TRANSITIVE GEOSTATISTICS AND STATISTICS PER INDIVIDUAL: A RELEVANT FRAMEWORK FOR ASSESSING RESOURCES WITH DIFFUSE LIMITS

Nicolas BEZ [*]

ABSTRACT

When assessing marine resources, inferring spatial models has to be performed from a unique realisation. The situations with repetitive surveys that can be considered as repetition of the same regionalized variable are (obviously) rare. In intrinsic geostatistics, this question is usually solved by a couple of key assumptions namely stationarity and ergodicity. Unfortunately, these assumptions and their consequences are often too strong with regards to the reality of fish survey data. It is especially unrealistic to assume that the spatial structure is independent from the geometry of field.

Transitive geostatistics has proven to be an operational alternative to intrinsic geostatistics and was the seed for the development of a framework called "statistics per individual". This article presents the rationale of the approach and sketches the main tools developed during the past few years with practical illustrations. Statistics per individual have the advantage to be simple and thus more robust than intrinsic approaches (robust in the sense that the properties of the estimator are based on fewer and checkable assumptions). On the one hand, "statistics per individual" allow for summarizing and describing series of spatial distributions into few quantitative features. On the other hand, as developed in the first ages of geostatistics, they allow for estimating global abundance with estimation variance thanks to the (transitive) covariogram and for interpolating between observations (transitive kriging). The price to pay for the simplicity of the method is that it leads to fewer possible applications than the intrinsic geostatistical approaches and that, as a design based approach, it is constrained to some specific sampling schemes (e.g. the regular, stratified regular or point process survey strategies).

Keywords : Single realisation, transitive geostatistics, covariogram, design-based estimation variance.

[*] IRD, Sète, France. nicolas.bez@ird.fr

RÉSUMÉ

En écologie halieutique, l'inférence des modèles spatiaux se fait le plus souvent à partir d'une réalisation unique du phénomène. En effet, les cas où plusieurs campagnes d'observations pourraient être considérées comme des répétitions d'un même phénomène sont rares voire inexistants. Cette carence de répétition est contournée en géostatistique intrinsèque par des hypothèses d'ergodicité et de stationnarité portant sur le modèle. Cependant, dans la pratique, ces hypothèses clefs apparaissent souvent trop fortes par rapport aux caractéristiques réelles des données. En particulier, il parait non fondé de supposer que la structure spatiale des individus d'une population donnée est indépendante de la position de ces individus par rapport aux frontières ou au cœur de la zone de présence de la population.

La géostatistique transitive est dorénavant reconnue, en écologie halieutique, comme une alternative opérationnelle à l'approche intrinsèque et a été à l'origine du développement de statistiques dites «par individus ». L'objectif de cet article est de rappeler les fondements théoriques de cette approche et de donner un aperçu illustré des développements méthodologiques réalisés ces dernières années. Les statistiques par individus présentent l'avantage d'être simples et robustes car fondées sur des hypothèses en faible nombre et falsifiables. Ces statistiques permettent d'une part, de résumer des séries de distributions spatiales à l'aide de quelques descripteurs et, d'autre part, de fournir, grâce au covariogramme, des variances d'estimations globales ou des cartes d'interpolation (krigeage transitif). En contrepartie de sa simplicité, la démarche transitive offre un spectre d'applications du modèle plus restreint qu'en géostatistique intrinsèque et une utilisation restreinte à certains types d'échantillonnage (e.g. régulier, aléatoire stratifié ou processus ponctuel).

Mots-clés : Géostatistique transitive, covariogramme, réalisation unique, variance d'estimation.

1. Introduction

Monitoring marine resources, and more specifically fish populations, puts forward two major estimation questions for which an appropriate methodological choice is needed: the estimation of total abundance and the definition and estimation of relevant summary statistics of key spatial features of these fish populations. Contrary to mining from which geostatistics originated, fish stock estimations cannot be confronted to field truth and one must be very cautious not to overpass modelling thresholds and parsimony principles. This is all the more required that statistical inference is to be performed from one single realisation of the study variable. The situations where repetitive surveys can be considered as repetition of the same regionalized variable are rare. Following Matheron recommendation on parsimony (1989), this should lead practionners to choose models based on as few as possible and as tractable as possible assumptions because this reduces the possibilities to observe discrepancies between the characteristics of the data and the assumptions on which the estimator is based (robustness). In fact, this is also the choice towards which the confrontation between fish survey data and random function models led part of the fisheries community.

Geostatistics often follows the so called *intrinsic* approach using random functions. In this framework, expectations (expected value, variance, variogram, etc) are theoretically considered over all possible realisations of the random function. Although this is possible in theory, this does not hold in practice when one single realisation is available. This problem is usually solved by a couple of key assumptions namely stationarity and ergodicity. Unfortunately, these assumptions and their consequences, are often too strong and are most likely not supported by the reality of fish survey data. It may well be unrealistic to assume that fish spatial distribution is independent from the geometry of its field and that large values can occur anywhere in the field. In addition the expectations and their practical translations in terms of averages are strongly influenced by the zero samples observed outside the area of presence of fish. While extending sampling beyond the area of presence of fish ought to insure that the whole targetted population has been sampled, the influence of zero density values is undesirable. This question extends to the influence of the very numerous low densities that are observed on large areas at the borders of fish populations. All together, this raises the question of the subjective delineation of a population field, and indicates how carefully we should use statistics that are affected by zeroes and more generally by low concentrations. Although some authors are suggesting more robust estimators for the variogram (Cressie 1991), the method itself, i.e. the intrinsic geostatistical approach, also might be regarded as based on too strong hypotheses in the particular case of fish data.

An alternative has been considered in fish applications with the objectives to allow pragmatic answers to the two questions of interest (global estimation and summary spatial statistics) and to cope with data characteristics. This alternative derives from the *transitive* approach, a design-based method developed by Matheron (1971) which requires fewer and more easily controllable hypotheses than the intrinsic one. Despite its simplicity, the transitive approach, first designed for regular samplings, has not been widely spred in the fisheries community. However, this method has recently been shown to be appropriate for the treatment of spatial data sets with numerous zeroes (Bez and Rivoirard 2001) and for global estimation variance in case of random stratified samplings (Bez *et al.* 1995; Bez 2002). This led the fisheries community to more widely endorse it. The first objective of the paper is to present in details the theory and the practical implementation of the transitive approach.

Scientific surveys objectives are not only to estimate stock abundances at a given time of the year. Yearly surveys feed time series whose trends (positive or negative) are of great influence on scientific diagnostics and management decisions, and help monitor any changes in the spatial distributions of various components of a given ecosystem. For this latter reason, time series of surveys need to be summarised with appropriate tools. Given the characteristics of fish data above mentioned, criteria for evaluating and choosing summary statistics must be revisited with considerations on field dependency, support effect, and sensitivity to the zero. Derived from the transitive approach, statistics per individual have been developed and applied to several survey series. The second objective of the paper is to present the general principles of this family of summary statisics and its practical implementation.

