A note on robust and non-robust variogram estimators

(Uma discussão sobre os estimadores robustos e não-robustos de variogramas)

Sueli Aparecida Mingoti

Ph.D em Estatística. Professora Associada do Departamento de Estatística da Universidade Federal de Minas Gerais-UFMG, Belo Horizonte, Minas Gerais
E-mail: sueliam@est.ufmg.br

Gilmar Rosa

Mestre em Estatística. Doutorando em Computação Aplicada
Instituto Nacional de Pesquisas Espaciais - INPE, São José dos Campos, São Paulo
E-mail: Gilmar@dpi.inpe.br

Resumo

Em 1998, Genton propôs, um estimador de variograma que seria robusto em relação à presença de valores discrepantes (*outliers*) e o comparou com os estimadores propostos por Matheron e Cressie-Hawkins. Lark (2000) estendeu os resultados, avaliando o desempenho dos estimadores na presença de não-normalidade. Entretanto, ambos os trabalhos trataram apenas do modelo de variograma esférico e com algumas limitações. Nesse artigo, quatro estimadores de variogramas, incluindo o de Genton, são comparados através de simulação de Monte Carlo. Dados sem e com outliers foram simulados, considerando os modelos de variograma esférico, exponencial e senoidal. Os resultados mostraram que os estimadores de Genton e o da Mediana são melhores para dados com *outliers*, enquanto que o de Matheron e o de Hastlett são melhores para dados sem *outliers*, sendo o último adequado apenas para o caso de análise de séries temporais.

Palavras-chave: Estatística espacial, estimadores robustos e não-robustos de variogramas, dados discrepantes.

Abstract

In 1998 Genton proposed a variogram estimator claimed to be robust against outliers and compared it to Matheron's and Cressie-Hawkins' variogram estimators. Lark (2000) extended the comparison evaluating the effects of nonnormality. However, the comparison was limited to the spherical variogram model. In this paper 4 variogram estimators are compared including Genton's by using Monte Carlo simulation. Data with and without outliers were simulated using the spherical, exponential and wave models. The results showed that Genton's and the Median estimators were the best choices for contaminated data, while those of Matheron and Haslett presented better results for non-contaminated date; this latter being appropriate only for time series analysis.

Keywords: *Spatial statistics, robust and non-robust variogram estimators, outliers.*

1. Introduction

Variogram is an important tool in Geostatistics because it is used in the kriging procedure (Marchant and Lark, 2004). Many variogram estimators can be found in literature using parametric and non-parametric methodologies (Chilès and Delfiner, 1999). The better known is Matheron's (1962) which is very affected by the presence of outliers in the data set. Other alternatives are: Cressie and Hawkins' (1980) which was build to be robust against outliers and nonnormality; Median's (Cressie, 1993) and Genton's (1998) which were supposed to be robust against outliers; and the estimator proposed by Haslett (1997) used in a time series context especially for non-stationary data. Genton (1998) showed that his estimator had good performance in comparison to that of Matheron's and Cressie & Hawkins'. However, only spherical variogram models were considered in his study and only one replicate was generated for each simulated model. In 2000, Genton's comparison was extended by Lark (2000) who included Dowd's variogram estimator in his study and showed that all the estimators, except Matheron's, were very affected by nonnormality. Both mentioned papers used only the spherical variogram model. In this paper, the authors extended Genton's and Lark's results in respect to the outliers problem for non-spherical model and explored the wave variogram model, which has not appeared very often in other published studies. The Hastlett´s (1997) variogram estimator was also included in the study.

2. Methods and materials

2.1 Geostatistics methodolody

Geostatistics methodology was initially formulated for geological data (Matheron,1962). Nowdays, it has been used in many other fields, even for variables that are not of the physical-chemistry nature (Cressie, 1993; Mingoti et. al, 2006). Let $\{Z(x),\ x \in \text{D}\}$ be the spatial intrinsically stationary stochastic process, i.e.

(i) $E\,[Z(x)] = \mu\ ,\ \forall\, x \in \text{D}$

(ii) $Var\,[Z(x_l) - Z\,(x_k)] = 2\gamma(x_l - x_k)$, where $x_l \neq x_k \in$ D.

The quantities $2\gamma(.)$ and $(.)$ are called, respectively, variogram and semivariogram of the process $\{Z(x),\ x \in \text{D}\}$, where D is the domain under investigation. When the variogram is only a function of the distance between the two coordinates, $\|x_l - x_k\| = h$, the process is also called isotropic. The variogram determines the weights given to each sample value in the prediction of unsampled locations and it is also used to estimate the kriging variance.

