2

Robustness Problems in the Analysis of Spatial Data

Marc G. Genton

ABSTRACT Kriging is a widely used method of spatial prediction, particularly in earth and environmental sciences. It is based on a function which describes the spatial dependence, the so called variogram. Estimation and fitting of the variogram, as well as variogram model selection, are crucial stages of spatial prediction, because the variogram determines the kriging weights. These three steps must be carried out carefully, otherwise kriging can produce noninformative maps. The classical variogram estimator proposed by Matheron is not robust against outliers in the data, nor is it enough to make simple modifications such as the ones proposed by Cressie and Hawkins in order to achieve robustness. The use of a variogram estimator based on a highly robust estimator of scale is proposed. The robustness properties of these three variogram estimators are analyzed by means of the influence function and the classical breakdown point. The latter is extended to a spatial breakdown point, which depends on the construction of the most unfavorable configurations of perturbation. The effect of linear trend in the data and location outliers on variogram estimation is also discussed. Variogram model selection is addressed via nonparametric estimation of the derivative of the variogram. Variogram estimates at different spatial lags are correlated, because the same observation is used for different lags. The correlation structure of variogram estimators has been analyzed for Gaussian data, and then extended to elliptically contoured distributions. Its use for variogram fitting by generalized least squares is presented. Results show that our techniques improve the estimation and the fit significantly. Two new SPLUS functions for highly robust variogram estimation and variogram fitting by generalized least squares, as well as a MATLAB code for variogram model selection via nonparametric derivative estimation, are available on the Web at http://www-math.mit.edu/~genton/.

1 Introduction

In statistics, the concept of robustness is usually defined as the lack of sensitivity of a particular procedure to departures from the model assumptions. For example, a proportion of 10–15% of contaminated observations, called outliers, can sometimes be found in real data sets (Hampel, 1973, Huber, 1977). These outlying values may be due to gross errors, measure-

M. Moore (ed.), *Spatial Statistics: Methodological Aspects and Applications*

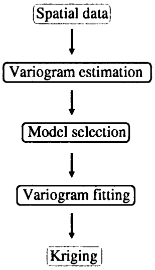


Figure 1. Illustration of the steps from the data to the spatial prediction called kriging. We want to robustify the variogram estimation, model selection, and variogram fitting steps.

ment mistakes, faulty recordings, and can seriously affect the results of a statistical analysis. Therefore, it is of prime interest to provide robust, i.e., reliable, estimators and methods. However, robust procedures are typically more difficult to develop in the spatial statistics context than in the classical one, because different types of outliers may occur. For instance, outliers can replace, or be added to, some observations of the underlying stochastic process. Furthermore, the configuration of spatial locations, where the contaminations occur, becomes important: isolated and patchy outliers can have different effects, and both have been observed in practice. In fact, the main problem is that estimators which take account of the spatial structure of the observations, are not invariant under permutation of the data, as in the case of estimators for independent and identically distributed (i.i.d.) observations. The analysis of spatial data often follows the steps summarized in Figure 1. From the spatial data, one usually models the spatial dependence structure between the observations, i.e. the so called variogram. The first step consists in estimating the variogram at various lag distances. Next, a valid variogram model has to be selected, and is fitted to the variogram estimates in the first step. Finally, this variogram is used in the spatial prediction procedure called kriging (Cressie, 1993). As one can see, the modeling of the variogram is an important stage of spatial prediction, because it determines the kriging procedure. Therefore, it is important to have a variogram estimator which remains close to the true underlying variogram, even if outliers are present in the data. Otherwise, kriging can produce noninformative maps. Of course, one might argue that any reasonable exploratory data analysis would identify and remove outliers in the data. However, this approach often contains a subjective aspect (Genton and Furrer, 1998a) that we would like to avoid, or at least to minimize.

In this paper, we bring together several ideas (Genton, 1998a,b,c, 1999, Gorsich and Genton, 1999) about the robustification of the three steps involved in the modeling of the variogram. First, in the next section, we use a highly robust estimator of the variogram and derive some of its properties.

Second, in Section 3, the issue of variogram model selection is addressed via nonparametric estimation of the derivative of the variogram. Finally, in Section 4, we use the correlation structure of variogram estimates to fit a variogram model by generalized least squares.

