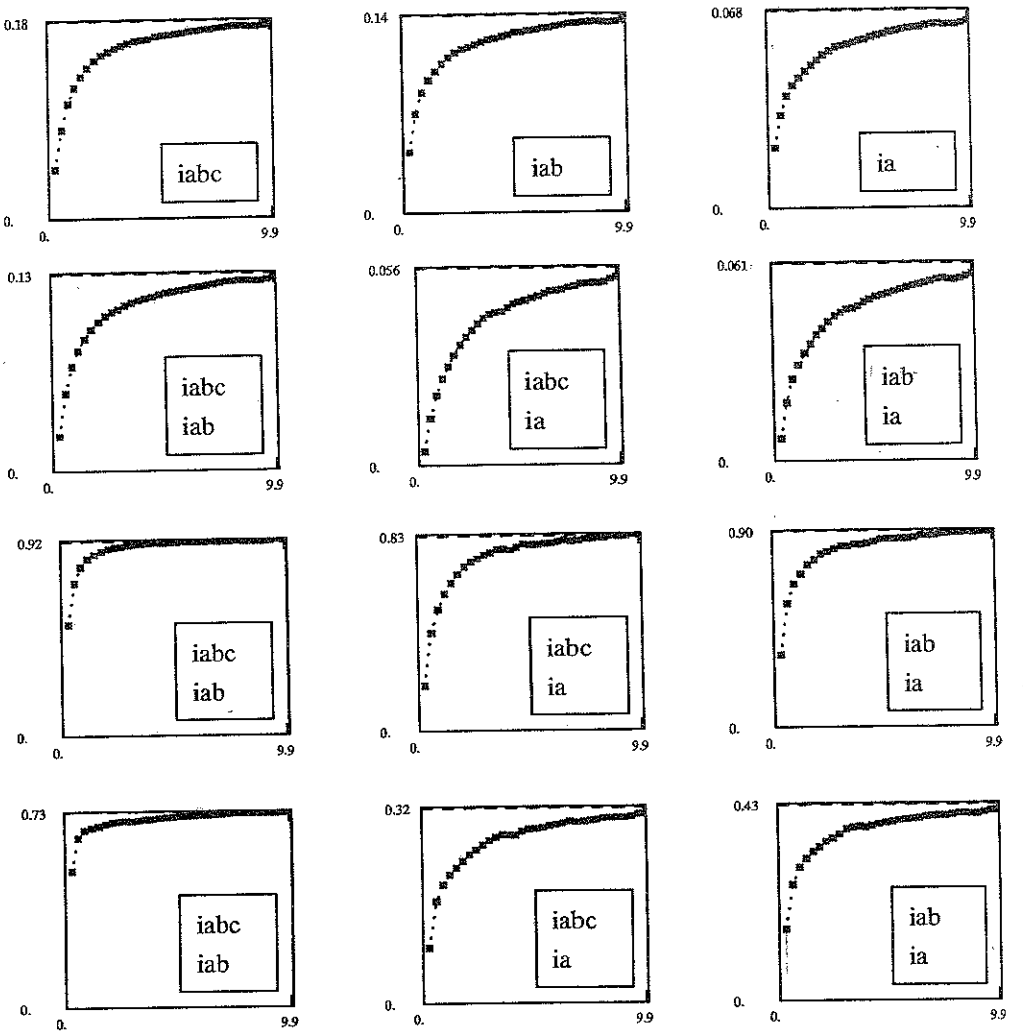

Figure 6 : cumulated rock facies: ABC, AB, A.
From top to bottom:
– variograms of indicators;
– cross variograms of indicators;
– conditional probabilities upwards: $P_h\ [\rightarrow AB \mid \rightarrow ABC]$, ...;
– conditional probabilities downwards: $P_h\ [ABC \rightarrow \mid AB \rightarrow]$, ...

Example 2

These rock facies data were provided by the Direction Production & Developpement Total. They correspond to increasing granulometries: shales (denoted D), shaly sandstones C, sandstones B, sandstones A, and appear in a fluvio-deltaic sedimentation with probabilities 77%, 6%, 10%, 7% (cumulated: 100%, 23%, 17%, 7%). The data are known along wells, along which the variograms of the cumulated indicators A, AB, ABC are calculated. They show beautiful structures (Figure 6).

The cross variogram AB–ABC is very close to the simple variograms of AB and ABC. This is because AB and ABC corresponds to close probabilities (17 and 23%). The facies C being little represented, the conditional probability to reach AB when entering ABC, or to leave ABC when leaving AB, increases rapidly with the distance. The cross variograms A–AB and A–ABC are more regular at the origin than the corresponding simple ones. The conditional probabilities increase regularly with the distance, showing well marked border effects within the domains AB and BCD. The facies B is a transition between A and CD. These data have been modelled by a diffusion-type model (thresholded gaussian model, see further).