2.2 Variogram estimators

In this section, we briefly present the variogram estimators, so-called sample or experimental variograms, which will be compared in this paper. We use the conventional notation for Geostatistics methodology. For all the estimators, N_h, is the cardinality of $N(h) = \{(x_i, x_j): \|x_i - x_j\| = h\}$ and $\{Z(x_1), Z(x_2), \ldots, Z(x_n)\}$is a sample of the spatial process $\{Z(x),\ x \in \text{D}\}$.

2.2.1 Matheron's classical variogram estimator

The estimator based on the method-of-moments proposed by Matheron (1962) is defined in (1) and is unbiased for the true theoretical values. It is the average of the squared differences between observations separated by the distance h.

$$2\hat{\gamma}_M(h) = \frac{1}{N_h} \sum_{N(h)} \left(Z(x_i) - Z(x_j)\right)^2 \tag{1}$$

This estimator is very affected by the presence of outliers and even a single discrepant datum can distort the final variogram estimates.

2.2.2 Cressie-hawkins robust variogram estimator

The estimator proposed by Cressie and Hawkins (1980) is given by

$$2\hat{\gamma}_{CH}(h) = \frac{1}{C_h}\left[\left(\frac{1}{N_h} \sum_{N(h)} \left| Z(x_i) - Z(x_j) \right|^{\frac{1}{2}}\right)^4\right] \tag{2}$$

where $C_h = \left(0.457 + \frac{0.494}{N_h} + \frac{0.494}{N_h^2}\right)$ is a correction factor for bias when the variable $Z(.)$ is normal. This estimator was built to be robust against outliers and nonnormality for distributions that are normal-like in the central region but heavier than normal in the tails.

2.2.3 Median variogram estimator

The median variogram estimator (Cressie and Hawkins, 1980) is defined as

$$2\hat{\gamma}_{Me}(h) = \frac{1}{B_h}\left[\left(med\left(\left|Z(x_i) - Z(x_j)\right|^{\frac{1}{2}} : (x_i, x_j) \in N(h)\right)\right)^4\right] \tag{3}$$

where ***med(.)*** denotes the median and B_h is a correction factor for bias when the variable ***Z(.)*** is normal. Asymptotically B_h=0.457. This estimator is also known as Dowd's (1984) and it is supposed to be robust against outliers.

2.2.4 Genton's highly robust variogram estimator

The estimator proposed by Genton (1998) is defined as:

$$2\hat{\gamma}(h) = (Q_{N_h})^2 \tag{4}$$

where $Q_{N_h} = 2.2191\left(\left|V_i(h) - V_j(h)\right| : i < j \right)_{(k)}$, $V(h) = Z(x+h) - Z(x)$, $k = \begin{pmatrix} [\frac{N_h}{2}] + 1 \\ 2 \end{pmatrix}$ is the kth order statistics of the differences $(V_i(h) - V_j(h))$ and ***[w]*** is the integer part of ***w***. The factor 2.2191 is a correction for bias when ***Z(.)*** has a normal distribution. Genton's estimator is based on Rousseeuw and Croux (1993) methodology. It is claimed to be highly robust against outliers.

2.2.5 Haslett's variogram estimator

The variogram proposed by Haslett (1997) in the univariate time series context is defined as:

$$2\hat{\gamma}_H(h) = \frac{1}{N_h - 1} \sum_{N(h)} (d_{hi} - \bar{d}_h)^2 \tag{5}$$

where $d_{hi} = (Z(x_i) - Z(x_j))$. Haslett's estimator is the sample variance of the differences $(Z(x_i) - Z(x_j))$and it was proposed as a tool to recognize ARMA stochastic processes in time series context. In his paper Haslett showed that his estimator had a good performance for non-stationary series.

