Linear and applied geostatistics

nicolas.bez@ird.fr

January 28, 2019

Contents

This report is a short note supporting my course on linear geostatistics. It is not meant to be a complete and comprehensive course by itself. It focuses on the preambles of the geostatistical technique: the meaning of variogram, the basic manipulation of variances, the stationary hypothesis which make random model operational, etc. For the kriging techniques and the associated geostatistical practice, readers can find detailed information in dedicated textbooks.

SUMMARY

- Geostatistics applies on variables whose spatial averages over different areas are consistent (spatial "additivity"). In particular, the mean over an area must be equal to the the mean of averages of regular sub-parts of this area.
- Except the mean, all statistics vary with the support of the observations: the variance decreases when the support increases (regularisation).
- The dispersion variance $s^2(v|V)$ quantifies the variability over a given area (V) at a given support (v). Dispersion variances are spatially compound following: $s^2(v|V) = s^2(v|v) + s^2(v|V)$
- The variance of a set of n values $z_i, i = 1, ..., n$ is also the (semi) average of the square differences between all pairs of values

$$s^2 = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (z_i - z_j)^2$$

.

- It is convenient to apply the above equation for neighboring data only (Geary's index) or for data h apart (the variogram).
- The variogram at distance h is the semi average of the square differences between observations h apart

$$\gamma(h) = \frac{1}{2n_p(h)} \sum_{\substack{i,j \\ x_i - x_j = h}} (z_i - z_j)^2$$

- An AR1 is a random process with a stationary spatial covariance $C(h)$ and we have $\gamma(h) = C(0) - C(h)$.
- A random walk is a random process where the variogram is stationary i.e. only depend on h.
- In general $var(\sum_i \lambda_i Z_i) = \sum_i \sum_j \lambda_i \lambda_j cov(Z_i, Z_j)$.
- For a Stationary Random Function (SRF), $var(\sum_i \lambda_i Z_i) = \sum_i \sum_j \lambda_i \lambda_j C(h_{i,j})$.
- For an Intrinsic random Function (IRF), $var(\sum_i \lambda_i Z_i) = -\sum_i \sum_j \lambda_i \lambda_j \gamma(h_{i,j})$ provided that $\sum_i \lambda_i = 0$.
- The best estimator of $Z(x_0)$ from $Z(x_1), ..., Z(x_n)$ is the expected value $E[Z(x_0)|Z(x_1), ..., Z(x_n)]$
- For a Gaussian RF, it is linear without any approximation $E[Z_0|Z_1, ..., Z_n] = \sum_i \lambda_i Z_i + \lambda_0$
- For other RF, the linear expression of the expected value is postulated and is sub-optimal.
- In Ordinary Kriging , the weights are constraint to $\sum_i \lambda_i = 1$ to abide the unbiasness. The estimation error is thus a linear combination where the weights sum to 0. The weights that insure minimum variance under this constraint are the kriging weights. They are defined by the variogram function through the following kriging system:

$$\left[\begin{array}{ccc|c} \ddots & & & \vdots \\ & \gamma_{i,j} & & 1 \\ & & \ddots & \vdots \\ \hline \dots & 1 & \dots & 0 \end{array}\right] \left[\begin{array}{c} \vdots \\ \lambda_i \\ \vdots \\ \hline \mu \end{array}\right] = \left[\begin{array}{c} \vdots \\ \gamma_{i,0} \\ \vdots \\ \hline 1 \end{array}\right]$$

The $n \times n$ matrix $[\gamma_{i,j}]$ represents the covariance between observations. The $n \times 1$ vector $[\gamma_{i,0}]$ represents the covariance between the observations and the target. The $n \times 1$ vector $[\lambda_i]$ represents the vector of the unknown kriging weights. The Lagrange parameter μ is a mathematical term that takes the constraint in charge.

1 Three selected references amongst plenty

Three references:

- **Chilès, J.-P. and Delfiner, P. 2012.** Geostatistics: Modeling Spatial Uncertainty. John Wiley & Sons, New York. 2nd edition. 731 p.
- **Cressie, N., 1991.** Statistics for Spatial Data. Wiley, New York, NY, 900 p.
- **Petitgas, P., Woillez, M., Rivoirard, J., Renard, D. and Bez, N. 2017**. Handbook of Geostatistics in R for fisheries and marine ecology. ICES Cooperative Research Report No. 338. 177 pp.
  http://www.ices.dk/sites/pub/PublicationReports/CooperativeResearchReport(CRR)/crr338/CRR_338_Final.pdf

Software

- RGeostatS : a dedicated R package; free download at http://rgeostats.free.fr/

2 Empirical properties of spatial data without any model

2.1 Basic, but nevertheless important, notions

2.1.1 Notation

- x denotes the position in space. This can be in 1D (distance from a staring point, time), in 2D (longitude and latitude), 3D (longitude, latitude, depth), or more .... There is no limitation in theory.
- z denotes the study variable. In multivariate geostatistics (not considered in this short note), one must specify the number of the variable $z^i(x)$.
- $z(x)$ is thus a regionalised variable.

The regionalised variable is sampled at some sampling stations x_i for $i = 1, ..., n$ so that the sample values are denoted $z(x_i)$ and more simply z_i.

2.1.2 Geographical support of the observations and/or of the target

Support refers to the geographical area associated to the recordings or to the target of the estimation.
Samples can be punctual measurements or can be considered as quasi-punctual at the study scale. However, in some cases, the observations have a support which is not punctual (e.g. pixels of satellite images, quadrats in agronomy).
The target of the estimation can be punctual (e.g. mapping) but it might not be punctual at all (e.g. in mine, samples consisted in holes while the estimation concerns blocks ; in epidemiology, samples consisted in individuals while the estimation concerns counties, etc).

Punctal supports are denoted x, and the corresponding regionalised variable $z(x)$.
Small supports are denoted v, and the corresponding regionalised variable $z(v)$.
Large supports are denoted V, and the corresponding regionalised variable $z(V)$.

2.1.3 "Additivity"

The mean value of z over a spatial domain V is the average of the punctual values at any points x of V :

$$z(V) = \frac{1}{V} \int_V z(x)dx$$

For a region V decomposed in non-overlaping sub-regions v_i, we get

$$z(V) = \frac{\sum_i v_i z(v_i)}{\sum v_i}$$

If all the sub-regions have the same surface area,

$$v_i = v_j = v \qquad \forall i, j$$

then

$$z(V) = \frac{\sum_i z(v_i)}{N}$$

where N is the number of sub-regions, $V = Nv$. This means that the regionalised variables used in geostatistics are supposed to be spatially "additive". Some variables are. Some variables are not. Amongst the variables that are spatially additive, one can mention:

- densities of individuals expressed for instance in $\frac{nbr}{m^2}$
- densities of pollutants expressed for instance in $\frac{g}{m^2}$
- altitude
- ...

Amongst the variables that are not spatially additive, we find:

- proportions (the mean of proportions is not the mean proportion)
- mean length of the individuals (the mean length of the population is not the mean of the mean lengths of its sub-populations)
- ...

As a matter of fact, if $p(v_i)$ represents the proportions of red balls in cells v_i, then the proportion of red balls in a larger area V is

$$p(V) = \frac{\sum_i q(v_i)p(v_i)}{\sum q(v_i)}$$

where $q(v_i)$ are the number of balls in cell v_i. It is not

$$p(V) \neq \frac{\sum_i v_i p(v_i)}{\sum v_i}$$

except in the very particular case where the number of balls are the same in all the cells.

The main difference between the "additive" and the "non additive" cases, is that the weights needed to get consistent overall means are the surface areas in the case of "additivity" but are another regionalized variable in the "non additivity" case (e.g. the biomass).

2.1.4 Field

The field is the geographical domain where the regionalized variable is positive. In ecology, this corresponds to the habitat of a species. In epidemiology, this is the infected area. etc
Its definition might not be straightforward.
In particular when crossing one variable whose field is bounded (e.g. the distribution of a wild species) with an explanatory variable whose field is much larger (e.g. air or water temperature); or when analyzing the interaction between two species with different habitats/fields.

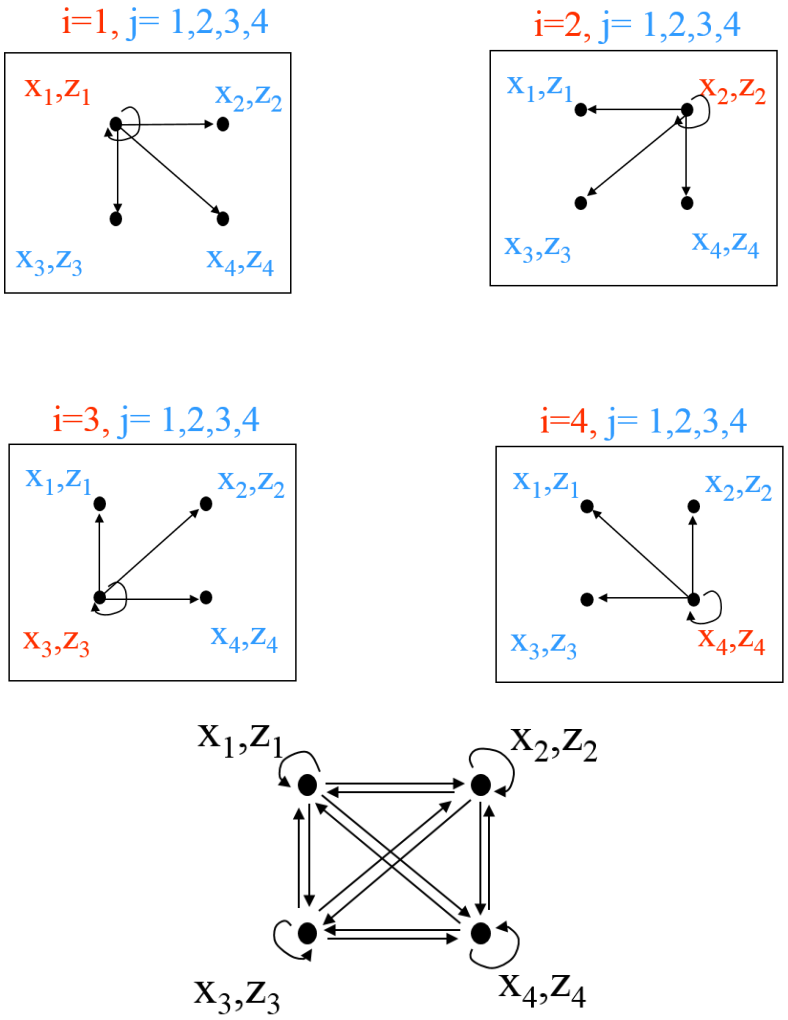


Figure 1: Decomposition of the variance so as to exhibit the double summation hidden in the algorithm. In the example we get four sample points and four sample values. This gives a total of $4^2 = 16$ pairs.

