Avaliando a Qualidade dos Estimadores de Variogramas (Variograma Experimental) em presença de normalidade e "Outliers".

Sueli Aparecida Mingoti, PhD em Estatística
Profa. Ajunta do Dep. de Estatística/ UFMG
sueli@est.ufmg.br

Gilmar Rosa
Doutorando em Computação Aplicada – INPE
Mestre em Estatística/UFMG
gilmar@dpi.inpe.br

Resumo

Neste artigo são analisadas as propriedades de robustez dos estimadores de variogramas mais amplamente utilizados, bem como novos estimadores, que recentemente tem surgido na literatura, alguns dos quais com características robustas em relação a outliers. Os resultados mostraram uma boa performance dos estimadores com características robustas na presença de outiliers, mas estes mesmos estimadores não apresentaram bons resultados na ausência de outliers. Este trabalho é uma extensão nos estudos apresentados por Genton (1988) em seu artigo "*Highly Robust Variogram Estimation*".

1 Introdução.

Muitos dos estimadores de variogramas propostos na literatura, não são robustos em relação a outiliers. Genton, 1998, propôs em seu artigo um novo estimador de variograma fundamentado nas idéias de estimação robusta apresentadas em Rousseeuw & Croux (1993), sendo um estimador de escala da classe M (Hampel e outros 1986). No entanto suas simulações restringiram-se apenas ao modelo teórico de variograma esférico e aos estimadores clássicos de Matheron (1962) e o proposto por Cressie & Hawkins(1980). Este trabalho extendeu o trabalho de Genton aos variogramas teóricos exponencial e senoidais, considerando, além dos estimadores de variogramas robusto (Cressie & Hawkins, 1980) e clássico de Matheron (1963), também os estimadores das diferenças (Haslett,1997) e o das medianas (Cressie, 1993).

2 Variogramas Teóricos

A metodologia de geoestatística é aplicada na análise de variáveis regionalizadas, entendendo-se como tais, variáveis cujos valores são relacionados de algum modo com a posição que ocupam no espaço (Chilés & Delfiner, 1999).

Basicamente, para analisar-se o comportamento da variável $Z(\bullet)$, duas suposições são necessárias: a estacionariedade intrínseca e a isotropia do processo $\{Z(x), x \in D\}$. Estas suposições são definidas como:

(i) $E[Z(x)] = \mu,\ x \in D$ e (ii) $Var\{Z(x_i) - Z(x_k)\} = 2\gamma(\| x_i - x_k \|),\ \forall x_i, x_k \in D$

onde $\| \bullet \|$ denota a distância Euclidiana. Neste caso, a variância das diferenças $(Z(x_i) - Z(x_k))$ é uma função apenas da distância $\| x_i - x_k \| = h$ entre as localizações, não dependendo da direção das localizações. Este é o conceito de isotropia.

3 Estimadores de Variogramas Experimentais

O estimador clássico de variograma proposto por Matheron (1963), fundamentado no método dos momentos (Cressie, 1993), é dado por:

$$2\hat{\gamma}(h)=\frac{1}{N_h}\sum_{N(h)}\left(Z(x_i)-Z(x_k)\right)^2,\quad h\in R^d\ ,$$

onde $N(h)=\left\{(x_i,x_k): \|x_i-x_k\|=h\ ,\forall x_i\neq x_k\right\}$ e N_h é a cardinalidade de $N(h)$.

Cressie & Hawkins (1980) propuseram um estimador tecnicamente menos sensível à presença de "outliers", denominado de variograma robusto, sendo dado por:

$$2\bar{\gamma}(h)=\frac{\left[\left(\frac{1}{N_h}\sum_{N(h)}\left|Z(x_i)-Z(x_k)\right|^{\frac{1}{2}}\right)^4\right]}{C_h},\quad h\in R^d,\ \text{onde:}$$

$N(h)=\left\{(x_i,x_k): \|x_i-x_k\|=h\right\},\ \forall x_i\neq x_k,\quad C_h=\left(0{,}457+\frac{0{,}494}{N_h}\right)$ e N_h é a cardinalidade de $N(h)$, sendo o denominador C_h um fator de correção para o vício do estimador de $2\bar{\gamma}(h)$ quando $Z(\bullet)$ tem distribuição normal.

