The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

2163-8

College on Soil Physics: Soil Physical Properties and Processes under Climate Change

30 August - 10 September, 2010

GEOSTATISTICS: BASIC ASPECTS FOR ITS USE TO UNDERSTAND SOIL SPATIAL VARIABILITY

Luis Carlos Timm
Federal University of Pelotas
Brazil

GEOSTATISTICS: BASIC ASPECTS FOR ITS USE TO UNDERSTAND SOIL SPATIAL VARIABILITY

TIMM, L.C.[1*]; AQUINO, L.S. [2]; RECKZIEGEL, N.L. [3]; REICHARDT, K.[4]; TAVARES, V.E.Q.[5]

[1]Regular Associate of the ICTP, Rural Engineering Department, Faculty of Agronomy, Federal University of Pelotas, CP 354, 96001-970, Capão do Leão-RS, Brazil.

*lctimm@ufpel.edu.br; lcartimm@yahoo.com.br.

[2]PhD Student, Soil Science Department, Faculty of Agronomy, Federal University of Pelotas, CP 354, 96001-970, Capão do Leão-RS, Brazil.

[3]Agronomist, Precision Agricultural Consultant, Teutônia-RS, Brazil.

[4]Soil Physics Laboratory, Center for Nuclear Energy in Agriculture, University of São Paulo, CP 96, 13400-970, Piracicaba-SP, Brazil.

[5]Rural Engineering Department, Faculty of Agronomy, Federal University of Pelotas, CP 354, 96001-970, Capão do Leão-RS, Brazil.

1 Introduction

Frequently agricultural fields and their soils are considered homogeneous areas without well defined criteria for homogeneity, in which plots are randomly distributed in order to avoid the eventual irregularity effects. Randomic experiments are planned and when in the data analyses the variance of the parameters of the area shows a relatively small residual component, conclusions are withdrawn about differences among treatments, interactions, etc. If the residual component of variance is relatively large, which is usually indicated by a high coefficient of variation (CV), the results from the experiment are

considered inadequate. The high CV could be a cause of soil variability, which was assumed homogeneous before starting the experiment.

In Fisher´s classical statistics, which is mainly based on data that follows the Normal Distribution, each sampling point measurement is considered as a random variable Z_i, independently from the other Z_j. The adjustment of a set of data to the normal distribution does not guarantee the independence among soil samples, but which can be verified by the autocorrelation function. The main reason for this is that in the normal distribution frequency calculation the position at which each sample was collected in the field is disregarded.

More detailed studies of soil variability reveal that Fisher's classical statistics methods have limitations. Generally the data independence and normality hypotheses are not tested and, more than that, data independence has to be assured in the sampling design.

If the spatial distribution of soil measurements is observed and taken into consideration in the analysis, in many cases it is possible to take advantage and make use of this spatial variability. This is another way of planning experiments, new in Agronomy, but which imported not recent concepts from Geostatistics and from Time/spatial Series Analysis.

It has been emphasized that adjacent observations of a certain soil attribute are not completely independent and that this spatial dependence should be considered in the data analysis. In view of this, statistical tools like autocorrelograms, crosscorrelograms, semivariograms, etc have been used to study the soil attribute spatial variability and can potentially lead to a better management and understanding of the soil-plant-atmosphere interaction processes (Reichardt and Timm, 2004). Therefore, the use of statistical tools that considers the spatial structure dependence among observations have contributed to the

adoption of better agricultural practice managements as well as to the impact caused by them on the environment. The concern with the soil attribute spatial variability has been expressed in several works connected to agronomy (Warrick and Nielsen, 1980; Vieira et al., 1983; Sousa et al., 1999; Webster and Oliver, 2001; Wendroth et al., 2001; Tominaga et al., 2002; Timm et al., 2003; Wendroth et al., 2003; Timm et al., 2004; Iqbal et al., 2005; Mzuku et al., 2005; Grego et al., 2006; Lamhamedi et al., 2006; Terra et al., 2006; Timm et al., 2006; Novaes Filho et al., 2007; Silva et al., 2007).

Vieira (2000) reported that since the beginning of the 20th century, soil attribute spatial variability studies have been the target of researchers related to Soil Science. Among them, the author mentions: Smith, in 1910, studied the plot arrangements in experimental fields of corn yield varieties; Montgomery, in 1913, studied nitrogen effects on wheat yield; Waynick, in 1918, studied the soil nitrification spatial variability; and Waynick and Sharp, in 1919, characterized soil carbon and total soil nitrogen spatial variability, in different sampling arrangements.

The spatial variability of soil attributes can occur at different levels, and can be related to several factors: parent material origin, climate, relief, organisms and time, i.e., soil formation genetic processes and/or the effect of different soil management techniques from different agricultural uses (McGraw, 1994).

In agronomic experimentation soil and plant or atmosphere sampling methodologies are of fundamental importance. In classical statistics random sampling is recommended, while the regionalized variable techniques require coordinates of sampled points which are used in the analyses. In this case, sampling is carried out along transects, in equidistant intervals; or in grids in equidistant and irregular intervals, but with known coordinates.

In spatial variability studies when sampling is made in a grid the analysis requires the use of geostatistics, which was originated in South of Africa, when Mr. Krige, in 1951, working with gold concentration data in mining, concluded that it was difficult to find the meaning of the variances, if the distance between samples was not taken into account. Matheron, in 1963 and 1971, based on these observations, developed the Regionalized Variable Theory, containing the Geostatistics fundaments (Journel and Huijbregts, 1978). This theory is based on the fact that the difference between the values of the variable taken in two field points depends on the distance between them. Thus, the difference of a variable between two near points should be smaller than the difference of this variable for distant points. Therefore, each variable value carries strong information from its neighborhood, indicating a spatial continuity.

The classical statistics and the regionalized variable techniques are complementary, i.e. one does not exclude the other (Reichardt and Timm, 2004). When Geostatistics tools are used for characterizing the structures of the spatial distribution of the considered variables, the intrinsic hypothesis is assumed. This hypothesis is that the variogram function (to be seen later) depends only on the separation vector h (modulus and direction) and not on the location x_i (Journel and Huijbregts, 1978). It means that the structure of the variability between two observed variable values $z(x_i)$ and $z(x_i+h)$ is constant and, thus, independent of x_i.

Geostatistics applied to the precision agriculture concept has the objective of identifying in a random order among samples a spatial correlation structure, to estimate variable values in no-sampled points based on some known variable values in the sample (Kriging interpolator), and to study the relation among soil properties collected in the space. It also allows studying adequate resampling patterns.

A data exploratory analysis (classical descriptive statistics like the calculation of the mean, median, and mode values; the sample variance and the coefficient of variation values; the histogram and box-plot diagrams, etc) made previously before applying Geostatistics tools, is very important. The data normality distribution should be verified checking the presence of outliers or the need of a data transformation to normality. Having this in mind, a basic review of classical statistics concepts is presented below.

2 Review of Classical Statistics Concepts

2.1 Statistical position measures of a data set

These are used to determine the position of the observed variable within a data set. They have the objective of representing the center of a set of measurements. The arithmetic mean (AM) is the usual measure of central tendency. Denoting AM by $\bar{z}$ it is calculated considering that all observations z_i have the same weight in its calculation. Then, if we have a set of n observations z_i, i=1, 2, …, n, we can calculate their arithmetic mean by

$$\bar{z} = \frac{\sum_{i=1}^{n} z_i}{n} \tag{1}$$

Nothing in the equation (1) dictates where the observations should be taken, i.e. the locations of the observations are disregarded (Nielsen and Wendroth, 2003).

The median (Md) is the central value of a set of data when the observations are ranked in two equal parts: 50% of the values are below its value and 50% are above There are two different ways calculate the median , however, in both cases, the first step is to rank the observations from the lowest to the highest values.

- 1st case: when the number of observations n is odd: then we have to determine the most central position value (p) of the ranked data set, as follows

$$p = \frac{n+1}{2} \tag{2}$$

In this case, the median of a data set is the value that occupies the p position, i.e. $Md=z_p$.