2.3 Monte Carlo Simulation

In order to compare the performance of the variogram estimators, 100 samples of size n=100 were simulated from each of the theoretical variograms presented in Figure 1. Basically, the data were generated according to the methodology proposed by Sharp (1982). The idea is to represent the semivariogram model in terms of an ARMA (p,q) stationary process (Box and Jenkins, 1976). If $Z(x)$, $x \in \Re$, follows an ARMA(p,q) stationary process then the theoretical variogram of ***Z(.)*** can be expressed as

$$2\gamma(h;\theta) = 2\sigma^2 (1 - \rho_h) \tag{6}$$

where σ^2 is the variance of the process, ρ_h is the autocorrelation between $Z(x)$ and $Z(x+h)$, and θ is the parameter vector of the theoretical variogram of the process. By using the relation (6) and changing the values of (p,q), it is possible to generate variogram models such as spherical, exponential and wave. For p=1 and q=0 the related variogram is the spherical. For p=1 and q=1 the exponential variogram is generated and finally for p=q=1 the related variogram is the wave. For all ARMA processes considered in this paper the random noise was generated according to a normal distribution with zero mean and fixed variance. All generated series have the same mean value. Data with and without outliers were simulated. The percentage of outlier contamination were 5,10 and 15%. The introduction of outliers was performed in three steps: (1) first a sample was simulated from the respective ARMA process without outliers, i.e, Z(x)=δ+ϕZ(x-1)-αe(x-1)+ e(x), with e(.) being the random error generated from a normal distribution; (2) then a second sample was simulated from the same ARMA model but with a constant added, i.e, Z(x)=δ+ϕ Z(x-1)-αe(x-1) + e(x) + δ* ; (3) a percentual of the first sample was then substituted for values from a second sample. By using this procedure, it was assured that the theoretical covariance and correlation structure from both ARMA models (generating data with and without outliers) were the same.

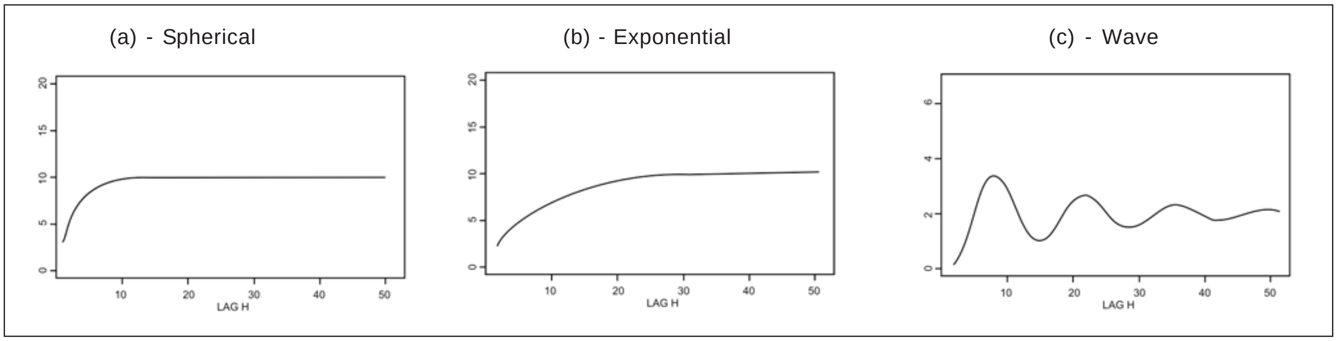


Figure 1 - Simulated theoretical variogram models.