This paper is a review of existing pieces of work for which references are given. Most of the figures used to illustrate the ideas described in the present review are borrowed to those references. The annex corresponds to new material.

2. Transitive approach

2.1. Theory

2.1.1. Regionalised variable, covariogram and, standardised covariogram

A notation inspired from one dimension conventions is chosen for simplicity: x represents a point in space. Contrary to intrinsic geostatistics where the fish density is considered as a realisation of a random function, in transitive geostatistics, the fish density, denoted $z(x)$, is considered as a deterministic function and is called a *regionalised variable*. This regionalised variable is often expressed in practice as the number of individuals per unit surface area (e.g., ind$\cdot$m^{-2}). The *covariogram* (Matheron 1971):

$$g(h) = \int z(x)z(x+h)\ dx$$

is the convolution product of the fish densities (h being the distance). In 2D case studies, it is expressed as the square number of individuals per unit surface area (e.g., ind$^2\cdot$m^{-2}). The total fish abundance Q is the first quantity of interest:

$$Q = \int z(x)dx$$

Using the density relative to the total abundance leads to the *standardised covariogram* (Bez and Rivoirard 2001).

$$\tilde{g}(h) = \frac{g(h)}{Q^2} = \frac{\displaystyle\int z(x)z(x+h)dx}{\left(\displaystyle\int z(x)dx\right)^2}$$

The standardised covariogram is homogeneous to the inverse of a surface area. It globally decreases from its maximum value $\tilde{g}(0)$ taken as an index of aggregation (as it will be justified in section 3) to 0 for long distance (Figure 1). The distance at which the covariogram reaches zero (strictly or approximately for exponentially decreasing covariograms) is called the *range*. It quantifies the maximal diameter of the population in the particular direction of concern.

The standardised covariogram is analogous to the correlogram used for random functions replacing space integrals by expected values. As a matter of fact, for centered random function, the correlogram is:

$$\rho(h) = \frac{E\big(Z(x)Z(x+h)\big)}{E\big(Z(x)^2\big)}.$$

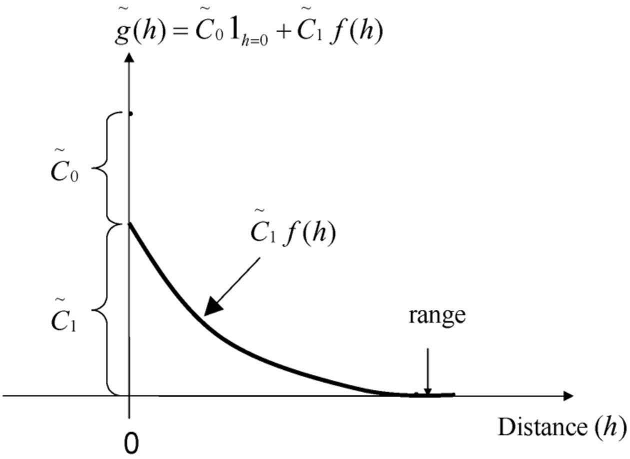


FIG 1. — Definition of a standardised covariogram model $\tilde{g}(h)$: nugget effect ($1_{h=0}$) with sill $\tilde{C}_0$ and continuous part ($f(h)$) with sill $\tilde{C}_1$ and range a_1.

Contrary to the correlogram, the range of a covariogram is a geometrical property of the field. As the latter is never circular in real studies, covariograms of fish densities will generally be anisotropic, although it is sometimes difficult to model in practice.

The behaviour of the covariogram near the origin is related to the spatial continuity of the fish density and is often reduced to a discontinuity. One usually considers two causes for this discontinuity: the small-scale variability of the fish distribution itself and measurement errors. The standardized covariogram for $z(x)$ can thus be written (Figure 1 and annex A) as

$$\tilde{g}(h) = \tilde{C}_0 \; 1_{h=0} + \tilde{C}_1 \; f(h)$$

where $\tilde{C}_0 \; 1_{h=0}$ corresponds to the discontinuous part (nugget effect) and $\tilde{C}_1 \; f(h)$ to the continuous part.

2.1.2. Global estimation for strictly regular sampling

Following the 1D notation, the origin of the sampling grid is denoted x_0, and the grid mesh interval, s. A sample point is then located at $x_0 + ks$, *i.e.*, the origin plus an integer multiple of the grid mesh interval. Several actual fish surveys do follow a regular sampling: e.g. most of acoustic surveys, the ICES triennial egg surveys (ICES 2003), cod and haddock survey in the Barents Sea (Jakobsen *et al.* 1997), halibut surveys in the North Pacific (Hoag *et al.* 1980), etc. The total fish abundance is estimated by:

$$Q^*(x_0) = s \sum_k z(x_0 + ks)$$

a deterministic quantity. Assuming that x_0 is the outcome of a random uniform variable over a grid cell, the estimator, now denoted $Q^*(X_0)$, becomes a random variable. Its bias is zero due to the uniform distribution of X_0:

$$\begin{aligned} E[Q^*(X_0)] &= \int_s Q^*(x)\ \frac{dx}{s} \\ &= \int_s s\sum_k z(x+ks)\ \frac{dx}{s} \\ &= \sum_k \int_s z(x+ks)dx \\ &= \int z(x)dx \\ &= Q \end{aligned}$$

After Matheron (1971), the estimated coefficient of variation is:

$$CV_E = \frac{\sqrt{\mathrm{var}(Q^*(X_0))}}{Q^*} = \sqrt{s\sum_k \tilde{g}(ks) - \int \tilde{g}(h)\ dh}$$

i.e. the square root difference between the exact integral of the standardised covariogram model and its discrete approximation at the grid spacing level. The latter quantity is never negative since

$$E[Q^*(X_0)^2] = s\sum_k g(ks) \quad \text{and} \quad E[Q^2] = Q^2 = \int g(h)\ dh$$

and since the covariogram function is positive definite (Matheron 1965, p 74). Directly from this definition, also comes that (i) the smaller the grid mesh interval, the smaller the estimation CV, (ii) the more irregular the spatial distribution, i.e., the larger the nugget effect, the larger the estimated CV (Figure 2). When a significant nugget effect exists, it explains nearly all the estimated CV which can then be approximated by (Bez 2002):

$$CV_E \approx \sqrt{s \cdot \tilde{C}_0}$$

The unbiasness of the estimators refers to all the estimations generated by all the possible grid origins. To avoid systematic errors over a series of annual surveys, one should change the origin of the grid from year to year. One alternative is to use a random stratified sampling.