- 2nd case: when the number of observations n is even: then we have to determine the two most central position values (p_1 and p_2 values) of the ranked data set, i.e.

$$p_1 = \frac{n+2}{2} \quad \text{and} \quad p_2 = \frac{n}{2} \tag{3}$$

Then, the median of a data set is the arithmetic mean of z_{p1} and z_{p2}:

$$Md = \frac{z_{p1} + z_{p2}}{2} \tag{4}$$

The mode (Mo) is the most typical value of a data set. It is the unique measure which cannot exist. If it exists it can occur more than once.

Quartiles, denoted by Q_i, i=1, 2 and 3, are three measurements which divide a ranked data set into four equal frequencies. The three measurements are:

- quartile first (Q_1): 25% of the values that are below and 75% are above of this measurement;
- quartile second (Q_2): 50% of the values are below and 50% are above of this measurement. This quartile corresponds to the median value, i.e. $Q_2=Md$;
- quartile third (Q_3): 75% of the values are below and 25% are above of this measurement.

To calculate the three quartiles, we have, initially, to rank the observations from the lowest to the highest values, and, after this, to calculate the p position value of the quartile in the ranked data set. There are two different cases to calculate the p position value of the quartile:

a) 1st case: when the number of observations n is odd:

- For Q_1 we have

$$p = \frac{n+1}{4} \tag{5}$$

- For Q_2 we have

$$p = \frac{2(n+1)}{4} \tag{6}$$

- For Q_3 we have

$$p = \frac{3(n+1)}{4} \tag{7}$$

b) 2nd case: when the number of observations n is even

- For Q_1 we have

$$p = \frac{n+2}{4} \tag{8}$$

- For Q_2 we have

$$p = \frac{2n+2}{4} \tag{9}$$

- For Q_3 we have

$$p = \frac{3n+2}{4} \tag{10}$$

2.2 Variation or dispersion measures of a data set

The dispersion measures of a data set describe the spread within the distribution of a set of measurements, and they are: range, interquartile range, variance, standard deviation and coefficient of variation.

The range (A) is the difference between the greatest and the smallest observations of a data set. Then, we have

$$A = Ls - Li \tag{11}$$

where Ls is the highest observation among z_i and Li is the lowest one.

The range measure is a less precise measurement, because it only uses extreme values of a data set in its calculation. For this reason, it is extremely influenced by the presence of outlier values in the data set (Piana and Machado, 2004).

The interquartile range, denoted by q, is the difference between the Q_3 and Q_1 quartiles, i.e.

$$q = Q_3 - Q_1 \tag{12}$$

Despite of being a less used dispersion measurement, q has an important characteristic which is that it is not influenced by the presence of outlier values in a data set.

The sample variance, denoted by s^2, is the most used dispersion measure to describe the spread of a set of measurements. The population variance σ^2 of a set of values is by definition given by

$$\sigma^2 = \frac{\sum (z_i - \bar{z})^2}{n} \tag{13}$$

where $\sum (z_i - \bar{z})^2$ is the square of the standard deviation σ. Bellow we shall replace the divisor (n) by (n-1) in equation (13), so that we can use the variance of a sample (s^2) to estimate σ^2, the population variance, without bias (Webster and Oliver, 2001). Like the mean, the sample variance is calculated based on all observations and only their magnitudes are involved in calculations, while their space coordinates are neglected (Nielsen and Wendroth, 2003).

The sample standard deviation (s) is the square root of s^2, i.e.

$$s = \sqrt{s^2} \tag{14}$$

It expresses the dispersion of the distribution in the same units as those in which the variable is measured, which facilitates its interpretation. In its calculation, the locations are also disregarded, providing a measure of the range or scatter of the observations within the undefined sample region.

The coefficient of variation (CV) is the most used dispersion measure when we are interested in comparing the variability of different data sets. In cases of comparison of variances of different data sets or when a studied soil property was measured in two different regions to give similar sample variance values with different means, the coefficient of variation can elucidate these cases (Webster and Oliver, 2001). The CV coefficient is calculated from the ratio between the standard deviation and the mean and it is usually presented as a percentage

$$CV(\%) = \frac{s}{\bar{z}}.100 \tag{15}$$

The CV is a relative dispersion measure because it is a standardized measurement of the sample variance. From this, it is useful for comparing the variation of different sets of

observations of the same property or for comparing the variation of sets of observations of the different properties (Nielsen and Wendroth, 2003). According to Timm et al. (2003a), the CV can also be used to describe the frequency distribution of the observations, and, when being large, it indicates that the arithmetic mean is not appropriate to characterize the set of data due to their high variability.

2.3 Moments, skewness and kurtosis coefficients of a data probability distribution

The moments, denoted by m_r, are measures which are calculated with the aim of studying data probability distribution behaviors. An r order moment centered on b value is given by

$$m_r = \frac{\sum (z_i - b)^r}{n} \tag{16}$$

when b is equal to $\bar{z}$, we have the r order moments centered on the $\bar{z}$ mean value and they can be represented by m_r. Then, we have:

- for r = 1

$$m_1 = \frac{\sum (z_i - \bar{z})^1}{n} \tag{16a}$$

- for r = 2

$$m_2 = \frac{\sum (z_i - \bar{z})^2}{n} \tag{16b}$$

- for r = 3

$$m_3 = \frac{\sum (z_i - \bar{z})^3}{n} \tag{16c}$$

- for r = 4

$$m_4 = \frac{\sum (z_i - \bar{z})^4}{n} \tag{16d}$$

which are used in the definitions given below.

The skewness coefficient (a_3) is a measure which characterizes the degree of asymmetry of a data distribution around its mean. It is calculated from the ratio between the third moment and the second one:

$$a_3 = \frac{m_3}{m_2 \sqrt{m_2}} \tag{17}$$

Based on its absolute value and its signal, the skewness coefficient is used for characterizing the symmetry or asymmetry of a distribution, as follows:

- when a_3 is less than zero this indicates a longer tail on the left hand side of a histogram with the median being larger than the mean ;
- when a_3 value is equal to zero ,the mean and median are coincident, i.e. the distribution is symmetric;
- when a_3 is higher than zero this indicates a longer tail on the right hand side of a histogram with the median is smaller than the mean .

The kurtosis coefficient (a_4) is a measure which characterizes the peakness or flatness of a data distribution. It is calculated from the ratio between the fourth moment and the second one (Webster and Oliver, 2001)

$$a_4 = \frac{m_4}{(m_2)^2} - 3 \tag{18}$$

Its significance relates mainly to the normal distribution, for which $a_4=0$. Distributions that are more peaked than normal have $a_4 > 0$; flatter ones have $a_4 < 0$.

2.4 Frequency distribution: histogram, normal plot and box-plot diagram

Any set of measurements may be divided into several classes, and we may count the number of individuals in each class. The resulting set of frequencies constitutes the frequency distribution (Piana and Machado, 2004). A soil property, for example, is a continuous variable across a given field being, therefore, an infinite population of a given area. To infer characteristics of this population, i.e., describe it within the area, a statistical approach becomes necessary, which is performed using a finite set of observations, which is assumed to represent the population. In this case we obtain an observed value of the sample mean $\bar{z}$ (equation 1), which is an estimate of the true or population mean (μ), which represents the whole population. For our particular case of a set of infinite elements, μ is never known. The intention of any statistical approach is to find out which theoretical distribution best adjusts to the sampled distribution, so that deductions can be made in relation to the true distribution.

The frequency distributions can graphically be represented by a histogram. To construct a histogram, observations have to be divided in classes according to their magnitude, and the number of observations of each class is counted. From this information a graph of bars is constructed, the height of each bar being proportional to the number of observations of the class. Superposing the theoretical curve of the normal distribution, we can verify visually how close the heights of the bars coincide with the theoretical line, and conclude how well the observations follow the normal distribution. Figure 1 illustrates a histogram of a clay content data set.