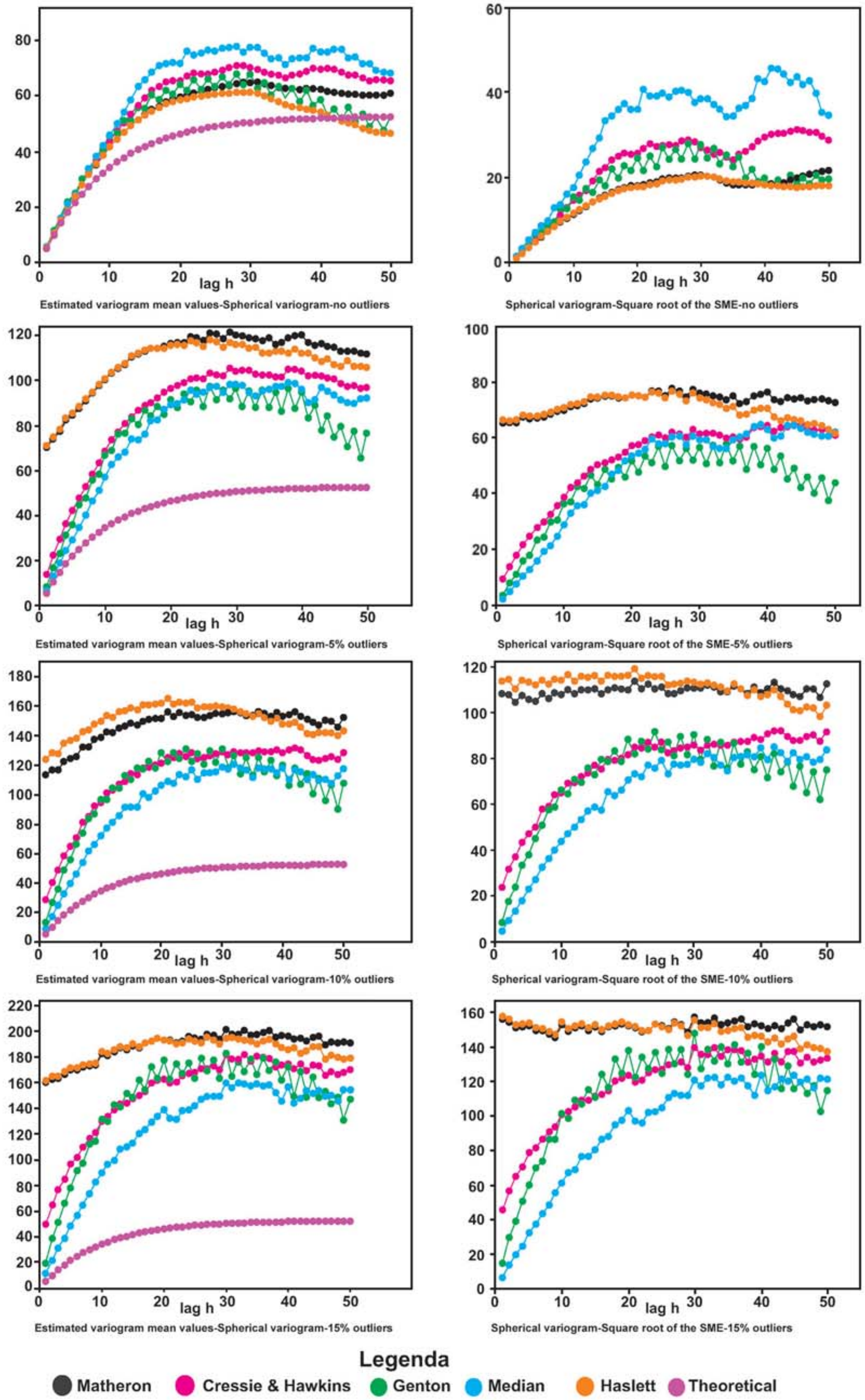
All the variogram estimators were calculated for each generated sample for $h=\{1,2,\ldots,[\frac{n}{2}]\}$. The value $[\frac{n}{2}]$ was chosen according to Journel and Huijbregts' rule (1997). For each value of h the estimated and the theoretical variogram values were compared by using the mean, the absolute, the relative and the squared mean errors. For each estimator, the average of the variogram estimates for each lag was calculated. They are presented in Figures 2 to 4 with the squared root of the square mean error (*SME*) values. In each graph the simulated variogram model is represented by a solid line. All mean errors presented a geometrical form similar to the squared root of the *SME* and will not be shown in this paper. The square mean error (*SME*) values for each variogram estimator are presented in Tables 1 and 2, for the first 4 lags, for data with and without outliers.

3. Results and discussion

For data with no outliers, the analysis of the results showed that for the spherical and exponential variogram models, Haslett's estimator achieved the best performance followed by Matheron's, Genton's and Cressie & Hawkins'. The Median estimator presented the highest number of errors and greatest dispersion. In the case of the wave variogram, for the initial lag h, all the estimators were able to recover the general mathematical form of the theoretical variogram model of the spatial process. However, for larger h, the errors increased significantly especially for values of h near to $[\frac{n}{2}]$. The best estimators for the wave variogram were Matheron's followed by Haslett's, Cressie & Hawkins' and Genton's. Median presented the worst results. For the spherical and exponential variogram models having 5, 10 and 15% of outliers, the Median estimator presented the best performance, followed by Genton's and Cressie & Hawkins'. Haslett's and Matheron's had the worse results. In general, for small and larger lags and regardless of the estimators used, the insertion of a larger amount of outliers in the data increased the value of the variogram estimates and the number of

Table 1 - Square Root of the Square Error Mean - Data without outliers.

