Math Geosci (2011) 43: 363–377
DOI 10.1007/s11004-010-9308-3

Combining Robustness with Efficiency in the Estimation of the Variogram

Hilário Miranda · Manuela Souto de Miranda

Received: 31 March 2009 / Accepted: 4 August 2010 / Published online: 24 September 2010

Abstract In the present paper, we propose a new method for the estimation of the variogram, which combines robustness with efficiency under intrinsic stationary geostatistical processes. The method starts by using a robust estimator to obtain discrete estimates of the variogram and control atypical observations that may exist. When the number of points used in the fit of a model is the same as the number of parameters, ordinary least squares and generalized least squares are asymptotically equivalent. Therefore, the next step is to fit the variogram by ordinary least squares, using just a few discrete estimates. The procedure is then repeated several times with different subsets of points and this produces a sequence of variogram estimates. The final estimate is the median of the multiple estimates of the variogram parameters. The suggested estimator will be called multiple variograms estimator. This procedure assures a global robust estimator, which is more efficient than other robust proposals. Under the assumed dependence structure, we prove that the multiple variograms estimator is consistent and asymptotically normally distributed. A simulation study confirms that the new method has several advantages when compared with other current methods.

Keywords Spatial statistics · Multiple variograms estimator · Robust estimator · Bounded influence function · Breakdown point

H. Miranda (✉)
Portucalense University, Infante D. Henrique, Rua Dr. António B. Almeida 541, 4200-072 Porto, Portugal
e-mail: hmiranda@upt.pt

M. Souto de Miranda
Department of Mathematics, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal
e-mail: manuela.souto@ua.pt

1 Introduction

The variogram plays an important role in spatially distributed random processes since it describes the dependence structure of the spatial process and it decisively influences the prediction of the process at unobserved locations (kriging). It is important then to estimate the variogram using estimators with good theoretical properties in order to assure accurate variogram estimates which produce good kriging results. The usual variogram estimation procedure consists of two steps. In the first one, the variogram is estimated directly from the process sample at specific lags, resulting in a finite set of discrete variogram estimates. In this step, the most commonly used estimator is the method of moments estimator proposed by Matheron (1962). This nonparametric estimator of the variogram has many nice properties, like unbiasedness and consistency in a pointwise sense. However, the Matheron estimator just provides discrete estimates for posterior fit of a parametric model of a valid variogram. In the traditional approach, the second step of the variogram estimation is performed using the least squares method. Since the nonparametric estimates of the variogram are correlated (because the same process observation is used to estimate the variogram at different lags) the generalized least squares (*GLS*) estimator is the most adequate method. Lahiri et al. (2002) proved that the *GLS* estimator is asymptotically efficient. Nevertheless, the *GLS* criterion is not feasible since the exact expression for the covariance matrix of the nonparametric estimator is very difficult to obtain, even for Gaussian processes. Furthermore, as Lahiri et al. (2002) mention, inversion of the covariance matrix of the nonparametric variogram estimator and minimization by the *GLS* criterion often proves to be computationally prohibitive. Given these difficulties, a solution can be attained by approaching the *GLS* by weighted least squares (*WLS*). There are several proposals for the weights. The classical and most commonly used was proposed by Cressie (1985).

The traditional procedures for the estimation of the variogram have good properties, but they are not robust. Genton (1998a) points out that the Matheron estimator has a null breakdown point and an unbounded influence function. Notice that nowadays the robustness of an estimator is often evaluated through its breakdown point and its influence function. The empirical breakdown point of an estimator is the limiting proportion of atypical data that the estimator can handle—a robust estimator has a non-null breakdown point. The influence function of an estimator measures the relative effect of each observation towards the value of the estimate, and it can be interpreted as a derivative of the estimator—a robust estimator has a bounded influence function (see, e.g., Hampel et al. 1986; Maronna et al. 2006). Thus, the classical nonparametric estimator of the variogram behaves poorly if the model assumptions are not valid. The *WLS* proposed by Cressie (1985) is also not robust.

The lack of robustness can have worse consequences in the estimation of the variogram than in other popular models because any single observation contributes to the computation of many increments of the process. This fact implies that the existence of just one contaminated observation might strongly affect the results. Therefore, the use of a robust nonparametric estimator of the variogram is essential and it also determines the robustness properties of the final estimator. If the nonparametric estimates are controlled in the first step of the estimation method, the second step can be performed with a nonrobust estimator without loosing robustness.

It is also possible to consider the use of robust procedures in both steps of the variogram estimation. However, when the degree of robustness is increased, generally the efficiency gets lower under Gaussian processes. Hence, the use of robust estimators in both steps of the variogram estimation is not recommended since the gain in robustness is low when compared with the efficiency that is lost. So, for improving both robustness and efficiency, we should maximize the robustness of the nonparametric estimator in the first step, and maximize the efficiency of the fitting estimator in the second step. To maximize the robustness of the nonparametric estimator, Genton (1998a) proposed a highly robust variogram estimator that will be denoted henceforth by Q_n. This Q_n estimator achieves the maximum breakdown point without loosing too much efficiency, and it seems to be the best robust choice for the first step of the estimation procedure. To improve the use of the Q_n estimator, we concentrated our attention on the maximization of the efficiency of the method used in the second step. The desirable procedure would be to use the *GLS* estimator for fitting the variogram model to the nonparametric estimates. But the form of the Q_n estimator does not allow the explicit computation of its covariance structure, even in the independent scenery. Assuming independent observations, Genton (1998b) concluded through a simulation study that the covariance structure of the Q_n estimator could be approached by the corresponding covariance structure of the Matheron estimator. This made it possible to use the weights in Cressie (1985) and the *WLS* estimator for fitting a valid variogram model in the second step. However, the *WLS* estimator is not efficient and in geostatistical processes the data is hardly ever independent. Therefore, the second step of Genton's proposal can still be improved.