2 Highly Robust Variogram Estimation

Variogram estimation is a crucial stage of spatial prediction, because it determines the kriging weights. The most widely used variogram estimator is certainly the one proposed by Matheron (1962), although it is highly nonrobust to outliers in the data (Cressie, 1993, Genton, 1998a,c). One single outlier can destroy this estimator completely. The main reasons for this popularity are its simple appealing formulation and unbiasedness property. If $\{Z(\mathbf{x}) : \mathbf{x} \in D \subset \mathbb{R}^d\}$, $d \geq 1$, is a spatial stochastic process, ergodic and intrinsically stationary, then Matheron's classical variogram estimator, based on the method-of-moments, is

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{N_{\mathbf{h}}} \sum_{N(\mathbf{h})} \left(Z(\mathbf{x}_i) - Z(\mathbf{x}_j)\right)^2, \quad \mathbf{h} \in \mathbb{R}^d, \tag{1}$$

where $Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n)$ is a realization of the process, $N(\mathbf{h}) = \{(\mathbf{x}_i, \mathbf{x}_j) : \mathbf{x}_i - \mathbf{x}_j = \mathbf{h}\}$, and $N_{\mathbf{h}}$ is the cardinality of $N(\mathbf{h})$. It is not enough to make simple modifications to formula (1), such as the ones proposed by Cressie and Hawkins (1980), in order to achieve robustness. In this section, we advocate the use of a highly robust variogram estimator (Genton, 1998a)

$$2\hat{\gamma}(\mathbf{h}) = (Q_{N_{\mathbf{h}}})^2, \quad \mathbf{h} \in \mathbb{R}^d, \tag{2}$$

which takes account of all the available information in the data. It is based on the sample $V_1(\mathbf{h}), \ldots, V_{N_{\mathbf{h}}}(\mathbf{h})$ from the process of differences $V(\mathbf{h}) = Z(\mathbf{x}+\mathbf{h}) - Z(\mathbf{x})$ and the robust scale estimator $Q_{N_{\mathbf{h}}}$, proposed by Rousseeuw and Croux (1992, 1993)

$$Q_{N_{\mathbf{h}}} = 2.2191 \{|V_i(\mathbf{h}) - V_j(\mathbf{h})|; i < j\}_{(k)}, \tag{3}$$

where the factor 2.2191 is for consistency at the Gaussian distribution, $k = \binom{[N_{\mathbf{h}}/2]+1}{2}$, and $[N_{\mathbf{h}}/2]$ denotes the integer part of $N_{\mathbf{h}}/2$. This means that we sort the set of all absolute differences $|V_i(\mathbf{h}) - V_j(\mathbf{h})|$ for $i < j$ and then compute its kth order statistic (approximately the $\frac{1}{4}$ quantile for large $N_{\mathbf{h}}$). This value is multiplied by the factor 2.2191, thus yielding $Q_{N_{\mathbf{h}}}$. Note that this estimator computes the kth order statistic of the $\binom{N_{\mathbf{h}}}{2}$ interpoint distances. At first sight, the estimator $Q_{N_{\mathbf{h}}}$ appears to need $O(N_{\mathbf{h}}^2)$ computation time, which would be a disadvantage. However, it can be computed using no more than $O(N_{\mathbf{h}} \log N_{\mathbf{h}})$ time and $O(N_{\mathbf{h}})$ storage, by means of the fast algorithm described in Croux and Rousseeuw (1992). An SPLUS function for the highly robust variogram estimator, denoted `variogram.qn`, is available on the Web at `http://www-math.mit.edu/~genton/` .

The variogram estimator (2) possesses several interesting properties of robustness. For instance, its influence function (Hampel et al., 1986), which describes the effect on the estimator of an infinitesimal contamination, is bounded. This means that the worst influence that a small amount of contamination can have on the value of the estimator is finite, in opposition to Matheron's classical variogram estimator and Cressie and Hawkins' proposal. Another important robustness property is the breakdown point ε^* of a variogram estimator, which indicates how many data points need to be replaced by arbitrary values to make the estimator explode (tend to infinity) or implode (tend to zero). The highly robust variogram estimator has an $\varepsilon^* = 50\%$ breakdown point on the differences $V(\mathbf{h})$, which is the highest possible value. On the contrary, Matheron's classical variogram estimator and Cressie and Hawkins' estimator both have only an $\varepsilon^* = 0\%$ breakdown point, which is the lowest possible value. More details about the use and properties of this estimator, including some simulation studies, are presented in Genton (1998a,c).

The breakdown point discussed in the previous paragraph is based on the process of differences $V(\mathbf{h})$. However, in spatial statistics, one is also interested in the breakdown point related to the initial process Z. In this case, the effect of the perturbation of a point located on the boundary of the spatial domain D, or inside of it, can be quite different and depends notably on the lag vector $\mathbf{h}$. Therefore, a concept of spatial breakdown point of variogram estimators is introduced by Genton (1998c). Denote by I_m a subset of size m of $\{1,\dots,n\}$, and let $\mathcal{Z} = \{Z(\mathbf{x}_1),\dots,Z(\mathbf{x}_n)\}$. The spatial sample breakdown point of a variogram estimator $2\widehat{\gamma}(\mathbf{h}) = (S_{N_\mathbf{h}})^2$, based on a scale estimator $S_{N_\mathbf{h}}$, is defined by