Um outro estimador proposto por Cressie (1993) denominado das Medianas é definido como:

$$2\tilde{\gamma}(h)=\frac{\left[med\left\{\left|Z(x_i)-Z(x_k)\right|^{\frac{1}{2}}:\|x_i-x_k\|=h\right\}\right]^4}{B(h)},\ \forall x_i\neq x_k,$$

onde $med\{\bullet\}$ denota a mediana da seqüência $\{\bullet\}$, e $B(h)$ a correção para o vício de $2\tilde{\gamma}(h)$ assumindo normalidade para a variável $Z(\bullet)$. Assintoticamente, $B(h)$=0,457**.**

O estimador altamente robusto proposto por Genton (1998), é definido como: $2\hat{\gamma}(h)=\left(Q_{N_h}\right)^2$, onde $Q_{N_h}=2{,}2191\left\{\left|V_i(h)-V_j(h)\right|;\ i<j\right\}_{(k)}$, sendo $V(h)=z(x+h)-z(x)$ e $k=\binom{[N_h/2]+1}{2}$ a k-ésima estatística de ordem das diferenças $\left(V_i(h)-V_j(h)\right)$ e $[N_h/2]$ denota a parte inteira de $(N_h/2)$. O fator de *2,2191* é uma correção para o vício do estimador de $2\hat{\gamma}(h)$ quando $Z(\bullet)$ tem distribuição normal.

O estimador proposto por *Haslett* (1997), denominado variograma experimental das diferenças é definido como: $2\tilde{\gamma}(h)=\frac{1}{N_h-1}\sum_{N(h)}\left(d_{hi}-\bar{d}_h\right)^2$, onde $d_{hi}=\left(Z(x_i)-Z(x_k)\right)$,

$N(h)=\left\{(x_i,x_k): \|x_i-x_k\|=h\ \ \forall x_i\neq x_k\right\}$, e N_h é a cardinalidade de $N(h)$.

4 Simulação de Monte Carlo e Métodos de Avaliação

Dentre a classe dos modelos de séries temporais estacionários, interessa-nos nesta dissertação a classe ARMA(p,q), sendo que as amostras oriundas destes modelos estão estreitamente relacionadas com os modelos de variogramas teóricos discutidos no Capítulo 2. Serão abordados os modelos

ARMA(0,1) ou MA(1), ARMA (1,0) ou AR(1), ARMA (2,0) ou AR(2) e ARMA(1,1). Algumas referências sobre estes modelos são Box & Jenkins (1976) e Brockwell & Davis (1991). A geração de amostras foi efetuada de acordo com a sugestão de Sharp (1982).
A geração de variogramas teóricos deriva-se da relação:

$$2\gamma(h;\theta) = 2\sigma^2(1-\rho_h)$$

onde as autocorrelações ρ_h, em $\Re^1$, são obtidas pelas relações de recursividade dos modelos de séries temporais do tipo ARMA, e θ é o vetor de parâmetros do variograma teórico do processo estocástico gerador dos dados amostrais. Em todas os casos a variância fixa e igual a 5.

Os modelos de variogramas teóricos simulados e os respectivos parâmetros estão apresentados na Tabela 1, a seguir As Figuras 5.2.1 à 5.2.18 mostram exemplos de variogramas teóricos gerados neste trabalho.

Tabela 1: Modelos de Variogramas Teóricos e respectivos parâmetros.

Modelo Teórico	Percentagem de "outliers" nas Amostras			
	0%	5%	10%	15%
Esférico	$\phi = 0{,}9$	$\phi = 0{,}9$	$\phi = 0{,}9$	$\phi = 0{,}9$
Exponencial	$\phi = 0{,}9$ $\theta = 0{,}3$	$\phi = 0{,}9$ $\theta = 0{,}3$	$\phi = 0{,}9$ $\theta = 0{,}3$	$\phi = 0{,}9$ $\theta = 0{,}3$
Senóide	$\phi_1 = 1{,}7$ $\phi_2 = -0{,}9$	$\phi_1 = 1{,}7$ $\phi_2 = -0{,}9$	$\phi_1 = 1{,}7$ $\phi_2 = -0{,}9$	$\phi_1 = 1{,}7$ $\phi_2 = -0{,}9$

5 Resultados

Na análise dos estimadores de variogramas com 0% de contaminação e considerando-se os modelos teóricos esféricos e exponenciais, verificou-se que os estimadores que produziram melhores resultados foram: das diferenças, Clássico de Matheron, de Genton, o Robusto de Cressie & Hawkins e o das Medianas, na ordem respectiva de desempenho.

Em nossas simulações, as estimativas produzidas forneceram valores sistematicamente maiores em relação ao teórico, concluindo-se que os estimadores superestimam os verdadeiros valores dos variogramas teóricos, resultado similar ao observado no artigo de Genton (1998).