In the present paper, we propose a method that increases the global efficiency of the robust variogram estimator by considering two additional steps. The new method emphasizes robustness or efficiency, depending on the main criterion used in each stage. The asymptotic efficiency of the estimator used in the second step can be increased using a result of Lahiri et al. (2002) that compares the asymptotic efficiency of different least squares estimators. Applying such a result, we recommend that only a few nonparametric variogram estimates have to be computed, precisely as many as the number of parameters of the variogram model. Afterwards, we fit the variogram model using ordinary least squares (*OLS*), knowing that the procedure has an asymptotic efficiency that is equivalent to the asymptotic efficiency attained with the *GLS* estimator. We repeat both stages several times with different nonparametric estimates. The final variogram estimate is then selected among the curve estimates. The four steps that constitute the new estimation method are summarized and discussed in the following section. The global procedure will be called multiple variograms (*MultV*) estimator, suggested by the third step of the method. The *MultV* estimator revealed good properties. Besides robustness, we proved that the estimator is consistent and asymptotically normally distributed. Furthermore, a simulation study showed that the *MultV* estimator performs better than the estimators that are commonly used by practitioners, either with clean or contaminated data.

The paper is organized as follows. In Sect. 2, we introduce some notation, we define the *MultV* estimator and we discuss the motivation behind this proposal. In Sect. 3, we prove the consistency and the asymptotic normality of the *MultV* estimator. The simulation study is detailed in Sect. 4. Finally, we present some concluding remarks.

2 The Multiple Variograms Estimator

In this section, we introduce some definitions and notation that will be necessary for presenting the *MultV* estimator. Consider a spatial stochastic process $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$, where the domain D is a subset of $\mathbb{R}^d, d \geq 1$. Assume that $Z(\mathbf{s})$ satisfies the hypothesis of intrinsic stationarity, which ensures that

$$\forall \mathbf{s}_i, \mathbf{s}_j \in D, \quad \mathrm{E}\big[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\big] = 0,$$

$$\forall \mathbf{s}_i, \mathbf{s}_j \in D, \quad \mathrm{Var}\big[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\big] = 2\gamma(\mathbf{s}_i - \mathbf{s}_j).$$

The function 2γ is called the variogram, and it is defined as the variance of the increments $Z(\mathbf{s}_i) - Z(\mathbf{s}_j)$. Let $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\}$ be a sample of the spatial process $Z(\mathbf{s})$. The classical nonparametric estimator of the variogram was proposed by Matheron (1962), and for a fixed $\mathbf{h} \in \mathbb{R}^d$, it is defined as

$$2\hat{\gamma}_M(\mathbf{h}) = \frac{1}{\#N(\mathbf{h})} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{h})} \big[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\big]^2,$$

where $N(\mathbf{h})$ is the set defined by

$$N(\mathbf{h}) = \big\{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j = \mathbf{h}; i, j = 1, \ldots, n\big\}.$$

This estimator is obtained by the method of moments since it results from equating the variance of the increments and their sample variance.

The Matheron estimator has good properties, such as unbiasedness and consistency, but it is not robust. Genton (1998a) stresses that the Matheron estimator has a null breakdown point and an unbounded influence function. The author proposed to estimate the variance of the increments with a robust scale estimator. The estimator of the standard deviation becomes

$$Q_{\#N(\mathbf{h})} = c \times \big\{\big|\big(Z(\mathbf{s}_i) - Z(\mathbf{s}_i + \mathbf{h})\big) - \big(Z(\mathbf{s}_j) - Z(\mathbf{s}_j + \mathbf{h})\big)\big| : i < j\big\}_{(k)},$$

where $Z(\mathbf{s}_i) - Z(\mathbf{s}_i + \mathbf{h})$ represents an increment of the process for a fixed vector $\mathbf{h}$, the index (k) stands for the kth order statistic with $k = \binom{[\#N(\mathbf{h})/2]+1}{2}$, and $c = 2.2191$ is a factor to achieve consistency for the Gaussian distribution. Therefore, for a fixed $\mathbf{h}$, the variogram can be estimated by

$$2\hat{\gamma}_Q(\mathbf{h}) = Q^2_{\#N(\mathbf{h})}.$$

This expression defines the Q_n estimator of the variogram. The estimator Q_n was adapted from Rousseeuw and Croux (1993) and possesses useful advantages. According to Genton (1998a), $2\hat{\gamma}_Q$ is consistent, it has a 50% breakdown point, it has a bounded influence function with gross error sensitivity $\gamma^* = 2.069$ for the standard Gaussian distribution, and it keeps a Gaussian asymptotic efficiency of 82%, which is close to the 100% of the Matheron estimator that is not robust. These properties show that the Q_n estimator is the nonparametric estimator of the variogram that better combines a high efficiency with the best robustness properties among the known

alternatives. Hence, this seems to be the best option for the robust nonparametric estimator of the variogram. But the covariance matrix of the Q_n estimator cannot be computed explicitly and this is a great disadvantage because it is impossible to fit a valid variogram model to the nonparametric estimates with the *GLS* estimator. The solution found in Genton (1998b) is an approximation of the *GLS* estimator, which induces a considerable loss of efficiency.