2.2 Revisiting the sample variance with spatial considerations

2.2.1 Rewritting the variance: a simple but a key step

The variance of the sample values z_i is defined as:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^2 = \frac{1}{n}\sum_{i}(z_i - \bar{z})^2$$

Note the simplification of the summation notation; this will be used systematically. So in each step i of the sum, the sample value z_i is confronted to the average value (squared). This means that each value is, in a way, confronted to the entire set of values contributing to the mean $\bar{z}$ (see Figure 1). So doing, the variance can be re-interpreted as (half) the double sum over the n^2 possible squared differences between sample values:

$$\boxed{s^2 = \frac{1}{n}\sum_{i}(z_i - \bar{z})^2 = \frac{1}{2n^2}\sum_{i}\sum_{j}(z_i - z_j)^2} \quad (1)$$

Proof:

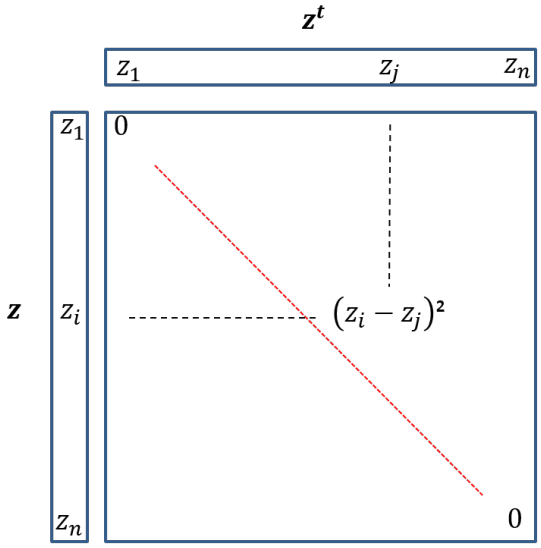


Figure 2: Another way of representing the variance.

$$
\begin{aligned}
\frac{1}{2n^2}\sum_i\sum_j (z_i - z_j)^2 &= \frac{1}{2n^2}\sum_i\sum_j (z_i^2 - 2z_i z_j + z_j^2) \\
&= \frac{1}{2n^2}\sum_i\sum_j z_i^2 - \frac{1}{n^2}\sum_i\sum_j z_i z_j + \frac{1}{2n^2}\sum_i\sum_j z_j^2 \\
&= \frac{1}{2n^2}\sum_i n z_i^2 - \frac{1}{n^2}\sum_i z_i(\sum_j z_i) + \frac{1}{2n^2}\sum_j n z_j^2 \\
&= \frac{1}{2n}\sum_i z_i^2 - \frac{1}{n}\sum_i z_i \bar{z} + \frac{1}{2n}\sum_j z_j^2 \\
&= \frac{1}{n}\sum_i z_i^2 - \bar{z}\bar{z} \\
&= \overline{z^2} - \bar{z}^2 \\
&= s^2
\end{aligned}
$$

So, the variance gets the four totally equivalent writings:

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_i^n (z_i - \bar{z})^2 && (2) \\
s^2 &= \overline{z^2} - \bar{z}^2 && (3) \\
s^2 &= \frac{1}{2n^2}\sum_i^n\sum_j^n (z_i - z_j)^2 && (4) \\
s^2 &= \frac{1}{2n_{pairs}}\sum_i^n\sum_j^n (z_i - z_j)^2 && (5)
\end{aligned}
$$

2.2.2 Geary Index

Based on Eq. 4, Geary (1954) suggested to decomposed the n^2 possible pairs into two groups, one for neighboring observations, and one distant observations. The definition of neighbors is user

and case specific. It corresponds to samples whose geographical distance is smaller than a given threshold d_0. Denoting $n(d_0)$ the number of such pairs of neighbors, we get:

$$s^2(d_0) = \frac{1}{2n(d_0)} \sum_i \sum_j (z_i - z_j)^2 \mathbf{1}_{d(i,j)<d_0}$$

where

$$\mathbf{1}_{d(i,j)<d_0} \begin{cases} 1 & \text{if } d(i,j) < d_0 \\ 0 & \text{if } d(i,j) > d_0 \end{cases}$$

This variance quantifies the variance that comes from neighboring observations. Geary compared it to the sample variance to get an evaluation of the existence of spatial autocorrelation. The Geary Index naturally derived from the above equation is

$$I_C = \frac{s^2(d_0)}{s^2}$$

It is small (respectively large) when the local variability is small wrt the overall variance (respectively large), that is when the spatial structure is strong (respectively small). The Geary index can be used to test statistically for the existence of short scale autocorrelation. However the sampling distribution of I_C is not straightforward and permutation test are often preferred.

2.2.3 The empirical variogram

Generalizing the decomposition suggested by Geary leads to the variogram. As a matter of fact, instead of decomposing the variance in two parts, i.e. between neighbors versus not neighbors, we can decompose the overall variance by distance class between pairs of samples.
In the previous example (Fig. 1), the 16 pairs of square differences are associated to either 0, 1 or $\sqrt{2}$ unit distance; the number of pairs being respectively 4, 8 and 4. We can thus define the empirical variogram as (half) the mean square differences between samples h distance apart:

$$\boxed{\gamma(h) = \frac{1}{2n(h)} \sum_i \sum_j (z_i - z_j)^2 \mathbf{1}_{d(i,j)=h}} \tag{6}$$

The variogram is nothing but the variance split into distance classes. **It is also (half) the average of the square differences between pairs of points h apart**. Of course, we can recover the variance by averaging the different variogram values :

$$s^2 = \frac{\sum n(h)\gamma(h)}{\sum n(h)}$$

In statistics, when considering a set data or the random variable from which these data are supposed to outcome, the usual practice is to distinguish between the mean and the expected value, both in wording (mean vs expected value) and notations (m vs E). However, this is not the case for the variance as one uses the same term (variance) but also the same notation (s^2) in both cases. The same is true in geostatistics as the variogram refers either to the empirical variogram of the data or to the variogram model of the Random Function. In both case, one uses the same notation $\gamma(h)$.

Here is a first example of empirical variogram

Let us consider a 1D example, where 10 data points are separated by 10 m intervals along a line. We measure the density of nano-particles in each of these sampling sites and find the following values:

$$2 - 3 - 1 - 1 - 2 - 1 - 1 - 2 - 3 - 4$$

With 10 data points, we get $10^2 = 100$ possible pairs of points. The distances associated to these 100 pairs range from 0 to 90 m. For 0 distance, the variogram is 0. It is (half) of the mean square difference between each sample value and itself. This happens 10 times. If we first consider distances of 10 m eastward, we get 9 pairs of observations. Applying the formula of the variogram, we get that:

$$\gamma(10) = \frac{(3-2)^2 + (1-3)^2 + (1-1)^2 + (2-1)^2 + (1-2)^2 + (1-1)^2 + (2-1)^2 + (3-2)^2 + (4-3)^2}{2 * 9}$$

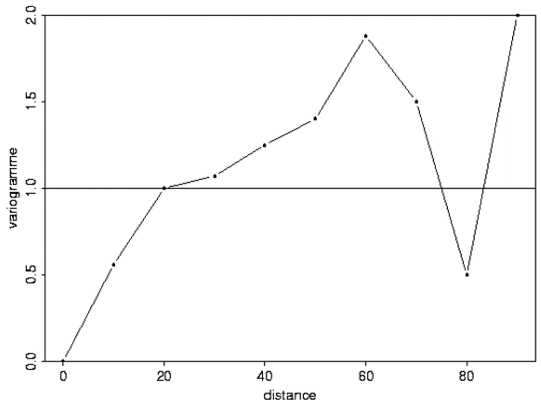


Figure 3: Empirical variogram.

$= 0.56$

Looking westward, we get the same result because $(z_i - z_j)^2 = (z_j - z_i)^2$. So $\gamma(-10) = \gamma(10)$. The next possible distance between samples is 20 m. We know have 8 pairs and:

$$\gamma(20) = \frac{(1-2)^2 + (1-3)^2 + (2-1)^2 + (1-1)^2 + (1-2)^2 + (2-1)^2 + (3-1)^2 + (4-2)^2}{2*8} = 1$$

The largest possible distance is 90 m for which only one pair of observations can be built:

$$\gamma(90) = \frac{(4-2)^2}{2*1} = 2$$

The graph of the variance as a function of the distance is the vario-gram. It is always represented in the same way:

- the x-axis represents h the geographical distance between sample points
- the y-axis represents $\gamma(h)$ the variance of the (sub) set of pairs of points h apart
- an horizontal line is added at the level of the sample variance (here $s^2 = 1$)
- a symbol, a number or a third axis can be added to represent the evolution of the number of pairs $n(h)$

Some key properties of the empirical variogram can already be listed:

- the variogram is (half) the average of the square differences between pairs of points h apart. Reminding that a variance is the mean of the squares minus the square of the mean (Eq. 3), **if the average of the differences between pairs of points h apart is null**, then **the variogram is the semi variance of the differences between pairs of points h apart**.
- because $(z_i - z_i) = 0$, the variance for distance 0 is 0: $\gamma(0) = 0$.
- being sums of squares, variogram values are positive: $\gamma(h) \geq 0, \forall h$.
- because the mean of the variogram values (weighted by the number of pairs) is equal to the sample variance, because variogram values are all positive, and because $\gamma(0) = 0$, there must be some variogram values larger than the sample variance. When the variogram is represented entirely, some points must be above the sample variance. This does *not* hold when representing a sub part of the variogram, either the variogram for distances smaller than half the dimension of the field as often recommended, either when looking to one particular geographical direction (see below).
- as $(z_i - z_j)^2 = (z_j - z_i)^2$, the variogram is symmetrical: $\gamma(h) = \gamma(-h)$.
- each variogram value is based on a fluctuating number of pairs of points $n(h)$. When this number increases, the corresponding variogram value represents the mean of a larger number of pairs of points. Points of the variogram based on the largest numbers of pairs of points are the points that explain most of the variance. The number of pairs of points usually decreases with the distance. Decrease is strict in 1D (see the example above). In 2D (see below), it first increases and then diminishes.
- one particular observation is involved in n pairs with many different distance. The variogram values share common observations (in probabilistic terms we will say that they are not independent).
- when permuting observations in space, their mean and variance do not change. The variogram at 0 distance does not change. But all the other variogram values change. The mean and variance are *not* spatial statistics. The variogram is a spatial statistics.

2.2.4 Dispersion variance and support effect

Let us start with an example.

We consider an exhaustive survey of a field denoted V of 6 m^2 based on the use of 1m x 1m quadrats (denoted $v_i, i = 1, ..., 6$). In each quadrat, we measure the density of insects expressed in ind/m^2. In the end of the survey, we get the following measures denoted $z(v_i), i = 1, ..., 6$:

1	3	2	4	2	6

One can check that the mean and variance are respectively 3 ind/m^2 and $\frac{8}{3}$ ind^2/m^4.

