DETRENDING NON STATIONARY DATA FOR GEOSTATISTICAL APPLICATIONS [1]

SIDNEY ROSA VIEIRA [2*]; JOSÉ RUY PORTO DE CARVALHO [3]; MARCOS BACIS CEDDIA [4]; ANTONIO PAZ GONZÁLEZ [5]

ABSTRACT

The use of geostatistics requires at least that the intrinsic hypothesis be satisfied. The presence of a trend in the data invalidates this hypothesis. One of the ways of solving this problem is by subtracting a function fitted to the original data and working with the residuals. This technique also represents a change to a smaller scale of the variability and surface roughness. This paper describes the detrending technique of subtracting a trend surface fitted by the least squares method and discusses the results using topographical data as examples. The objective is to show how the detrending technique works for different scales and degrees of trend and how to interpret the results. It is shown that the simplest the surfaces fitted that does the work of removing the trend the best are the results obtained. The use of jack knifing is proved useful to validate the resulting semivariograms. For most of the applications and depending upon the scale, a linear or a parabolic surface works reasonably well. The back transformation of the data afterwards is very easily done by adding back the subtracted trend surface.

Key words: Semivariograms, stationarity, topography, scale of variation.

RESUMO

REMOÇÃO DE TENDÊNCIA DE DADOS NÃO ESTACIONÁRIOS PARA APLICAÇÕES DE GEOESTATÍSTICA

O uso de geoestatistica exige que pelo menos a hipótese intrínseca seja satisfeita. A presença de tendência invalida esta hipótese. Uma das maneiras mais práticas de resolver esta questão é subtrair dos dados originais uma função ajustada por mínimos quadrados, trabalhando com a função residual resultante. Esta técnica representa uma mudança de escala da variabilidade e da rugosidade da superfície para uma escala menor. Este trabalho descreve a técnica da remoção da tendência pela subtração de uma superfície ajustada por mínimos quadrados e discute os resultados usando dados topográficos como exemplos. O objetivo é mostrar como usar esta técnica para remover tendências de superfícies de diferentes escalas e com diferentes graus de superfície e como interpretar os resultados. Foi mostrado que quanto mais simples for a superfície ajustada que resolve o problema de tendência, melhores são os resultados obtidos. O uso da validação cruzada provou ser útil para validar os semivariogramas resultantes. Para a maioria das aplicações e dependendo da escala, uma superfície linear ou parabólica funciona muito bem. A transformação de retorno nos dados é bastante fácil e é feita somando de volta a superfície subtraída.

Palavras-chave: erro reduzido, estacionaridade, topografia, variação de escala.

1. INTRODUCTION

The use of Geostatistics in the analysis of spatial variability has grown significantly during the last decades. It was introduced as a science by the French school of mathematics involved in mining industry in 1963 (MATHERON, 1963) based on some field observations in the South African mining reported by KRIGE (1951). The basic principles of Geostatistics were designed to mathematically handle the idea of spatial data sets that show regions of large values and other regions of smaller values as this was the common situation in mining. It was then given the name of Geostatistics and its foundations are in the Theory of Regionalized Variables (MATHERON, 1963). From there on its application have evolved to many fields of crop and soil science (BURGESS and WEBSTER, 1980; VIEIRA et al., 1981, VIEIRA et al., 1983, MCBRATNEY and WEBSTER, 1986; VIEIRA, 2000; VIEIRA et al., 2002).

Because the Theory of Regionalized Variables (MATHERON, 1963) is based on random functions and the measurements are assumed to be a realization of a particular random function for a given position, some restrictions have to be made on the data. These restrictions are called stationarity hypothesis (JOURNEL and HUIGBREGTS, 1978). The order of the stationarity hypothesis will depend on the order of the statistical moments required to be stationary. Thus, when second order stationarity is required, at least the first and second order moments (mean, variance and covariance) must be stationary. Second order stationarity is too restrict and difficult to be verified in practice (VIEIRA, 2000). Alternatively a condition weaker, simpler and easier to verify was proposed by MATHERON (1963) which he called the intrinsic hypothesis. Therefore, in order to make proper use of Geostatistics, at least the intrinsic hypothesis must be fulfilled (JOURNEL and HUIGBREGTS, 1978). The intrinsic hypothesis requires that the mean and the semivariance depend strictly on the separation distance between samples and not on the coordinate position of the data. When the intrinsic hypothesis can not be satisfied it is because the data has some trend which must be removed before the data can be adequately analyzed through Geostatistics (VIEIRA et al., 1983). One very simple way to remove a trend is by fitting a trend surface by least squares and subtracting it from the original data generating a new variable called

residuals with the difference. The presence or not of a trend in the data can be easily verified by the existence of a sill in the semivariogram at approximately the a priori variance value (VIEIRA, 2000).