4. USING THE RESULTS : THE CHOICE OF TECHNIQUES

We have seen the descriptive tools that are the variograms of indicators, simple and cross, as well as the conditional probabilities. Let us now look at the estimation or simulation, according the different types encountered in 2. (see Rivoirard 1990 for more details on estimation). Of course the indicators are not independent: if $Z(x) \geq 3$, then $Z(x) \geq 2$. The dependence between $I[Z(x) = 2]$ and $I[Z(x) = 3]$ also holds between neighboring points x and x+h, as generally Z(x+h) has less chance to be ≥ 3 if Z(x) is itself < 2.

The mosaic model with independent values

We have already seen how to build this model from a mosaic (random partition of the space). Let $\varrho(h)$ be the probability for two points h apart, x and x+h, to belong to the same compartment. If x and x+h belong to the same compartment, then Z(x) = Z(x+h). If not, Z(x) and Z(x+h) are independent. So, whatever the functions f and g, we have

$$E[\, f(Z(x))\, g(Z(x+h))] = E[f(Z(x))\, g(Z(x))]\, \varrho(h) + E[f(Z)]\, E[g(Z)]\, (1-\varrho(h))$$

hence a covariance

$$Cov\, [f\,(Y(x)),\, g\,(Y(x+h))] = Cov\, [f\,(Y(x)),\, g\,(Y(x))]\, \varrho\,(h)$$

which is proportional to $\varrho(h)$. Thus the simple (take $f=g$) or cross covariances between any two functions of Z(x) are proportional to $\varrho(h)$ (to $1-\varrho(h)$ for

variograms). In particular this is so for the indicators, which are intrinsically correlated (Matheron 1965).

The consequence for the estimation is the following. Cokriging the indicators is reduced to their separate kriging, with weights that are the same for all indicators. Moreover, as any function f of Z(x) (taking the values $f_1, f_2, \ldots$ when $Z(x) = 1, 2, \ldots$) can be written as a linear combination of the indicators

$$f[Z(x)] = f_1\, I(Z(x) = 1) + f_2\, I(Z(x) = 2) + \ldots$$

the resulting estimation of $f[Z(x)]$, i.e. its disjunctive kriging, is nothing but its kriging.

To sum up this is the model in which the separate kriging of indicators, as recommended by Journel (1982), finds its theoretical justification. Basically the reason for any indicator to be estimated separately comes from the elementary property: knowing $I(Z(x) = 2)$, no matter the exact value of Z(x) when one wants to estimate $I(Z(x + h) = 2)$.

The model with orthogonal indicator residuals

Its construction comes directly from the properties seen in section 2. Starting from independent RS A_1, A_2, A_3, A_4, we put $Z(x) = 1$ if $x \in A_1$; $Z(x) = 2$ if $x \notin A_1$ and $x \in A_2$; $Z(x) = 3$ if $x \notin A_1$, $x \notin A_2$ and $x \in A_3$; $Z(x) = 4$ if $x \notin A_1$, $x \notin A_2$ and $x \notin A_3$. The indicator

$$I(Z(x) \geq j + 1) = \prod_{1}^{j} I(x \notin A_i) = I(Z(x) \geq j) \quad I(x \notin A_j)$$

depends only on the A_i up to j. Geologically this process makes one think of successive and partial erosions. After each of them the eroded parts are filled in with a material, the value of which is smaller and smaller (other hierarchical values can of course be imagined).

For the estimation, the indicators can be factorized with the residuals of the regressions between successive indicators (hence the name): Rivoirard 1989. These residuals need only to be kriged to obtain the cokriging of the indicators and then the disjunctive kriging of any function of Z(x). In this hierarchical model, the estimation of $I(Z(x) \geq j)$ depends only on the indicators for cut−offs $\leq j$ (the cut−offs higher in the hierarchy).

The way to build the model is similar to the technique proposed by Alabert (1987) to sequentially simulate the indicators $I[Z(x) < z_k]$ corresponding to a thresholded RF Z(x). Alabert simulates each indicator using the values of this indicator, either known at the data points, or deduced at points from previously simulated indicators (if $Z(x) < 1$, then $Z(x) < 2$ for this point x). Although the order of the sequence is important

and may generate bias, Alabert has no direct criteria for choosing this order. As a matter of fact, a sequential simulation of indicators $I[Z(x) < z_k]$ in the order z_{i_1}, z_{i_2}, ..., z_{i_n} corresponds to a building from independent RS A_1, A_2, ... with

$$I(Z(x) \geq z_{i_p}) = I(Z(x) \geq z_{i_{p-1}}) \quad I(x \notin A_p)$$

or

$$I(Z(x) < z_{i_p}) = I(Z(x) < z_{i_{p-1}}) \quad I(x \notin A_p)$$

according to whether z_{i_p} is larger or smaller than $z_{i_{p-1}}$. It is an indicator residual model with a particular hierarchy. We expect, for $p < q$, to have a cross variogram between $I(Z(x) < z_{i_p})$ and $I(Z(x) < z_{i_q})$ proportional to the variogram of $I(Z(x) < z_{i_p})$. This gives a criterion to choose such a sequential processing of indicators.