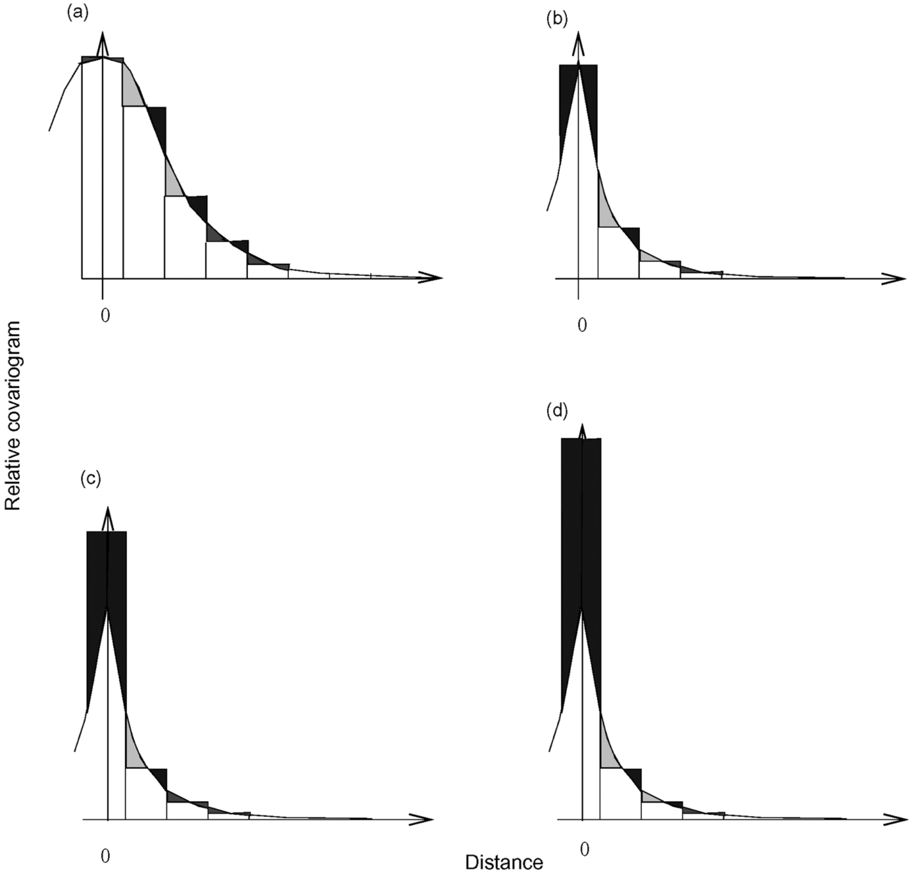


FIG 2. — Global estimation coefficient of variation and behavior of the standardised covariogram near the origin. The estimation CV corresponds graphically to the difference between the black and the grey areas. It increases when the spatial local heterogeneity of the fish density, i.e., C_0, increases : (a) No nugget effect and parabolic behavior, very small estimation CV, (b) No nugget effect and linear behavior, small estimation CV, (c) Reasonable nugget effect and linear behavior, large estimation CV, (d) Large nugget effect and linear behavior, very large estimation CV.

2.1.3. Global estimation for random stratified sampling

In a regular stratified sampling, each sample point X_k, for $k \in [1, N]$ where N is the number of samples, is random uniform in its grid cell s_k. Many surveys use this sampling design: e.g. the International Bottom Trawl Surveys in the North Sea (ICES 1997), the snow crab survey in Canada (Conan *et al.* 1988), and the Moroccan cephalopod surveys (Faraj and Bez, submitted). The estimator of the abundance is a random variable, function of a set of i.i.d. random variables:

$$Q^*(\{X_k\}) = s \sum_k z(X_k)$$

As the X_k are random uniform over s_k, the estimator is unbiased:

$$\begin{aligned} E\big[Q^*(\{X_k\})\big] &= s \sum_k E[z(X_k)] \\ &= s \sum_k \int_{s_k} z(x)\, \frac{dx}{s} \\ &= Q \end{aligned}$$

Due to the independence of the X_k, Matheron (1989) showed that the estimation CV

$$CV_E = \sqrt{s\left(\tilde{g}(0) - \overline{\tilde{g}(s)}\right)}$$

only depends on the behaviour of the covariogram at distances smaller than the grid mesh size. $\overline{\tilde{g}(s)}$ is the mean value of the standardised covariogram between two points x and y located independently in a grid cell s :

$$\overline{\tilde{g}(s)} = \frac{1}{s^2} \int_s \int_s \tilde{g}(x-y) dx\ dy$$

2.1.4. Transitive kriging

Transitive kriging is a linear interpolation procedure where each unknown value $z(x)$ is estimated by a weighted average of the neighbouring sample values:

$$z^*(x) = \sum_i \lambda_i z(x + h_i)$$

The weights are chosen so that if the kriging configuration made of the points x and $x + h_i$ were translated in all possible locations, the sum of squared errors between true and estimated values would be minimised. This sum can be written as a function of the covariogram (Bez *et al.* 1997):

$$\int \big(z(x) - z^*(x)\big)^2 dx = g(0) - 2\sum_i \lambda_i g(h_i) + \sum_i \sum_j \lambda_i \lambda_j g(h_i - h_j)$$

The covariogram model can then be used to choose the weights that minimise this sum (Matheron 1989, 119-122 pp). Transitive kriging is thus, in an algorithm perspective, similar to stationary kriging with mean equal to 0. The unbiasness of the interpolation procedure is obtained by assuming that the total abundance is recovered when running the kriging configuration over the entire field. This provides a supplementary condition on the weights similar to ordinary kriging: $\sum_i \lambda_i = 1$.

2.2. Practical implementation

2.2.1. Covariogram estimation

In case of regular grids, which are preferred, the standardised covariogram is estimated for any distance and direction in the grid. Most logical directions correspond to the two main directions of the grid where sampling density is highest, but all the diagonal directions can be looked at. The standardised covariogram for a distance equal to a multiple l number of grid intervals is estimated by the sum of the products of pairs of densities separated by l grid nodes (Matheron 1971):

$$\tilde{g}^*(ls) = \frac{\sum_k z(x_0 + ks) \cdot z(x_0 + ks + ls)}{s \cdot \left(\sum_k z(x_0 + ks) \right)^2}$$

This assumes that the fish density is zero beyond the sampling area.