Reports on trend removal are more commonly found in time series analysis (WU et al., 2007; SÁFADI, 2004). Some papers report on the presence a spatial trend in environmental pollution related problems but do not deal with trend removal (BOSSI et al., 2005; JOHANNESSON and CRESSIE, 2004; FAUSS-KESSLER at al., 1999). Few papers report on two dimensional trend removal (VIEIRA et al., 1983; VIEIRA et al., 2002, BLACKMORE et al., 2003) but did not do a thorough discussion on the interpretation of results for different scales and sampling density.

This paper describes the detrending technique of subtracting a trend surface fitted by the least squares method and discusses the results using topographical data as examples. The objective is to show how the detrending technique works for different scales and degrees of trend and how to interpret the results.

2. MATERIAL AND METHODS

Data

The five data sets used were topographical heights measured with a high precision optical engineering level from: a square field of 90m on each side named Wang sampled on a 2m grid with a total of 2500 points; a triangular field named Paddock of 110m by 220m, sampled on a 10m square grid, with a total of 164 points; an approximately rectangular field named Recuperação measuring 90x250m, sampled on trapezoidal grid of 5m, with a total of 383 points; an approximately rectangular field named Variabilidade, measuring 120x160m sampled on a 10m square grid with 302 data points; a circular field named Planalto of 77ha sampled on a square grid of 50m, at every one of the 322 points.

Table 1 shows a summary about the five fields with grid information. Topography was chosen as a variable to be analyzed because its form is easily verified in the field and its surface does not change with time allowing for the field validation of the results if necessary. These specific fields were chosen because they represent a very wide range of scales (from 0.81 to 77 hectares), of grid

Table 1. Area, grid, number, number of sample (N/ha) and kind of trend surface used for the five fields

Site	Location	Area (ha)	Grid (m)	N	N/ha	Trend
Wang	Ottawa	0.81	2	2500	3086.42	Parabolic
Recuperação	Pindorama	2.25	5	383	170.22	Cubic
Paddock	Ottawa	1.21	10	164	135.54	Parabolic
Variabilidade	Campinas	3.00	10	301	100.33	Parabolic
Planalto	Angatuba	77.00	50	322	4.18	Linear

sampling spacing (from 2 to 50m), of number of values (from 2500 to 164 values) and consequently of number of samples per hectare (from 3086.42 to 4.18).

Intrinsic hypothesis

A set of N values $Z(x_i)$, where x_i denotes a geographical position, will be intrinsic if it follows two conditions:

The expected value $E\{Z(x_i)\}$ exists and does not depend on the position x_i. This is mathematically written as:

$$E\{Z(x_i)\} = m \tag{1}$$

The variance of the increment $[Z(x_i) - Z(x_i+h)]$ is finite and does not depend on the position x_i. This can be written as

$$VAR[Z(x_i) - Z(x_i + h)] = E[Z(x_i) - Z(x_i + h)]^2 \tag{2}$$

Notice that the presence of any kind of trend would not agree with any of the two above conditions because the mean value, m, would be dependent on the position x_i, and the quantity $VAR[Z(x_i)-Z(x_i+h)]$ can not be guaranteed to be finite and it would also depend on the position in the field. A field showing a trend is easily imagined in practice, as we can think that a trend would represent an increase in variable values in some direction. In this condition, the mean value, m, and also the variance of the increment $[Z(x_i)-Z(x_i+h)]$ would not be independent of the position x_i.

Semivariogram

The semivariogram is, by definition:

$$\gamma(h) = 1/2\, E\{Z(x_i) - Z(x_i + h)\}^2 \tag{3}$$

And can be estimated by:

$$\gamma^*(h) = \frac{1}{2\,N(h)} \sum_{i=1}^{N(h)} [Z(x_i) - Z(x_i + h)]^2 \tag{4}$$

where N(h) is the number of pairs of measured values $Z(x_i)$, $Z(x_i+h)$, separated by a vector h (JOURNEL and HUIJBREGTS, 1978). The graph of $\gamma^*(h)$ versus the corresponding values of h, called semivariogram, is a function of the vector h, and therefore it depends on both magnitude and direction of h. When the semivariogram is the same for all directions it is called isotropic. Many variables show anisotropic semivariograms depending on the dimensions of the field and of the nature of the variability. There are ways of transforming an anisotropic semivariogram (JOURNEL and HUIJBREGTS, 1978; BURGESS and WEBSTER, 1980) in order to reflect the variability in different directions. Jack knifing procedure can also be used to verify the distance range over which

a semivariogram can be used before anisotropic effects may affect the results (VIEIRA, 2000).