A remedy to the problem is the use of the *MultV* estimator that we describe. The *MultV* estimator is supported by an important property that was demonstrated in Lahiri et al. (2002). Adjusting the notation, write $2\gamma(\mathbf{h}, \boldsymbol{\theta}) = 2\gamma(\mathbf{h})$ to denote a valid variogram model with parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$ and assume that the model will be fitted to the vector of estimates $2\hat{\boldsymbol{\gamma}} = (2\hat{\gamma}(\mathbf{h}_1), \ldots, 2\hat{\gamma}(\mathbf{h}_H))$. In Lahiri et al. (2002), the authors showed that if the dimension q of the parameter $\boldsymbol{\theta}$ is equal to the dimension H of the vector $2\hat{\boldsymbol{\gamma}}$, then the *OLS*, the *WLS* and the *GLS* are all asymptotically efficient under the Gaussian model. Therefore, if we consider a number of nonparametric estimates of the variogram equal to the dimension of $\boldsymbol{\theta}$, that is, $H = q$, then we can use the *OLS* estimator for fitting the valid variogram model, since the *OLS* estimator turns to be asymptotically efficient. Hence, in computing the *MultV* estimator, we will always consider $H = q$.

For the most popular variogram models, the number of parameters is very small, and therefore only a few nonparametric estimates will be used in the fitting procedure. This fact requires special attention for the identifiability of the parameter. It is necessary to assure that the lags where the nonparametric estimates are computed result in identifiability conditions of $\boldsymbol{\theta}$ and, therefore, in the existence of a unique solution for the *OLS* estimator. To illustrate the issue, consider an isotropic variogram model (which depends on $\mathbf{h}$ only through its norm, e.g., the spherical variogram model). One needs to estimate the three traditional parameters (range, nugget effect, and sill) and so we will use just three nonparametric estimates of the variogram to fit the spherical model. It is possible to guarantee that the parameters are identifiable, taking estimates such that two lags are smaller than the range and the other one is greater than or equal to the range. If we consider fewer than two nonparametric estimates with lag smaller than the range, then the parameters of the spherical model become non-identifiable and the solution of the *OLS* estimator is non-unique. From now on, we will assume that the parameter $\boldsymbol{\theta}$ is identifiable.

Summarizing the main ideas of the procedure until this point, we conclude that the first step of the new estimator is the estimation of the variogram in a number of lags equal to the number of parameters of the valid variogram model that will be fitted. Those lags are randomly selected according to the identifiability conditions stated above. The nonparametric estimates are computed with the Q_n estimator. Thus, they are safe against outliers. Notice that this step is mainly concerned with the robustness of the procedure. Next, we fit a valid variogram model using *OLS*.

The *OLS* estimator is very sensitive to each individual point that contributes to the fit. When we fit the variogram model to the $H = q$ nonparametric estimates by *OLS*, the *OLS* estimate $\hat{\boldsymbol{\theta}}$ becomes strongly dependent on the few vectors $\mathbf{h}_1, \ldots, \mathbf{h}_H$ where the estimates are computed. Therefore, the obtained curve estimate $\hat{\boldsymbol{\theta}}$ gives a poor picture of the variogram, while it should be precise. The question is particularly important near the origin, where the dependence structure is stronger. Actually, the

estimate of the nugget effect is crucial since it describes the regularity of the spatial process; it also gives a measure of the microscale variation. To deal with this problem, we propose constructing a set of *OLS* estimates $\{\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_B\}$ with the same sample. This is motivated by the continuity of the spatial index of the geostatistical processes and it allows repeating the fitting procedure, varying the lags where the nonparametric estimates of the variogram are computed. To obtain different sets of lags, we repeat the random selection of $\mathbf{h}_1, \ldots, \mathbf{h}_H$. It is not possible to assure that there exist observed increments for each new selected $\mathbf{h}_i$. Therefore, we follow a procedure that is similar to the one used by practitioners in the computation of the classical estimates. The Q_n estimates are then computed with the increments contained in a neighborhood of each $\mathbf{h}_i$, for $i = 1, \ldots, H$. The constructed set of *OLS* estimates provides a better picture of the underlying $Z(\mathbf{s})$ dependence structure. Like the second step, this stage also aims at increasing the efficiency of the procedure. Finally, the *MultV* estimate is defined as a central estimate of the obtained set $\{\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_B\}$. We suggest the use of the median as a central measure since the efficiency of the method was assured in the two previous steps. This set of estimates of the variogram parameters define the final variogram estimate. In this stage, robustness is the main criterion.