While the mean over the field is

$$z(V) = \frac{1}{n}\sum_{i=1}^{n=6} z(v_i) = 3$$

we will denote the variance of the small support data in the field in the following way:

$$s^2(v|V) = \frac{1}{n}\sum_{i=1}^{n=6} \left(z(v_i) - z(V)\right)^2 = \frac{8}{3}$$

We now consider that the same survey is done with another support for the measurements. The densities for double size quadrats of 2 m^2 denoted with a larger letter $v_i, i = 1, ..., 3$ would have been:

2	3	4

While the mean over the field is unchanged,

$$z(V) = \frac{1}{n}\sum_{i=1}^{n=3} z(\mathcal{v}_i) = 3$$

the variance of the large support data in the field is smaller:

$$s^2(\mathcal{v}|V) = \frac{1}{n}\sum_{i=1}^{n=3} \left(z(\mathcal{v}_i) - z(V)\right)^2 = \frac{2}{3}$$

It happens though that:

$$s^2(\mathcal{v}|V) < s^2(v|V)$$

The diminution of the sample variance comes from the absorption of the local variability that exists between two neighboring quadrats. To quantify this local variance let us consider, the three variances of two small supports data into their union:

$$s^2((v_1, v_2)|\mathcal{v}_1) = \frac{1}{2}\sum_{i=1}^{n=2} \left(z(v_i) - z(\mathcal{v}_1)\right)^2 = 1$$

$$s^2((v_3, v_4)|\mathcal{v}_2) = \frac{1}{2}\sum_{i=3}^{n=4} \left(z(v_i) - z(\mathcal{v}_2)\right)^2 = 1$$

$$s^2((v_5, v_6)|\mathcal{v}_3) = \frac{1}{2}\sum_{i=5}^{n=6} \left(z(v_i) - z(\mathcal{v}_3)\right)^2 = 4$$

The mean of these local variances is $s^2(v|v) = 2$. And we get, the following combination of the variances

$$s^2(v|V) = s^2(v|\mathcal{v}) + s^2(\mathcal{v}|V)$$

the variance of the small support in the field is the variance of the small supports in the medium ones + the variance of the medium support in the field (the size of the notation being important here).

This equation is indeed general and explains the effect of the support on the variances: the latter decreases when the former increases.

Practical consequences:

- one must not mix data with different supports (e.g. counties with regions monitoring data, large with small quadrat counts, large with small pixel satellite data, etc)
- looking for properties at a support larger than the sample support needs a model to quantify the variance reduction

Note: $s^2(v|V)$ is not sensitive to permutation of quadrats, but $s^2(v|v)$ and $s^2(v|V)$ are and can be considered as spatial statistics.

2.2.5 Conclusions

Without any model, i.e. without specifying any probability distribution for the data, we have been able to set some relevant properties of raw statistics, and in particular of the variance of regionalized variables. The following chapters will transport observations into a probabilistic framework useful to go further.

3 Variance of linear combinations of random variables (reminder)

This subsection is a reminder on general formula in statistics. We now consider random variables. We no longer consider observations, i.e. data, if they exist, are considered as outcomes of a random variable(s).
Let us start by considering two random variables Z_1 and Z_2.

$$var(Z_1 + Z_2) = var(Z_1) + var(Z_2) + 2cov(Z_1, Z_2)$$

$$cov(Z_1, Z_2) = E(Z_1 Z_2) - E(Z_1)E(Z_2) = cov(Z_2, Z_1)$$

$$cov(Z_1, Z_1) = var(Z_1)$$

$$cov(\lambda_1 Z_1, \lambda_2 Z_2) = \lambda_1 \lambda_2 cov(Z_1, Z_2)$$

$$var(\lambda_1 Z_1) = \lambda_1^2 var(Z_1)$$

so

$$var(Z_1 + Z_2) = cov(Z_1, Z_1) + cov(Z_1, Z_2) + cov(Z_2, Z_1) + cov(Z_2, Z_2)$$

and the variance of the sum of Z_1 and Z_2 is the double sum of the covariances between all possible pairs between (Z_1, Z_2) and itself

$$var(Z_1 + Z_2) = \sum_{i=1}^{2} \sum_{j=1}^{2} cov(Z_i, Z_j)$$

This can be generalized into the very important following equation which is the core of kriging:

$$\boxed{var\left(\sum_{i=1}^{n} \lambda_i Z_i\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j cov(Z_i, Z_j)} \tag{7}$$

This equation is totally general. It makes a covariance function a particular mathematical function namely a *positive definite* function.

4 Random processes with stationary covariance or variogram

In this section, we build two different random processes (in 1D). These two random processes are meant to show examples of random processes getting a stationary variogram. The first one is an auto-regressive process of order 1, with a stationary spatial covariance. The a second one, a random walk, introduces random processes without stationary spatial covariance but with stationary variogram. From now one, we will considered that $Z(x)$ is a Random Variable located at x in the geographical space. A **Random Function** is a set of such variables located everywhere in space: $Z(x), x \in \mathbb{R}^d$. Characteristics of $Z(x)$ that do not depend on x are said to be **stationary**. The stationarity is a property of the Random Function.

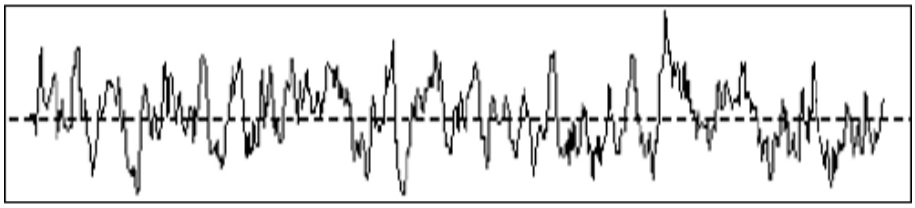


Figure 4: Auto-regressive process of order 1. Simulation of one realisation of such random process showing that there is a tendency not to move to much away from the mean.

4.1 AR1

Auto-regressive processes of order 1 are such that

$$Z(x+1) = \rho Z(x) + \sqrt{1-\rho^2}U_{x+1} \qquad \rho \in]0,1[$$

where U_i are independent and identically distributed (i.i.d.) random variables or, at least, independent random variables with identical expected values and variances. Here we will just specify that $E(U_i) = 0$ and that $var(U_i) = 1$.
Setting $Z(1) = U_1$, we get $E(Z(1)) = E(U_1) = 0$. If we assume that $E(Z(x-1)) = 0$, then

$$E(Z(x)) = \rho E(Z(x-1)) + \sqrt{1-\rho^2}E(U_x) = 0$$

So by induction,

$$E(Z(x)) = 0, \forall x$$

Similarly, $var(Z(1)) = 1$. If we assume that $var(Z(x-1)) = 1$, then, $Z(x)$ and $U(x)$ being independent, we get that

$$var(Z(x)) = \rho^2 var(Z(x-1)) + (1-\rho^2)var(U_x) = 1$$

So by induction,

$$var\left(Z(x)\right) = 1, \forall x$$

The expected value (moment of order 1) and the variance (moment of order 2) are constant in space and independent of x. They are thus stationary.
The spatial covariance can also be deduced by recursion

$$\begin{aligned} cov(Z(x), Z(x+1)) &= cov\left(Z(x), \rho Z(x) + \sqrt{1-\rho^2}U_{x+1}\right) \\ &= \rho \cdot cov\left(Z(x), Z(x)\right) + \sqrt{1-\rho^2}cov\left(Z(x), U_{x+1}\right) \\ &= \rho \cdot var(Z(x)) + 0 \\ &= \rho \end{aligned}$$

Assuming that $cov(Z(x), Z(x+h)) = \rho^h$, then it comes that

$$\begin{aligned} cov(Z(x), Z(x+h+1)) &= cov\left(Z(x), \rho Z(x+h) + \sqrt{1-\rho^2}U_{x+h+1}\right) \\ &= \rho \cdot cov\left(Z(x), Z(x+h)\right) + \sqrt{1-\rho^2}cov\left(Z(x), U_{x+h+1}\right) \\ &= \rho \cdot \rho^h + 0 \\ &= \rho^{h+1} \end{aligned}$$

So, $cov(Z(x), Z(x+h)) = \rho^h, \forall h$. The spatial covariance is thus stationary. It does only depend on the distance h between the two points and not on the location x. The AR1 is stationary random process of order 2, that is:

$$\begin{aligned} E(Z(x)) &= E(Z(x+h)) \\ var(Z(x)) &= var(Z(x+h)) \\ cov(Z(x), Z(x+h)) &= C(h) \end{aligned}$$

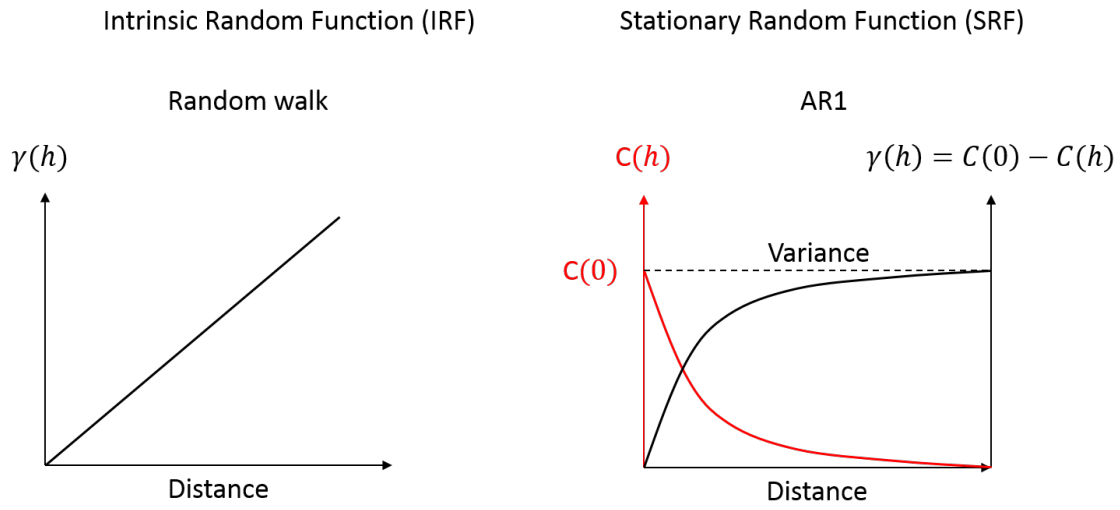


Figure 5: Covariance and variogram for random walks and auto-regressive processes.

The variogram is

$$\begin{aligned}\gamma(h) &= \frac{1}{2}var\big(Z(x) - Z(x+h)\big) \\ &= \frac{1}{2}\Big(var\big(Z(x) + Z(x+h) - 2cov(Z(x), Z(x+h))\big)\Big) \\ &= C(0) - C(h)\end{aligned}$$

An AR1 random process is thus a stationary random function SRF where

$$cov\big(Z(x), Z(x+h)\big) = C(h)$$

and

$$\gamma(h) = C(0) - C(h)$$

Note that the pdf of U and by consequence of Z has *not* been specified. In case $Z(x)$ is Gaussian, the knowledge of its expected value and its spatial covariance fully characterizes the model. This is not the case for all other pdf.