For random stratified sampling, things are more complicated, because when the first point of a pair sweeps its grid cell, the other one does not. A solution suggested by Bez *et al.* (1995) is based on the use of the surfaces of influence of sample points (Voronoï polygons). Given the links between the covariogram and the covariance, they suggested the following weighted procedures:

$$\tilde{g}^*(h) = \frac{1}{\left(\sum_k z_k S_k \right)^2} \left(\sum_k z_k S_k \cdot \frac{1}{2} \left(\frac{\sum_{l \approx k+h} z_l S_l}{\sum_{l \approx k+h} S_l} + \frac{\sum_{l \approx k-h} z_l S_l}{\sum_{l \approx k-h} S_l} \right) \right)$$

2.2.2. Numerical layout and units

When computing covariogram, fish density and grid mesh surface area should be expressed with compatible units. However there are cases where the fish density gets units that do not simplify with those of a surface area (*i.e.*, kg·h^{-1}). A standardised covariogram is expressed as the inverse of a surface area whatever the units of the fish density which makes it more practical:

$$\text{units of } \left(\frac{g(h)}{Q^2} \right) = \frac{(\text{units of } z)^2 \cdot (\text{units of } s)}{((\text{units of } s) \cdot (\text{units of } z))^2} = \frac{1}{(\text{units of } s)}$$

2.2.3. Reference system

Computation of distances between points often requires the projection of data points in a Euclidean reference system. Regular sampling might be no longer regular after projection due to some cosine operations.

2.2.4. Covariogram fitting

The covariogram models must be symmetrical, bounded, positive or null when the fish density is positive or null, and must be positive definite to ensure that the estimation variances are always positive (or null). The range is either the distance beyond which the function is null (model with finite support, e.g. spherical model) or below 5 % of the value at the origin (model with infinite support, e.g. exponential model). Given that fish distributions do not extend to infinity, models with finite true range are recommended.

In intrinsic geostatistics, anisotropies are generally zonal or geometrical (Chilès and Delfiner 1999). Models with zonal anisotropy cannot be used for covariogram as they become negative after some distance and for some directions. The type of anisotropy that is left is thus the geometrical anisotropy based on two parameters: the direction in which the range is given, and the coefficient by which the range has to be divided to get the range in the orthogonal direction; the range for an intermediate direction corresponding to the radius of the ellipse defined by the maximum and minimum range in the appropriate directions. When a covariogram is made of several functions, the anisotropy of each function can be different from the other one.

Classical hole effect models (e.g. Bessel model) cannot be used either as they are alternatively positive and negative around their sill, *i.e.* around 0. However, fish distributions often exhibits few large aggregations and hole effect models are needed to endorse these situations. By construction, convolution products are positive definite functions. This is used here to build a hole effect model based on the convolution of a regionalised variable equal to two bigaussian like aggregations (see annex B).

Fitting consists in estimating the model parameters by ad hoc means (manual, least squares, etc).

3. Statistics per individual and summary statistics

3.1. Theory

3.1.1. Random individuals

Individuals (fish, larvae or eggs) at a given location are all the more numerous when the density at this location is larger. So if we consider an individual I taken at random in the whole population, $I \in [1, Q]$, the probability density function (p.d.f.) of its location x_I is:

$$\tilde{z}(x) = \frac{z(x)}{Q}.$$

Knowing the p.d.f. of the random variable x_I, one can derive several statistics. The mean location of a random individual

$$\overline{x_I} = \int x \cdot \tilde{z}(x)dx = \frac{\displaystyle\int x \cdot z(x)\ dx}{\displaystyle\int z(x)\ dx}$$

is the center of mass of the whole population and is usefully expressed in geographical coordinates (e.g. longitude and latitude). The trace of the variance-covariance matrix of x_I corresponds to the inertia of the population and measures its spatial dispersion:

$$tr\big(\text{var}(x_I)\big) = \frac{\displaystyle\int \|x - \overline{x_I}\|^2 \cdot z(x)dx}{\displaystyle\int z(x)dx}$$

In 2D cases, inertia is homogenous to a surface area (e.g. square nautical miles). In 2D cases, a weighted principal component analysis (PCA) of the inertia provides two orthogonal factors equal to linear combination of the two input coordinates, i.e., the spatial directions that explain respectively most and least of the fish spatial distribution. The ratio of the smallest axis over the largest axis is a proxy for quantifying isotropy. When the individuals are distributed in all directions with no preference, the two axes get the same length and the index is 1. On the contrary when individuals are organised along a straight line, the index is 0.

Being a convolution product, the standardised covariogram represents the p.d.f. of the random vector $H = x_I - x_{I'}$ where I and I' denote two individuals taken at random and independently in the population. As a function of H, the expected value of square distance between two random individuals $||H||^2$ can be written as follows:

$$E\big(||H||^2\big) = \int ||h||^2 \cdot \tilde{g}(h)dh = 2 \times tr(\text{var}(x_I)).$$

So the inertia of the population can also be defined as the average (semi square) distance between two individuals taken at random in the field:

$$tr\big(\text{var}(x_I)\big) = \frac{1}{2}\int ||h||^2 \cdot \tilde{g}(h)dh.$$

Meanwhile, $\tilde{g}(0)$ is the density of probability for two random individuals to be at the same location, that is, aggregated in the same location, hence the use of $\tilde{g}(0)$ as an aggregation index (Bez and Rivoirard 2001). In practice, densities are measured on a non punctual support and aggregation must be considered at that support.

3.1.2. Global and local indices of collocation (GIC and LIC)

Let us consider two populations: $z_1(x)$ and $z_2(x)$. When coefficients of correlation (and further cross variograms or cross covariances) are concerned, the domain on which they are computed cannot be relevant simultaneously for the two species. Imagine that species 1 has a wider distribution than population 2. Taking the area of presence of population 1 for the computation introduces a set of undesirable zero values when computing population 2 statistics. On the contrary, the area of presence of population 2 does not allow one to take all the positive samples of population 1 for the computation. Using statistics per individual allows turning into a domain free version analysis of variance type of criteria and coefficient of correlation (Bez and Rivoirard 2000a).

Let us consider ΔCG the distance between the centers of gravity of populations 1 and 2 and I_1 and I_2 their inertia. Comparing the (square) mean distance between a random individual of population 1 and a random individual of population 2 ($||\Delta CG||^2$), and that between two individuals taken at random and independently from any of the two populations ($||\Delta CG||^2 + I_1 + I_2$), leads to the global index of collocation (GIC) .

$$GIC = 1 - \frac{||\Delta CG||^2}{||\Delta CG||^2 + I_1 + I_2}.$$

This index ranges between 0, in the extreme case where each population is concentrated on a single but different location (inertia equal 0), and 1, when the two centres of gravity are confounded. By way of illustration, GICs were computed for simplistic situations (isotropic Gaussian fish density with fish density being set to zero for densities below the quantile 5%). From this, a GIC between 0.6 and 0.8 is considered as a low value and 0.8 a threshold for good and poor collocations (Figure 3).