$$\varepsilon_n^{Sp}(2\widehat{\gamma}(\mathbf{h}),\mathcal{Z}) = \max\left\{\frac{m}{n} \;\middle|\; \sup_{I_m}\sup_{\mathcal{Z}(I_m)} S_n(\mathcal{Z}(I_m)) < \infty \text{ and } \inf_{I_m}\inf_{\mathcal{Z}(I_m)} S_n(\mathcal{Z}(I_m)) > 0\right\}, \tag{4}$$

where $\mathcal{Z}(I_m)$ is the sample of size n, obtained by replacing m observations of $\mathcal{Z}$, indexed by I_m, by arbitrary values. This definition takes into account the configuration, i.e. the spatial location, of the perturbation, and seeks its worst spatial configuration for a fixed amount of perturbation and fixed lag vector $\mathbf{h}$. The spatial breakdown point of Matheron's classical variogram estimator, as well as the one of Cressie and Hawkins, is still zero. Figure 2 shows the spatial breakdown point of the highly robust variogram estimator, represented by the black curve, for each lag distance h and a sample of size $n = 100$. Upper and lower bounds on the spatial breakdown point are computed in Genton (1998c), and represented by light grey curves in Figure 2. The interpretation of this figure is as follows. For a fixed h, if the percentage of perturbed observations is below the black curve, the esti-

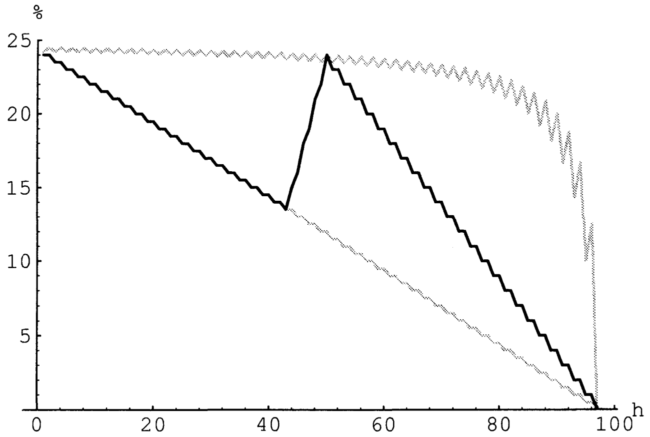


Figure 2. The spatial breakdown point (in black) as a function of the lag distance h, for the highly robust variogram estimator. The upper and lower bounds are drawn in light grey. Reprinted from *Mathematical Geology 30*(7), 853–871, with permission of the International Association for Mathematical Geology.

mator is never destroyed. If the percentage is above the black curve, there exists at least one configuration which destroys the estimator. This implies that highly robust variogram estimators are more resistant at small lags h or around $h = n/2$, than at large lags h or before $h = n/2$, according to Figure 2. The spatial breakdown point is a theoretical tool, indicating the worst-case behavior of the variogram estimator for each lag h. It allows to judge the resistance of the variogram estimates, and consequently their respective reliability. This is a local concept. However, in practice, one is generally confronted with a fixed configuration of perturbed data, which does not change with the lag h. Applied geostatisticians are usually concerned about the global effects (i.e., at all lags h) of a given configuration of perturbations on the estimation of the variogram. For that reason, Genton (1998c) carried out some simulations on the global effects of perturbations, and supported them by theoretical results. Table 1 presents the simulation (over 1000 replications) of the average percentage $\bar{p}$ of perturbed differences when m/n percent of observations of a regular grid of size $n = 10 \times 10 = 100$ are perturbed. For example, if $m/n = 20\%$ of observations Z of the grid are perturbed, then, on average, $\bar{p} = 36\%$ of differences V will be perturbed. Therefore, if a highly robust variogram estimator is used, with 50% breakdown point (on differences V), then it will have a global resistance to roughly 30% of outliers in the initial observations Z. Moreover, it turns out that this result is the same for every lag h. We note that if m/n is small, $\bar{p}$ equals approximatively $2m/n$, whereas it is slightly smaller if m/n is large. This decrease is due to differences taken between two perturbed observations. Simulations carried out on irregular grids showed similar behavior.

Table 1. Simulation (over 1000 replications) of the average percentage $\bar{p}$ of perturbed differences when m/n percent of observations of a regular grid of size $n = 10 \times 10 = 100$ are perturbed.

m/n	$\bar{p}$
5	10
10	19
15	28
20	36
25	44
30	51
40	64
50	75
60	84

Cressie (1993) investigated the effect of a linear trend on the estimation of the variogram. He considered a spatial stochastic process $Z(x)$ in $\mathbb{R}^1$, defined by $Z(x) = S(x) + \epsilon(x - (n+1)/2)$, $x = 1, \ldots, n$, where $S(x)$ is a zero-mean, unit variance, second-order stationary process, and ϵ is the degree of contamination. He showed that for Matheron's classical variogram estimator (1):

$$2\widehat{\gamma}_Z(h) \cong 2\widehat{\gamma}_S(h) + \epsilon^2 h^2, \tag{5}$$

in probability, where $2\widehat{\gamma}_Z(h)$ and $2\widehat{\gamma}_S(h)$ are the variograms of Z and S respectively. Thus, the effect of a linear trend contamination term of magnitude ϵ is that of an upward shift of magnitude $\epsilon^2 h^2$. This effect does not happen when using Q_{N_h}. Effectively, for the highly robust variogram estimator (2):

$$2\widehat{\gamma}_Z(h) = 2\widehat{\gamma}_S(h). \tag{6}$$

The highly robust variogram estimator is not affected by the linear trend contamination, because it is based on differences of the process $V(h)$.