4.2 Random walk

Let us consider a random variable U taking values +1 or -1 with equal probability. We will consider a series of independent and identically distributed (i.i.d.) such variables $U_i, i = 1, ...n$. They will be used to simulate a random walk. Starting at point 0, a walker gets up or down at each step of his walk according to realizations of the U_i.

$$E(U_i) = \sum_{k=1}^{2} p_k u_k = (1/2).(1) + (1/2).(-1) = 0$$

$$var(U_i) = \sum_{k=1}^{2} p_k u_k^2 = (1/2).(1)^2 + (1/2).(-1)^2 = 1$$

The altitude at which the walker is at point x depends on the number of up and down moves. Let us denote this altitude as $Z(x) = \sum_{i=1}^{x} U_i$. This is a random process in 1D.

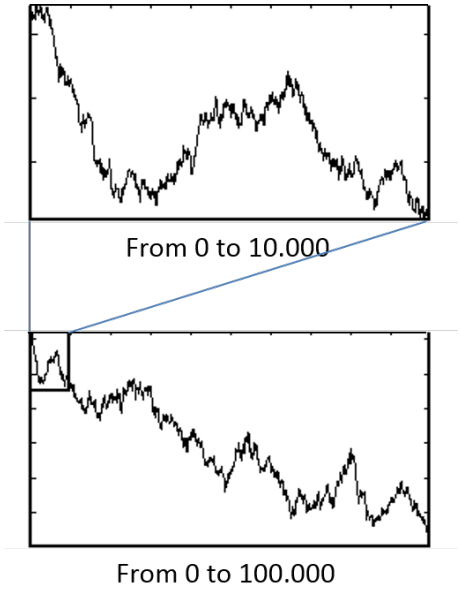


Figure 6: Random walk. Representation of one realization of a +1/-1 random walk over the first 10 000 and 100 000 steps of the walk.

The expectation is a linear operator. So

$$E(Z(x)) = E(\sum_{i=1}^{x} U_i) = \sum_{i=1}^{x} E(U_i) = 0$$

which means that on average, after x steps, the walker will be at floor 0. This is true whatever x. The expected value does *not* depend on x and is thus stationary.
Let us now compute the variance of the random walk. The random walk is the sum of i.i.d. random variables. So in the following equation

$$var(Z(x)) = var(\sum_{i}^{x} U_i) = \sum_{i}^{x}\sum_{j}^{x} cov(U_i, U_j)$$

all the cases where i and j are different correspond to the covariance between two independent variables, which is equal to 0. The only cases where the covariance is not null is when the two indices are the same which happens x times and which corresponds to $cov(U_i, U_i) = var(U_i)$. The U_i being identically distributed they have the same variance $var(U_i) = 1, \forall i$. Finally we get

$$var(Z(x)) = x$$

This means that the altitude reached by the walker after x steps is as variable as x increases. The variance of $Z(x)$ depends on x. It is thus *not* stationary.
Let considers the random walk at points x and $x + h$. As the random walk is build step by step, $Z(x)$ and $Z(x + h)$ get the segment $[0, x]$ in common, and we can write:

$$Z(x + h) = Z(x) + U_{x+1} + U_{x+2} + ... + U_{x+h} = Z(x) + \sum_{i=x+1}^{x} U_i$$

The U_i being iid, this means that

$$\begin{aligned} cov(Z(x), Z(x + h)) &= cov\left(Z(x), Z(x) + \sum_{i=x+1}^{x} U_i\right) \\ &= cov(Z(x), Z(x)) + \sum_{i=x+1}^{x} cov(Z(x), U_i) \\ &= var(Z(x)) = x \end{aligned}$$

So the spatial covariance is *not* stationary. But

$$\begin{aligned} var(Z(x) - Z(x+h)) &= var(Z(x)) + var(Z(x+h)) - 2cov(Z(x), Z(x+h)) \\ &= x + x + h - 2x \\ &= h \end{aligned}$$

indicating that the variance of the increment $var(Z(x) - Z(x+h)) = h$ does not depend on x and is thus stationary.

While

$$E(Z(x) - Z(x+h)) = 0$$

we have

$$var(Z(x) - Z(x+h)) = E\left((Z(x) - Z(x+h))^2\right)$$

and finally

$$var\big(Z(x) - Z(x+h)\big) = 2\gamma(h)$$

so that the random walk gets a stationary variogram.

The random walk is an instance of an intrinsic random function (IRF). An IRF is a random function whose increments are order 2 stationary, i.e. whose expected value and variance are stationary:

$$E(Z(x) - Z(x+h)) = 0 \quad \text{independent of x}$$

$$var(Z(x) - Z(x+h)) = 2\gamma(h) \quad \text{independent of x}$$

Surprisingly, while the process gets no drift ($E(Z(x)) = 0, \forall x$), the particular realization represented if Figure 6 shows a clear decreasing trend.

In practice, this means that:

- The choice of a random function model should be based on the behavior of several real (seldom possible) or virtual (normal case) realizations of the model.
- A real trend in the data, is *not* incompatible with a model without drift.

4.3 Stationarity makes inference possible in case of single realization

The two above paragraphs start from known models and look at their properties. In practice, this is reverse: one gets data and try to infer a possible model compatible with these observations. However, one usually gets only one realization and inference become impossible. Ideally, the variogram should be obtained from several realizations of the random function $z_i(x), i = 1, ..., n_{realization}$ by taking the average of the square differences obtained over several realisations at two particular points h apart:

$$\gamma(h) = \frac{1}{2n_{realization}} \sum_{i=1}^{n_{realization}} \big(z_i(x) - z_i(x+h)\big)^2$$

But this is not applicable in practice. Assuming that the data are outcomes of one of the two random processes above, we can argue that the mean square difference between data does not depend on their locations but only on their distance. The **pairs of observations h apart everywhere in space are thus also repetitions of what is expected between x and $x+h$ for a given point in space** x in the model:

$$\left\{\big(z_i(x), z_i(x+h)\big), i = 1..., n_{realization}\right\} \Longleftrightarrow \left\{(z_i, z_j), d(i,j) = h, , i = (1, ..., n), j = (1, ..., n)\right\}$$

Thanks to the stationary assumption, the estimation of the variogram becomes again possible in practice following Eq. 6.

5 (Intrinsic) Random Functions: a probabilistic framework to model and use variograms

5.1 Definitions

A random function (RF) $Z(x)$ is an infinite family of random variables. It is mathematically defined by its spatial probability law, that is the generalization of the density function of random variables:

$$F_{x_1,x_2,...,x_n}(z_1,z_2,...,z_n) = P\left(Z(x_1) < z_1, Z(x_2) < z_2, ..., Z(x_n) < z_n\right), \forall n$$

This spatial law includes all the monovariate probability density distributions i.e. the traditional pdf

$$F_x(z) = P(Z(x) < z)$$

but also all the bivariate probability density distributions

$$F_{x_1,x_2}(z_1,z_2) = P(Z(x_1) < z_1, Z(x_2) < z_2)$$

and all the upper ones i.e. the trivariates distributions, the quadrivariates distributions, and so on. A spatial law is extremely rich and is not operational in practice when one must parameterize it from data. In order to be able to use RF in practice, one must restrict the definition down to some reasonable level of parametrization. First, we will consider only moments of order 1 and 2 of the bivariate pdf that is, $E(Z(x))$ and $cov(Z(x), Z(x+h))$. Note that these two elements are far from being equivalent to the full bivariate distribution. Second, we will consider that these two moments are stationary, i.e. that they do not depend on x:

$$E\big(Z(x_1)\big) = E\big(Z(x_2)\big) = E\big(Z(x)\big) = m$$

and

$$cov\big(Z(x), Z(x+h)\big) = C(h)$$

Such an RF is called a stationary RF of order 2 and is hereafter referred to as a stationary random function (SRF). In the particular case of a SRF, the covariance only depends on the distance h between x and $x+h$, the general formula 7, reduces to the very important following equation

$$\boxed{\text{SRF:} var\left(\sum_i \lambda_i Z_i\right) = \sum_i \sum_j \lambda_i \lambda_j C(h_{i,j})} \tag{8}$$

We have seen that there are cases where the variance is not stationary but the variogram is stationary (e.g. random walk). These cases correspond to intrinsic random function (IRF), that is RF such that

$$E(Z(x) - Z(x+h)) = 0$$

and

$$\boxed{var(Z(x) - Z(x+h)) = 2\gamma(h)}$$

An increment $Z(x) - Z(x+h)$ is a linear combination with weights $+1$ for the first variable and -1 for the second. The sum of the weight is thus 0 and the variogram allows computing the variance of this linear combination. This can be generalized to any linear combination whose sum of weights is 0. The variogram allows computing the variance of any linear combination of an IRF provided that $\sum_i \lambda_i = 0$

$$\boxed{\text{IRF: } var(\sum_i \lambda_i Z_i) = -\sum_i \sum_j \lambda_i \lambda_j \gamma(h_{i,j}), \text{if} \sum_i \lambda_i = 0} \tag{9}$$

The variogram model is the engine to process the variance of any linear combination of data (provided that the weights sums to 0). It only needs to know the geographical distance between points.

An SRF is an IRF, but an IRF is not necessarily a SRF : $SRF \Rightarrow IRF$ but $IRF \nRightarrow SRF$. The kriging equations will thus be developed with variograms which is more general than covariances.

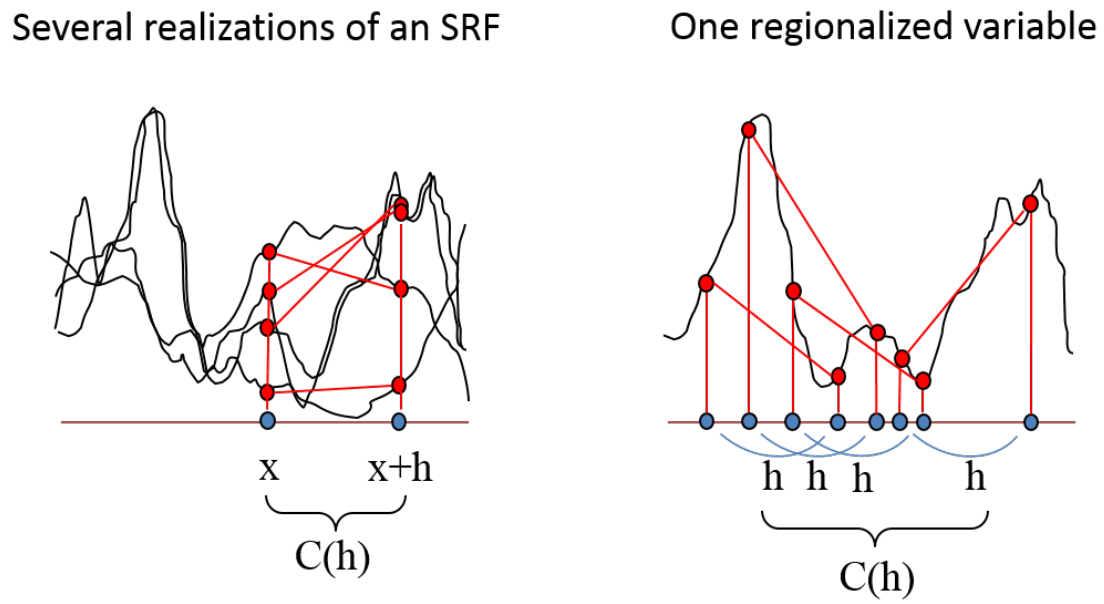


Figure 7: Stationarity: repetitions and statistics over several realizations are replaced by realizations over space over one particular realization, i.e. over the regionalized variable. Here, we illustrate the stationarity of the order 2, i.e. the stationarity of the covariance of the RF.