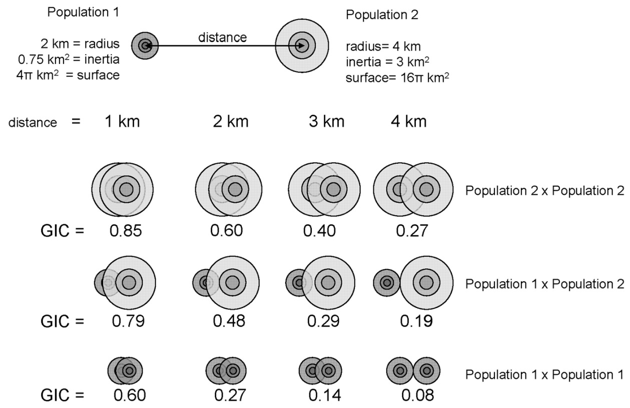

Population 1
2 km = radius
0.75 km² = inertia
4π km² = surface
distance
Population 2
radius= 4 km
inertia = 3 km²
surface= 16π km²
distance = 1 km 2 km 3 km 4 km
Population 2 x Population 2
GIC = 0.85 0.60 0.40 0.27
Population 1 x Population 2
GIC = 0.79 0.48 0.29 0.19
Population 1 x Population 1
GIC = 0.60 0.27 0.14 0.08

FIG 3. — Global Indices of Collocations (GICs) for simplistic situations. Fish distributions are considered to be isotropic and distributed according to a Gaussian distribution with fish density being set to zero for densities below the quantile 5%. Two types of fish populations are concerned (patchy or spred). Several possible distances between the centers of mass are concerned.

Replacing averages by sums in the coefficient of correlation leads to the local index of collocation (LIC)

$$LIC = \frac{\int z_1(x)z_2(x)dx}{\sqrt{\int z_1^2(x)dx}\sqrt{\int z_2^2(x)dx}}$$

As it is based on sums, this index is not impacted by the zeroes contrary to the usual coefficient of correlation. When $z_1(x)$ and $z_2(x)$ are all identical or proportional, this index is 1. On the other hand, when no individuals of the two species are found simultaneously in any sample, it is equal to 0. Hence this index measures local collocation between the populations. This index is not affected by a permuation of samples values in space.

3.1.3. Crossing fish density and environmental parameters

We now consider an environmental parameter, e.g. water temperature, denoted $t(x)$. The value of this regionalised variable at the location of a random individual is a random variable, $t(x_I)$, function of x_I, called the temperature per individual. By definition, the expected value and the variance (or inertia) of the temperature per individual are

$$E[t(x_I)] = \bar{t}_I = \int \tilde{z}(x)t(x)dx$$
$$\mathrm{var}[t(x_I)] = \int \tilde{z}(x)\big(t(x) - \bar{t}_I\big)^2 dx$$

Similarly to the use of toroidal shifts to test for associations between point processes, Bez and Rivoirard (2000b) developed the concept of inertiogram to test for the association between a fish distribution with fuzzy, limited, and unknown geographical extension and an environmental variable with unlimited field. The very nature of the fish distributions is aggregative while environmental variables are spatially smooth. A large part of the populations concentrate themselve in narrow intervals of temperature values exhibiting a mode in their relationship while they are independent. The inertiogram is defined as the graph of the inertia of the parameter when the fish distribution is translated in all possible directions and distances h

$$E_{t_I}(h) = \int \tilde{z}(x)t(x+h)dx$$
$$I_{t_I}(h) = \int \tilde{z}(x)\big(t(x+h) - E_{t_I}(h)\big)^2 dx$$

A minimum in the inertiogram for h=0 means that the inertia increases when the population is translated and suggests that the match between the fish density and the parameter was largest for the actual situation. Inertia, and then inertiogram, can be developed for multi-parameters situations.

3.2. Practical layout

3.2.1. Estimations

For regular samplings, the estimators are straightforward:

$$\overline{x_I}^* = \frac{\sum_k x_k z_k}{\sum_k z_k} \qquad tr(\mathrm{var}(x_I))^* = \frac{\sum_k (x_k - \overline{x_I}^*)^2 z_k}{\sum_k z_k}$$

$$\overline{t(x_I)}^* = \frac{\sum_k t_k z_k}{\sum_k z_k} \qquad \mathrm{var}(t(x_I))^* = \frac{\sum_k (t_k - \overline{t(x_I)}^*)^2 z_k}{\sum_k z_k}$$

The zero values sampled (or assumed) on the grid outside the area of presence of fish do not contribute to these statistics. If the origin of the grid is random uniform, these estimators are unbiased.

For irregular sampling designs, weighting by the areas of influence is required:

$$\overline{x_I}^* = \frac{\sum_k x_k z_k S_k}{\sum_k S_k} \qquad tr(\mathrm{var}(x_I))^* = \frac{\sum_k (x_k - \overline{x_I}^*)^2 z_k S_k}{\sum_k z_k S_k}$$

$$\overline{t(x_I)}^* = \frac{\sum_k t_k z_k S_k}{\sum_k z_k S_k} \qquad \mathrm{var}(t(x_I))^* = \frac{\sum_k (t_k - \overline{t(x_I)}^*)^2 z_k S_k}{\sum_k z_k S_k}$$

3.2.2. Reference system

Distances between points must be computed in a Euclidean reference system. For the center of gravity and the ellisposoid of inertia whose final objective is often to be represented on a geographical map, back transformation is required. However, while the two factors of the PCA are orthogonal in the Euclidean system, this is no longer true after backtransformation.

4. Discussion

4.1. Summarizing series of distributions

Together with the transitive approach that allows estimating fish abundances with estimation variances and interpolating fish densities, statistics per individual lead to efficient summary statistics for summarizing series of surveys.

For instance, in the case of the cephalopod (*Octopus vulgaris*) off Morocco, trawl surveys are regularly performed twice a year by the National Institute of Fisheries Research (INRH). Since 1998, a random stratified survey is used with a 11 nautical miles by 11 nautical miles grid cell (Figure 4) targeting both juveniles and mature females. Based on the use of the above mentioned statistics per individual Faraj and Bez (submitted) indicated that juveniles are more coastal (center of gravity), less spatially dispersed (inertia), more anisotropically distributed (index of isotropy), and more patchy (index of aggregation) than the mature female. Transitive krigings allowed the authors getting a series of spatial distributions.

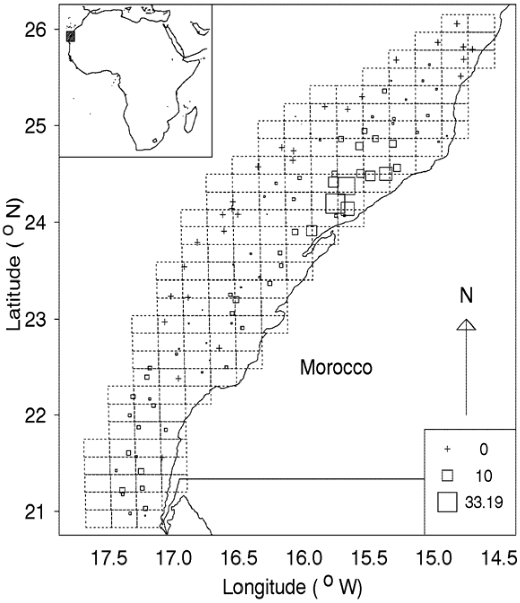


FIG 4. — Cephalopod (*Octopus vulgaris*) trawl surveys. Representation of the stratified random sampling: location of the trawl hauls at random in 11 n.mi. x 11 n.mi. cells. Courtesy of National Institute of Fisheries Research (INRH) – Morroco.