Another interesting issue pointed out by Peter Guttorp during the workshop is the robustness of variogram estimators towards misspecification of the location of the data, i.e. towards location outliers. For instance, consider the very simple situation where the observations $Z(x_1), \ldots, Z(x_n)$ have a unidimensional and regular support. Assume that one location is misspecified, resulting in the exchange of the data at two locations x_i and x_j, i.e. new values $Z^*(x_i)$ and $Z^*(x_j)$. What is the effect on the estimation of the variogram? First, note that the set $N(h)$ will be modified for most lags h, depending on the spatial locations of x_i and x_j with regard to the border of the domain D. The resulting set $N^*(h)$ may be quite different from $N(h)$. Therefore, the final effect on variogram estimations depends on the sensitivity of the variogram estimator to changes in the set $N(h)$. Because Matheron's classical variogram estimator has no robustness properties, it is more sensitive than the highly robust variogram estimator. As an illustrative example, consider the five locations given in Figure 3, with

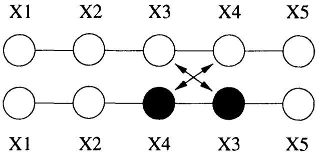


Figure 3. Illustration of one location outlier at x_3, resulting in the exchange of the two data at locations x_3 and x_4.

a location outlier at x_3. The locations x_3 and x_4 have been exchanged, resulting in the following sets $N(h)$ and $N^*(h)$:

$$\begin{aligned}
N(1) &= \{(x_1,x_2),(x_2,x_3),(x_3,x_4),(x_4,x_5)\}\\
N(2) &= \{(x_1,x_3),(x_2,x_4),(x_3,x_5)\}\\
N(3) &= \{(x_1,x_4),(x_2,x_5)\}\\
N(4) &= \{(x_1,x_5)\}\\
N^*(1) &= \{(x_1,x_2),(x_2,x_4),(x_3,x_4),(x_3,x_5)\}\\
N^*(2) &= \{(x_1,x_4),(x_2,x_3),(x_4,x_5)\}\\
N^*(3) &= \{(x_1,x_3),(x_2,x_5)\}\\
N^*(4) &= \{(x_1,x_5)\}
\end{aligned}$$

Although $N(4)$ and $N^*(4)$ are the same, only half of the elements of $N(1)$ (respectively $N(3)$) are the same as $N^*(1)$ (respectively $N^*(3)$). The worst case is for $h = 2$, where $N(2) \cap N^*(2) = \varnothing$. As a consequence, the estimation of the variogram at $h = 2$ could be seriously biased, depending on the robustness of the variogram estimator. Of course, these effects may be even worse depending on the number of location outliers, the intensity of the perturbation of locations, the dimension $d > 1$ of the spatial domain D, and the irregularity of the locations. Further research is needed to characterize the effects of such situations.

3 Variogram Model Selection

The variogram estimates obtained in the previous section cannot be used directly for kriging because they are not necessarily valid, i.e. conditionally negative definite. Therefore, a valid parametric variogram model must be chosen and fitted to the variogram estimates. The choice of the variogram model is important, since it affects the kriging procedure. Unfortunately, no selection technique can be found in the literature, apart from a priori knowledge about the underlying process, and the user's subjectivity. Figure 4 presents an example where the choice among a spherical, exponential, or Gaussian variogram is unclear, although the true underlying variogram of the data is exponential. In order to reduce the subjectivity in the choice