To conclude the presentation of the *MultV* estimator, we summarize the procedure in the following *MultV* estimator algorithm:

1. Compute $H = q$ nonparametric estimates of the variogram with the Q_n estimator, obtaining $2\hat{\boldsymbol{\gamma}}_Q = (2\hat{\gamma}_Q(\mathbf{h}_1), \ldots, 2\hat{\gamma}_Q(\mathbf{h}_H))$. The vectors $\mathbf{h}_1, \ldots, \mathbf{h}_H$ must be randomly selected among those which assure that the parameters of the variogram model are identifiable. All the increments contained in a neighborhood of $\mathbf{h}_i$ contribute to the computation of $2\hat{\gamma}_Q(\mathbf{h}_i)$ in a similar way to the classical estimates.
2. Fit the variogram model $2\gamma(\mathbf{h}, \boldsymbol{\theta})$, with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$, by *OLS*, to the nonparametric estimates of the variogram obtained in the step above, obtaining $\hat{\boldsymbol{\theta}}_b = (\hat{\theta}_{b,1}, \ldots, \hat{\theta}_{b,q})$.
3. Repeat the steps above B times, varying the lags $\mathbf{h}_1, \ldots, \mathbf{h}_H$ where the nonparametric estimates are computed, thus obtaining the set of estimates $\{\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_B\}$ of the model parameters. B is the minimum integer such that the final variogram estimate becomes unchanged.
4. The *MultV* estimate is defined by $2\gamma(\mathbf{h}, \tilde{\boldsymbol{\theta}}_B)$, where

$$\tilde{\boldsymbol{\theta}}_B = \big(\text{Median}\{\hat{\theta}_{1,1}, \ldots, \hat{\theta}_{B,1}\}, \ldots, \text{Median}\{\hat{\theta}_{1,q}, \ldots, \hat{\theta}_{B,q}\}\big).$$

Notice that the first and last steps aim at improving the robustness of the procedure. On the other hand, Step 2 and Step 3 improve the efficiency of the global estimator.

3 Properties of the *MultV* Estimator

In this section, we will see that the *MultV* estimator is robust in the sense that it has a bounded influence function and a positive breakdown point. Furthermore, we shall demonstrate that under mild conditions, the *MultV* estimator is consistent and asymptotically normally distributed. Analyzing the robustness of the *MultV* estimator, we can observe that if the nonparametric estimator of the variogram is not robust, then a

single outlier in the data can destroy all the nonparametric estimates of the variogram. Since a single outlier contaminates all nonparametric estimates of the variogram, the contamination becomes so strong that it cannot be removed during the next steps of the *MultV* estimator, even if we utilize highly robust estimators in those steps. Hence, the global estimator becomes nonrobust. On the other hand, if the nonparametric estimator of the variogram is robust, then it will control the outliers that might exist in the data. Therefore, the nonparametric estimates of the variogram will be clean of outliers. In this way, the next step of the *MultV* estimator can be performed with efficient estimators that are not necessarily robust, maintaining the robustness of the global estimator. Thus, the robustness of the *MultV* estimator is determined by the robustness of the nonparametric estimator of the variogram which is used in the first step, that is, the robustness properties of the *MultV* estimator are determined by the robustness properties of the Q_n estimator.

Using results from Rousseeuw and Croux (1993), Genton (1998a) stated that the Q_n estimator of the variogram has a bounded influence function and a 50% breakdown point. However, these two results were stated for the process of increments of $Z(\mathbf{s})$, that is, $\{Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h}), \mathbf{h} \in \mathbb{R}^d, \mathbf{s} \in D \text{ such that } \mathbf{s}+\mathbf{h} \in D\}$, with a fixed vector $\mathbf{h}$, on which the Q_n estimator is applied. But in Geostatistics, we are much more interested in the computation of the breakdown point that is associated with the initial spatial process $Z(\mathbf{s})$, which is called the spatial breakdown point. Genton (1998c) studied this problem. The author affirmed that the spatial breakdown point of the Q_n estimator cannot be computed exactly because it is a very difficult numerical problem. However, he also concluded that the number of initial data from $\{Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)\}$ which can be perturbed without destroying the Q_n estimates is roughly 30%. Genton (1998c) confirmed this result by simulations and theoretical results. Nevertheless, the author advises that there exist particular configurations of perturbation, which he called most unfavorable configurations, for which the maximal number of initial data which can be perturbed is much lower than 30%. Since the breakdown point of the *MultV* estimator coincides with the breakdown point of the Q_n estimator, we can conclude that the *MultV* estimator has a positive breakdown point, which is roughly 30%.

To establish the asymptotic properties of the *MultV* estimator, it is convenient to analyze separately the four different steps of the method. In the first step, the nonparametric estimates of the variogram are computed with the Q_n estimator. Its properties are already studied in Genton (1998a) under the same conditions considered herein, thus we use them directly. In that paper, the author confirms that the Q_n estimator is consistent in a pointwise sense. That is, for any fixed vector $\mathbf{h}$,

$$2\hat{\gamma}_Q(\mathbf{h}) \xrightarrow{\mathcal{P}} 2\gamma_0(\mathbf{h}),$$

where $\mathcal{P}$ means convergence in probability and $2\gamma_0(\mathbf{h})$ represents the true variogram value at $\mathbf{h}$. On the other hand, we know from Rousseeuw and Croux (1993) that the Q_n estimator is asymptotically normally distributed.

The properties of the method in the second step are a consequence of the results stated in Lahiri et al. (2002). The authors showed that, under some regularity conditions, the properties of the estimator used in the first step are reflected in the *OLS*

estimator. Therefore, assuming that the Q_n estimator is consistent and it has a normal asymptotic distribution, the *OLS* estimator is also consistent and it has also a normal asymptotic distribution. Using mathematical notation, one has

$$\hat{\boldsymbol{\theta}}_b \xrightarrow{\mathcal{P}} \boldsymbol{\theta}_0,$$

where $\boldsymbol{\theta}_0$ is the true parameter of the variogram, and

$$\hat{\boldsymbol{\theta}}_b \xrightarrow{\mathcal{L}} N_q(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_b),$$

where $\mathcal{L}$ represents convergence in law, N_q stands for the q-dimensional normal distribution, and the covariance matrix $\boldsymbol{\Sigma}_b$ has a specific form that can be found in Lahiri et al. (2002), which depends on the covariance matrix of the estimator used in the first step.