A Gaussian random function (GRF) gets gaussian pdf, bigaussian distributions, trigaussian laws, etc. It is a very particular model which gets very particular properties. In particular a stationary GRF is fully defined by its spatial covariance. SGRF is THE only case where knowing the first 2 moments amounts to knowing the entire model.

5.2 Stationarity of the RF and statistics over a single realization

The stationarity is a property of the random function. It states properties possibly observed under several realizations of the RF. In particular one could verify that over several independent realizations of the same RF, the averages obtained at points x and $x+h$ are the same, or that the covariance computed between realization at points x and $x+h$ is similar than that between y and $y+h$.
However, once the stationarity is stated (or speculated), it means that the statistics computed over realizations at a given point or for a given pair of points can be access to through repetitions over space. This is particularly important since in the real world, no RF $Z(x)$ exists but only a single regionalized variable $z(x)$, and even more, most of time, only a discrete version of it $z_i, i = 1, ..., n$. Based on a regionalized variable considered as a realization of a RF, stationarity assumptions are thus key to access to the parameters of an SRF or an IRF through spatial averages.

Note: Pooling all the data into a single histogram and fitting a pdf to it, amounts to a full stationarity hypothesis. As a matter of fact, not only moments of order 1 and 2 are considered stationary, but the full distribution. This is a much stronger hypothesis than the usual one made in linear geostatistics where only the first two moments are concerned by the stationarity hypotheses.

5.3 Variogram properties

We can now update the list of the variogram characteristics already listed in subsection 2.2.4.

5.3.1 Positive definiteness

The variance is a positive quantity. The mathematical function used to model variogram must then be such that Equation 9 never generates negative values. A variogram is thus a mathematical function such that $-\gamma(h)$ is conditionally positive definite.

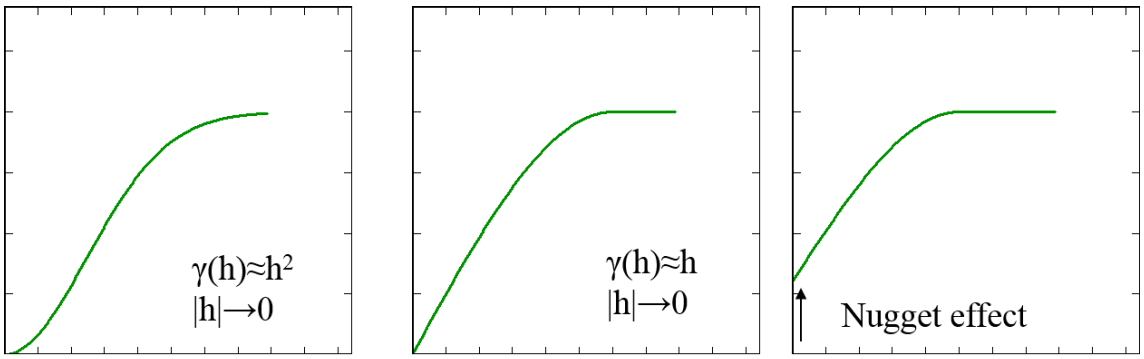


Figure 8: relationship between the behavior of the variogram near the origin and the spatial continuity of the random function.

Not all the functions fulfill this condition and a variogram must be selected amongst a family of allowed functions.

5.3.2 Behavior at the origin

The mathematical behavior of the variogram at the origin is connected to the behavior of the the RF $Z(x)$

$$Z(x) \text{ continuous and differentiable } \Leftrightarrow \gamma(h) \approx h^2 \text{ when } |h| \to 0$$

$$Z(x) \text{ continuous and not differentiable } \Leftrightarrow \gamma(h) \approx h \text{ when } h \to 0$$

$$Z(x) \text{ not continuous and not differentiable } \Leftrightarrow \gamma(h) \nrightarrow 0 \text{ when } h \to 0 \text{ (Nugget effect)}$$

The choice of the behavior of the variogram at the origin should be made with considering the physics of the regionalized variable considered as a realization of the random function.

The behavior of the variogram at the origin is also the part of the model that most impacts the kriging.

5.3.3 Measurement errors and nugget effect

In case the regionalized variable $z(x)$ is sampled with *unsystematic* measurements errors, the RF associated to it could be the following RF

$$Y(x) = Z(x) + \epsilon(x)$$

where $\epsilon(x)$ is a white noise independent of $Z(x)$ (i.e. the measurement errors are not auto-correlated and are not correlated with the target variable).
The variogram of $Y(x)$, for $h \neq 0$, would be

$$\begin{aligned}
\gamma_Y(h) &= \frac{1}{2} var(Y(x) - Y(x+h)) \\
&= \frac{1}{2} var(Z(x) + \epsilon(x) - Z(x+h) - \epsilon(x+h)) \\
&= \frac{1}{2} var(Z(x) - Z(x+h)) + \frac{1}{2} var(\epsilon(x)) + \frac{1}{2} var(\epsilon(x+h)) \\
&= \gamma_Z(h) + \sigma^2_\epsilon
\end{aligned}$$

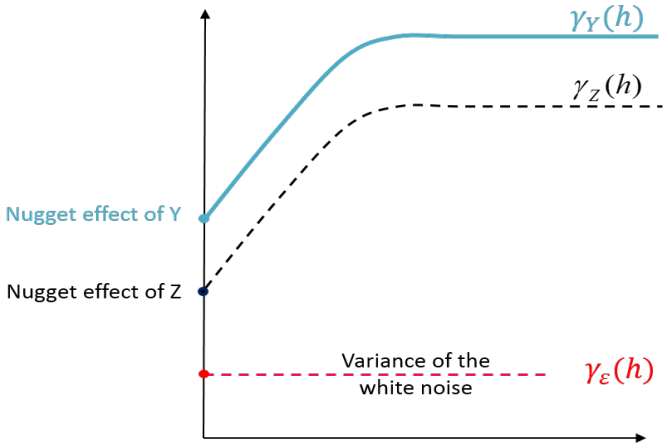


Figure 9: Impact of a random unsystematic measurement error on the variogram. If the regionalized variable $z(x)$ is measured with an unsystematic errors not related to the value of $z(x)$ itself and independent from one location to the other, the new regionalized variable is indeed $y(x) = z(x) + \epsilon(x)$. The nugget effect of $Y(x)$ includes the nugget effect of $Z(x)$ (if it exists) and the variance of the measurement error. The variogram of the white noise is a pure nugget effect.

- The variogram of a noisy version of $Z(x)$ is thus nothing but the variogram of $Z(x)$ plus a nugget effect equal to the variance of the measurement errors.
- The nugget effect is thus a not disentangling mixture of measurement errors *and* spatial structures that exist at small scales.
- The nugget effect is a key parameter of the behavior of the variogram near the origin.
- Any time the variable is noisy, a nugget effect should be integrated in the model.

5.4 Variogram fitting

In the same way that the choice of a pdf is a crucial point in statistics, the choice of a variogram model is central in geostatistics. Despite all the cautious required to variogram fitting, this subsection is relatively succinct.
By default, one proceed by minimizing square differences between the empirical variogram and the model:

$$\varepsilon_i = \gamma_{empirical}(lag_i) - \gamma_{model}(lag_i)$$

$$argmin\left(\frac{1}{\text{nb of lags}}\sum_{i=1}^{\text{nb of lags}}\varepsilon_i^2\right)$$

This can be weighted by i) the number of pairs associated to the values of the empirical variogram ($n(lag_i)$) and ii) the distance from the origin 0 (lag_i):

$$\omega_i = \frac{n(lag_i)}{lag_i}$$

$$argmin\left(\frac{\sum_{i=1}^{\text{nb of lags}}\varepsilon_i^2\omega_i}{\sum_{i=1}^{\text{nb of lags}}\omega_i}\right)$$

Some (non exhaustive) elements to keep in mind:

- the behavior at the origin is key. Nugget effect is expected anytime there are measurement errors and/or small scale structures that are not resolved at the sampling scale
- a trend in the data does not necessarily translate into a drift in the model
- a structure in the variogram should be based on 4 points at least
- in case of directional empirical variogram, model must be based on all directions together
- given a model, one can simulate several realizations under this model and compute the variogram of each realization. The fluctuations of these variograms around the model can be large (as large as the range is large wrt to the size of the simulated field)
- only use functions that are allowed
- the validity of the model is linked to the number of lags used in the fitting. This is in interaction with the futur use of the model i.e. local versus global estimation.
- statistical testing of the fittings is confronted to the fact that the variogram values are not independent. Alternatively cross validation procedure can help choosing bieween several models.

6 Estimation and kriging

6.1 Limits of the classical method

In sampling theory, estimating the mean from samples that are considered as N outcomes of n random variables independent and identically distributed (iid) leads to an estimation variance of the form

$$\sigma_E^2 = \frac{\sigma^2}{N}$$

The precision of the estimate increases when the estimation variance decreases, i.e. when the variance of the random process decreases and/or when the number of samples increases. This makes sense.
However, this formula is only accessible when the samples can be considered as iid. This amounts to consider the three following things:

- data are outcomes of random variables. This is not straightforward. Random variables are mathematical entities. They do no exist in the field. Data are real.
- the random variables are independent. This is a very strong constraint which can be achieved when the sampling scheme is strictly random. However there exists many cases where this is not the case (e.g. systematic, regular, and stratified schemes; see figure 10).
- the random variable have the same pdf. This is also a very strong constraint. This corresponds to a strict stationarity assumption, i.e. all moments are stationary (and not only the first two ones).

If autocorrelations exist which is the vast majority of the practical cases, this gets two opposite consequences:

- First, the variance of the sum increases due to the covariances that must now be considered $var(X+Y) = var(x)+var(Y)+2cov(X,Y)$ and redundancy in the data reduces the effective number of samples in hand. This is detrimental to the precision of the estimate.
- Second, autocorrelation means that some links exists between the data. Such links enable a better interpolation between the data than when no structure between the data exists. The structure is an additional information useful for the inference. This goes towards an increase of the precision of the estimate.