Notice that the equation of the definition of the semivariogram (equation 3) is nothing more than the right hand side of equation 2, multiplied by ½. Therefore, the maximum value that an experimental semivariogram of an intrinsic variable could reach is the *a priori* variance of that variable. When this condition fulfills, the semivariogram has a clearly defined and stable sill at the value of the *a priori* variance. Therefore, if the semivariogram does not stabilizes at the *a priori* variance value, this is a clear indication that the variable under study has a trend somewhere in the field which must be dealt with before further geostatistical analysis is done (VIEIRA, 2000).

Models

Experimental semivariograms contain a set of discrete data points of distance and semivariance. A model must be fit to the experimental data with the objective of having semivariances available for every distance needed (GOTWAY, 1991). In order to be used to properly describe the spatial variability of any variable, one of the requirements on the model is that the function used must be conditional positive definite (MACBRATNEY and WEBSTER, 1986). This condition will guarantee that the variances calculated will be positive. The main models which satisfy that condition and are adequate for use in geostatistical calculations are the spherical, the exponential and the gaussian. On the equations bellow, C_0, C_1, and a represent the nugget effect, the structural variance and the range, respectively.

For the spherical model, usually symbolized as $Sph(C_0, C_1, a)$, the equation is:

$$\gamma(h) = C_0 + C_1\left(1.5\frac{h}{a} - 0.5\left(\frac{h}{a}\right)^3\right),\ \ h \le a \tag{5}$$
$$\gamma(h) = C_0 + C_1,\ h > a$$

The exponential model, symbolized as $Exp(C_0, C_1, a)$, the equation is:

$$\gamma(h) = C_0 + C_1\left(1 - EXP\left(\frac{-3h}{a}\right)\right) \tag{6}$$

The gaussian model, symbolized as $Gau(C_0, C_1, a)$, the equation is:

$$\gamma(h) = C_0 + C_1\left(1 - EXP\left(-3\left(\frac{h}{a}\right)^2\right)\right) \tag{7}$$

With the parameters fitted to the semivariogram the dependence ratio *(DR)* can be calculated according with ZIMBACK (2001)

$$DR = \frac{C_1}{(C_0 + C_1)} \times 100 \qquad (8)$$

The dependence ratio *(DR)* represents the proportion of the semivariance which is structured. The smallest the *DR* value the weakest is the spatial dependence.

Detrending

A very simple and practical way of removing a trend is done by fitting a trend surface to the original values by minimum least squares and then subtracting the value of the trend surface function from the original values thus constructing a new residual variable. Because the topographical data used in this study showed very strong trend for all fields as it was verified by the absence of a sill in the experimental semivariograms, the trend was removed with a trend surface according to Vieira (2000). The degree of the trend surface used in each case is listed in the last column of Table 1, with one surface linear, three parabolic and one cubic. The presence of a trend is detected when the semivariogram does not have a stable sill. This condition violates the intrinsic hypothesis (equations 1 and 2) as it represents a field for which the mean value and the variance of the increment $[Z(x_i)–Z(x_i+h)]$ depend on the spatial position. The trend removal technique used as described in Vieira (2000) consists of fitting a three dimensional surface to the data by the least squares and subtracting its values from the originals. A parabolic trend surface is estimated by

$$Z^*(x,y) = A_0 + A_1 X + A_2 Y + A_3 X^2 + A_4 XY + A_5 Y^2 \qquad (9)$$

where $Z^*(x,y)$ is the estimated trend surface, X and Y are the coordinate positions and A_0, A_1, A_2, A_3, A_4 and A_5 are the regression parameters estimated by the least squares method. This surface is then subtracted from the originals generating a new variable which we are calling Residuals.

$$Z_{res}(x,y) = Z(x,y) - Z^*(x,y) \qquad (10)$$

The criteria for the choice of a degree for the detrending surface is the simplest surface that will produce a semivariogram with a stable sill. Thus, if a linear surface solves the starionarity problem producing a semivariogram with a sill, there is no need to look for any other degree of a surface. Usually it is advisable to start the detrending procedure with the smallest degree (linear), examining the coefficient of determination between the estimated and the original surfaces. If the coefficient is high and significant, as it can be tested with the available F value, then the semivariogram of the residuals should be calculated. If the resulting semivariogram shows a sill, then there is no need to seek any further degree of surface. If, on the other hand, the semivariogram produced did not show a sill,

the degree of the surface should be increased and the procedure started over (Vieira et al., 2002). Notice that the objective is not to produce the highest coefficient of determination but rather to produce a residual surface whose semivariogram has a sill.