4.2. Support effect

Compared to the survey domain, the support on which fish density is measured is generally small enough to be considered as a point. But support may change from survey to survey and needs to be considered. Let v be the support, $v(x)$ this support when centered at point x, and $|v|$ the absolute size of the support. The fish density on support v is the following moving average:

$$z_v(x) = \frac{1}{|v|}\int_{v(x)} z(x+h)dh$$

The regularisation affects the covariogram. In particular, the value of the standardised covariogram at the origin, and so the index of aggregation, decrease with regularisation:

$$\tilde{g}_v(0) = \frac{\displaystyle\int z_v(x^2)dx}{Q^2} \leqslant \tilde{g}(0) = \frac{\displaystyle\int z(x)^2dx}{Q^2}$$

This is why the same support should be used when comparing the spatial properties of two populations.

4.3. Design-based versus model-based geostatistical approaches

Within a given probabilistic framework, the quality of an estimator is quantified by its bias, its convergence and its precision. However, the sole use of these three quality parameters could be misleading when choosing between estimators that are not based on the same assumptions. In particular, when choosing between a design-based and a model-based estimator, one should consider the number of hypotheses of the respective approaches and our ability the control their adequacy to field data in practice.

In the transitive approach, which is a design-based approach, the major assumption concerns the randomness of either the origin of the sampling grid, or the location of sampling points in the grid cells. Such assumptions are easy to control in practice. In the classical model-based geostatistical approach the stochastic part of the model and thus the constituent hypotheses concern the fish density itself considered as a realisation of a random function. Assumptions concern the stationarity of some aspects of the random process (e.g., the expected value and the variance for simple cases) and are much more difficult to control in practice.

4.4. Area-based versus area-free approaches

All the samples, including the zero data, are considered. Zero data do not influence the results because the method refers to spatial integrals and because sums are unaffected by the addition of zero data. The method does not require the delineation of the transition zone between inside and outside the population. This explains the term transitive and makes the technique area-free. Still, the zeroes are not ignored and represent crucial information. Their presence tells us whether or not the whole population has been sampled and, thus, allows for the meaningful use of the method. In practice zeroes must have been observed or assumptions are to be made for unsampled (generally assumed empty).

4.5. Infra versus supra support spatial statistics

Most spatial summary statistics are area-based and are relevant or meaningful in a given predefined geographical field. While they originated from the statistical framework (e.g. clumping index, Lloyd's index of patchiness, overdispersed index), they quantify what happens at scale smaller than the support of the information (typically the swept area in trawl data) and are not affected by permutation of the sample values in space. This remains true for some of the statistics per individual like the index of aggregation or the local index of collocation.

The rest of them, i.e., the center of gravity, the inertia, the global index of collocation, are affected by permutation of the sample values in space but are unaffected by spatial re-allocation of fish at scale smaller than the support.

The former are infra support statistics. The latter are supra support ones.

5. Conclusions

Statistics per individual benefit from the properties of transitive approach: they have a strong descriptive power, insensitive to zero densities, weakly sensitive to outliers, most suited for regular samplings but also operational for random stratified ones.

The transitive approach, a spatially explicit technique, allows for a design-based global estimation of abundance with an estimation variance for regular or regular stratified sampling. The theory makes relatively few assumptions (randomness of either the origin of the sampling grid or the location of data points in grid cells) which are easily controllable in practice. Together with the low number of parameters to be estimated, this ensures robust results. Such transitive geostatistics are in operation to give abundance assessments in several fisheries institutes.

6. References

Bez N., J. Rivoirard and Ph. Guiblin (1997). Covariogram and related tools for structural analysis of fish survey data. Kluwer Academic Publisher, E.Y. Baafi and N.A. Schofield (eds), Geostatistics Wollongong'96, Volume 2, 1316-1327.

Bez N. (2002). Global fish abundance estimation from regular sampling: the geostatistical transitive method. Canadian Journal of Fisheries and Aquatic Sciences, **59**: 1921-1931.

Bez N., and Rivoirard J. (2000a). Indices of collocation between populations. In : Checkley, D.M., J.R. Hunter, L. Motos, and C.D. van der Lingen (eds). Report of a workshop on the use of Continuous Underway Fish Egg Sampler (CUFES) for mapping spawning habitat of pelagic fish. GLOBEC Report 14, 1-65 pp.

Bez N., and Rivoirard J. (2000b). On the role of sea surface temperature on the spatial distribution of early stages of European mackerel (1989) using inertiograms. *ICES Journal of Marine Science*, **57**: 383-392.

Bez N. and Rivoirard J. (2001). Transitive geostatistics to characterize spatial aggregations with diffuse limits: an application on mackerel ichtyoplankton. Fish. Res. **50**: 41-58.

Bez N., Rivoirard J., and Poulard J.C. (1995). Approche transitive et densités de poissons. Compte-rendu des journées de Géostatistique, 15-16 juin 1995, Fontainebleau, France. Cahiers de Géostatistique **5** 161-177.

Chilès J.P., and Delfiner P. (1999). Geostatistics, modeling spatial uncertainty. John Wiley and Sons, New York.

Conan G.Y., Moriyasu M., Wade E., and Comeau M. (1988). Assessment and spatial distribution surveys of snow crab stocks by geostatsistics. ICES C.M. 1988/K :10.

Cressie N.A.C. (1991). Statistics for spatial data. Wiley, New York.

Faraj A. and Bez N. (2007). Spatial pattern of the *Octopus vulgaris* life cycle in the Southern North Atlantic off Morocco. ICES Journal of Marine Science, in press.

Hoag S., Williams G., Myhre R., and Mcgregor R. (1980). Halibut assessement data: setline surveys in the North Pacific ocean, 1963-1966 and 1976-1979. International pacific halibut commission, technical report $n\circ 18, 42p$.

ICES (2003). Report of the working group on mackerel and horse mackerel egg surveys. International Council for the Exploration of the Sea, Lisbon 2003, ICES CM 2003/G:7, 60 p.

ICES (1997). Report of the International Bottom Trawl Survey Working Group. ICES CM/1997H:6.

Jakobsen T., Korsbrekke K., Mehl S., and Nakken O. (1997). Norwegian combined acoustic and bottom trawl surveys for demersal fish in the Barents Sea during winter. ICES CM Y:17, 25 pp.

Matheron G. (1965). Les variables régionalisées et leur estimation. Masson et Cie, Paris.