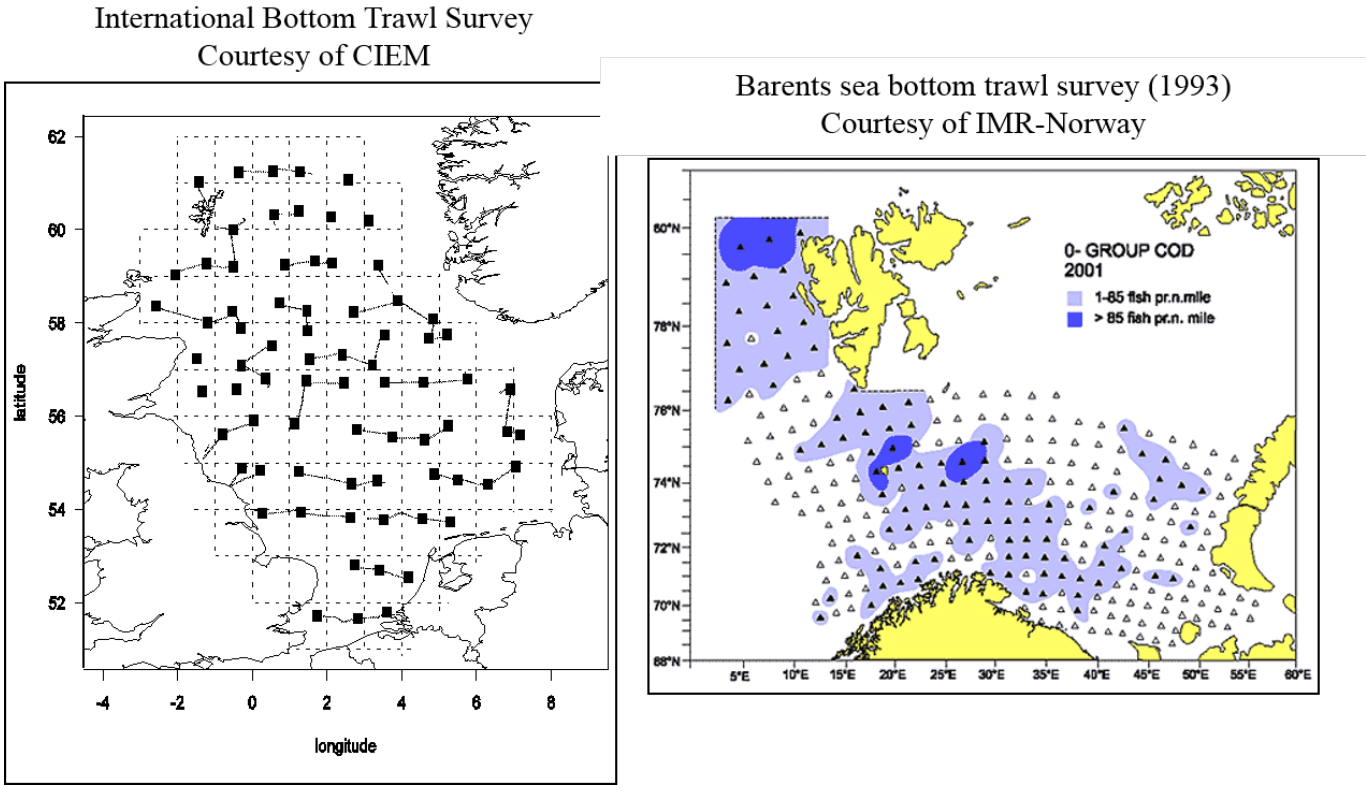


Figure 10: Examples of surveys where the samples are *not* collected under a strict random protocol. Left: A random stratify survey where one sample is located at random in each sampling square independently from the other squares. All together the samples are not located at random. Right: A systematic survey design where samples are located on a regular grid (the projection produce a deformation). Samples are not located at random.

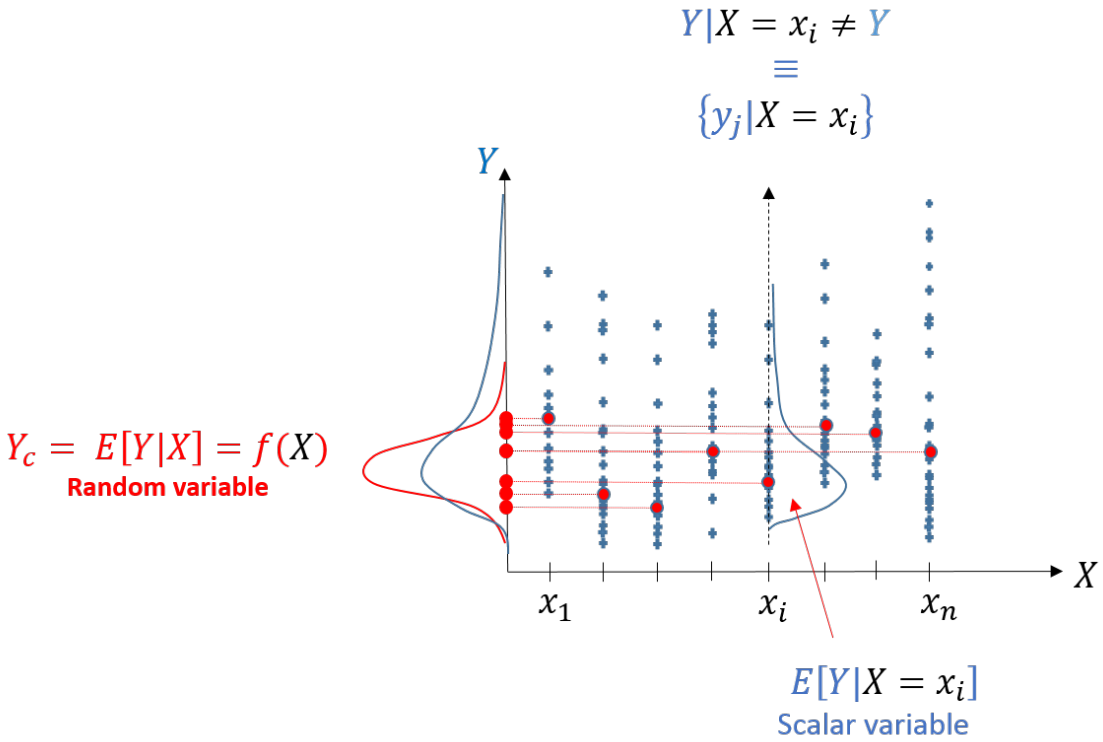


Figure 11: Correlation cloud, conditional and marginal pdf, expected value. In black, the explanatory variable X. In blue, the target variable Y. In red, conditional properties.

So, due to autocorrelations, one gets less information than the one provided by N independent observations (the effective sample size is smaller than N) but one gets a spatial structure to help interpolating in between the data. The balance between these two contradictory aspect in terms of gain or loss in precision of the estimation is not known in advance. It depends on the strength of the spatial structure and on the geographical location of the samples and of the target of the estimation.

6.2 Conditional expectation and linear regressions

6.2.1 Definitions

In statistics, the analysis of the relationship between two random variables leads to concept the conditional expectation (figure 11). One must distinguish between:

- the random variable Y which gets its own pdf f_Y (blue marginal distribution in the scatter plot)
- the random variable $Y|X = x_i$ which gets a particular pdf ($f_{Y|X=x_i}$) and whose average is expectation of Y conditional to the fact that $X = x_i$: $E[Y|X = x_i]$
- the random variable $E[Y|X]$ called the conditional expectation of Y knowing X. It corresponds to the former expression when X is randomized according to its pdf (f_X). This random variable gets its own pdf which is different from the pdf of Y (red in the figure). Despite its name, i.e. conditional *expectation*, this is *not* a real value. This is a random variable.
- the average value of the conditional expectation is the expected value of Y (the blue and the red pdf get the same average): $E\left(E(Y|X)\right) = E(Y)$

Parallel to this graphical definition, the conditional expectation is key in the theory of estimation. As a matter of fact, the best approximation of Y by a function of X, best in terms of the mean square, is the $E(Y|X)$. This means that

$$E((Y-[Y|X])^2) \leqslant E(Y-\phi(X))^2), \forall \phi$$

Assuming no bias, in both case, this means that

$$var((Y-[Y|X])) \leqslant var(Y-\phi(X))), \forall \phi$$

In the very particular case of Gaussian random variables (X Gaussian, Y Gaussian, and (X,Y) biGaussian), the conditional expectation is linear.

$$E(Y|X) = \lambda_1 X + \lambda_0 \qquad \text{(Gaussian case)}$$

In Gaussian cases, the best approximation of Y by a function of X is linear without approximation; but only in Gaussian cases. In all other cases, the linear expression of the condition expectation is an approximation, but can still be used

$$E(Y|X) \approx \lambda_1 X + \lambda_0 \qquad \text{(non-Gaussian case)}$$

which can be generalized to several explanatory variables

$$Y^* = E(Y|X_1, ..., X_n) = \sum_i \lambda_i X_i + \lambda_0 \qquad \text{(Gaussian case)}$$

$$Y^* = E(Y|X_1, ..., X_n) \approx \sum_i \lambda_i X_i + \lambda_0 \qquad \text{(non-Gaussian case)}$$

Even though the linear expression of the conditional expectation is an approximation, we can still search for the Best Linear Unbiased Estimator (BLUE). We know that it is not the optimal one in the general case. But it will be the best one in the linear framework.

One must not mix up the estimator $E(Y|X_1, ..., X_n)$, a random variable, and the estimation $E(Y|X_1 = x_1, ..., X_n = x_n)$, a real value.

6.2.2 Reminder on the parameters estimation

Let us start with the monovariate case.
We observe independent pairs of (x_i,y_i). Unknows are then *chosen* to minimise the mean square difference between the observations and their estimations:

$$\frac{1}{n}\sum_i (y_i - y_i^*)^2 = \frac{1}{n}\sum_i (y_i - \lambda_1 x_i - \lambda_0)^2 = F(\lambda_0, \lambda_1)$$

which is a function of the two unknown parameters.
Minimization is obtained when the two partial derivatives equal 0:

$$\frac{\partial F(\lambda_0, \lambda_1)}{\partial \lambda_0} = \frac{\partial F(\lambda_0, \lambda_1)}{\partial \lambda_1} = 0$$

After some developments, we get the following two well known equations solving the system with the two unknowns:

$$\begin{aligned} \lambda_0 &= \bar{y} - \lambda_1 \bar{x} \\ \lambda_1 &= \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{cov(X,Y)}{var(X)} \end{aligned}$$

In the bivariate case, the developments become a bit more complex but, for λ_1 we get that

$$\lambda_1 = \frac{\sigma_{X_1}}{\sigma_Y} \frac{\rho_{Y,X_1} - \rho_{X_1,X_2}\rho_{Y,X_2}}{1 - \rho^2_{X_1,X_2}}$$

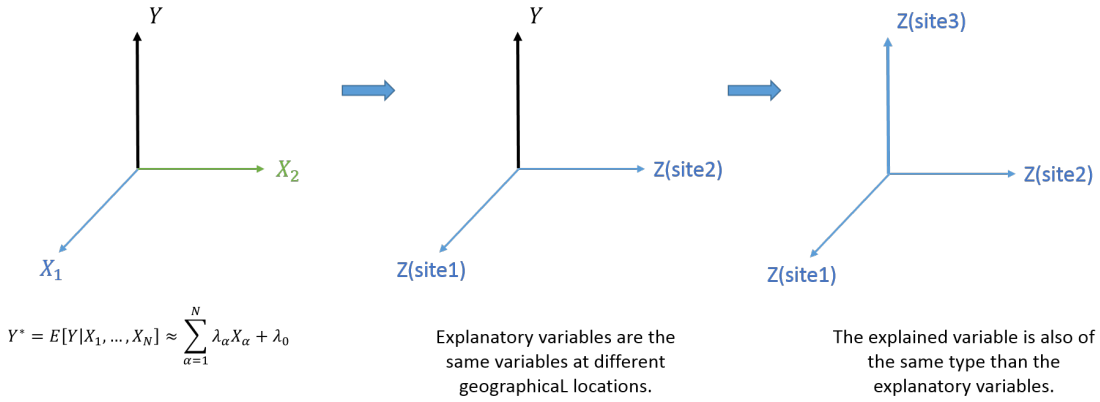


Figure 12: Revisiting the nature of the variables used for a regression.