Notice that removing a trend by the method described above represents a change in the scale for the data. Because the trend surface describes the overall spatial variation for the variable, the operation written in equation (10) will produce a new surface which will contain simply the remaining surface roughness and its values will be a fluctuation above and below zero (0). If the degree of the trend surface is adequate, invariably the mean value for residuals is close to zero (0) (Vieira et al., 1983).

Detrending the variable and producing a residual variable is a requirement for non stationary variables in order to be under the intrinsic hypothesis. After detrending, the Geostatistical analysis follows its normal way with the residual variable with the calculation of the semivariogram, fitting a model to it, jack knifing in order to validate, and then do kriging estimation of the residual variable adding back the trend surface using the parameters fitted.

3. RESULTS AND DISCUSSION

A summary information about the sampling from all five fields is shown in table 1. The last column shows the degrees of the trend surfaces used to remove the trend. Notice that the degrees of the surface show an increase as the sampling density increases. This is due to the detailed characterization of the field with samples close to one another for the higher density sampling, thus needing a higher degree surface in order to remove the trend. In other words, a field intensively sampled such as the Wang, shows a much higher resolution on the view of the surface roughness.

The descriptive statistics for the original data from all five fields is shown in table 2. Except for the field Planalto, all others approach a normal distribution as their coefficients of skewness and kurtosis are close to zero (0). The field Planalto was also the one for which the range of values was the largest (exceeded 22m). The smallest range of values is for the Wang field, with only 68cm. Therefore, in terms of sampling and spatial resolution of the data, the examples used in this study represent the reverse of what it should have been, because the field Planalto had the largest range of elevations and also de lowest sampling density whereas the field Wang had the lowest elevation range and the highest sampling density. All of these surveys were done sometime ago and we did not have any influence on the choices of sampling spacing, except for fields Variabilidade and Recuperação for which we made

the choices of both sampling spacing and number of samples.

Table 3 shows the summary descriptive statistics for the residual values for the five fields. Because of what is left after the trend removal is simply a fluctuation under and above the fitted surface, the mean values are very small (less than 10^{-3}). This is an indication that trend was effectively removed as the mean values would be higher if the surface underestimated them and would be lower if the surface overestimated them. Similar results have already been reported for other field scales (VIEIRA et al., 2002). The ranges of values for the residuals are almost symmetrical with respect to zero and the frequency distributions all depart from normal distribution as seen by the coefficients of skewness and kurtosis.

The parameters for the models fitted to the semivariograms of residual data are shown in table 4. There were two spherical models, two Gaussian and one exponential, but all of them with very low nugget effects and very high Dependence Ratio (DR), which means the surface left after the trend removal is a very continuous one. WOLLENHAUPT et al., (1997) report on similar results for crop yield data. The ranges of spatial dependence are all small as compared to the field dimensions which means that the trend removal decreased the scale of the variability as reported by VIEIRA (2000).

Figures 1 to 5 show the graphs of the semivariograms for the original values along with the ones for the residuals on figures 1a to 5a. The graphs of the semivariograms for the residuals with the models fitted are in figures 1b to

Table 2. Summary descriptive statistics for the original data

Name	Unit	N	Mean	Var.	Min.	Max.	Skew.	Kurt.
Wang	cm	2500	42.06	87.74	0.00	68.00	-0.660	0.148
Recuperação	m	340	93.98	18.19	84.94	104.70	0.071	-0.847
Paddock	cm	164	77.13	651.10	0.00	122.00	-0.929	0.889
Variabilidade	m	301	638.10	31.58	622.50	644.90	-1.168	0.097
Planalto	m	322	656.20	26.06	638.10	664.10	-1.013	1.048

Table 3. Summary descriptive statistics for the residual data

Name	Unit	N	Mean	Var.	Min.	Max.	Skew.	Kurt.
Wang	cm	2500	0.000000535	40.38	-36.71	20.31	-0.582	1.935
Recuperação	cm	340	0.00002962	0.02	-0.44	0.29	-0.722	0.812
Paddock	cm	164	-0.0000107	272.40	-54.55	68.69	0.045	3.225
Variabilidade	m	301	-0.0001931	6.40	-12.79	7.14	-0.780	3.950
Planalto	m	322	-0.0001192	6.03	-9.41	3.31	-1.186	1.142