Remark : Even with a continuous distribution, the indicator residual model is theoretically made of compartments with constant values. Then it is also a mosaic model (in the distinction made by Matheron 1989, page 309), but where the value of a compartment is not independent of its size, nor of the neighboring compartments.

The diffusion-type models

In probability, physics, or earth science, diffusion processes correspond to phenomena with gradual variations. There is a transition through neighboring values (but this may be rapid, in particular for some ranges of values). In these models the RS obtained by thresholding are closely linked. Whereas estimating an indicator uses only the values of this indicator at the data points in the mosaic model, and in the residual model it uses also the values of the hierarchically higher indicators, here it depends on the values for all indicators.

In the multigaussian model for instance, the most common diffusion model, the estimation of $I(Z(x) < z)$ depends on the kriging of the variable Z(x) itself, then on all the indicators at the data points. The fact that we know the multivariate and then the conditional distributions, and that the residuals of the regressions are independent from the conditioning values, makes this model suited for estimation, simulation, and above all for conditional simulation. As the variable under study is rarely normal, this is usually considered, for this model, as being transformed from a normal variable. Thresholding a gaussian RF is a particular transformation.

Besides the gaussian model, there are other diffusion models, based on a given statistical distribution (e.g. gamma, Hu 1988), or built on empirical distributions (Matheron 1984, Lajaunie et Lantuéjoul 1989). These models are suited, with the flexibility of an additional transformation, to various distributions (discrete, or with large peaks). As multivariate and conditional distributions are not workable, the estimation is performed through the disjunctive kriging technique (which needs only bivariate hypotheses and is also available in the gaussian case).

Remark: It is possible to imagine models with border effects which are not strictly diffusive (transition by neighboring values). Example: a mosaic where the value is the average, within each compartment, of a given RF.

5. CONCLUSION

Looking at the cross as well as simple variograms of indicators is a convenient way to study the arrangement between the RS obtained by thresholding a RF with stationary and symmetric bivariate distributions. The conditional probabilities that can be deduced can help choosing a mosaic model without border effects, an indicator residual model (no border effects upwards or downwards), or a diffusion-type model (border effects upwards and downwards).

These are simple models, and more sophisticated ones may be needed sometimes. For instance a geological environment resulting from combined diffusion and erosion (mixed diffusion–residual models). Or a sedimentary process where the granulometry gradually decreases, but sometimes increases by jumps (there is no symmetry in h vertically).

References

Alabert, F. (1987) "Stochastic imaging of spatial distributions using hard and soft information", M. Sc. Thesis, Stanford University.
Hu, Lin–Ying (1988) "Mise en oeuvre du modèle gamma pour l'estimation de distributions spatiales", Thèse de Doctorat de l'Ecole des Mines de Paris.
Journel, A. (1982) "The indicator approach to estimation of spatial distributions", Proceedings of the 17th APCOM, T. B. Johnson and R. J. Barnes Editors, New York, pp. 793–806.
Lajaunie, Ch. et Lantuéjoul, Ch. (1989) "Setting up the general methodology for discrete isofactorial models", Proceedings, 3rd International Geostatistics Congress, Avignon, 5–9 Sept. 1988 : "Geostatistics", M. Armstrong Editor, Kluwer Academic Publ., Dordrecht, Holland.
Matheron, G. (1965) Les variables régionalisées et leur estimation, Thèse de Doctorat d'Etat, Ed. Masson, Paris.
Matheron, G. (1976) "A simple substitute for conditional expectation: the disjunctive kriging", Proceedings, NATO ASI : "Advanced Geostatistics in the Mining Industry", Rome, Oct. 1975, pp. 221–236.
Matheron, G. (1982) "La destructuration des hautes teneurs et le krigeage des indicatrices", Centre de Géostatistique, Ecole des Mines de Paris.
Matheron, G. (1984) "Une méthodologie générale pour les modèles isofactoriels discrets", Sciences de la Terre n°21, pp. 1–78.
Matheron, G. (1989) "Two classes of isofactorial models", Proceedings, 3rd International Geostatistics Congress, Avignon, 5–9 Sept. 1988 : "Geostatistics", M. Armstrong Editor, Kluwer Academic Publ., Dordrecht, Holland.
Petitgas, P. (1991) "Contributions géostatistiques à la biologie des pêches maritimes", Thèse de Géostatistique, Ecole des Mines de Paris – 211p.
Rivoirard, J. (1989) "Models with orthogonal indicator residuals", Proceedings, 3rd International Geostatistics Congress, Avignon, 5–9 Sept. 1988 : "Geostatistics", M. Armstrong Editor, Kluwer Academic Publ., Dordrecht, Holland.
Rivoirard, J. (1990) "Introduction to disjunctive kriging and nonlinear geostatistics", Centre de Géostatistique, Ecole des Mines de Paris, 89 p. : bibliogr.