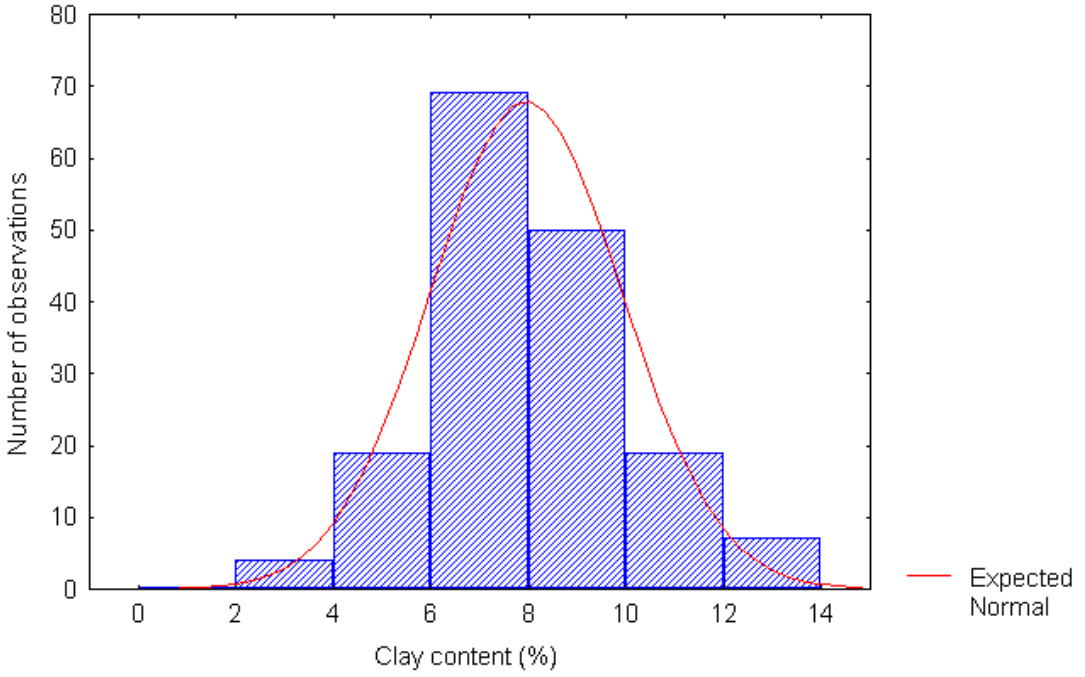


Figure 1 – Illustration of a histogram constructed from a clay content experimental data set measured in an Entisol Quartzipsamment cultivated with irrigated grapevines in Northeast Brazil.

To construct a normal plot, observations are arranged into ascending magnitudes and their logarithms are plotted in relation to their cumulative probability values. The better the observed values fit to the straight line the more we can consider that the data set follows the normal distribution. The linearity of the normal plot can be quantified using the Kolmogorov-Smirnov (K-S) test (Landim, 2003). An illustration of a normal plot is shown in Figure 2.

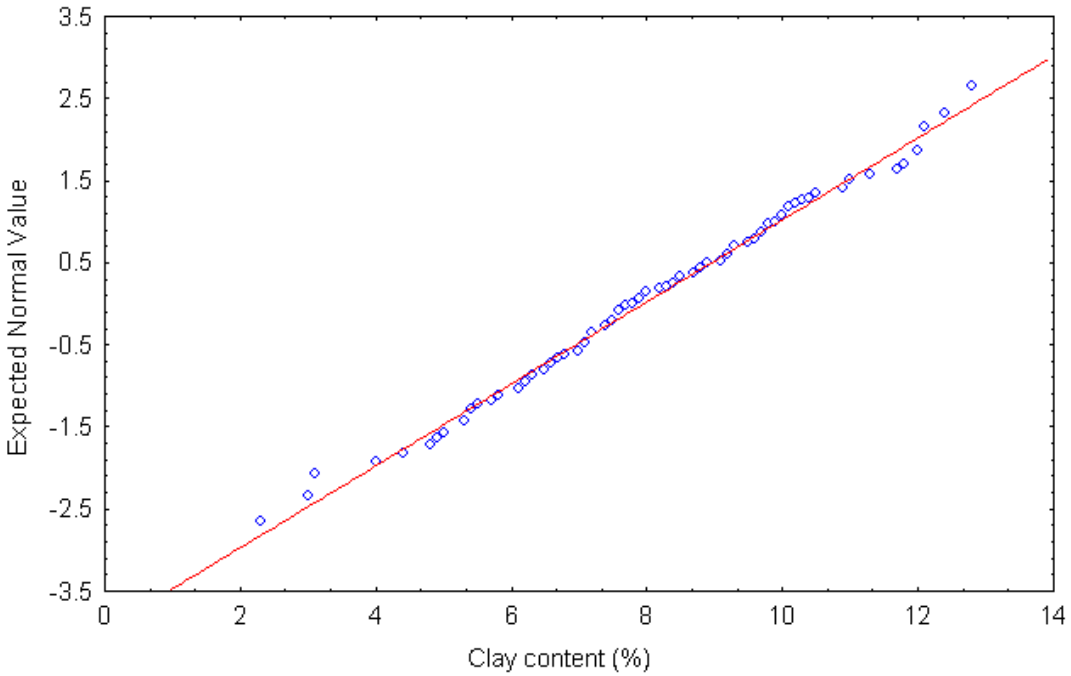


Figure 2 – Illustration of the normal plot constructed from a clay content experimental data set measured in an Entisol Quartzipsamment cultivated with irrigated grapevines in Northeast Brazil.

Since the clay content observations fall relatively nicely on the straight line in the normal plot (Figure 2), we can visually consider that their values are normally distributed. Had they not formed a straight line, the observations would not be normally distributed (Nielsen and Wendroth, 2003). On a statistical basis the normality is judged by the K-S test, as mentioned before.

The box-plot diagram has a box enclosing the interquartile range, a line showing the median, and lines extending from the limits of the interquartile range to the extremes of the data (Webster and Oliver, 2001). Figure 3 illustrates a box-plot diagram and its components.

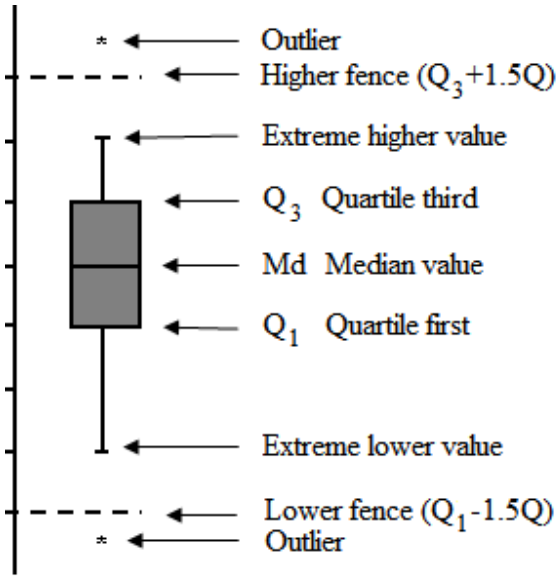


Figure 3 – Illustration of a box-plot diagram and its components (figure extracted from Piana and Machado, 2004).

The central position of the data set is represented by the median (Md, equations 2 or 4) and the dispersion by the interquartile range (q, equation 12). The outlier values have an individual representation by using a letter or a symbol.

Both the histogram and the box-plot allow us to analyze the data distribution, to see how it arranges around the mean or median, and identify extreme values.

Data sets might contain outlier observations. Special attention has to be taken in these cases because it is difficult to judge if an outlier is a wrong measurement or it is a true one belonging to a long tailed non-symmetric distribution. To identify outliers within a data set, two measures are used which are called lower fence (LF) and higher fence (HF). LF is calculated subtracting from the quartile first (Q_1) 1.5 times the interquartile range (q) and HF is calculated adding to Q_3 1.5 times q, i.e.

$$LF = Q_1 - 1.5q \qquad \text{and} \qquad HF = Q_3 + 1.5q \tag{19}$$

The observations falling out of this interval (between LF and HF) are considered outliers. When an observation appears as an outlier, its origin should be investigated. With the integrity of such an observation being questioned, the investigator tries to learn whether some kind of a mistake was made during its observation. Without the exact locations of the observations being known, it is not possible to resample the location of the suspect observation (Nielsen and Wendroth, 2003).

2.5 The normal distribution

The normal distribution is central to the statistical theory. A large diversity of natural sets, such as those of the soil, is distributed in a way that approximates the normal probability distribution, being therefore widely used in statistical analyses (Webster and Oliver, 2001). This distribution is bell shaped and has a maximum that coincides with the mean, being continuous and symmetrical. The hypothesis of normality is the basis for the adoption of this distribution for which most of the statistical models have been developed.