In the third step, we obtain the set $\{\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_B\}$ of the parameter estimates. These *OLS* estimates are obtained in an independent manner because the lags where the nonparametric estimates of the variogram are computed are selected independently. Finally, it is necessary to investigate the properties of the results obtained in the fourth step. In this step, the median of the obtained *OLS* estimates is computed, that is, we compute

$$\tilde{\boldsymbol{\theta}}_B = \big(\text{Median}\{\hat{\theta}_{1,1}, \ldots, \hat{\theta}_{B,1}\}, \ldots, \text{Median}\{\hat{\theta}_{1,q}, \ldots, \hat{\theta}_{B,q}\}\big).$$

The consistency and the asymptotic normal distribution of the median estimator is well known when the variables are independent and identically distributed (*i.i.d.*). However, that is not the case. Recall that for any $i = 1, \ldots, q$, the random variables $\hat{\theta}_{1,i}, \ldots, \hat{\theta}_{B,i}$ are independent, and that when n is sufficiently large, they converge to the normal distribution. Nevertheless, they are not identically distributed since they have different variances. Notice that in every loop of the procedure, the lags of the nonparametric estimates are different, thus implying that the variance of the successive *OLS* estimators varies.

The following results are convenient generalizations of the *i.i.d.* scenery. We assume that the sample size n is sufficiently large for applying the convergence in law cited above.

Theorem 1 Assume that $\hat{\boldsymbol{\theta}}_b = (\hat{\theta}_{b,1}, \ldots, \hat{\theta}_{b,q})$, $b = 1, \ldots, B$ are independent random vectors and let $\tilde{\boldsymbol{\theta}}_B = (\text{Median}\{\hat{\theta}_{1,1}, \ldots, \hat{\theta}_{B,1}\}, \ldots, \text{Median}\{\hat{\theta}_{1,q}, \ldots, \hat{\theta}_{B,q}\})$. If for all $b = 1, \ldots, B$, $\hat{\boldsymbol{\theta}}_b \sim N_q(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_b)$, then

$$\tilde{\boldsymbol{\theta}}_B \xrightarrow[B\to\infty]{\mathcal{P}} \boldsymbol{\theta}_0.$$

Proof For each fixed $i = 1, \ldots, q$, let

$$\bar{F}_{B,i} = \frac{1}{B} \sum_{b=1}^{B} F_{b,i},$$

where $F_{b,i}$ is the probability distribution function of the random variable $\hat{\theta}_{b,i}$. Since $\hat{\boldsymbol{\theta}}_b \sim N_q(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_b)$, then $\hat{\theta}_{b,i} \sim N(\theta_{0,i}, \sigma^2_{b,i})$ and the symmetry of the distribution assures that

$$\bar{F}_{B,i}(\theta_{0,i}) = \frac{1}{B}\sum_{b=1}^{B} F_{b,i}(\theta_{0,i}) = \frac{1}{2}.$$

On the other hand, as $F_{b,i}, b = 1, \ldots, B$, are all strictly increasing functions for fixed b, then $\bar{F}_{B,i}$ is also a strictly increasing function. Therefore, for all $\varepsilon > 0$, we have $1/2 < \bar{F}_{B,i}(\theta_{0,i} + \varepsilon) < 1$ and $0 < \bar{F}_{B,i}(\theta_{0,i} - \varepsilon) < 1/2$. As a consequence,

$$\sqrt{B}\left(\bar{F}_{B,i}(\theta_{0,i} + \varepsilon) - \frac{1}{2}\right) \underset{B}{\longrightarrow} \infty$$

and

$$\sqrt{B}\left(\frac{1}{2} - \bar{F}_{B,i}(\theta_{0,i} - \varepsilon)\right) \underset{B}{\longrightarrow} \infty.$$

The two above conditions, in addition to the independence of the $\hat{\boldsymbol{\theta}}_b$, for $b = 1, \ldots, B$, assure the necessary conditions for applying the result of Theorem 1 in Mizera and Wellner (1998) and conclude that

$$\text{Median}\{\hat{\theta}_{1,i}, \ldots, \hat{\theta}_{B,i}\} \underset{B}{\xrightarrow{\mathcal{P}}} \theta_{0,i}.$$

The consistency of the sample median was verified for all $i = 1, \ldots, q$. Thus, we can conclude that $\tilde{\boldsymbol{\theta}}_B$ converges in probability to $\boldsymbol{\theta}_0$ as $B \to \infty$. The result presented in Theorem 1 confirms that the *MultV* estimator is consistent under general regularity conditions. Now we state the asymptotic distribution of the *MultV* estimator.