Table 4. Semivariogram parameters for the models fitted

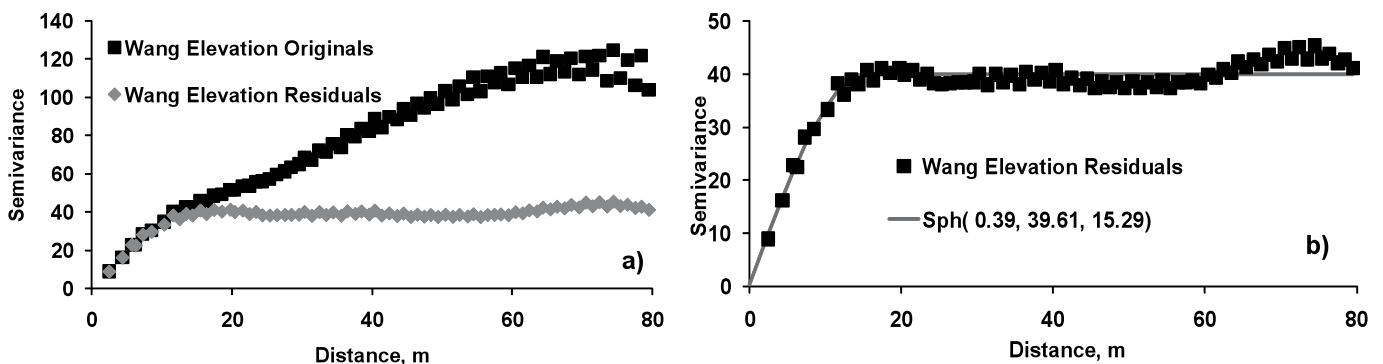| Variable	Model	C_0	C_1	a	r^2	RMSE	DR
Wang	Spherical	0.39	39.61	15.29	0.8842	0.02632	99.03
Recuperação	Exponential	0.00	0.016	26.13	0.9716	0.00015	100.0
Paddock	Gaussian	8.24	224.34	48.80	0.8778	0.93938	96.46
Variabilidade	Spherical	0.02	0.29	50.67	0.8672	0.00034	92.13
Planalto	Gaussian	0.00	2.19	323.69	0.8908	0.00573	100.0

Figure 1. Semivariograms for the Wang field: a) original values together with residuals; b) residuals with model fitted.
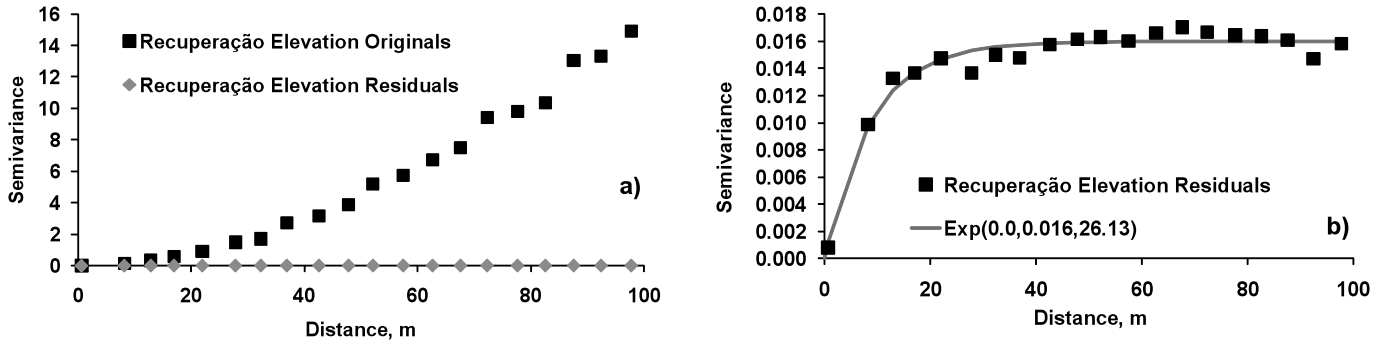
Figure 2. Semivariograms for the Recuperação field: a) original values together with residuals; b) residuals with model fitted.
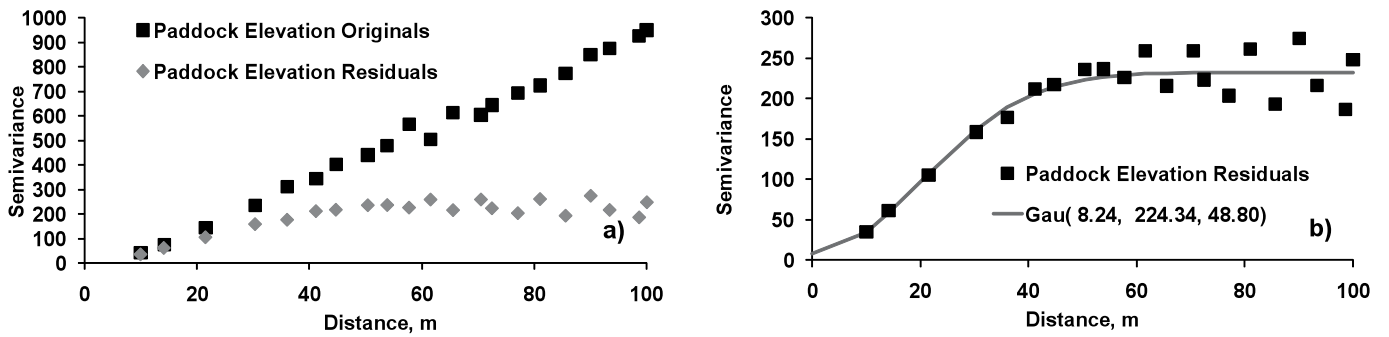