The normal distribution is defined for a continuous random variable Z in terms of the probability density function, f(z), as follows

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{(z-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < Z < +\infty \tag{20}$$

where μ is the mean of the distribution and σ^2 is the variance. For each μ and σ values (remembering that σ is the population standard deviation) there is a different normal distribution. From this, the calculation of the area below the normal curve (i.e. the probability of the distribution) should be always made for each μ and σ specific values which becomes a hard work. To avoid this, the normalized normal distribution (also called

standard normal distribution in the literature) is defined, which allows us to study any normally distributed variable for any μ and σ values. By definition it is the normal distribution of a Z variable which has the mean equal to zero (μ=0) and the standard deviation equal to one (σ=1). For a Z continuous random variable, its probability density function, f(z), is

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}, \quad -\infty < Z < +\infty \qquad (20a)$$

The equation (20a) facilitates the calculation of the area below the curve which can be found in the normalized normal distribution tables, available in many statistics books [for example, Landim´s book (2003), among others].

The normal distribution has the following properties:

- The peak of the normal distribution is at the mean of the distribution, i.e. the z = μ;
- It is symmetrical in relation to its mean where μ = Md =Mo;
- It has two points of inflexion, one on each side of the mean at a distance σ, i.e. μ - σ and μ + σ;
- The ordinate f(z) at any given value of z is the probability of the density at z and the total area under the normal curve is equal to 1 or 100%, i.e. the total probability of the distribution (Webster and Oliver, 2001);
- More than two-thirds (68.25%) of the probability distribution lies within one standard deviation of the mean, i.e. between μ - σ and μ + σ; 95.44% lies in the range μ - 2σ to μ + 2σ; and 99.74% lies within three standard deviations of the mean, i.e. between μ - 3σ and μ + 3σ (Webster and Oliver, 2001).

Assuming that the distribution of frequencies of a property Z is approximately normal, the arithmetic mean $\bar{z}$ is taken as a good estimator of the central position of the values of the population. In this way, this mean is taken as the estimate of the property in locations not sampled, making it is necessary to identify the precision of this mean as an estimator. To do this, the parameters that quantify the dispersion of the data around this mean have been used, like s, CV or confidence limits. However, in many instances distributions are far from normal, and these departures from normality give rise to unstable estimates that interfere in the interpretation, making it less certain. In this situation, we can be in some doubt as to which measure of centre is to taken if the distribution is skewed. Perhaps more seriously, statistical comparisons between means of observations are unreliable if the variable is skewed because the variances are likely to differ substantially from one set to another (Webster and Oliver, 2001).

3 Statistical tools most commonly used to analyze and characterize the spatial variability

3.1 Autocorrelation Function ACF

Spatial (or temporal) series can be studied as being the realization of a particular stochastic process, based on probability laws. The correlation that exists between adjacent observations frequently limits the application of classical statistics methods which are based on the fact that observations should be independent and identically distributed.

After sampling a variable Z along a transect, for example, its mean and variance are calculated to reflect the sampled population, assuming that the set is representative and obtained randomly. In many cases the observations are not independent of each other, and it is possible to calculate an autocorrelation coefficient, which plotted as a function of the

distance between observations will indicate their level of auto-dependence. For stationary processes (those in which the static properties are independent of space or time), the covariance between the observations is a function of the number of lags h (h=0, the very same point; h=1, the first neighbor; h=2, the second neighbor; h=h, the h^{th} neighbor) between their sampling points. Time series are collected along time at intervals of α minutes, hours, months, etc and space series along transects (or grids) at spacings of α ($x_i - x_{i-1} = \alpha$), in cm, m, km, etc. The covariance C(h) between such variables at different lag distances h, given by Salas et al. (1988) is

$$C(h) = \frac{1}{n-h}\sum_{i=1}^{n-h}\left[z(x_{i+h}) - \bar{z}\right]\left[z(x_i) - \bar{z}\right] \tag{21}$$

If C(h) is normalized dividing it by the variance s^2 of the sample, we obtain the coefficient r(h) of the autocorrelation function (we say auto- because it is a correlation between value of the same variable Z, but measured at different positions):

$$r(h) = \frac{C(h)}{s^2} \tag{22}$$

which manifests values between +1 and -1. It is important to note that for the calculation of r(h), the observations z_i of the random variable Z have to be collected at regularly spaced intervals α. The values of r(h) for h = 0, which represents the correlation between $z(x_i)$ and $z(x_i)$ is obviously equal to 1. For the first neighbor pairs $z(x_i)$ and $z(x_{i+1})$ for a distance of one lag α (h = 1), a value of r(1) can be obtained using equations (21) and (22). The same procedure is used for second neighbor pairs [$z(x_i)$ and $z(x_{i+2})$], and further neighbors (h = 3, 4, ...) obtaining a r(h) value for each h. Plotting r as a function of h we obtain the autocorrelogram of the variable Z. Figure 4 illustrates an autocorrelogram plot of a soil water content data set extracted from Timm et al. (2006).

The next step is the calculation of the fiducial intervals of r, to recognize if they are significant or not, and in this way define the length interval αh in which the spatial dependence of the variable is significant. One way to measure the autocorrelation confidence interval CI is using the accumulated probability function (e.g., ± 1.96 for a 95% probability level) for the normalized distribution function (Davis, 1986), and the number of observations (n-h). Therefore,

$$CI = \pm \frac{p}{\sqrt{n - h}} \tag{23}$$

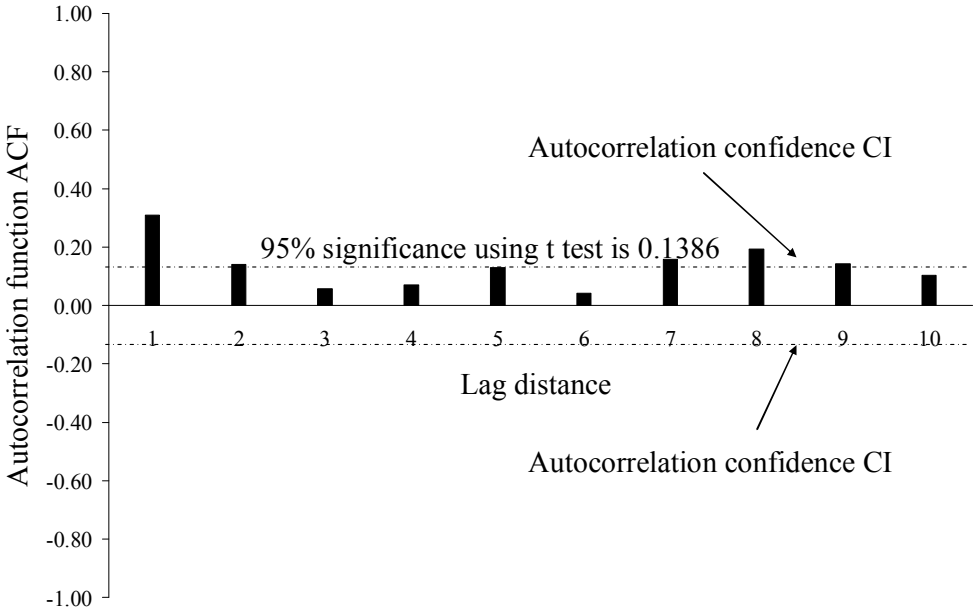


Figure 4 – Illustration of an autocorrelogram plot of a soil water content data set (figure extracted from Timm et al., 2006).

3.2 Crosscorrelation function CCF

Having two sets of variables $Z(x_i)$ and $W(x_i)$ observed at the same locations x_i (or same times t_i), their spatial crosscorrelation structure can be analyzed calculating coefficients of crosscorrelation. Although each variable has its own autocorrelogram, an analysis of their crosscorrelation indicates to which distance (or time interval) one is related to the other. The coefficient r_c of the crosscorrelation function will be also a function of h, and describes the degree of linear association between both variables (Davis, 1986; Shumway, 1988; Wendroth et al., 1997).