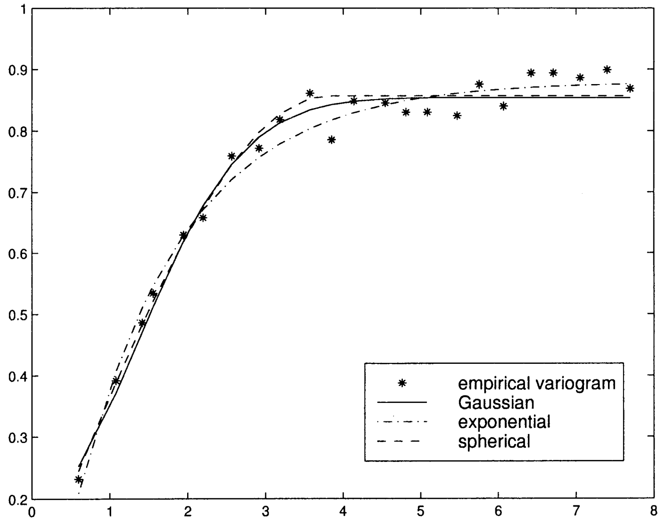


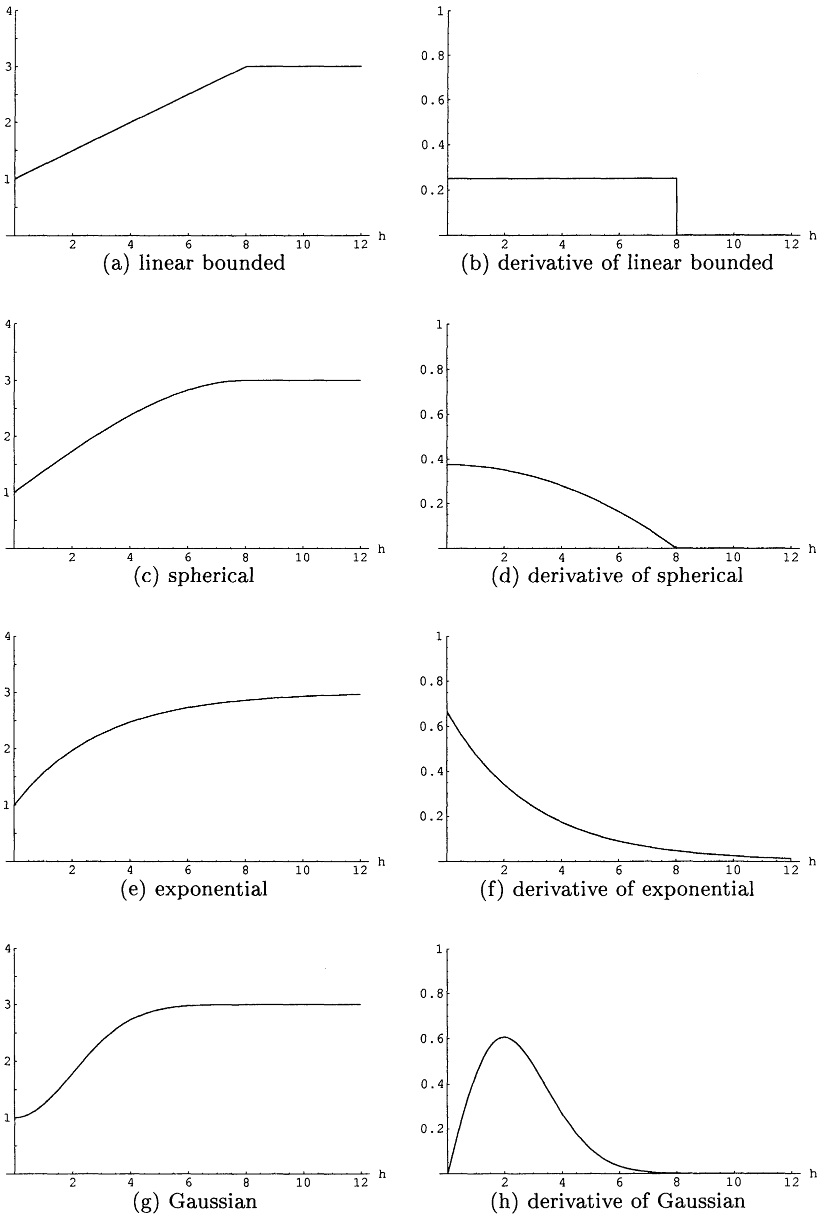
Figure 4. Example of the subjectivity involved in the selection of a variogram model: spherical, exponential, or Gaussian models are fitted equally well, although the true underlying variogram of the data is exponential.

of a variogram model, Gorsich and Genton (1999) have suggested to estimate the derivative of the variogram in a nonparametric way as a help for the selection. Effectively, although many variogram models appear very similar, their derivatives with respect to the lags are not, as is shown in Figure 5.

In order to estimate the derivative without assuming a prior model, a nonparametric variogram estimator which guarantees its conditional negative definiteness is needed (Shapiro and Botha, 1991, Cherry et al., 1996, Cherry, 1997). It is based on the spectral representation of positive definite functions (Bochner, 1955), and consists in finding by nonnegative least squares, positive jumps $p_1, \dots, p_n$ corresponding to the nodes $t_1, \dots, t_n$ in

$$2\hat{\gamma}(h) = 2\sum_{j=1}^{n} p_j\big(1 - \Omega_r(ht_j)\big), \tag{7}$$

where $\Omega_r(x) = (2/x)^{(r-2)/2}\Gamma(r/2)J_{(r-2)/2}(x)$. Here Γ is the gamma function, and J_v is the Bessel function of the first kind of order v. The parameter r controls the smoothness of the fit, and $r \geq d$ is required in order to maintain positive definiteness, where d is the dimension of the spatial domain D. Note that the nugget effect, range, and sill are not well defined in the nonparametric fit. The derivative of the variogram can now be estimated by differentiating the estimator (7) or by using finite differences. It turns out that the first approach does not work well because the basis Ω_r is causing aliasing problems. The second approach is more appropri-