<table>
<tr><th colspan="6">Spherical Variogram</th></tr>
<tr><th>Lag(h)</th><th>Matheron</th><th>Cressie-H</th><th>Genton</th><th>Median</th><th>Haslett</th></tr>
<tr><td>1</td><td>0.8034</td><td>0.9935</td><td>10.503</td><td>14.089</td><td>0.8227</td></tr>
<tr><td>2</td><td>19.634</td><td>22.847</td><td>27.761</td><td>30.901</td><td>20.129</td></tr>
<tr><td>3</td><td>33.191</td><td>39.047</td><td>41.226</td><td>50.996</td><td>33.983</td></tr>
<tr><td>4</td><td>46.550</td><td>54.796</td><td>65.026</td><td>68.990</td><td>47.663</td></tr>
<tr><th colspan="6">Exponential Variogram</th></tr>
<tr><td>1</td><td>35.912</td><td>36.797</td><td>33.997</td><td>37.970</td><td>35.431</td></tr>
<tr><td>2</td><td>24.040</td><td>26.932</td><td>32.882</td><td>32.881</td><td>24.646</td></tr>
<tr><td>3</td><td>59.628</td><td>62.213</td><td>67.707</td><td>72.892</td><td>61.130</td></tr>
<tr><td>4</td><td>93.087</td><td>95.355</td><td>111.677</td><td>113.774</td><td>94.964</td></tr>
<tr><th colspan="6">Wave Variogram</th></tr>
<tr><td>1</td><td>26.658</td><td>41.787</td><td>43.520</td><td>72.689</td><td>26.830</td></tr>
<tr><td>2</td><td>97.376</td><td>164.102</td><td>198.830</td><td>266.239</td><td>98.185</td></tr>
<tr><td>3</td><td>188.116</td><td>331.809</td><td>343.898</td><td>555.262</td><td>189.798</td></tr>
<tr><td>4</td><td>265.146</td><td>511.283</td><td>604.818</td><td>887.817</td><td>267.636</td></tr>
</table>



Estimated variogram mean values-Spherical variogram-no outliers

Spherical variogram-Square root of the SME-no outliers

Estimated variogram mean values-Spherical variogram-5% outliers

Spherical variogram-Square root of the SME-5% outliers

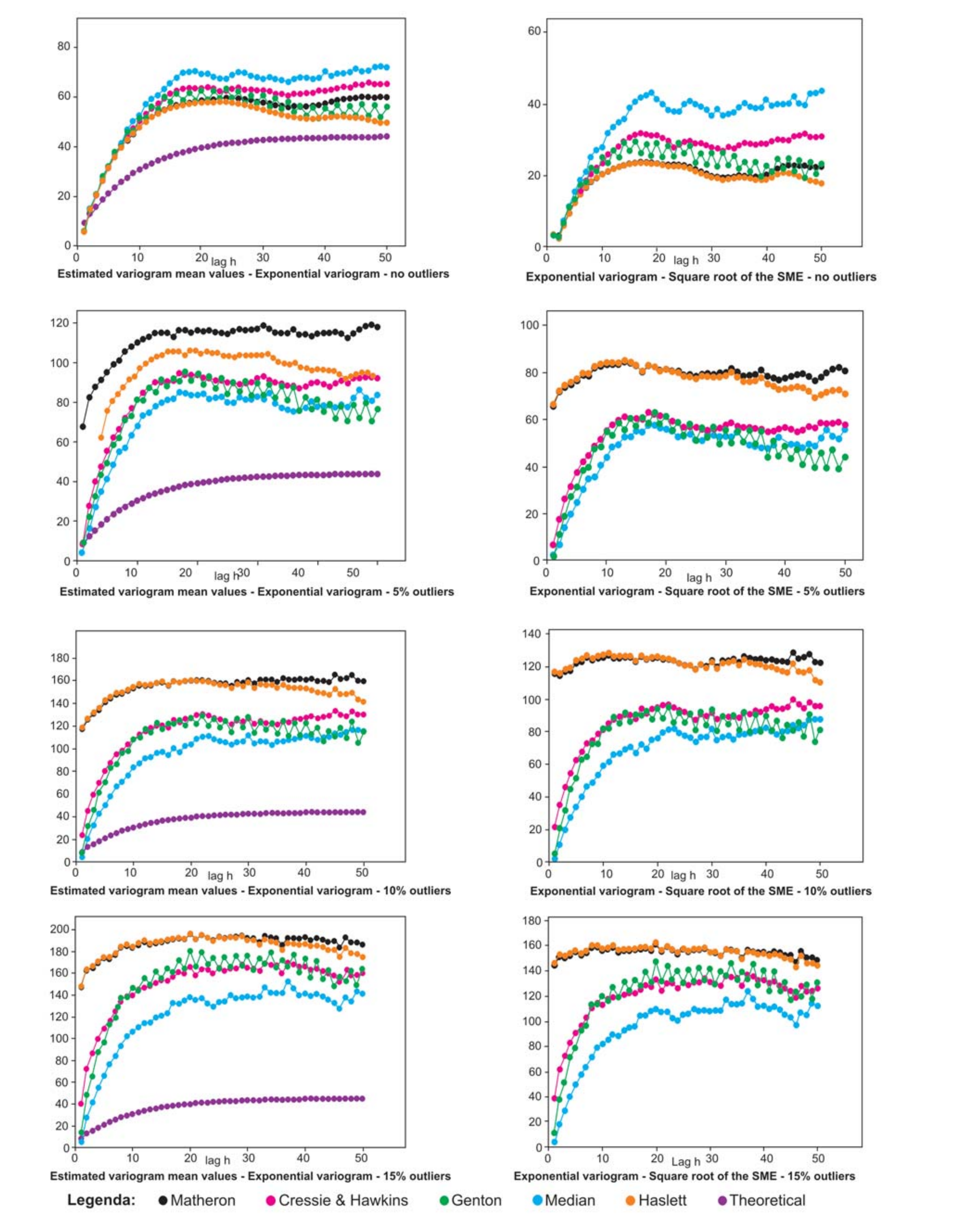
Estimated variogram mean values-Spherical variogram-10% outliers