The coefficients of the crosscorrelation function $r_c(h)$, between the variables Z and W, separated by distances αh, or by a lag number h, are calculated by:

$$r_c(h) = \frac{cov_{ZW}(h)}{s_Z \times s_W} \tag{24}$$

where

$$cov_{ZW}(h) = \frac{1}{n-h} \sum_{i=1}^{n-h} \left[z(x_i) - \bar{z} \right] \left[w(x_{i+h}) - \bar{w} \right] \tag{25}$$

and s_Z^2 is the variance of Z

$$s_Z^2 = \frac{1}{n} \sum_{i=1}^{n} \left[z(x_i) - \bar{z} \right]^2 \tag{26}$$

and s_W^2 is the variance of W

$$s_W^2 = \frac{1}{n} \sum_{i=1}^{n} \left[w(x_i) - \bar{w} \right]^2 \tag{27}$$

A plot of r_c as a function of h represents the crosscorrelogram. An illustration of a crosscorrelogram plot is shown in Figure 5. For h = 0 (observations taken at the same position x_i), the value $r_c(0)$ given by equation (24) is the linear regression coefficient obtained through classical statistics. For the first neighbor pairs $[z(x_i), w(x_{i+1})]$ collected at a distance α in one direction (h = 1), we obtain the coefficient r_c (1), and for the other direction (h = -1) the coefficient $r_c(-1)$. This is because in the case of two variables, each of them has different neighbors for each direction, i.e., we have two pairs – (z_i, w_{i+1}) and (z_i, w_{i-1}). The same procedure is used for more distant neighbors, obtaining values of $r_c(h)$ and $r_c(-h)$. A crosscorrelogram indicates how far two different observations are spatially related (Wendroth et al., 1997).

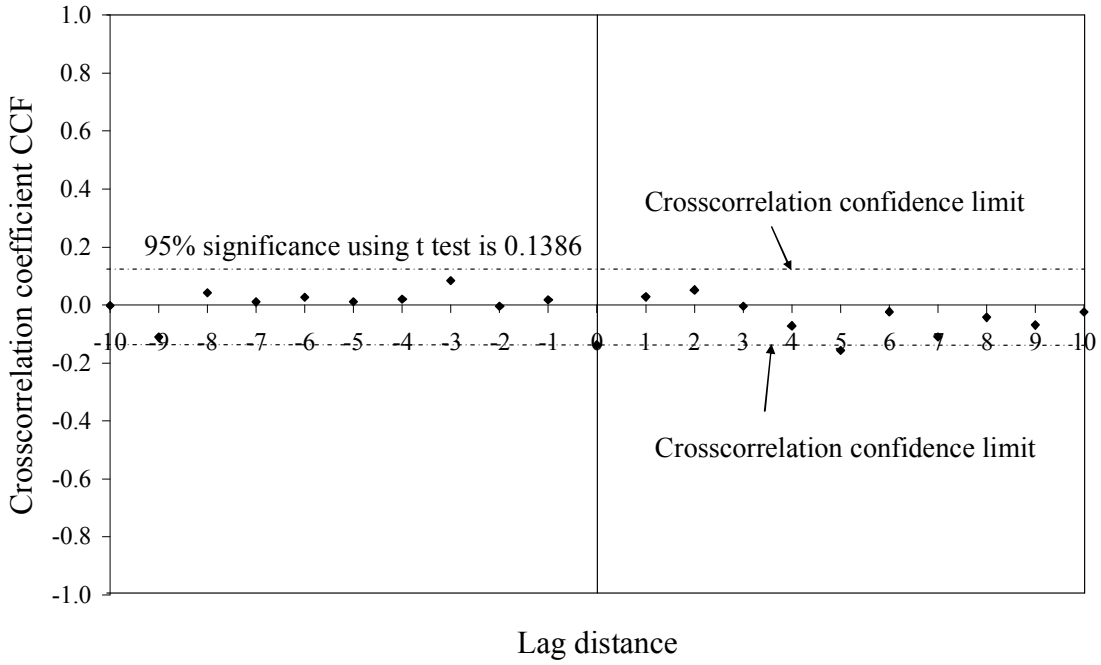


Figure 5 – Crosscorrelogram plot between soil water content and soil bulk density data sets (figure extracted from Timm et al., 2006).

According to Nielsen and Wendroth (2003), it is more difficult to estimate the significance of $r_c(h)$ as compared to r(h). Significance tests like the t test are usually based on the assumption that the observed values of $z(x_i)$ and $w(x_i)$ are normally distributed and independent among themselves. Taking this into consideration, the significance level of r_c is, in general, given by

$$t = \sqrt{\frac{(n-h)-2}{1-r_c^2}} \tag{28}$$

where (n-h) is the number of pairs used for the calculation of r_c. The level of significance of the test is obtained by comparing the value of t in equation (28) with critical values of t for (n-2) degrees of freedom. The crosscorrelation function is, in general, not symmetric, i.e., $r_c(h) \neq r_c(-h)$. Note that in the case of the autocorrelation there is symmetry, r(h) = r(-h). When there is a physical relation between Z and W, the crosscorrelogram will tend to symmetry (Nielsen and Wendroth, 2003).

3.3 Semivariogram

To evaluate if there is spatial dependence between the values of the variable samples in a grid the semivariogram can be used, which describes the structure of the spatial dependence among neighbor points. The experimental semivariogram is a graph which represents the estimative of data semivariances [$\gamma(h)$] as a function of the vector h that separates them. Here the h vector indicates a vector of modulus $|h|$ and two-dimensional coordinates (h_x, h_y). The semivariance estimative can be obtained by the following equation:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(x_i) - z(x_i + h)]^2 \quad (29)$$

where $\gamma(h)$ is the estimated semivariance between two experimental measures [$z(x_i)$ and $z(x_i+h)$] at any two points separated by the vector h and N(h) is the number of experimental pairs [$z(x_i)$,$z(x_i+h)$] of data separated by the vector h (Journel and Huijbregts, 1978).

Then, the experimental semivariogram is adjusted to the best fitted mathematical model using a chosen statistical measure (for example, residual sums of squares, r^2 coefficient, etc). The adjusted mathematical model is called theoretical semivariogram model. Within the distance over which pairs of observations remain spatial correlated the geostatistics could be applied efficiently.

The choice of the adjusted mathematical model to the experimental semivariogram is of great importance, because it influences further results. The adjusted model should describe the phenomenon in the field, and the best fitted theoretical model to the experimental semivariogram can be performed by crossed validation, for example.

Figure 6 illustrates a semivariogram (and its parameters) with characteristics close to the ideal. Its pattern represents what, intuitively, is expected from the field data, i.e., the [$z(x_i)$ - $z(x_i + h)$] differences increase once h lag distance increases.

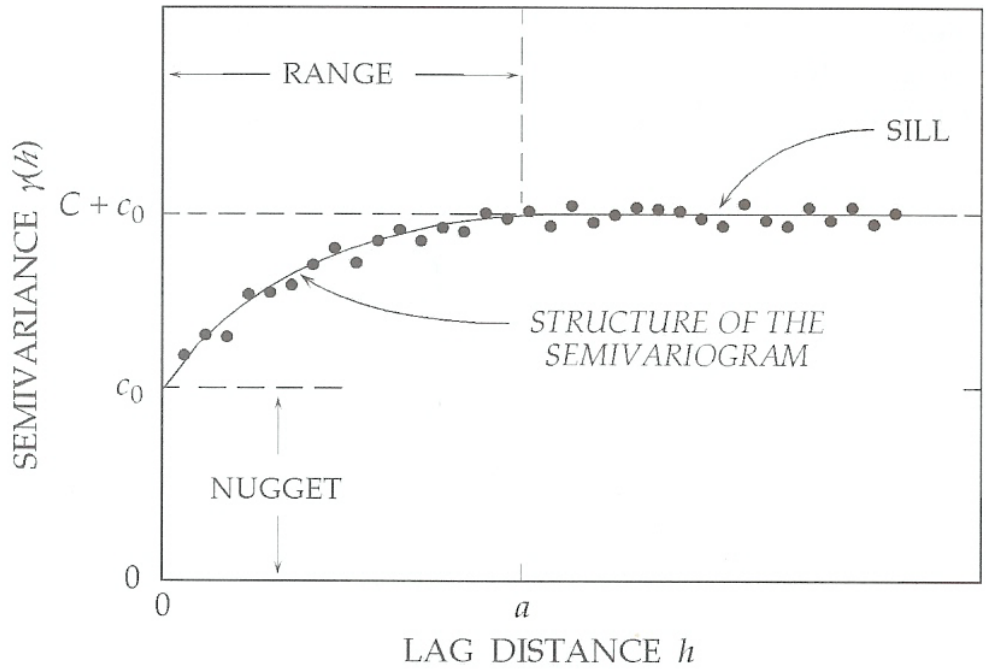


Figura 6 - Parameters describing a semivariogram (figure extracted from Nielsen and Wendroth, 2003).