Beside the exact formula, what is key is that:

- the parameters of the BLUE estimation of Y^* are only dependent on the knowledge of all the covariances between explanatory variables and of all the covariances between the target variable and the explanatory ones.
- the parameters of the BLUE estimation of Y^* does not need any assumption on the pdf of X and Y. No parametric assumptions are required to estimate λ_i. They are only needed to perform statistical testings (not considered here).
- X_1 and X_2 do not need to be independent (their correlation is incorporated in the above formula). However, the realizations $(x_{1,i}, x_{2,i}, y_i)$ have to be independent outcomes.

6.2.3 The spatial re-interpretation of the multivariate linear regression: the kriging

Traditionally, X_1 and X_2 represent two different random explanatory variables or covariates. However, nothing precludes from considering that they correspond to the same (regionalized) variable measured in two different locations. Without loss of generality, they can be denoted $Z(x_1)$ and $Z(x_2)$ (see figure 12). This can be generalized one step further by considering that the Y-variable of the regression is also concerning the RF Z but in a third location $Z(x_3)$. So doing, we look at a regression explaining $Z(x_3)$ by a linear combination of $Z(x_1)$ and $Z(x_2)$.

The traditional way of representing a regression in a Cartesian space can then be revisited by implementing the axes of the regression in the geographical space (see figure 13). While an observation $(x_{1,i}, x_{2,i}, y_i)$ is represented by a point in the 3D Cartesian space, it is now represented by a triangle $(z(x_1), z(x_2), z(x_3))$.

Traditionally, one gets several observations on which the regression is based (see figure 14). In spatial statistics, we usually gets only one realization to play with. The covariances used in the regression (see above) are thus not available in practice and one will need a model of spatial covariance to replace them.

The situation is even worth. As a matter of fact, the kriging objective is not to estimate the parameters that insure the BLU-Estimate of $Z(x_3)$ but to insure the BLU-Estimate of an unsampled location, denoted x_0 to underline the difference between a sample point and a point where we want to make the estimation (see figure 15). In other words, there is no observation y_i (respectively z_0) to be compared with y_i^* (respectively z_0^*); and thus no possible square error to minimize. Here again, a model of spatial covariance is needed to replace the empirical covariance used in the regression.

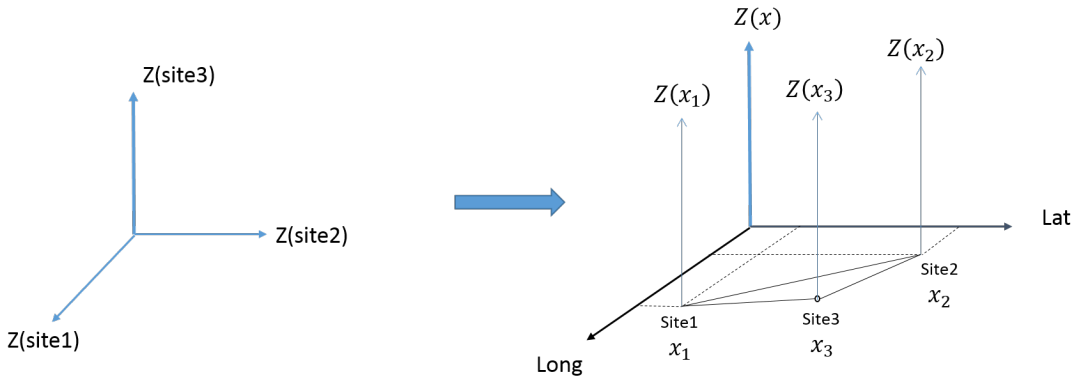


Figure 13: From a Cartesian to a geographical representation of a (bivariate) regression.

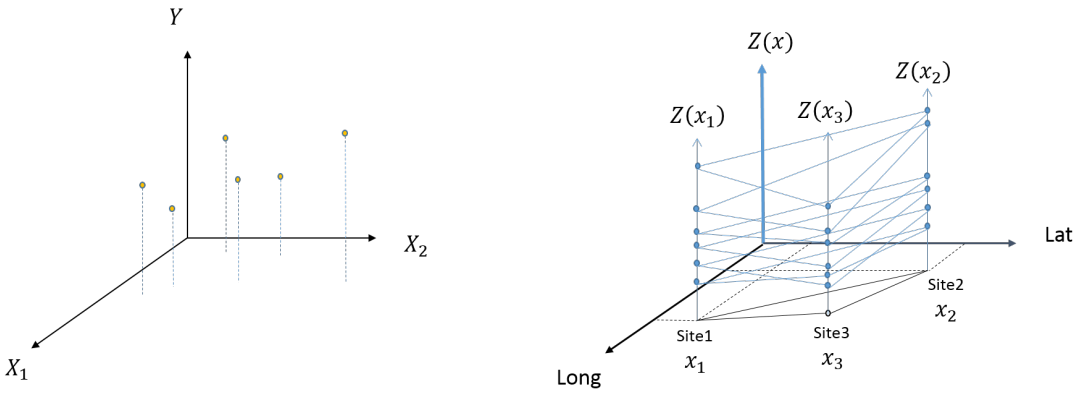
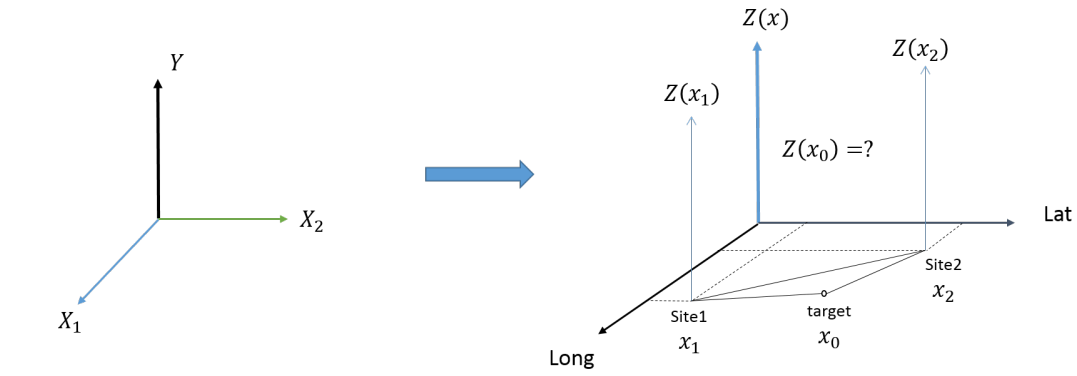


Figure 14: Graphical representation of the sample data on which the regression could be based in a Cartesian and in a geographical representations.



$$Y^* = E[Y|X_1, \ldots, X_N] \approx \sum_{\alpha=1}^{N} \lambda_\alpha X_\alpha + \lambda_0 \qquad Z_0^{KRIGING} = E[Z_0|Z_1, \ldots, Z_N] \approx \sum_{\alpha=1}^{N} \lambda_\alpha Z_\alpha + \lambda_0$$

Figure 15: Regression: estimating the parameters of the regression to best explain Y by a linear combination of X_1 and X_2. Kriging: choosing the parameters of the linear combination of $Z(x_1)$ and $Z(x_2)$ that allows a BLUE estimation of $Z(x_0)$

- One key difference between the regression in the Cartesian space and the kriging is geographical space is the number of available realizations.
- While Z_0 is also not observed, the regression in praxi can not be built from the minimization of the square difference between Z_0^* and Z_0. However, the principle of minimization of the variance to get a BLU-Estimate of Z_0 by a linear combination of $Z(x_i)$ applies, and kriging is based on a model which provides the (spatial) covariances between all pairs of covariates and between covariates and the target.
- The key step of a geostatistical analysis is thus the definition of a model of spatial covariance (or equivalently of variogram). Then after, kriging is nothing but regression techniques.

6.3 (Ordinary) Kriging in theory

All the sections above are, I thing, not really available in usual textbooks. On the contrary, there exist plenty of textbook presenting kriging. I thus restrict myself to the basics.

6.3.1 Ponctual (ordinary) kriging

We denoted x_0 the point where the estimation is performed. To get a kriging map, one has to replay the procedure for each point of a regular grid.
Kriging is the BLU-estimate of $z(x_0) = z_0$ based on the values of the RF at some known points $Z(x_1) = z(x_1), ..., Z(x_N) = z(x_N)$. The kriging estimator is:

$$Z_0^K = \sum_{i=1}^{n} \lambda_i Z_i$$

without loss of generality the intercept of the regression is not considered here. This allows to get homogeneous notation where all 0-subscripts concerns the target point.
When all the sample values are used (n=N), this is called kriging with unique neighborhood. When only the values of the samples belonging to a restricted neighborhood around x_0 are used, this is called kriging with moving neighborhood.
While the framework of IRF is more general than that of SRF (the variogram is more general than the spatial covariance), we solve the equations for IRF.