(a) linear bounded

(b) derivative of linear bounded

(c) spherical

(d) derivative of spherical

(e) exponential

(f) derivative of exponential

(g) Gaussian

(h) derivative of Gaussian

Figure 5. Some typical variogram models with their corresponding derivatives.

ate and can be used to select a variogram model based on its derivative. More details about this method, the aliasing problem, the choice of the smoothness parameter r, and some simulations, can be found in Gorsich and Genton (1999). A graphical user interface (GUI) in MATLAB for the nonparametric estimation of the variogram's derivative is available to users at `http://www-math.mit.edu/~gorsich/`.

4 Variogram Fitting by Generalized Least Squares

Once a variogram model has been selected, it must be fitted to the variogram estimates. Variogram fitting is another crucial stage of spatial prediction, because it also determines the kriging weights. Variogram estimates at different spatial lags are correlated, because the same observation is used for different lags. As a consequence, variogram fitting by ordinary least squares is not satisfactory. This problem is addressed by Genton (1998b), who suggests the use of a generalized least squares method with an explicit formula for the covariance structure (GLSE). A good approximation of the covariance structure is achieved by taking into account the explicit formula for the correlation of Matheron's classical variogram estimator in the Gaussian independent case. Simulations were carried out with several types of underlying variograms, as well as with outliers in the data. Results showed that the GLSE technique, combined with a robust estimator of the variogram, improves the fit significantly.

Recently, Genton (1999) has extended the explicit formula for the correlation structure to elliptically contoured distributions (Fang et al., 1989, Fang and Zhang, 1990, Fang and Anderson, 1990). This is a general class of distributions whose contours of equal density have the same elliptical shape as the multivariate Gaussian, but which contains long-tailed and short-tailed distributions. Some important elliptically contoured distributions are the Kotz type, Pearson type, multivariate t, multivariate Cauchy, multivariate Bessel, logistic, and scale mixture. For a subclass of elliptically contoured distributions with a particular family of covariance matrices, the correlation structure depends only on the spatial design matrix of the data, i.e. it is exactly the same as for the multivariate Gaussian distribution. This result allows to extend the validity of the GLSE method of variogram fitting.

Consider an omnidirectional variogram estimator $2\widehat{\gamma}(h)$ for a given set of lags $h_1,\ldots,h_k$, where $1 \leq k \leq K$ and K is the maximal possible distance between data. Denote further by $2\widehat{\gamma} = (2\widehat{\gamma}(h_1),\ldots,2\widehat{\gamma}(h_k))^T \in \mathbb{R}^k$ the random vector with covariance matrix $\mathrm{Var}(2\widehat{\gamma}) = \Omega$. Suppose that one wants to fit a valid parametric variogram $2\gamma(h,\boldsymbol{\theta})$ to the estimated points $2\widehat{\gamma}$. The method of generalized least squares consists in determining the estimator $\widehat{\boldsymbol{\theta}}$ which minimizes

$$G(\boldsymbol{\theta}) = (2\widehat{\gamma} - 2\gamma(\boldsymbol{\theta}))^T\Omega^{-1}(2\widehat{\gamma} - 2\gamma(\boldsymbol{\theta})), \tag{8}$$

where $2\gamma(\boldsymbol{\theta}) = (2\gamma(h_1, \boldsymbol{\theta}), \ldots, 2\gamma(h_k, \boldsymbol{\theta}))^T \in \mathbb{R}^k$ is the vector of the valid parametric variogram, and $\boldsymbol{\theta} \in \mathbb{R}^p$ is the parameter to be estimated. Note that $2\gamma(h, \boldsymbol{\theta})$ is generally a nonlinear function of the parameter $\boldsymbol{\theta}$. Journel and Huijbregts (1978) suggest to use only lag vectors h_i such that $N_{h_i} > 30$ and $0 < i \leq K/2$. This empirical rule is often met in practice. The GLSE algorithm is the following:

(1) Determine the matrix $\Omega = \Omega(\boldsymbol{\theta})$ with element Ω_{ij} given by

$$\mathrm{Corr}(2\widehat{\gamma}(h_i), 2\widehat{\gamma}(h_j))\gamma(h_i, \boldsymbol{\theta})\gamma(h_j, \boldsymbol{\theta})/\sqrt{N_{h_i} N_{h_j}}.$$

(2) Choose $\boldsymbol{\theta}^{(0)}$ and let $l = 0$.