The semivariogram parameters can be observed in the figure above:

- Range (a): is the distance in which the sampled variable values are spatially correlated;
- Sill ($C+c_0$): is the maximum semivariance value calculated for a transitional or bounded semivariogram. From this value, there is no spatial dependence between the sampled variable values, because the difference between the pairs of sample variances $\{Var[Z(x_i) - Z(x_i+h)]\}$ becomes constant with the increase of h lag distance and very close to the variance of independent variables;
- Nugget effect (c_o): by definition, $\gamma(h=0) = 0$. However, in the practice, as h lag distance tends to zero, $\gamma(h)$ tends to a positive value called nugget effect (c_0). This value reveals the discontinuity of the semivariogram for shorter distances than the shortest

distance among samples. Part of this discontinuity can be due to measurement errors, but it is difficult to quantify if the higher contribution is from measurement errors or from the spatial variability in small scale which is not shown by sampling.

- Contribution (C): is the difference between the sill ($C+c_0$) and the nugget effect (c_0). It is the structural variance of the data set.

3.3.1 Mathematical models used to adjust experimental semivariograms

The adjustment of a mathematical model to the experimental semivariogram is a very important aspect in the Regionalized Variable Theory applications and it can be considered one of the highest sources of uncertainties and controversies in this approach. All further geostatistics calculations depend on this stage (Vieira et al., 1983; Guimarães, 2004). From this, if the adjusted model is incorrectly adjusted, all further calculation will fail.

Nowadays, there are commercial softwares (for example, GS+ software developed by Gamma Design Software, 2004) which have different ways of adjusting a mathematical model to the experimental semivariogram. As a rule the user should choose the easiest adjusted mathematical model. The essential condition to adjust a mathematical model is that it represents the trend of $\gamma(h)$ in relation to h lag distance increases and that $\gamma(h)$ is a positive-definite function, i.e. $\gamma(h) \geq 0$ and $\gamma(-h) = \gamma(h)$, for any h lag distance increases (Journel and Huijbregts, 1978; Isaaks and Srisvatava, 1989; Webster and Oliver, 2001).

The main mathematical models used in the literature to adjust experimental semivariograms are (Nielsen and Wendroth, 2003):

a) Transitional or bounded models

- Pure nugget model

$$\gamma(h) = \begin{cases} 0 & \text{h} = 0 \\ C & \text{h} > 0 \end{cases} \tag{30}$$

- Linear model

$$\gamma(h) = \begin{cases} \dfrac{\text{C h}}{\text{a}} & 0 \le \text{h} \le \text{a} \\ C & \text{h} > \text{a} \end{cases} \tag{31}$$

- Spherical model

$$\gamma(h) = \begin{cases} C\left[\dfrac{3h}{2a} - \dfrac{1}{2}\left(\dfrac{h}{a}\right)^3\right] & 0 \le \text{h} \le \text{a} \\ C & \text{h} > \text{a} \end{cases} \tag{32}$$

- Exponential model

$$\gamma(h) = C\left[1 - \exp\left(-h/a\right)\right] \quad , \quad \text{h} \ge 0 \tag{33}$$

- Gaussian model

$$\gamma(h) = C\left\{1 - \exp\left[-\left(h/a\right)^2\right]\right\} \quad , \quad \text{h} \ge 0 \tag{34}$$

b) Nontransitional or unbounded models

- Linear model

$$\gamma(h) = mh \quad , \qquad h \geq 0 \tag{35}$$

- Power model

$$\gamma(h) = mh^{\alpha} \quad , \quad h \geq 0; 1 < \alpha < 2 \tag{36}$$

Bounded semivariograms occur wherever the variance of all of observations remains constant throughout the sampled domain. Unbounded semivariograms are manifested when the variance of all observations within a domain is not constant (Nielsen and Wendroth, 2003).

3.4 Ordinary kriging – a geostatistical interpolation method

An interpolation method has a function of estimating unknown values within a domain based on some already known values (Nielsen and Wendroth, 2003). There are deterministic interpolation methods, such as: polygon methods, inverse distance weighting method, among others. These methods, however, do not estimate the error associated to each interpolated value, which can be obtained by a geostatistics interpolator method called kriging (Journel and Huijbregts, 1978; Webster and Oliver, 2001).

The semivariogram is the geostatistical tool which allows describing the structure of the spatial dependence of a studied variable. Using the spatial variance structure available in a semivariogram, the kriging interpolator provides the best linear unbiased estimate of an unmeasured value calculated from values measured in a local neighborhood (Journel and Huijbregts, 1978; Nielsen and Wendroth, 2003).

The kriging interpolator is considered the best linear interpolator because it produces unbiased estimatives with minimum estimated variance (Webster and Oliver,

2001). In the linear kriging the estimates are data weighted linear combinations. These weights vary as a function of the separation distance among the locations of the variable ue to be estimated and the location of the observed variable involved in the estimative. This unmeasured variable is, therefore, calculated by solving a kriging system of equations (Journel and Huijbregts, 1978; Isaaks and Srivastava, 1989).

To apply the ordinary kriging for interpolation it is assumed that: $z(x_1)$, $z(x_2)$,..., $z(x_n)$ are measured values of a random variable Z at positions x_1, x_2,...,x_n ; and that the theoretical semivariogram of the studied variable has already been determined. Therefore, the objective is to estimate values of z* at positions x_0 $[z^*(x_0)]$, assuming that its value is the linear function of the known values $z_i(x_i)$,

$$z^*(x_0) = \sum_{i=1}^{N} \lambda_i z_i(x_i) \tag{37}$$

where N is the number of measurements of the variable Z involved in the $z^*(x_0)$ estimative and λ_i are the associated weights to each measured value $z_i(x_i)$ of Z.

The best $z^*(x_0)$ estimation is given if:

a) the expected error is zero, i.e., the estimation is unbiased

$$E\left[z^*(x_0) - z(x_0)\right] = 0 \tag{38}$$

b) the estimated variance is minimum, i.e.,

$$Var\left[z^*(x_0) - z(x_0)\right] = \text{minimum} \tag{39}$$

To ensure that the z* estimation is unbiased, Webster and Oliver (2001), assume:

$$\sum_{i=1}^{N} \lambda_i = 1 \tag{40}$$

To obtain a minimum estimated variance under the constraint of equation (40), the Lagrange multiplier is applied for solving the following kriging system equations:

$$\sum_{i=1}^{N} \lambda_i \gamma(x_i, x_j) + \psi = \gamma(x_i, x_0) \quad , \quad \text{i} = 1 \text{ a N} \tag{41}$$

where ψ is the Lagrange multiplier. Values of the semivariance between $z(x_i)$ and $z(x_j)$ [$\gamma(x_i,x_j)$] and those between $z(x_i)$ and $z(x_0)$ [$\gamma(x_i,x_0)$] are obtained from the theoretical semivariogram model. The (N+1) equations in (40) and (41) are solved for the N+1 unknowns λ_1, λ_2,… λ_n and ψ (Nielsen and Wendroth, 2003). The associated minimum kriging variance of each estimative is calculated by the following expression:

$$\sigma_E^2 z^*(x_0) = \psi + \sum_{i=1}^{N} \lambda_i \gamma(x_i, x_0) \tag{42}$$

The kriging equations can be represented in a matrix from. For ordinary kriging they are

$$[A][\lambda] = [b] \tag{43}$$

where

$$[A] = \begin{bmatrix} \gamma(x_1,x_1) & \gamma(x_1,x_2) & \dots & \gamma(x_1,x_n) & 1 \\ \gamma(x_2,x_1) & \gamma(x_2,x_2) & \dots & \gamma(x_2,x_n) & 1 \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \gamma(x_n,x_1) & \gamma(x_n,x_2) & \dots & \gamma(x_n,x_n) & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}; [b] = \begin{bmatrix} \gamma(x_1,x_0) \\ \gamma(x_2,x_0) \\ . \\ . \\ . \\ \gamma(x_n,x_0) \\ 1 \end{bmatrix}; [\lambda] = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ . \\ . \\ . \\ \lambda_n \\ \mu \end{bmatrix} \tag{44}$$

where matrix A is inverted, and the weights and the Lagrange multiplier are obtained as

$$[\lambda] = [A]^{-1}[b] \tag{45}$$

The σ_E^2 kriging variance, in matrix notation, is given by:

$$\sigma_E^2 z^*(x_0) = [\lambda]^t [b] \quad (46)$$

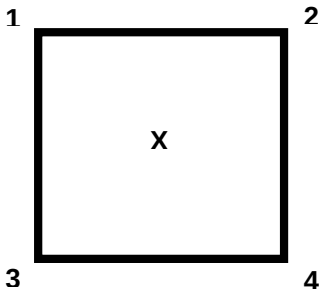
where $[\lambda]^t$ matrix is the transposed matrix of $[\lambda]$.