The bias is:

$$E(Z_0^K - Z_0) = E\left(\sum_{i=1}^{n} \lambda_i Z_i - Z_0\right) = m\sum_{i=1}^{n} \lambda_i - m = m\left(\sum_{i=1}^{n} \lambda - 1\right)$$

where m is the unknown expected value of the IRF. As a matter of fact, an IRF is such that $E(Z(x) - Z(x+h)) = 0$, which amounts to consider that $E(Z(x)) = m$ but with undefined value for m. Note: in the vast majority of cases, the objective of the estimation is to estimate m. So, considering it is unknown, is (more than) relevant.
To insure no bias whatever the value of m, one must insure that

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

So the kriging estimator can be rewritten is the following manner:

$$Z_0^K = \sum_{i=1}^{n} \lambda_i Z_i, \quad \text{with} \sum_{i=1}^{n} \lambda_i = 1$$

The weights are to be chosen to insure the minimum estimation variance:

$$\sigma_E^{2\,K} = var(Z_0^K - Z_0) = var\left(\sum_{i=1}^{n} \lambda_i Z_i - Z_0\right) = var\left(\sum_{i=0}^{n} \lambda_i Z_i\right)$$

with

$$\sum_{i=1}^{n} \lambda_i = 1 \quad \text{and} \quad \lambda_0 = -1 \quad \text{so that} \quad \sum_{i=0}^{n} \lambda_i = 0$$

The linear combination in the estimation variance is a linear combination where the sum of the weights sum to 0. So Equation 9 applies and we get

$$\sigma_E^{2\,K} = -\sum_{i=0}^{n}\sum_{j=0}^{n} \lambda_i \lambda_j \gamma_{i,j} \quad \text{with} \quad \sum_{i=0}^{n} \lambda_i = 0$$

In mathematics, minimizing the estimation variance under the constraint that $\sum_{i=1}^{n} \lambda_i = 1$, amounts to minimizing

$$\sigma_E^{2\,K} - 2\mu(\sum_{i=1}^{n} \lambda_i - 1)$$

where μ is called the Lagrange parameter. As in the regression, this is done by getting the partial derivatives equal to 0. In the end, the kriging weights are solution of the following system of $n+1$ equations:

$$\begin{aligned} &\sum_{k=1}^{n} \lambda_k \gamma_{i,k} + \mu = \gamma_{i,0}, \quad i = 1, ..., n \\ &\sum_{k=1}^{n} \lambda_k = 1 \end{aligned}$$

The matrix version of this system is

$$\left[\begin{array}{ccc|c} \ddots & & & \vdots \\ & \gamma_{i,j} & & 1 \\ & & \ddots & \vdots \\ \hline \dots & 1 & \dots & 0 \end{array}\right] \cdot \left[\begin{array}{c} \vdots \\ \lambda_i \\ \vdots \\ \hline \mu \end{array}\right] = \left[\begin{array}{c} \vdots \\ \gamma_{i,0} \\ \vdots \\ \hline 1 \end{array}\right]$$

This presentation makes it clear that a kriging system involves the spatial structure for:

- distances between data points $\gamma_{i,j}$ (a n x n matrix)
- distances between the data points and the target point $\gamma_{i,0}$ (a n x 1 matrix)

Finally, the kriging weights are obtained by inverting the left hand side matrix:

$$\begin{bmatrix} \lambda_i^K \\ \mu^K \end{bmatrix} = \begin{bmatrix} \gamma_{i,j} & 1 \\ 1 & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \gamma_{i,0} \\ 1 \end{bmatrix}$$

and the kriging estimator is simply

$$Z_0^K = \sum_i \lambda_i^K Z_i$$

with the corresponding estimate

$$z_0^K = \sum_i \lambda_i^K z_i$$

and the estimation variance is

$$\sigma_E^{2\,K} = \sum_i \lambda_i^K \gamma_{i,0} - \mu^K$$

Note$_1$: while the representation of a kriging map is pixelized, what is estimated and represented is indeed a point value, not the mean of the regionalized variable over each pixel. This later case corresponds to block kriging (see below). While for small grid cells or pixels (small wrt the field size) this is not a major issue, one must not confuse the two.

Note$_2$: the range of the distances over which the model is used are determined by the size of the neighborhood. A model is, for sure, not solicited at distances larger than the largest diagonal of the neighborhood. So the quality of the model matters only for $h \in [0, \max(\text{diagonal}(\text{neighborhood}))]$. For small neighborhood, only the behavior of the spatial covariance/variogram at the origin matters. For unique neighborhood, the model is solicited over all the distances. The quality of the model fitting must be considered accordingly.

Note$_3$: Filtering out the mean m leads to a constraint on the weights and thus on the fact that we have to estimate n+1 parameters ($\lambda_1, ..., \lambda_n$ and μ). This necessarily means that the estimation variance is larger than the one we would have obtained without the constraint on the weights if we had knew the mean.

6.3.2 Non-ponctual kriging (block/polygon/global kriging)

When the objective is to estimate the regionalized variable over a non-punctual area, say $z(v)$ where v can be any geographical zone, the previous equations get a direct generalization. We still consider a linear estimator based on the observed points

$$Z(v)^K = \sum_i \lambda_i Z_i, \quad \text{with } \sum_i \lambda_i = 1$$

This means that the left member of the kriging system is unchanged. The only modification concerns the right side of the kriging system where one considers the variogram between observations and the target. The target being a polygon, the $\gamma_{i,0}$ are replaced by $\gamma_{i,v}$ which is the mean value of the variogram between the point x_i and (all the points of) v

$$\gamma_{i,v} = \frac{1}{v} \int_v \gamma_{i,x} dx$$

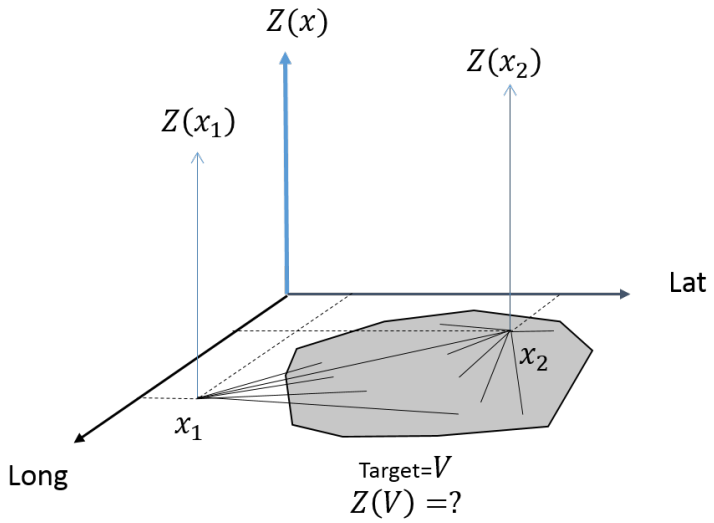
the kriging system is thus

$$\left[\begin{array}{ccc|c} \ddots & & & \vdots \\ & \gamma_{i,j} & & 1 \\ & & \ddots & \vdots \\ \hline \dots & 1 & \dots & 0 \end{array}\right] \cdot \left[\begin{array}{c|c} \vdots & \\ \lambda_i & \\ \vdots & \\ \hline \mu & \end{array}\right] = \left[\begin{array}{c|c} \vdots & \\ \gamma_{i,v} & \\ \vdots & \\ \hline 1 & \end{array}\right]$$

and the solutions are now

$$\begin{bmatrix} \lambda_i^K \\ \mu^K \end{bmatrix} = \begin{bmatrix} \gamma_{i,j} & 1 \\ 1 & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \gamma_{i,v} \\ 1 \end{bmatrix}$$

In practice, the $\gamma_{i,v}$ are evaluated by the discretization of v. The finer the resolution, the better the numerical approximation.



$$Z_V^{KRIGING} = E[Z_V|Z_1, \dots, Z_N] \approx \sum_{\alpha=1}^{N} \lambda_\alpha Z_\alpha + \lambda_0$$

Figure 16: Non-ponctual kriging: block, polygon or global kriging. To perform such kriging, we need to know all the spatial covariances between the sample points and the target area.

7 Supplementary materials

7.1 Impact of the 0 and of field delineation on the variance

Very often the vector of the observations gets 0, and eventually many 0, modifying the mean and the variance of the observations. This happens when surveying a wider area than the one where the phenomenon exists. This is the normal situation in ecology and any time you do not know in advance the area of presence.
It is convenient to split the vector of the n observations into the n_0 null data and the n_+ positive ones: $z = c(0, z_+)$ with $n = n_0 + n_+$. In the double sum of the variance, one can distinguish the cases when the two points get null data, when one is null and when the two are non-null (Figure 17):

$$s^2 = var(c(0, z_+)) = \frac{0 + 2n_0 \sum_i z_{+,i}^2 + \sum_{i,j}(z_{+,i} - z_{+,j})^2}{2(n_0 + n_+)^2}$$

If we denote s_+^2 the variance of the positive data we get:

$$s^2 = \frac{n_0 n_+ \bar{z_+^2} + n_+^2 s_+^2}{(n_0 + n_+)^2}$$

Given that $s_+^2 = \overline{z_+^2} - m_+^2$, this becomes:

$$s^2 = \frac{n_0 n_+ (s_+^2 + m_+^2) + n_+^2 s_+^2}{(n_0 + n_+)^2}$$

It is convenient to introduce two ratio, the ration between n_0 and n_+ (r_0) and the coefficient of variation of the positive data $CV_+ = s_+/m_+$. In most natural system, there exists a mean to variance relationships showing that the variance increases together with the mean. It is thus relevant to survey the various possible cases with variable coefficient of variations. The above equation finally writes:

$$s^2 = s_+^2 \frac{r_0 + (1 + r_0)CV_+^2}{CV_+^2(1 + r_0)^2}$$

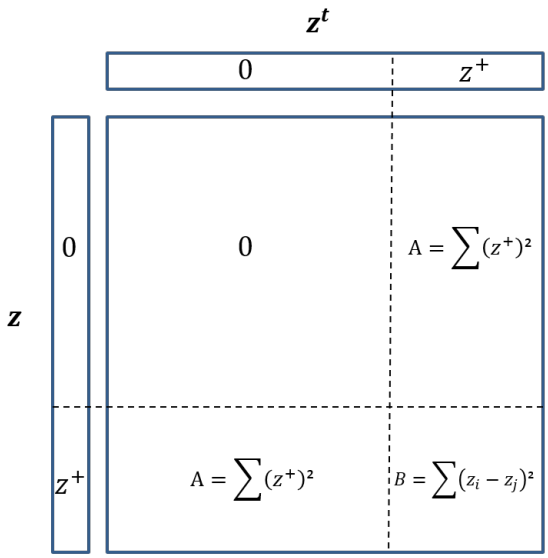


Figure 17: Impact of the 0 when computing the variance.

which indicates how the variance of data with 0 observations changes with the number of 0 (with regards to the number of non-null data and for some given coefficient of variation of the positive data).

The result (Fig. 18) is, somehow, counter intuitive: For data with high variability wrt the mean (i.e. with large CV), which is the standard situation in practice, adding 0 decreases the variance monotonically. On the contrary, when the CV of positive data is small, adding null data first increases the variability, but after some sufficient number of null data, the variance goes down towards 0.

This result holds whatever the geographical location of the 0, either in a sub-region homogeneously full of 0 or distributed everywhere in space.

Non available data are not 0 data!

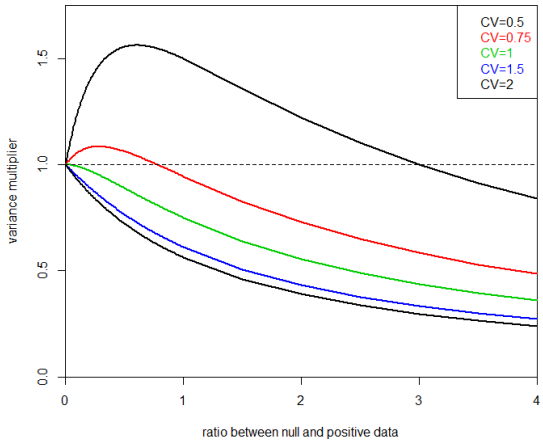


Figure 18: Impact of the null data on the variance. The ratio between the number of null and positive data is $r_0 = \frac{n_0}{n_+}$. The proportion of null data in the data set is 33%, 50% and 66% for $r_0 = 0.5$, $r_0 = 1$, and $r_0 = 2$ respectively.