(3) Compute the matrix $\Omega(\boldsymbol{\theta}^{(l)})$ and determine $\boldsymbol{\theta}^{(l+1)}$ which minimizes

$$\left(2\widehat{\gamma} - 2\gamma(\boldsymbol{\theta})\right)^T \Omega(\boldsymbol{\theta}^{(l)})^{-1}\left(2\widehat{\gamma} - 2\gamma(\boldsymbol{\theta})\right).$$

(4) Repeat (3) until convergence to obtain $\widehat{\boldsymbol{\theta}}$.

In step (1), an element of the matrix Ω is given by

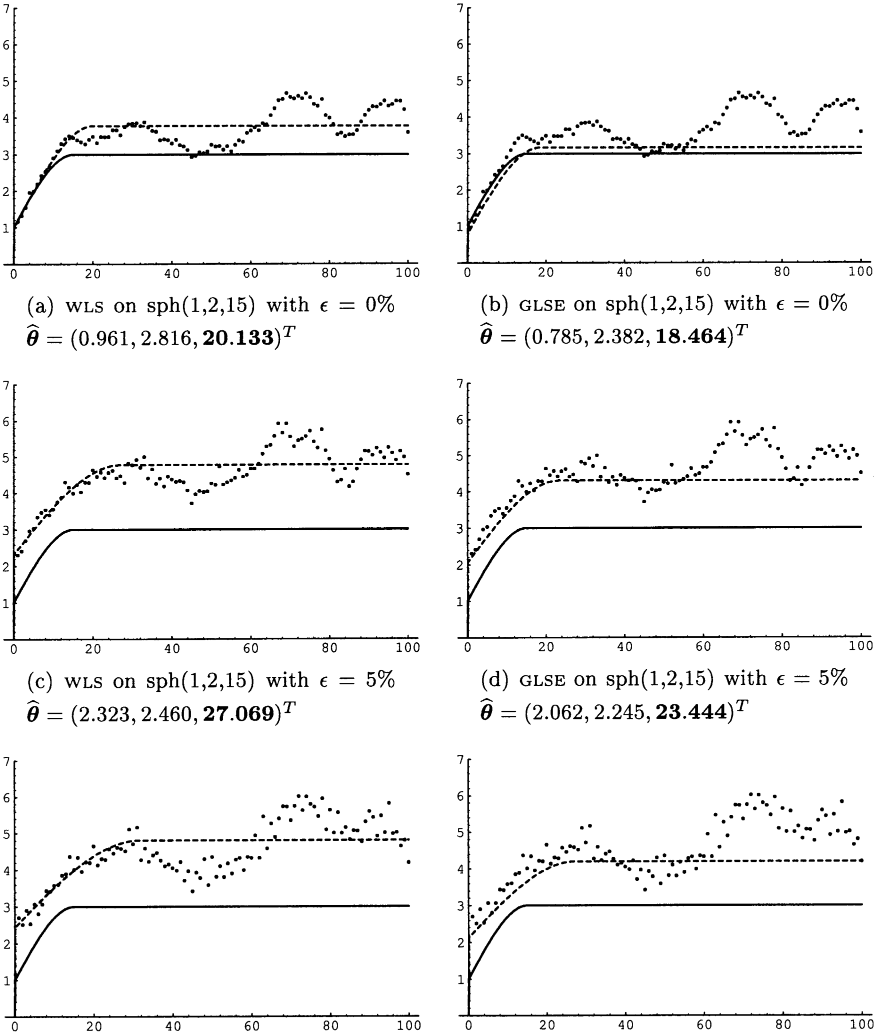
$$\begin{aligned}\Omega_{ij} &= \mathrm{Cov}\left(2\widehat{\gamma}(h_i), 2\widehat{\gamma}(h_j)\right) \\ &= \mathrm{Corr}\left(2\widehat{\gamma}(h_i), 2\widehat{\gamma}(h_j)\right)\sqrt{\mathrm{Var}\left(2\widehat{\gamma}(h_i)\right)\mathrm{Var}\left(2\widehat{\gamma}(h_j)\right)}. \qquad (9)\end{aligned}$$

The correlation $\mathrm{Corr}\left(2\widehat{\gamma}(h_i), 2\widehat{\gamma}(h_j)\right)$ can be approximated by the one in the independent case. An explicit formula can be found in Genton (1998b), which depends only on the lags h_i and h_j, as well as on the sample size n. The variances in equation (9) are replaced by their asymptotic expressions (Genton, 1998b), yielding the formula given in step (1). In step (2), the initial choice $\boldsymbol{\theta}^{(0)}$ can be carried out randomly, or with the result of a fit by ordinary least squares (OLS) or by weighted least squares (WLS). An SPLUS function for variogram fitting by generalized least squares (GLSE), denoted `glse.fitting`, is available on the Web at `http://www-math.mit.edu/~genton/`.