No caso de amostras, considerando-se o modelo teórico esférico, com 5, 10 e 15% de "outliers", verificamos que o estimador que apresentou melhor desempenho foi o das Medianas, seguido pelo estimador de Genton e o Robusto de Cressie & Hawkins e os que apresentaram piores desempenhos foram o das Diferenças e o Clássico. Para o modelo isotrópico exponencial, tem-se em linhas gerais, as mesmas conclusões que para o caso do modelo isotrópico esférico.

De uma maneira geral, a inserção de maior quantidade de "outliers" apenas aumenta o valor das estimativas para os estimadores da classe robusta (Medianas, Genton e Robusto de Cressie & Hawkins). De certa forma, estes estimadores conseguem retratar a forma geométrica dos isotrópicos esférico e exponencial, algo que os estimadores, clássico de Matheron e das Diferenças não conseguiram.

No caso do modelo senoidal, os estimadores conseguiram retratar a forma geométrica do modelo teórico. Os estimadores da classe robusta apresentam melhores resultados nos lags iniciais, lag=1,2 e 3, em relação aos estimadores não robustos. No geral, os estimadores da classe não robusta, especialmente o Clássico, fornecem as melhores estimativas em relação ao erro quadrático médio.

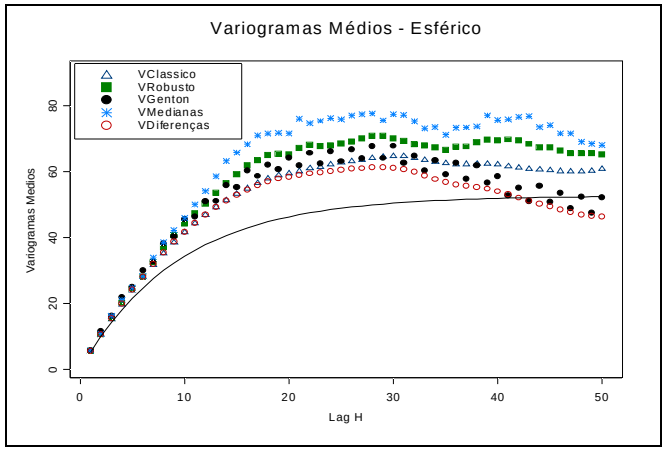

Gráfico 1: variogramas esféricos médios – A linha Contínua é o variograma teórico.

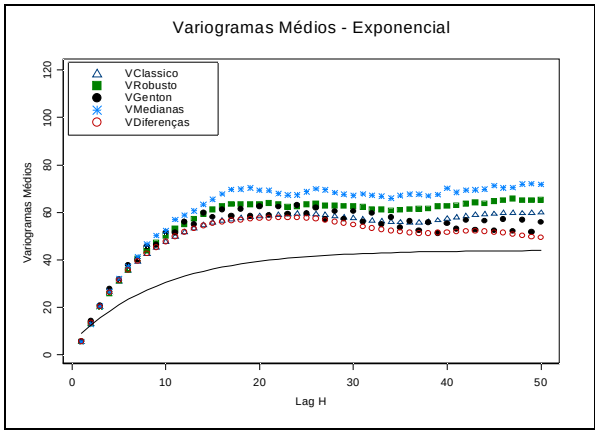

Gráfico 2: Variogramas exponenciais médios– A linha Contínua é o variograma teórico.

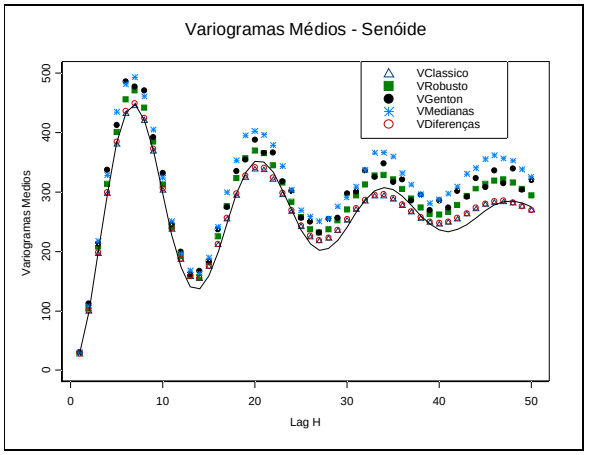

Gráfico 3: Variogramas senoidais Médios – A linha Contínua é o variograma teórico.