Spherical variogram-Square root of the SME-10% outliers

Estimated variogram mean values-Spherical variogram-15% outliers

Spherical variogram-Square root of the SME-15% outliers

Figure 2 - Performance results for variogram estimators - spherical model.



Estimated variogram mean values - Exponential variogram - no outliers

Exponential variogram - Square root of the SME - no outliers

Estimated variogram mean values - Exponential variogram - 5% outliers

Exponential variogram - Square root of the SME - 5% outliers

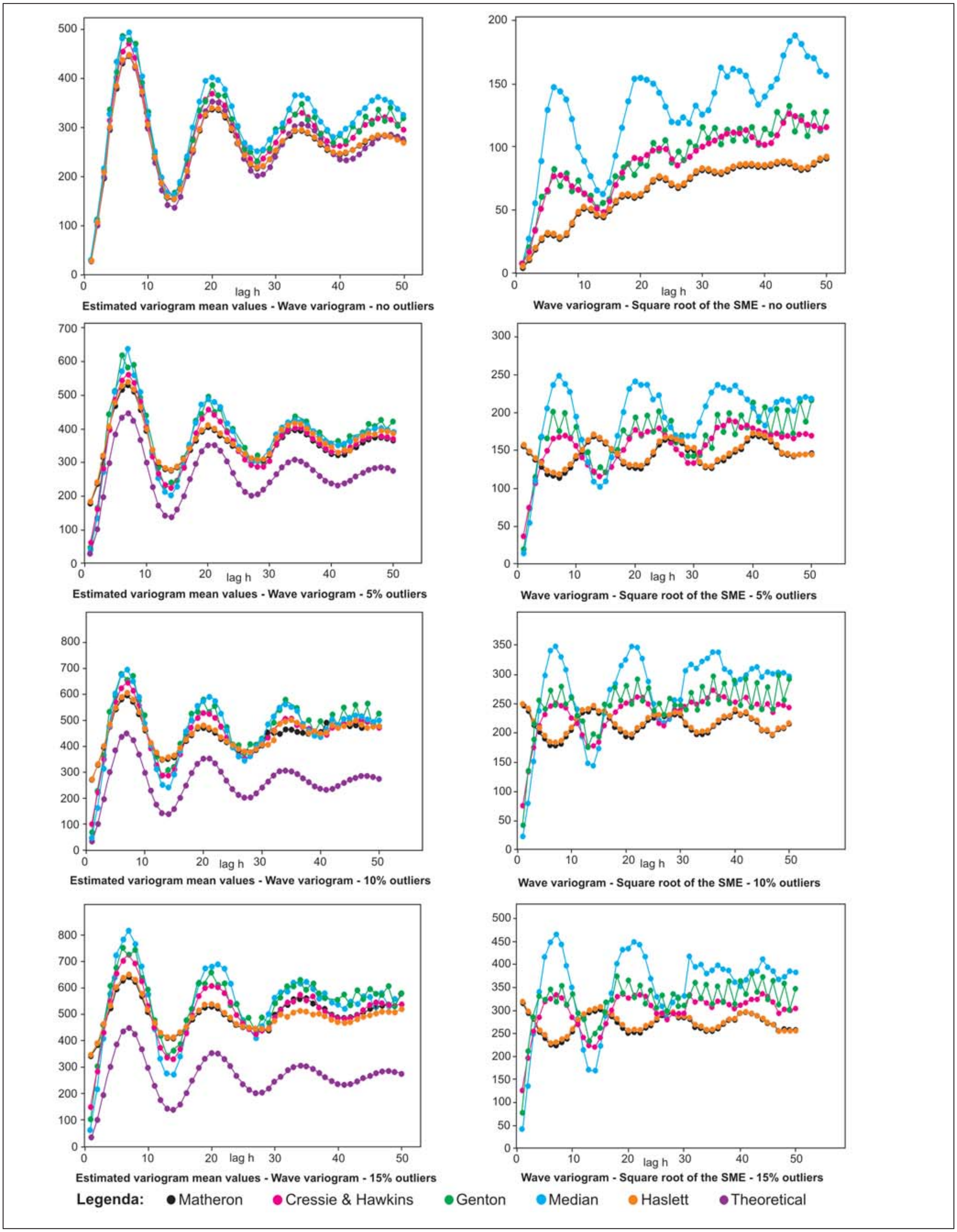
Estimated variogram mean values - Exponential variogram - 10% outliers