Theorem 2 Assume that the conditions of Theorem 1 hold and suppose that $\lim_{B\to\infty} \frac{1}{B}\sum_{b=1}^{B} f_{b,i}(\theta_{0,i})$ is a positive number, where $f_{b,i}$ is the normal probability density function of $\hat{\theta}_{b,i}$, for all $i = 1, \ldots, q$. Then,

$$\sqrt{B}\big(\text{Median}\{\hat{\theta}_{1,i}, \ldots, \hat{\theta}_{B,i}\} - \theta_{0,i}\big) \underset{B}{\xrightarrow{\mathcal{L}}} N\left(0, \left(2 \lim_{B\to\infty} \frac{1}{B}\sum_{b=1}^{B} f_{b,i}(\theta_{0,i})\right)^{-2}\right).$$

Proof According to the assumptions of the theorem, for a fixed index $i = 1, \ldots, q$, the random variables $\hat{\theta}_{1,i}, \ldots, \hat{\theta}_{B,i}$ are independent and follow a normal distribution, with mean $\theta_{0,i}$ and variance $\sigma^2_{b,i}$. The proof will be complete taking into account a result in Koenker (2005), namely, Theorem 4.1 of that book. This theorem assures that for any quantile τ $(0 < \tau < 1)$, the quantile regression estimator follows an asymptotic normal distribution, even when the errors of the quantile regression model are heteroscedastic. In the present proof, we make use of Koenker's result with $\tau = 0.5$, letting $x_{b,i} = 1$ for all $b = 1, \ldots, B$. Proceeding in this way, for each fixed component i of the parameter $\boldsymbol{\theta}$, we obtain the model

$$\hat{\theta}_{b,i} = \theta_{0,i} + \epsilon_{b,i},$$

where $\epsilon_{b,i}$ are the model errors and $Q_{0.5}(\hat{\theta}_{b,i}|x_{b,i}) = \theta_{0,i}$, for all $b = 1, \ldots, B$. If the above model satisfies the conditions of Koenker's Theorem 4.1, then the mentioned result concludes the present proof. Therefore, the present proof consist of showing that the conditions of Koenker's Theorem 4.1 hold. Condition A1 holds since the random variables $\hat{\theta}_{b,i}$ follow a normal distribution that is absolutely continuous and verifies $0 < f_{b,i}(\theta_{0,i}) < \infty$. Conditions A2 also hold because their original matrix form becomes a positive number. Thus,

1. $\lim_{B\to\infty} \frac{1}{B} \sum_{b=1}^{B} x_{b,i} x_{b,i}^{\mathrm{T}} = 1 > 0$ (T stands for the transpose matrix);
2. $\lim_{B\to\infty} \frac{1}{B} \sum_{b=1}^{B} f_{b,i}(\theta_{0,i}) x_{b,i} x_{b,i}^{\mathrm{T}} = \lim_{B\to\infty} \frac{1}{B} \sum_{b=1}^{B} f_{b,i}(\theta_{0,i}) > 0$; and
3. $\max_{b=1,\ldots,B} \frac{\|x_{b,i}\|}{\sqrt{B}} = \frac{1}{\sqrt{B}} \underset{B\to\infty}{\longrightarrow} 0.$

Since all the conditions of Koenker's theorem hold, the proof is complete. This result confirms the asymptotic normal distribution of the *MultV* estimator.

4 Simulation Study

To investigate the performance of the *MultV* estimator, we carried out a simulation study. We compared the *MultV* estimator with the traditional estimator and with a robust alternative. More precisely, we compared the *MultV* estimator with the following alternatives: first, the classical estimator uses the Matheron estimator to obtain the nonparametric estimates of the variogram and the *WLS* estimator with the weights in Cressie (1985) for fitting the valid variogram model. Second, Q_n with *WLS* uses the Q_n estimator to compute the nonparametric estimates of the variogram and the *WLS* estimator with the weights in Cressie (1985) for fitting the valid variogram model.

We simulated samples from Gaussian geostatistical processes. Following the work of Genton (1998a), the Gaussian samples were simulated from a process with an isotropic spherical variogram, given by

$$\gamma\big(\|\mathbf{h}\|;\boldsymbol{\theta}\big) = \begin{cases} 0, & \text{if } \|\mathbf{h}\| = 0, \\ \tau^2 + \sigma^2\left[\frac{3\|\mathbf{h}\|}{2\phi} - \frac{\|\mathbf{h}\|^3}{2\phi^3}\right], & \text{if } 0 < \|\mathbf{h}\| \le \phi, \\ \tau^2 + \sigma^2, & \text{if } \|\mathbf{h}\| > \phi, \end{cases}$$

where $\boldsymbol{\theta} = (\tau^2, \sigma^2, \phi)$ and $\phi > 0$. All the processes were generated with the following parameters: $\phi = 15$ (range); $\tau^2 = 1$ (nugget effect); and $\tau^2 + \sigma^2 = 3$ (sill). Since the chosen variogram model is isotropic, the samples were generated in $\mathbb{R}^1$, as in Genton (1998a). The sample locations were randomly selected in a line segment with 200 units length. We considered two different sample sizes in order to investigate the performance of the estimators as the sample size increases. We simulated samples with 50 observations and with 200 observations.

To study how the different estimators behave in the presence of contamination, we considered several cases. In each case, we randomly replaced $\epsilon\%$ of the data by new values. These values were generated as a random sample from a Gaussian distribution $N(0, \sigma^2)$. The following situations were considered:

A.1 Sample without contamination;
A.2 Sample 10% contaminated with $\sigma = 5$;
A.3 Sample 20% contaminated with $\sigma = 5$;
A.4 Sample 30% contaminated with $\sigma = 5$;
A.5 Sample 10% contaminated with $\sigma = 10$; and
A.6 Sample 10% contaminated with $\sigma = 20$.