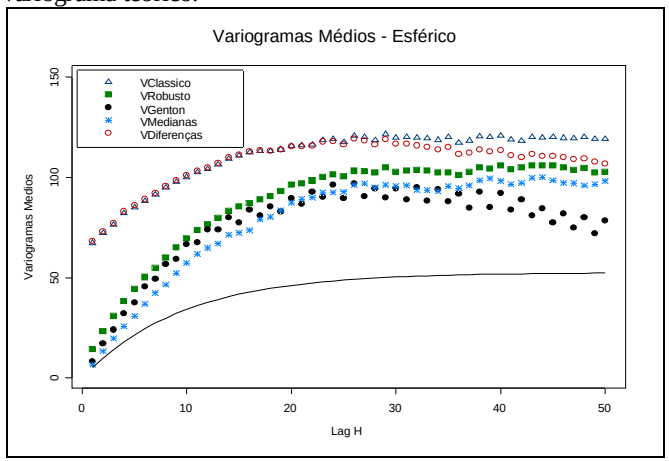

Gráfico 4: Variogramas Médios – 5% *outliers*- A linha continua é o variograma teórico.

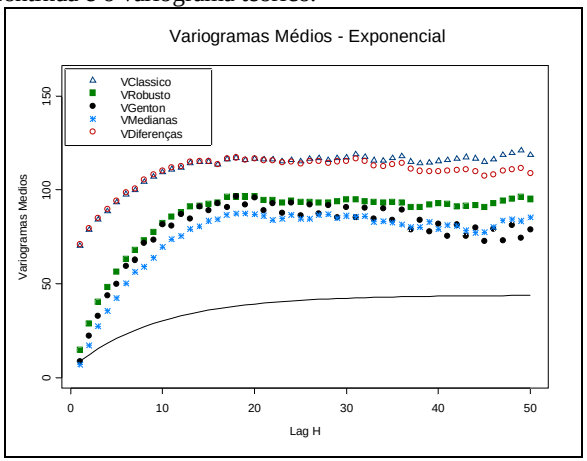

Gráfico 5: Variogramas Médios – 5% *outliers*- A linha continua é o variograma Exponencial

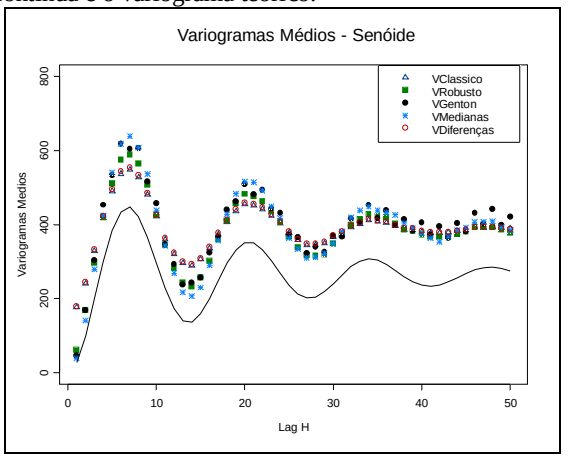

Gráfico 6: Variogramas Médios – 5% *outliers*- A linha - Continua é o variograma Senóide.

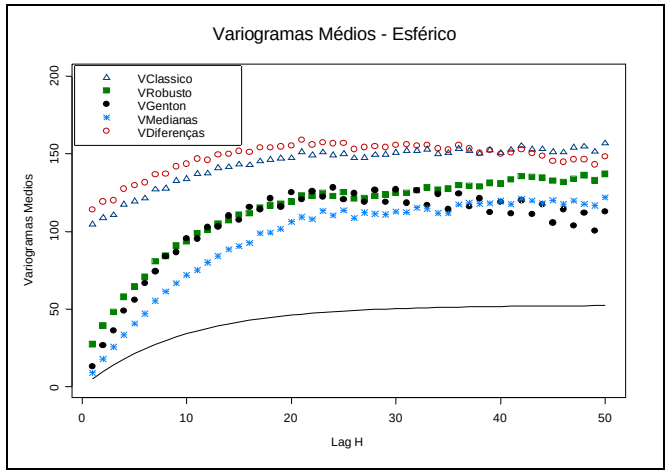

Gráfico 7: Variogramas Médios – 10% *outliers*- A linha continua é o variograma teórico.