For a better understanding of the use of the kriging interpolator method, two illustrative examples follow:

- **Example 1**: proposed by Landim (2003)

In a hypothetical situation in which 4 copper ore deposit thicknesses were measured at locations with coordinates (x_i,y_i) as shown in the table below, we want to estimate the value of z at position (15,15). The following linear semivariogram model was previously adjusted:

Linear semivariogram model: $\gamma = 5h$



locations	x_i (Km)	y_i (Km)	z_i (m)
1	0	30	500
2	30	30	450
3	0	0	550
4	30	0	490
x	15	15	?

The square grid (30 km x 30 km) was established (figure below) and the distance between the measured variable values are:

d(1-2) = d(1-3) = d(2-4) = d(3-4) = 30 km;

d(1-4) = d(2-3) = 42.43 km;

d(1-x) = d(2-x) = d(4-x) = 21.21 km;

Using the linear semivariogram model, the distances correspond to the following calculated semivariances:

21.21 km = 106.05 km^2

30.00 km = 150.00 km^2

42.43 km = 212.15 km^2

From this, we can construct the ordinary kriging system equations:

$$\begin{bmatrix} 0 & 150 & 150 & 212.15 & 1 \\ 150 & 0 & 212.15 & 150 & 1 \\ 150 & 212.15 & 0 & 150 & 1 \\ 212.15 & 150 & 150 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \mu \end{bmatrix} = \begin{bmatrix} 106.5 \\ 106.5 \\ 106.5 \\ 106.5 \\ 1 \end{bmatrix}$$

$$[A] \qquad [\lambda] \qquad [B]$$

which is solved by

$$[\lambda] = [A]^{-1}[B]$$

$$[A]^{-1} = \begin{bmatrix} -0.00520 & 0.00285 & 0.00285 & 0.00049 & 0.25000 \\ 0.00285 & -0.00520 & -0.00049 & 0.00285 & 0.25000 \\ 0.00285 & -0.00049 & -0.00520 & 0.00285 & 0.25000 \\ -0.00049 & -0.00520 & 0.00285 & -0.00520 & 0.25000 \\ 0.25000 & 0.25000 & 0.25000 & 0.25000 & -128.03750 \end{bmatrix}$$

$$[\lambda] = \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \\ -21.987 \end{bmatrix}$$

As expected from a regular distribution of the observed variable points, each one of them has a 0.25 weight for estimating z* at location (15,15). Then

$$z^*(x) = 0.25 \times (500) + 0.25 \times (450) + 0.25 \times (550) + 0.25 \times (450) = 497.50m$$

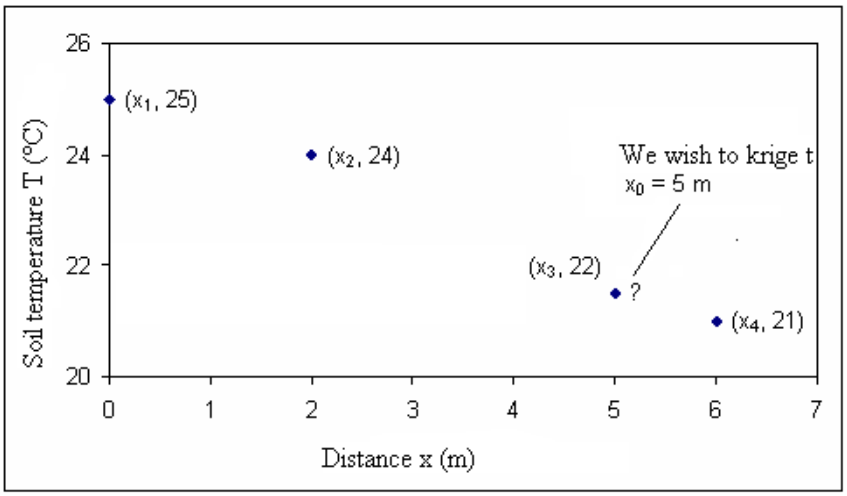
The associated kriging variance to this estimative is:

$$\sigma_k^2 z^*(x) = 0.25 \times (106.05) + 0.25 \times (106.05) + 0.25 \times (106.05) + 0.25 \times (106.05) - 21.9875 =$$
$$\sigma_k^2 z^*(x) = 84.063\, m^2$$

Assuming that the estimated values of Z have a normal distribution and that 95% of this data distribution is in the ± 1.96 x standard deviation intervals, we have that the fiducial limits are of the order of ± 9.169 x 1.96 = ± 18 m. Then, the estimative of z value at location (15,15) is: 497.50 m ± 18 m.

- **Example 2**: proposed by Nielsen and Wendroth (2003)

Soil temperature measurements were taken at four locations along a spatial transect, 2 m apart from each other, as illustrated in the figure below (extracted from Nielsen and Wendroth, 2003). We want to estimate the soil temperature value at a distance of 5 m from the beginning of the transect. Soil temperature measurements were 25, 24, 22 and 21 ^{o}C at locations x_1, x_2, x_3 and x_4, respectively.



The linear semivariogram model ($\gamma = 1.125h$) was previously adjusted. The distance between the soil temperature measurement pairs are:

$d(x_3\text{-}x_4) = 2$ m;

$d(x_1\text{-}x_3) = d(x_2\text{-}x_4) = 4$ m;

$d(x_1\text{-}x_4) = 6$ m.

Using the linear semivariogram model, the distances correspond to the following calculated semivariances:

γ (h=2 m) = 2.25 m^2; γ (h=4 m) = 4.5 m^2; γ (h=6 m) = 6.75 m^2.

From this, we can construct the ordinary kriging system equations:

$$\begin{bmatrix} 0 & 2.25 & 4.50 & 6.75 & 1 \\ 2.25 & 0 & 2.25 & 4.50 & 1 \\ 4.50 & 2.25 & 0 & 2.25 & 1 \\ 6.75 & 4.50 & 2.25 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \\ \mu \end{bmatrix} = \begin{bmatrix} 5.625 \\ 3.375 \\ 1.125 \\ 1.125 \\ 1 \end{bmatrix}$$

$$[A] \qquad [\lambda] \quad [B]$$

which is solved by

$$[\lambda] = [A]^{-1}[B]$$

$$[A]^{-1} = \begin{bmatrix} -0.2222 & 0.2222 & -4.11E^{-17} & -3.29E^{-17} & 0.5 \\ 0.2222 & -0.4444 & 0.2222 & 0 & -1.97E^{-16} \\ 0 & 0.2222 & -0.4444 & 0.2222 & 0 \\ -2.47E^{-17} & 0 & 0.2222 & -0.2222 & 0.5 \\ 0.5 & 0 & -1.67E^{-16} & 0.5 & -3.375 \end{bmatrix}$$

$$[\lambda] = \begin{bmatrix} 0.00 \\ 0.00 \\ 0.50 \\ 0.50 \\ -3.38 \end{bmatrix}$$

This means that x_1 and x_2 locations have the same associated weights and equal to zero on estimating x_0; x_3 and x_4 locations have the same associated weights and equal to 0.5 on estimating x_0. Substituting these values as well as the measured soil temperatures into equation 37, we obtained

$$z^*(x_0) = 0 \times (25) + 0 \times (24) + 0.5 \times (22) + 0.5 \times (21) = 21.5°C$$

The associated kriging variance to this estimative is:

$$\sigma_k^2 z^*(x_0) = 0 \times (5.625) + 0 \times (3.375) + 0.5(1.125) + 0.5(1.125) + 0 = 1.125°C^2$$

Then, the estimated soil temperature at 5 m location is of 21.5°C with a standard deviation of 1.060 °C. It is possible to note that the values of λ_i weight equally for the two positions close to x_0 and contribute with equal weights to the estimate while those for greater distances are null.