Exponential variogram - Square root of the SME - 10% outliers

Estimated variogram mean values - Exponential variogram - 15% outliers

Exponential variogram - Square root of the SME - 15% outliers

Figure 3 - Performance results for variogram estimators - exponential model.



Estimated variogram mean values - Wave variogram - no outliers

Wave variogram - Square root of the SME - no outliers

Estimated variogram mean values - Wave variogram - 5% outliers

Wave variogram - Square root of the SME - 5% outliers

Estimated variogram mean values - Wave variogram - 10% outliers

Wave variogram - Square root of the SME - 10% outliers

Estimated variogram mean values - Wave variogram - 15% outliers

Wave variogram - Square root of the SME - 15% outliers

Figure 4 - Performance results for variogram estimators - wave.

Table 2 - Square Root of the Square Error Mean- Data with outliers.

Spherical Variogram - 5% outliers					
Lag(h)	**Matheron**	**Cressie-H**	**Genton**	**Median**	**Haslett**
1	654.445	92.983	33.770	20.120	661.554
2	653.371	137.552	78.191	48.475	660.820
3	654.517	178.439	109.140	74.404	662.282
4	671.959	215.220	156.547	101.247	680.150
Spherical Variogram - 10% outliers					
1	1.085.309	239.243	83.459	43.988	1.140.997
2	1.078.840	318.031	176.450	91.820	1.149.541
3	1.048.934	372.010	236.223	134.933	1.106.994
4	1.076.420	433.446	331.939	181.427	1.143.266
Spherical Variogram - 15% outliers					
1	1.563.961	456.329	149.521	68.490	1.580.291
2	1.548.231	567.293	295.837	139.458	1.564.785
3	1.516.532	653.019	392.787	200.122	1.532.945
4	1.520.011	705.257	509.807	247.440	1.536.659
Exponential Variogram - 5% outliers					
1	656.070	65.617	14.249	23.514	663.533
2	713.765	174.251	109.808	67.798	722.099
3	734.944	262.287	186.249	139.898	743.573
4	747.219	315.131	268.705	196.350	755.985
Exponential Variogram - 10% outliers					
1	1.152.196	217.412	51.002	21.652	1.167.444
2	1.143.526	351.862	206.261	103.604	1.159.428
3	1.166.803	458.268	315.021	194.810	1.181.755
4	1.176.336	541.807	447.674	271.093	1.192.505
Exponential Variogram - 15% outliers					
1	1.447.039	383.122	110.299	35.470	1.462.512
2	1.515.220	615.008	373.650	177.080	1.531.612
3	1.505.296	722.597	512.435	283.309	1.521.783
4	1.523.626	828.245	710.080	393.094	1.540.404
Wave Variogram - 5% outliers					
1	1.554.921	352.822	191.793	139.841	1.573.291
2	1.465.806	735.577	732.720	534.613	1.490.297
3	1.372.436	1.069.226	1.144.364	1.096.379	1.405.224
4	1.287.531	1.347.030	1.663.621	1.678.248	1.328.201
Wave Variogram - 10% outliers					
1	2.469.558	755.574	414.729	228.535	2.497.187
2	2.381.877	1.332.707	1.351.223	787.117	2.415.639
3	2.160.840	1.749.147	1.874.147	1.499.643	2.201.810
4	2.047.295	2.128.884	2.550.874	2.374.310	2.096.445
Wave Variogram - 15% outliers					
1	3.173.478	1.254.706	772.100	401.619	3.208.233
2	2.949.795	1.956.174	2.122.154	1.350.985	2.989.318
3	2.796.696	2.538.375	2.844.294	2.481.067	2.844.066
4	2.523.909	2.852.847	3.309.492	3.385.333	2.578.106