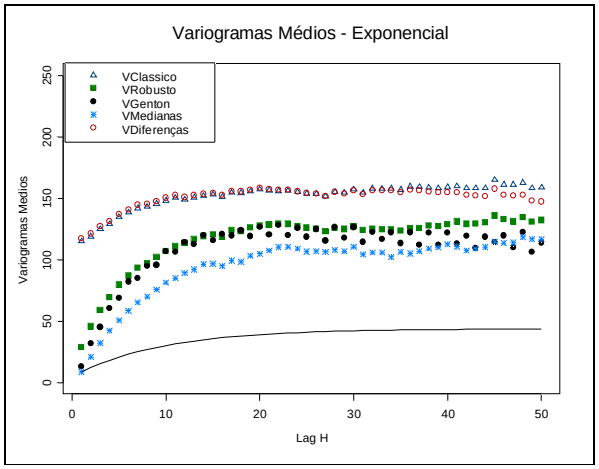

Gráfico 8: Variogramas Médios – 10% *outliers*- A linha continua é o variograma Exponencial,

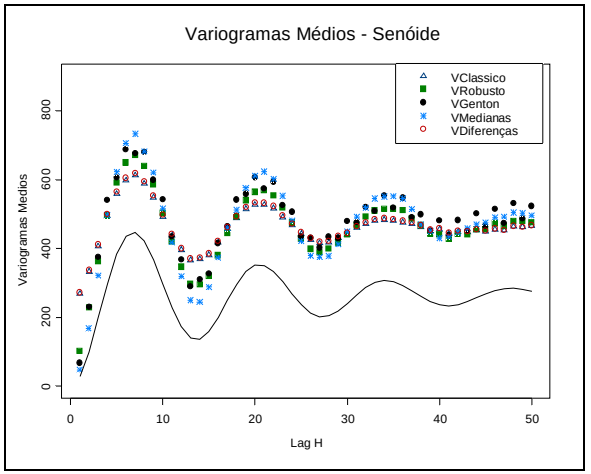

Gráfico 9: Variogramas Médios – 10% *outliers*- A linha
Continua é o variograma Senóide.

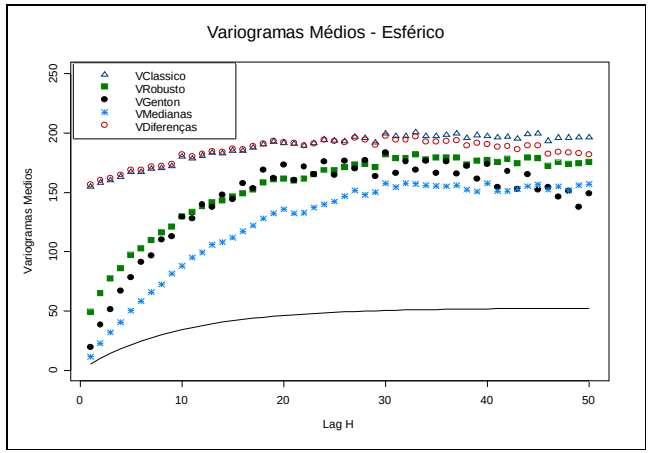

Gráfico 10: Variogramas Médios – 15% *outliers*- A linha continua é o variograma teórico.

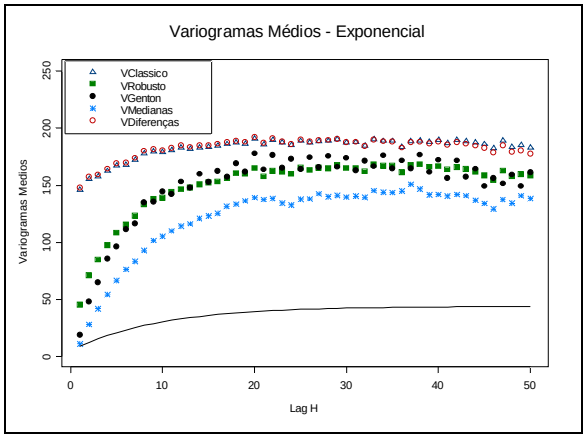

Gráfico 11: Variogramas Médios – 15% *outliers*- A linha continua é o variograma Exponencial.