4 ACKNOWLEDGMENTS

To CNPq and FAPERGS for financial support.

5 CITED LITERATURE:

DAVIS, J.C. Statistics and data analysis in geology. 2nd ed. New York: Wiley and Sons, 1986.

GAMMA DESIGN SOFTWARE. GS+: Geostatistics for the Environmental Sciences. Plainwell: Gamma Design Software, 2004.

GREGO, C.R.; VIEIRA, S.R.; ANTONIO, A.M.; DELLA ROSA, S.C. Geostatistical analysis for soil moisture content under the no tillage cropping system. Scientia Agricola, Piracicaba, v. 63, n.4, p. 341-350, 2006.

GUIMARÃES, E.C. Geoestatística básica e aplicada. Uberlândia: Faculdade de Matemática-Universidade Federal de Uberlândia, 2004. 77p. (Apostila).

IQBAL, J.; THOMASSON, J.A.; JENKINS, J.N., OWENS, P.R.; WHISLER, F.D. Spatial variability analysis of soil physical properties of Alluvial soils. Soil Science Society of America Journal, Madison, v. 69, n. 4, p. 1338-1350, 2005.

ISAAKS, E.H.; SRIVASTAVA, R.M. An introduction to applied geostatistics. New York: Oxford University Press, 1989. 561p.

JOURNEL, A.G.; HUIJBREGTS, CH. J. Mining geoestatistics. New York: Academic Press Inc., 1978. 600p.

LAMHAMEDI, M.S.; LABBÉ, L.; MARGOLIS, H.A.; STOWE, D.C.; BLAIS, L.; RENAUD, M. Spatial variability of substrate water content and growth of white spruce seedlings. Soil Science Society of America Journal, Madison, v. 70, n. 1, p. 108-120, 2006.

LANDIM, P.M.B. Análise estatística de dados geológicos. São Paulo: Editora UNESP, 2003. 253p.

MCGRAW, T. Soil test level variability in Southern Minnesota. Better Crops with Plant Foods, v.78, p.24-25, 1994.

MZUKU, M.; KHOSLA, R.; REICH, R.; INMAN, D.; SMITH, F.; MACDONALD, L. Spatial variability of measured soil properties across site-specific management zones. Soil Science Society of America Journal, Madison, v. 69, n. 5, p. 1572-1579, 2005.

NIELSEN, D.R.; WENDROTH, O. Spatial and temporal statistics: sampling field soils and their vegetation. Reiskirchen: Catena Verlag, 2003. 398p.

NOVAES FILHO, J.P.; COUTO, E.G.; OLIVEIRA, V.A.; JOHNSON, M.S.; LEHMANN, J.; RIHA, S.S. Variabilidade espacial de atributos físicos de solo usada na identificação de classes pedológicas de microbacias na Amazônia Meridional. Revista Brasileira de Ciência do Solo, Viçosa, v. 31, n.1, p. 91-100, 2007.

REICHARDT, K.; TIMM, L.C. Solo, Planta e Atmosfera: conceitos, processos e aplicações. São Paulo: Editora Manole, 2004. 478p.

SALAS, J.D.; DELLEUR, J.W.; YEVJEVICH, V.; LANE, W.L. 1988. Applied modeling of hydrologic time series. Littleton, CO: Water Resources Publications, 1988.

SHUMWAY, R. H. Applied statistical time series analyses. Englewood Cliffs (New York): Prentice Halll, 1988.

SILVA, F.M.; SOUZA, Z.M.; FIGUEIREDO, C.A.P.; JÚNIOR, J.M.; MACHADO, R.V. Variabilidade espacial de atributos químicos e de produtividade na cultura do café. Ciência Rural, Santa Maria, v.37, n.2, p. 401-407, 2007.

SOUSA, J.R.; QUEIROZ, J.E.; GHEYI, H.R. Variabilidade espacial de características físico-hídricas e de água disponível em um solo aluvial no semi-árido paraibano. Revista Brasileira de Engenharia Agrícola e Ambiental, v.3, p.140-144, 1999.

TERRA, J.A.; SHAW, J.N.; REEVES, D.W.; RAPER, R.L.; VAN SANTEN, E.; SCHWAB, E.B.; MASK, P.L. Soil management and landscape variability affects field-scale cotton productivity Soil Science Society of America Journal, v. 70, n. 1, p. 98-107, 2006.

TIMM, L.C.; PIRES, L.F.; ROVERATTI, R.; ARTHUR, R.C.J.; REICHARDT, K.; OLIVEIRA, J.C.M.; BACCHI, O.O.S. Field spatial and temporal patterns of soil water content and bulk density changes. Scientia Agricola, v. 63, n.1, p. 55-64, 2006.

PIANA, C.F.B.; MACHADO, A.A. Estatística básica. Pelotas: Instituto de Física e Matemática, Universidade Federal de Pelotas, 2004. 193p. (apostila – versão preliminar).

TIMM, L.C.; REICHARDT, K.; OLIVEIRA, J.C.M.; CASSARO, F.A.M.; TOMINAGA, T.T.; BACCHI, O.O.S.; DOURADO-NETO, D. Sugarcane production evaluated by the state-space approach. Journal of Hydrology, v.272, p.226-237, 2003.

TIMM, L.C.; REICHARDT, K.; OLIVEIRA, J.C.M.; CASSARO, F.A.M.; TOMINAGA, T.T.; BACCHI, O.O.S.; DOURADO-NETO, D. State-space for evaluating the soil-plant-atmosphere system. In: ACHYUTHAN, H. (ed.) Soils and Soil Physics in continental environment. Chennai (India): Allied Publishers Private Limited, 2003a. p. 23-81.

TIMM, L.C.; REICHARDT, K.; OLIVEIRA, J.C.M.; CASSARO, F.A.M.; TOMINAGA, T.T.; BACCHI, O.O.S.; DOURADO-NETO, D.; NIELSEN, D.R. State-space approach to evaluate the relation between soil physical and chemical properties. Revista Brasileira de Ciência do Solo, v.28, p.49-58, 2004.

TOMINAGA, T.T.; CASSARO, F.A.M.; BACCHI, O.O.S.; REICHARDT, K.; OLIVEIRA, J.C.M.; TIMM, L.C. Variability of soil water content and bulk density in a sugarcane field. Australian Journal of Soil Research, Collingwood, v.40, p.605-614, 2002.

VIEIRA, S.R. Geoestatística em estudos de variabilidade espacial do solo. In: NOVAIS, R.F.; ALVAREZ, V.H.; SCHAEFER, C.E.G.R. (eds.). Tópicos em ciência do solo. Viçosa: Sociedade Brasileira de Ciência do Solo, 2000. v.1. p.1-54.

VIEIRA, S.R.; HATFIELD, T.L.; NIELSEN, D.R.; BIGGAR, J.W. Geostatistical theory and application to variability of some agronomical properties. Hilgardia, Oakland, v. 51, n.1, p. 1-75, 1983.

WARRICK, A.W.; NIELSEN, D.R. Spatial variability of soil physical properties in the field. In: HILLEL, D. (ed.) Applications of soil physics. New York: Academic Press, 1980. cap. 2. p.319-344.

WEBSTER, R.; OLIVER, M.A. Geostatistics for environmental scientists. Chichester (England): John Wiley & Sons Ltd., 2001. 271p.

WENDROTH, O.; REYNOLDS, W. D.; VIEIRA, S. R.; REICHARDT, K.; WIRTH, S. Statistical approaches to the analysis of soil quality data. In: GREGORICH, E.G.; CARTER, M. R. (eds.) Soil quality for crop production and ecosystem health. Amsterdam: Elsevier, 1997. p. 247-276.

WENDROTH, O.; JÜRSCHIK, P.; KERSEBAUM, K.C.; REUTER, H.; VAN KESSEL, C.; NIELSEN, D.R. Identifying, understanding, and describing spatial processes in agricultural landscapes – four case studies. Soil and Tillage Research, v.58, p.113-127, 2001.

WENDROTH, O.; REUTER, H.I.; KERSEBAUM, K.C. Predicting yield of barley across a landscape: a state-space modeling approach. Journal of Hydrology, v.272, p.250-263, 2003.