For each of the six situations, we simulated 1000 samples. Notice that situations A.2, A.3, and A.4 are useful for investigating the performance of the variogram estimators as the degree of contamination increases. In those situations, the contaminated observations were generated from the same distribution. On the other hand, situations A.2, A.5, and A.6 are useful for investigating the behavior of the variogram estimators as we increase the variance of the distribution that generates the contamination. In those situations, the number of contaminated observations remains the same. The three variogram estimators were computed taking into account the practical rules recommended by Journel and Huijbregts (1978). Therefore, the nonparametric estimates were only computed at lags smaller than a half of the maximum distance between the sample locations; the nonparametric variogram estimates were computed only for the cases with at least 30 increments. The *MultV* estimator was calculated using $B = 250$ multiple variogram estimates, that is, the loop that iterates the first two steps was repeated 250 times.

To evaluate the performance of the estimators, we used the mean square error. Hence, for every situation under study, we computed the empirical mean square error

$$\mathrm{EMSE}(\bar{\theta}_i^*) = \frac{1}{1000}\sum_{j=1}^{1000}\left(\bar{\theta}_i^{*(j)} - \theta_i\right)^2, \quad i = 1, 2, 3,$$

where $\bar{\theta}_i^{*(j)}$ is the ith estimate of the variogram parameter that was obtained in the jth sample, and θ_i is the true value of the ith variogram parameter, which was used in the simulation of the processes. Every computation was performed using the *R* software (R Development Core Team 2008). We also used some additional *R* packages, namely, *geoR* and *robustbase*. The *geoR* package was constructed by Ribeiro and Diggle (2001). Along with this work, the *geoR* was helpful in the simulation of the Gaussian samples and for the computation of the classical variogram estimates. The package *robustbase* is devoted to robust methods, and it includes a function that was used for computing the Q_n estimates. We will provide the *R* macros to anyone who might be interested. The results that were obtained are shown in Table 1 ($n = 50$) and in Table 2 ($n = 200$).

From both tables, one can see that the *MultV* estimator performs better than the other two variogram estimators in almost every situation. Notice that the *MultV* estimator gives better results than the classical estimator, even when the sample is not

Table 1 Empirical mean square errors of the variogram estimators for samples with 50 observations

Spherical model with $\phi = 15$, $\tau^2 = 1$ and $\sigma^2 = 2$ ($n = 50$)				
Contamination	Variogram estimator	EMSE($\hat{\phi}$)	EMSE($\hat{\tau}^2$)	EMSE($\hat{\tau}^2 + \hat{\sigma}^2$)
A.1	Math. *WLS*	402.615	0.318	1.375
	Q_n *WLS*	376.581	0.380	1.717
	Mult. variog.	141.389	0.253	1.113
A.2	Math. *WLS*	476.080	7.590	12.667
	Q_n *WLS*	415.593	1.732	6.355
	Mult. variog.	166.621	1.489	5.549
A.3	Math. *WLS*	428.559	24.019	31.948
	Q_n *WLS*	387.574	6.060	16.966
	Mult. variog.	214.242	6.124	17.098
A.4	Math. *WLS*	352.776	52.121	66.025
	Q_n *WLS*	413.158	19.210	40.767
	Mult. variog.	231.812	19.287	41.931
A.5	Math. *WLS*	467.990	98.230	216.365
	Q_n *WLS*	478.496	2.885	19.395
	Mult. variog.	222.453	2.527	14.027
A.6	Math. *WLS*	407.576	1522.922	3773.975
	Q_n *WLS*	526.673	3.956	35.813
	Mult. variog.	238.036	3.721	24.565

contaminated. This is surprisingly good because the classical estimator should perform better in the non-contaminated samples. However, the use of the *WLS* to approximate the *GLS* makes the classical estimator to lose much efficiency. That is probably one of the main reasons which justify the fact that the *MultV* estimator performs better than the classical estimator under non-contaminated samples. Besides, the *MultV* estimator also performs better than the Q_n estimator with *WLS*, especially in the estimation of the range of the variogram model. According to Genton (1998b), the range is the most important parameter of the spherical model since it is the only parameter that influences the kriging weights. Thus, the results show that the *MultV* estimator is better than the Q_n estimator with the *WLS*. The same comments are valid for both sample sizes and for different degrees of contamination. Other studies with spatial samples of isotropic models simulated in the plane (not considered in the present paper) led to identical conclusions. In conclusion, the *MultV* estimator performed better than the alternatives considered here, either with contaminated samples or even without contamination. Notice that we just dealt with simple models of the variogram which are isotropic, but these models are used most of the time. Therefore, the simulation study confirmed that the *MultV* estimator is an interesting solution for the robust estimation of the variogram.

Table 2 Empirical mean square errors of the variogram estimators for samples with 200 observations

Spherical model with $\phi = 15$, $\tau^2 = 1$ and $\sigma^2 = 2$ ($n = 200$)				
Contamination	Variogram estimator	EMSE($\hat{\phi}$)	EMSE($\hat{\tau}^2$)	EMSE($\hat{\tau}^2 + \hat{\sigma}^2$)
A.1	Math. *WLS*	289.850	0.103	0.539
	Q_n *WLS*	369.682	0.117	0.663
	Mult. variog.	40.021	0.043	0.462
A.2	Math. *WLS*	395.172	7.104	6.506
	Q_n *WLS*	284.331	0.714	2.619
	Mult. variog.	34.998	0.574	2.206
A.3	Math. *WLS*	459.959	26.146	22.218
	Q_n *WLS*	207.917	4.168	9.085
	Mult. variog.	56.484	3.903	8.844
A.4	Math. *WLS*	457.074	56.074	47.923
	Q_n *WLS*	264.367	14.796	23.327
	Mult. variog.	95.673	14.560	24.859
A.5	Math. *WLS*	529.829	98.081	109.556
	Q_n *WLS*	238.009	1.421	7.410
	Mult. variog.	43.019	1.150	6.411
A.6	Math. *WLS*	266.372	783.580	1883.447
	Q_n *WLS*	312.207	2.069	15.237
	Mult. variog.	68.785	1.694	13.087