Figure 3. Semivariograms for the Paddock field: a) original values together with residuals; b) residuals with model fitted.
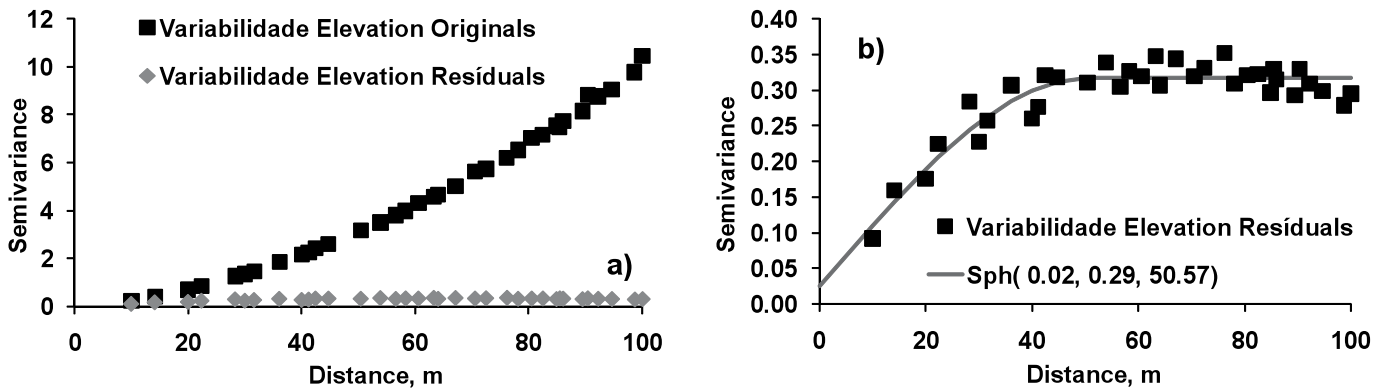

Figure 4. Semivariograms for the Variabilidade field: a) original values together with residuals; b) residuals with model fitted.
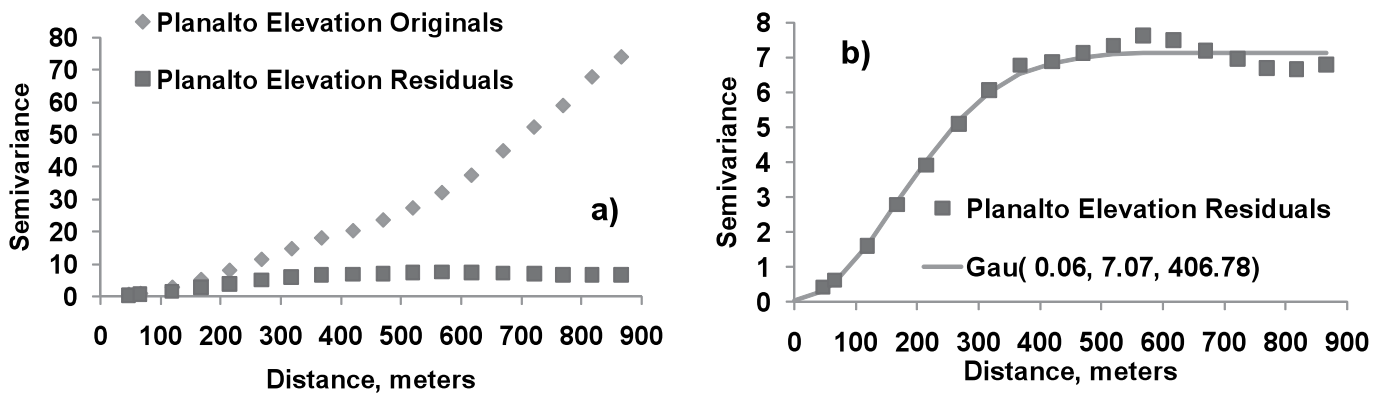