Matheron G. (1971). The theory of regionalized variables and its applications. Les Cahiers du Centre de Morphologie Mathématique **5**.

Matheron G. (1989). Estimating and chosing: an essay on probability in practice. Springer, Berlin.

Annex A: Nugget effects

To account for measurements errors, the regionalised variable $z(x)$ can be interpreted as a white noise $\omega_V(x)$ superimposed to a regionalized variable $y(x)$

$$z(x) = y(x) + \omega_V(x)$$

The white noise is defined as the restriction to the field V of a realization of a random function with a pure nugget covariance. The means of $\omega_V(x)$ and $y(x)$ over the field are denoted m_ω and m_y respectively. Considering that (i) the field is large enough for ergodicity to apply for the white noise and (ii) that the two components are independent in the transitive sense (Matheron, 1965, p 97), *i.e.* that

$$g_{\omega,y}(h) = \int \omega_V(x) y(x+h) dx = m_\omega m_y K(h)$$

the covariogram of $z(x)$ is

$$g_z(h) = g_y(h) + 2 m_\omega m_y K(h) + \sigma_\omega^2 K(0) \cdot 1_{h=0}$$

where $1_{h=0}$ represents the indicator function equal to 1 for distance $h = 0$ and 0 otherwise and $K(h)$ the geometrical covariogram of the field:

$$K(h) = \int 1_{z(x)\neq 0} 1_{z(x+h)\neq 0} dx$$

The above formula indicates that the contribution of a white noise is potentially twofold: it contributes to the nugget effect by a term equal to the variance of the white noise times the surface of the field; but it also modifies the continuous part of the structure by a quantity equal to the geometrical covariogram times the product of the global means of $y(x)$ and $\omega_V(x)$. This happens to be fundamentally different from the impact of a white noise in intrinsic geostatistics.

Unsystematic measurement errors ($m_\omega \approx 0$) add only to the nugget effect of $y(x)$. In practice, the respective contributions of $y(x)$ and $\omega_V(x)$ to the overall nugget effect are unknown, and splitting the discontinuity of the covariogram into terms for measurement errors and small-scale structures is impossible. The covariogram for $z(x)$ can thus be written (Figure 1):

$$g(h) = g_y(h) + \sigma_\omega^2 K(0) \cdot 1_{h=0}$$

Let us considered that the covariogram of $y(x)$ is made of a discontinuous part (nugget effect) and a continuous part so that

$$g_y(h) = C_0 \ 1_{h=0} + C_1 \ f(h)$$

Then, we have:

$$\begin{aligned} g(h) &= g_y(h) + \sigma_\omega^2 K(0) \cdot 1_{h=0} \\ &= C_0 \ 1_{h=0} + C_1 \ f(h) + \sigma_\omega^2 K(0) \cdot 1_{h=0} \\ &= (C_0 + \sigma_\omega^2 K(0)) \cdot 1_{h=0} + C_1 \ f(h) \\ &= \mathcal{C}_0 \ 1_{h=0} + C_1 \ f(h) \end{aligned}$$

which leads to the following standardised covariogram

$$\tilde{g}(h) = \tilde{C}_0 \ 1_{h=0} + \tilde{C}_1 \ f(h).$$

Annex B: A single hole effect model

By construction, convolution products of a function by its transposed value $(f^* \tilde{f})$ are positive definite. The idea is thus to explicit the covariogram obtained for a regionalized variable of the following form (Figure 5):

$$\begin{aligned} z(x,y) &= g_1(x,y) + g_2(t_x - x, t_y - y) \\ &= C_1 g(x,y,a_1) + C_2 g(t_x - x, t_y - y, a_2) \end{aligned}$$

where $g(x,y,a_1)$ is the probability density function (pdf) of a bigaussian vector of random variables without correlation ($\rho = 0$) and with the same standard deviation (a_1):

$$g(x,y,a_1) = \frac{1}{2\pi a_1^2} \ e^{-\frac{x^2+y^2}{2a_1^2}}.$$

Parameters a_1 and a_2 are expressed in distance units. In the bigaussian pdf they correspond to the standard deviation of each variable. In the present context they correspond to the spatial extension of each dome of the regionalization. Parameters C_1 and C_2 are expressed in the variable units. They quantify the level of each dome. The difference between the centers of the two bigaussian is $\Delta = (t_x, t_y)$.

The computation of the covariogram amounts to the following:

$$\begin{aligned} g(h_x,h_y) &= \iint z(x,y) z(x+h_x, y+h_y) dxdy \\ &= \iint \big[g_1(x,y) + g_2(t_x - x, t_y - y)\big] \cdot \\ &\qquad \big[g_1(x+h_x, y+h_y) + g_2(t_x - x - h_x, t_y - y - h_y)\big] dxdy \\ &= \begin{cases} \iint g_1(x,y) g_1(x+h_x, y+h_y) dxdy \\ + \iint g_2(t_x - x, t_y - y) g_2(t_x - x - h_x, t_y - y - h_y) dxdy \\ + \iint g_1(x,y) g_2(t_x - x - h_x, t_y - y - h_y) dxdy \\ + \iint g_2(t_x - x, t_y - y) g_1(x+h_x, y+h_y) dxdy \end{cases} \\ &= g11 + g22 + g12 + g21 \end{aligned}$$

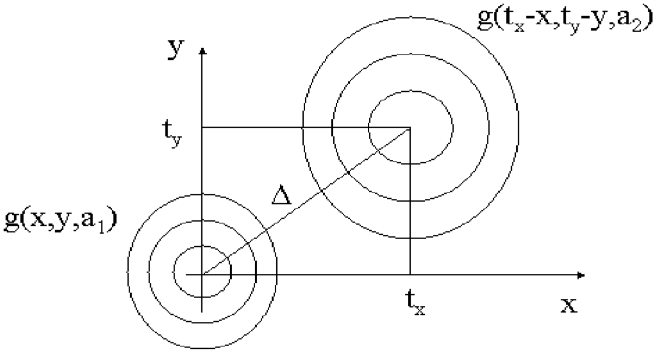


FIG 5. — Map a regionalized variable made of two bigaussian probability density functions. The circles represent isoprobability lines. Definition of the notations.