estimates errors. For contaminated data all the robust estimators were able to recover the geometric form of the spherical, exponential and wave theoretical variograms. Matheron's and Haslett's estimators were only able to recover the geometrical form of the wave model. The estimators in the robust class presented better results for the wave variogram in the initial lags in comparison to the non-robust ones. The simulations also show that in general the variogram estimators overestimate the true values of the theoretical variograms. These results were similar to those observed in Genton' paper (1998) for the spherical model. In relation to the Journel and Huijbregts' rule (1997), Matheron, Cressie-Hawkins and Genton's estimators presented more stable values around the reference number $\left[\frac{n}{2}\right]$, contrary to the Median estimator, which was very unstable in this neighborhood, and to the Haslett's one,which had the tendency to degenerate before the value $\left[\frac{n}{2}\right]$. Since in general the sample variogram values for $h=1,2,\ldots,\left[\frac{n}{2}\right]$ are used to fit the parameters of the true spatial model, the results presented in this paper are important because they show that Journel and Huijbregts' rule was not suitable for Haslett's and Median estimators.

4. Conclusions

The results showed that in the presence of outliers, Genton's and Median variogram estimators should be preferred for the estimation of all the theoretical variogram models presented in this paper. For data without outliers, Matheron's and Haslett's had better performance except for the wave variogram model. Therefore, a previous analysis for the presence of outliers in the data set is crucial in order to decide which variogram estimator is more appropriate. The Journel and Huijbregts' rule (1997) did not work very well for the Haslett's and Median estimators.

5. Acknowledgements

The authors are thankful to CAPES and CNPq for financial support.

6. References

BOX, G., E., P.JENKINS, G., M. ***Time series analysis: forecasting and control***. San Francisco: Holden Day, 1976. 575 p.

CHILÈ, J-P.DELFINER, P. ***Geostatistics: modeling spatial uncertainty***. New York: John Wiley, 1999. 695 p.

CRESSIE, N. ***Statistics for spatial data***. New York: John Wiley, 1993. 900 p.

CRESSIE, N., HAWKINS, M. Robust estimation of the variogram I. ***Mathematical Geology***, v. 12, n.2, p.115-125,1980.

GENTON, M. G. Highly robust variogram estimation. ***Mathematical Geology***,v. 30, n.2, p.213-221, 1998.

HASLETT, J. On the sample variogram and the sample autocovariance for non-stationary time series. ***The Statistician***, n. 46, p.475-485, 1997.

LARK, R.M. A comparison of some robust estimators of the variogram for use in soil survey. ***European Journal of Soil Sciences***, n.51, p. 137-157, 2000.

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

MARCHANT, B. P., LARK, R. M. Estimating variogram uncertainty. ***Mathematical Geology***, v.36, n. 8, p. 868-898, 2004.

MATHERON, G. Traité de géostatistique appliquée, Tome I. ***Mémoires du bureau de recherches géologiques et minières principles***. Paris: Editions Technip, 1962. n. 14. 333p.

MINGOTI, S., GUIMARÃES, A. L., ROSA, G. Describing the total number of diagnosed cases of aids by means of geostatistics. ***Revista de Matemática e Estatística***, v. 24, n.1, p.61-76, 2006.

ROUSSEEUW, P. J., CROUX. C. Alternatives of the median absolute deviation. ***Journal of American Statistics Association***, v. 88, n. 424, p.1273-1283, 1993.

SHARP, W. E. Stochastic simulation of semivariogram. ***Mathematical Geology***, v.14, n.5, p.445-457, 1982.

Artigo recebido em 26/02/2007 e aprovado em 27/07/2007.