The evaluation of time efficiency of the *MultV* estimator must take into account that the algorithm is a compound method, which uses the Q_n estimator in each iteration. The total computation time depends on the number of iterations used. Thus, clearly the *MultV* estimator consumes much more time than its alternatives, but the tradeoff between time consumption and accuracy of the robust estimates is valuable. Moreover, the *MultV* estimates are easily computed in most personal computers. For illustration purposes, we present the times that were needed to compute the estimates of the variogram in a real data set. We considered the *soil250* data set that can be found in the *geoR* package. The data set contains 250 observations of several chemical soil properties measured on a 10 times 25 regular grid with squares whose sides measure 5 meters (for more information on the *soil250* data set, see Bassoi 1994). We estimated the variogram for the potassium content using the three estimators studied above. On an Intel(R) Core(TM)2 Duo CPU with 2.66 GHz and 4.00 GB of RAM memory, the classical estimator consumed 0.30 seconds, the Q_n estimator with *WLS* consumed 2.76 seconds, and the *MultV* estimator with $B = 250$ repetitions of the algorithm consumed 17.86 seconds. However, we must advocate that the obtained times for the Q_n estimator with *WLS* and for the *MultV* estimator can still be improved with the development of a more efficient function for each estimator.

5 Conclusions

In this paper, we presented a new method for the estimation of the variogram function, called the multiple variograms (*MultV*) estimator. The estimator is constructed along four steps, combining priority criteria of robustness with efficiency. The method has several advantages in each of its stages. In the first place, it uses a robust procedure for computing discrete estimates of the variogram. This prevents against the action of atypical observations whose influence would be magnified by their contribution to different increments of the process. Besides, computational costs are less when robust techniques are used in an initial step. The procedure gains efficiency when compared with other robust procedures, and it is very simple for interpretation and computation. The reduced number of points used in the second step of the method is compensated by the computation of multiple variogram estimates, obtained independently with the same procedures. This step was inspired by the continuity of the variogram function and contributes to the reduction of the variance of the estimators of the parameters of the model. The last step just selects the central tendency of the variogram function estimates already obtained with great efficiency. The use of the median in this stage is consistent with robust and resistant concerns.

Statistical properties of the *MultV* estimator were proven taking into account its several stages. In spite of the dependence structure, we established that under mild conditions the *MultV* estimator is robust in the sense that it has a bounded influence function and a positive breakdown point. The *MultV* estimator is also consistent for estimating the true variogram function and it converges in law to the normal distribution.

From a computational evaluation, it should be noted that the estimates were obtained with free software, particularly using the programme *R* and several specific packages which are quite tested over the world. A simulation study confirmed that the results were very satisfactory when the process is Gaussian and also in the presence of contamination. The study was conducted considering five scenarios of contamination and different sample sizes. In every situation, the performance of the *MultV* estimator was superior when compared with the alternatives. In fact, the proposed method revealed significant advantages in the estimation of the range and nugget effect. The quality of these estimates is essential since they strongly influence the shape of the variogram near the origin. Finally, we conclude that the *MultV* estimator is a promising method for combing robustness with efficiency in the estimation of the variogram.

Acknowledgements This work was partially supported by the R&D unit CIDMA, via FCT and the EC fund FEDER/POCI 2010.

References

Bassoi L (1994) Nitrato no solo e acumulo de N pelo milho (Zea mays L) fertirrigado. PhD Thesis, University of São Paulo, Brazil

Cressie N (1985) Fitting variogram models by weighted least squares. J Int Assoc Math Geol 17:693–702

Genton M (1998a) Highly robust variogram estimation. Math Geol 30(2):213–221

Genton M (1998b) Variogram fitting by generalized least squares using an explicit formula for the covariance structure. Math Geol 30(4):323–345
Genton M (1998c) Spatial breakdown point of variogram estimators. Math Geol 30(7):853–871
Hampel F, Ronchetti E, Rousseeuw P, Stabel W (1986) Robust statistics: the approach based on influence functions. Wiley, New York
Journel A, Huijbregts C (1978) Mining geostatistics. Academic Press, London
Koenker R (2005) Quantile regression. Cambridge University Press, Cambridge
Lahiri S, Lee Y, Cressie N (2002) On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters. J Stat Plan Inference 103:65–85
Maronna R, Martin R, Yohai V (2006) Robust statistics—theory and methods. Wiley, London
Matheron G (1962) Traite de geostatistique appliquee, vol I. Memoires du bureau de recherches geologiques et minieres, vol 14. Editions Technip, Paris
Mizera I, Wellner J (1998) Necessary and sufficient conditions for weak consistency of the median of independent but not identically distributed random variables. Ann Stat 26(2):672–691
R Development Core Team (2008) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
Ribeiro P Jr., Diggle P (2001) geoR: A package for geostatistical analysis. R-News 1(2):15–18
Rousseeuw P, Croux C (1993) Alternatives to the median absolute deviation. J Am Stat Assoc 88(424):1273–1283