After developments, we get:

$$g11 = \frac{C_1^2}{4\pi a_1^2}\ e^{\frac{-||h||^2}{4a_1^2}}$$

$$g22 = \frac{C_2^2}{4\pi a_2^2}\ e^{\frac{-||h||^2}{4a_2^2}}$$

$$g12 = \frac{C_1C_2}{2\pi(a_1^2+a_2^2)}\ e^{\frac{-||\Delta||^2+||h||^2-2\langle\Delta,h\rangle}{2(a_1^2+a_2^2)}}$$

$$g21 = \frac{C_1C_2}{2\pi(a_1^2+a_2^2)}\ e^{\frac{-||\Delta||^2+||h||^2+2\langle\Delta,h\rangle}{2(a_1^2+a_2^2)}}$$

where $\langle\Delta, h\rangle$ represents the scalar product of Δ and h. Finally, it comes:

$$g(h) = \frac{1}{4\pi}\left[\frac{C_1^2}{a_1^2}\ e^{\frac{-||h||^2}{4a_1^2}} + \frac{C_2^2}{a_2^2}\ e^{\frac{-||h||^2}{4a_2^2}}\right] + \frac{C_1C_2e^{-\frac{||\Delta||^2+||h||^2}{2(a_1^2+a_2^2)}}}{2\pi(a_1^2+a_2^2)}\left[e^{\frac{\langle\Delta,h\rangle}{a_1^2+a_2^2}} + e^{-\frac{\langle\Delta,h\rangle}{a_1^2+a_2^2}}\right]$$

In particular:

$$g(0) = \frac{1}{4\pi}\left[\frac{C_1^2}{a_1^2}+\frac{C_2^2}{a_2^2}\right] + \frac{C_1C_2e^{\frac{-||\Delta||^2}{2(a_1^2+a_2^2)}}}{\pi(a_1^2+a_2^2)}$$

When $||\Delta||^2$ is large enough for the two domes not to overlap (or nearly so):

$$g(0) \approx \frac{1}{4\pi}\left[\frac{C_1^2}{a_1^2}+\frac{C_2^2}{a_2^2}\right] \quad \text{and} \quad g(\Delta) \approx \frac{C_1C_2}{2\pi(a_1^2+a_2^2)}$$

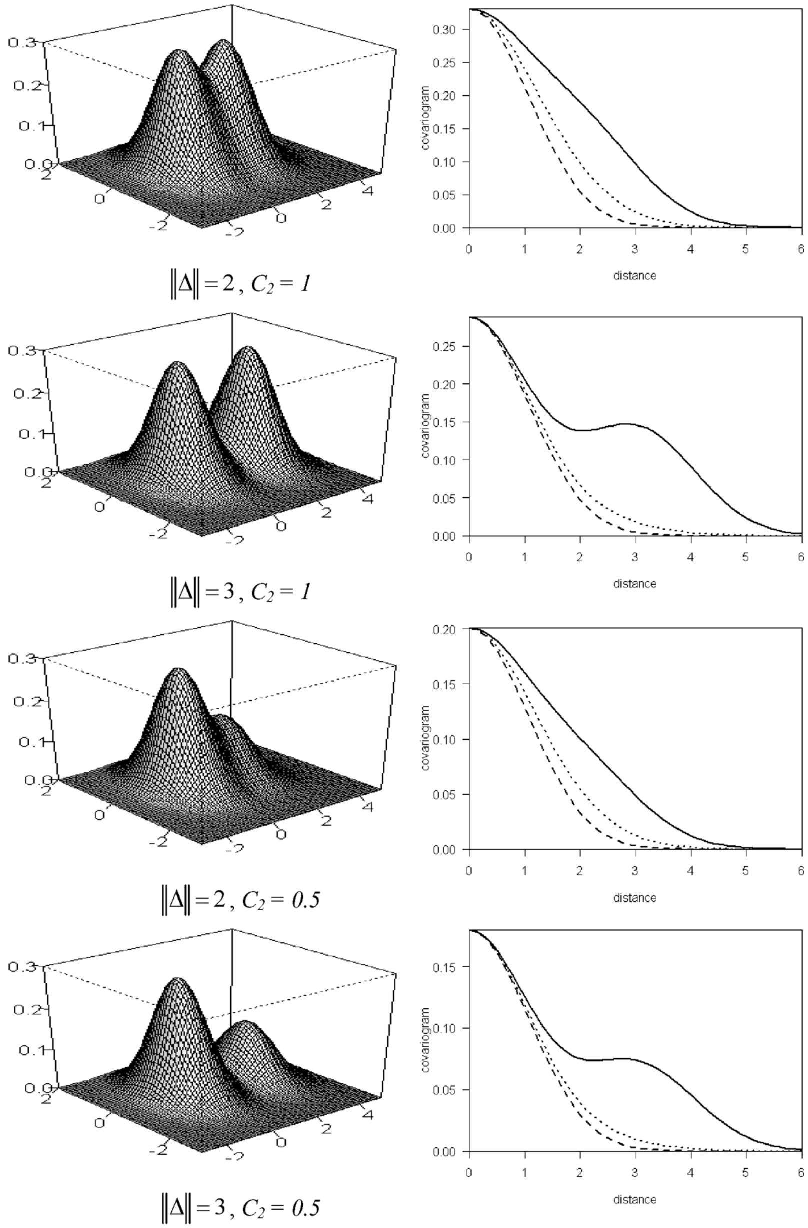
Interestingly, given that for gaussian distributions, 95% of the data belong to the interval centered on the mean $+/-$ twice the standard deviation, we can define a practical range equals to $||\Delta|| + 2 \cdot (a_1 + a_2)$ in the direction defined by the two maxima of the two bumps. So defined, the "double-gaussian" hole effect function gets a parabolic shape at the origin. Hence it can be considered for regular regionalized variables. The anisotropy is quite specific: full hole effect in the direction parallel to the line joining the two maxima of the regionalized variable; gaussian covariogram in a direction perpendicular to it.

This hole effect function is defined by 6 parameters:

- 2 parameters defining the spatial extension of each dome (a_1, a_2)
- 2 parameters defining the level of each dome (C_1, C_2)
- 2 parameters defining the distance and orientation between the two maxima (Δ)

An even more parameterized model can be considered when each bigaussian distribution corresponds to a vector of random variable with correlation $(\rho \neq 0)$ and with different standard deviations $(a_{1,1}, a_{1,2}, a_{2,1}, a_{2,2})$. In this case, the model gets 9 parameters.

Figure 6 represents situations for $a_1 = a_2 = 0.75$ and $C_1 = 1$ with two different distances between domes (Δ) and two levels for the second dome (C_2). When the domes are explicit but not enough apart one from each other, the hole effect might not be visible in the corresponding covariogram. The covariograms obtained for $||\Delta|| = 2$ do not get a clear hole effect despite the fish distributions.



$\|\Delta\| = 2$, $C_2 = 1$

$\|\Delta\| = 3$, $C_2 = 1$

$\|\Delta\| = 2$, $C_2 = 0.5$

$\|\Delta\| = 3$, $C_2 = 0.5$

Fig 6. — Hole effect. Regionalised variables and corresponding covariograms for $a_1 = a_2 = 0.75$ and $C_1 = 1$ with two different distances between domes (Δ) and two levels for the second dome (C_2). Covariograms are given for three directions: $x = 0$ (continuous line), $y = 0$ (dashed line) and $x = y$ (dotted line).