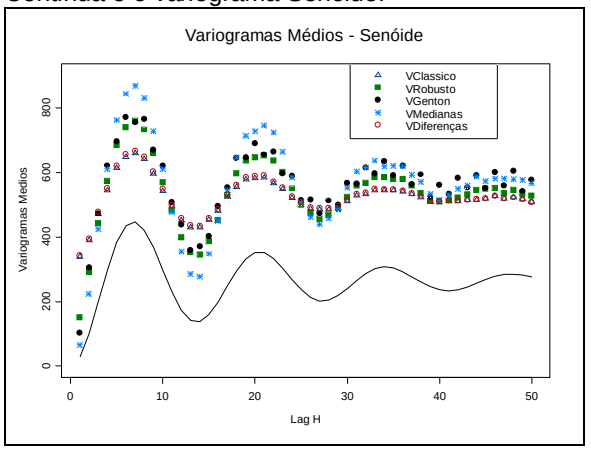

Gráfico 12: Variogramas Médios – 15% *outliers*- A linha continua é o variograma teórico.

6 Conclusão

Verifica-se que os estimadores de variograma da classe considerados robusta fornecem bons resultados nas situações em que há presença de "outliers" . Em situações nas quais os "outliers" não estão presentes nos dados, os estimadores da classe não robusta são preferíveis, isto é, os estimadores, Clássico de Matheron (1963) e das Diferenças de Hanslett (1997). O estimador proposto por Genton (1998) é uma boa alternativa considerando os estimadores da classe robusta pois os estudos mostraram que o estimador é robusto em relação a "outliers" . Nas simulações sem a presença de "outliers" e dentre os estimadores da classe robusta, foi o que apresentou melhor desempenho.

7 Referências

[1] BOX, G., E., P., JENKINS, G., M., *Time series analysis: Forecasting and Control*, 2nd ed. Holden-Day, San Francisco, 1976.

[2] CHERRY, S., BANFIELD, J., QUIMBY, W. F. An evaluation of non-parametric method of estimating semi-variograms of isotropic spatial process. *Journal of Applied Statistics*, 23, 4, 435-449,1996.

[3] CHILÉS, J-P, DEFINER, P. *Geostatistics: modelling spatial uncertainty*. New York: John Wiley, 1999. 695 p.

[4] CRESSIE, N. *Statistics for spatial data*. New York: John Wiley & Sons, 1993.

[5] CRESSIE, N.; HAWKINS, M. Robust estimation of the variogram:I. *Mathematical Geology*, 12 (2): 115-125,1980.

[6] DELAY, F., MARSILY, G. The integral of the semivariogram: a powerful method for adjusting the semivariogram in geostatistics, *Mathematical Geology*, 26,3,301-321, 1994.

[7] GENTON, M. G., GORSICH, D., J., Nonparametric variogram and covariogram estimation UIT Fourier-Bessel matrices. *Computacional Statistics & Data Analysis*, 41, 47-57, 2002.

[8] GENTON, M. G. Highly robust variogram estimation. *Mathematical Geology*, 30, 2, 213-221,1998.

[9] HAMPEL, F., R., RONCHETTI, E., M., ROUSSEEUW, P., J., STAHEL, W., A., Robust statistics, the approach based on influence functions: New York: John Wiley & Sons, 1986.

[10] HASLETT, J. On the sample variogram and the sample autocovariance for non-estacionary time series, *The Statistician*, v.46, pp 475-485, 1997.

[11] JOURNEL, A.G., HUIJBREGTS,Ch.J. *Mining geostatistics*. New York, Academic. Press, 1978.

[12] LAMOREY, G., JACOBSON, E. Estimation of semivariogram parameters and evaluation of the effects of data sparsity. *Mathematical Geology*, 27, 3, 327-358, 1995.

[13] KITANIDIS, P. K., "Minimum-variance unbiased quadratic estimation of covariances of regionalized variables, *"Mathematical Geology*, 17(2), 195-208, 1985

[14] KRIGE, D. G. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of Chemical, Metallurgical and Mining Society of South Africa*, 52, 119-139,1951.

[15] MATHERON, G. Principles of geostatistics. *Economic Geololy*, 58,1246-1266, 1963.

[16] MCBRATNEY, A., B., WEBSTER, R., Choosing functions for semivariograms of soil properties and fitting them to sampling estimates. *Journal of Soil Science*, 37, 617-639, 1986.

[17] OMRE, H., HALVORSEN, K. B. The *Bayesian bridge between simple and universal kriging. Mathematical* Geology, 21, 7, 767-786,1989.

[18] ROUSSEEUW, P. J., AND CROUX. C., 1993, *Alternatives o the median absolute deviation:* Jour. Am. Stat. Assoc., v. 88, nº 424, p. 1273-1283.

[19] SHARP, W. E. Stochastic simulation of semivariogram. *Mathematical Geology*, 14, 5, 445-457, 1982.