Figure 5. Semivariograms for the Planalto field: a) original values together with residuals; b) residuals with model fitted.
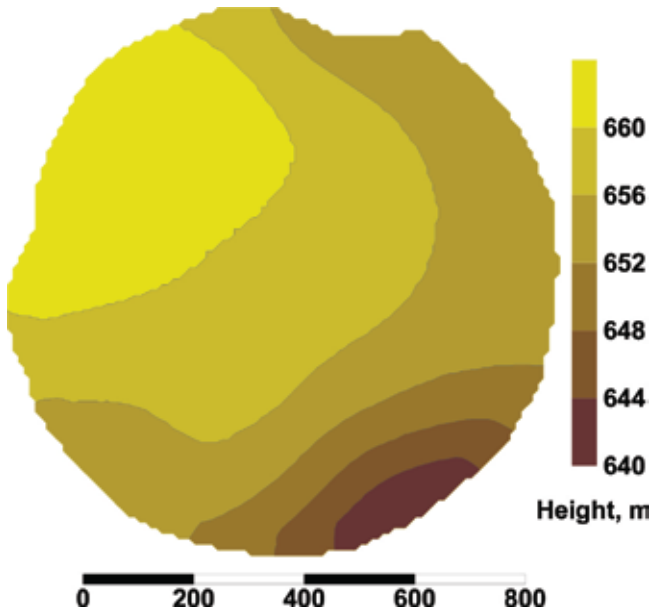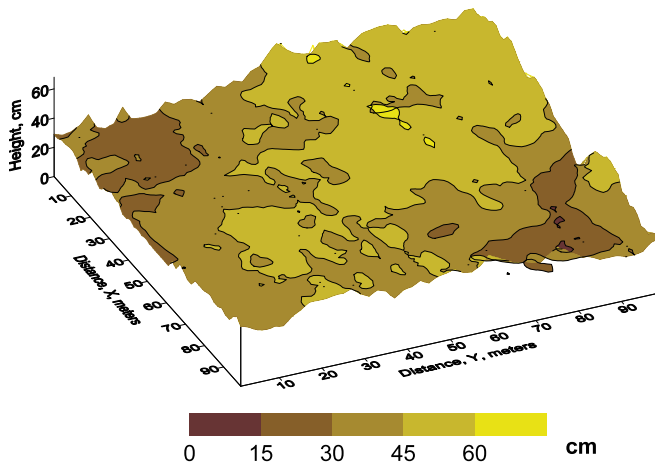
Figure 6. Topography map for the Planalto field.

Figure 7. Topography map for the Wang field.
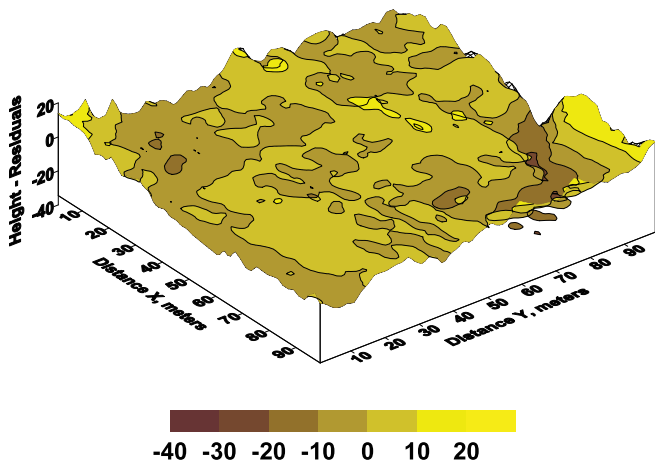

Figure 8. Topography for the Wang field after removal of the parabolic trend.

5b. It is quite clear that before the removal of the trend (FIGURE 1a to 5a) all the semivariograms had no limits and did not stabilize in any sill values. In this situation the field was not large enough to encompass all the variability it expressed (Vieira et al., 2002). On the other hand, the semivariograms on figures 1b to 5b clearly show a sill value which means the trend removal technique worked well.

The topographical map of the Planalto field is shown in figure 6. It can be seen on this map the reason why the original data had such a strong trend as the field shows a very pronounced slope. A comparison between linear and parabolic trend surface removing was made for this field. Because the coefficient of determination did not increase significantly and because the semivariogram of the residuals for linear trend showed a clear sill (Figure 5b) the linear was chosen for its simplicity.

Figure 7 shows the topographical tridimensional map for the Wang field before the removal of the parabolic trend and figure 8 shows the same field after the removal of the trend. On figure 7, the presence of higher elevations across the diagonal of the field and lower regions on the left and right corners reveals the reason for the parabolic trend. On figure 8, on the other hand, it can be seen that the sizes of the regions of low and high values decreased with respect to the original values. That is the reason for the smaller value for the range of spatial dependence and it has been found by Vieira et al., (1983).

4. CONCLUSION

The trend surface was shown to efficiently removing the trend from surfaces of different scales and different sampling densities. The semivariograms produced showed a very clear sill at the variance value, which means they do follow the intrinsic hypothesis after the trend removal. Even after the trend removal, the residuals for topographical data showed very continuous semivariograms near the origin.

REFERENCES

BLACKMORE, S.; GODWIN, R.; FOUNTAS, S. The analysis of spatial and temporal trends in yield map data over six years. **Biosystems Engineering**, v.84, p.455-466, 2003.

BOSSI, R.; RIGET, F.; DIETZ, R. Temporal and Spatial Trends of Perfluorinated Compounds in Ringed Seal (*Phoca hispida*) from Greenland. **Environmental Science and Technology**, v.39, p.7416-7422, 2005.

BURGESS, T.M.; WEBSTER, R. Optimal interpolation and isarithmic mapping of soil properties. I. The semivariogram and punctual kriging. **Journal of Soil Science**, v.31, p. 315-331, 1980

FAUS-KESSLER, T.; DIETL, C.; TRITSCHLER, J.; PEICHL, L. Temporal and spatial trends of metal contents of Bavarian mosses *Hypnum cupressiforme*. **The Science of the Total Environment**, v.232, p.13-25,1999.

GOTWAY, C.A. Fitting semivariogram models by weighted least squares. **Computers Geoscience**, v.17, p.171-172, 1991.

JOHANNESSON, G.; CRESSIE N. Finding large-scale spatial trends in massive, global, environmental datasets. **EnvironMetrics**, v.15, p.1-44, 2004.

JOURNEL, A.G.; HUIJBREGTS, C.H.J. **Mining geostatistics**. London: Academic Press, 1978. 600p.

KRIGE, D.G. A statistical approach to some basic mine evaluation problems on Witwatersrand. **Journal of Chemical Metallurgy of the Mining Society of South Africa**, v.52, p.119-139, 1951.

MATHERON, G. Principles of geoestatistics. **Economic Geology**, v.58, p.1246-1266, 1963.

McBRATNEY, A.B.; WEBSTER, R. Choosing functions for the semivariograms of soil properties and fitting them to sample estimates. **Journal of Soil Science**, v.37, p.617-639, 1986.

SÁFADI, T. Uso de séries temporais na análise de vazão de água na represa de Furnas. **Ciência e Agrotecnologia**, v.28, p.142-148, 2004.

VIEIRA, S.R. Geoestatística em estudos de variabilidade espacial do solo. In: NOVAIS, R.F.; ALVAREZ, V.H.;

SCHAEFER, G.R. (Ed.). **Tópicos em Ciência do solo**. Viçosa: Sociedade Brasileira de Ciência do solo, 2000. vol.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**, v.51, p.1-75, 1983.

VIEIRA, S.R.; MILLETE, J.; TOPP, G.C.; REYNOLDS, W.D. Handbook for geostatistical analysis of variability in soil and climate data. In: ALVAREZZ, V.H.; SCHAEFER, C.R.; BARROS, N.F.; MELLO, J.W.V.; COSTA, L.M. (Ed.). **Tópicos em Ciência do solo. Viçosa:** Sociedade Brasileira de Ciência do solo, 2002. vol.2, p.1-45, 2002.

VIEIRA, S.R.; NIELSEN, D.R.; BIGGAR, J.W. Spatial variability of field-measured infiltration rate. **Soil Science Society American Journal**, v.45, p.1040-1048, 1981.

WOLLENHAUPT, N.C.; MULLA, D.J.; GOTWAY CRAWFORD, C.A. Soil sampling and interpolation techniques for mapping spatial variability of soil properties, In: The site specific management for agricultural systems, ASA-CSSA-SSSA, 1997. p.19-53.

WU, Z.; HUANG, N.E.; LONG, S.R.; PENG, C.K. On the trend, detrending, and variability of nonlinear and nonstationary time series. **Proceedings of the National Academy of Sciences USA**, v.104, p.14889-14894, 2007.

ZIMBACK, C.R.L. **Análise especial de atributos químicos de solo para o mapeamento da fertilidade do solo**. 2001. 114p. (Livre-Docência) - UNESP/FCA, Botucatu.