We present the estimation and fitting of a variogram for a simulated data set, having a spherical underlying variogram given by:

$$\gamma(h, a, b, c) = \begin{cases} 0 & \text{if } h = 0, \\ a + b(\frac{3}{2}(\frac{h}{c}) - \frac{1}{2}(\frac{h}{c})^3) & \text{if } 0 < h \leq c, \\ a + b & \text{if } h > c, \end{cases}$$

with parameter $\boldsymbol{\theta} = (a, b, c)^T = (1, 2, 15)^T$. Figures 6 and 7 present the effects of outliers on estimation with Matheron's classical variogram estimator or the highly robust one, and fitting by weighted least squares



(a) WLS on sph(1,2,15) with $\epsilon = 0\%$
$\widehat{\boldsymbol{\theta}} = (0.961, 2.816, \mathbf{20.133})^T$

(b) GLSE on sph(1,2,15) with $\epsilon = 0\%$
$\widehat{\boldsymbol{\theta}} = (0.785, 2.382, \mathbf{18.464})^T$

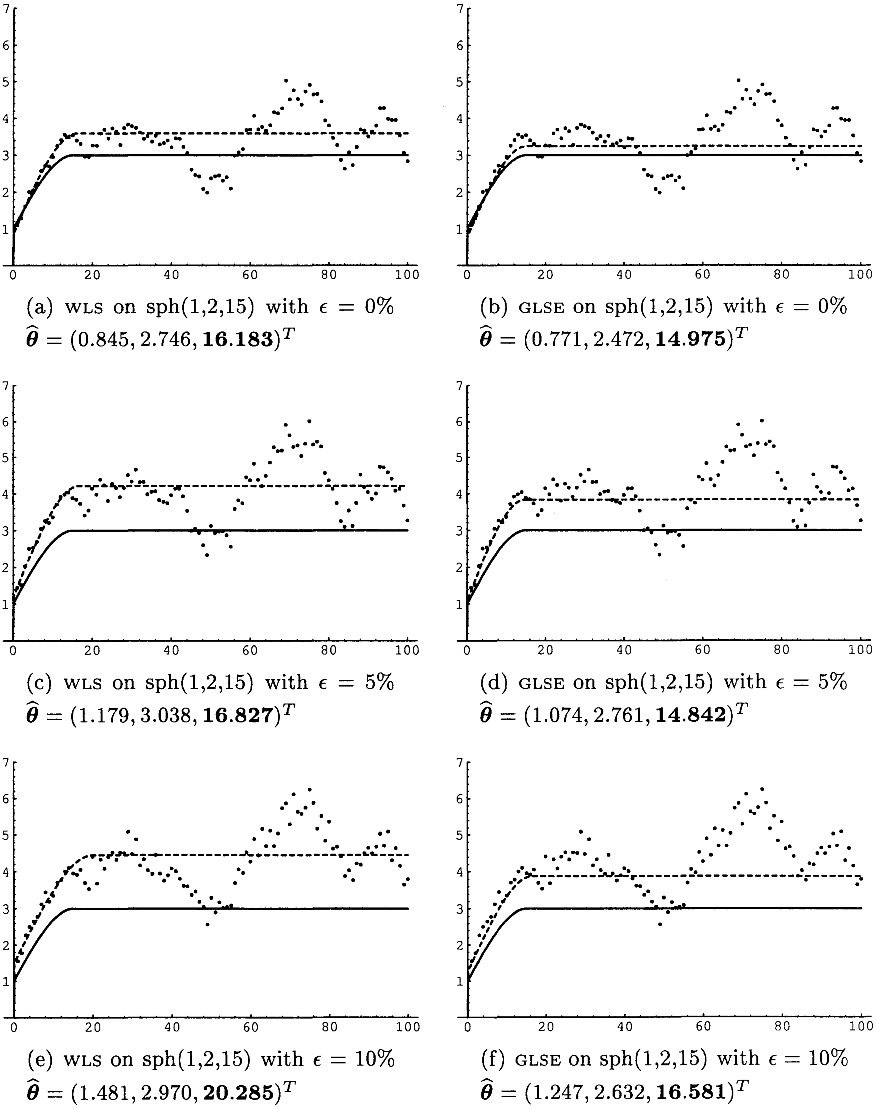
(c) WLS on sph(1,2,15) with $\epsilon = 5\%$
$\widehat{\boldsymbol{\theta}} = (2.323, 2.460, \mathbf{27.069})^T$

(d) GLSE on sph(1,2,15) with $\epsilon = 5\%$
$\widehat{\boldsymbol{\theta}} = (2.062, 2.245, \mathbf{23.444})^T$

(e) WLS on sph(1,2,15) with $\epsilon = 10\%$
$\widehat{\boldsymbol{\theta}} = (2.436, 2.375, \mathbf{31.755})^T$

(f) GLSE on sph(1,2,15) with $\epsilon = 10\%$
$\widehat{\boldsymbol{\theta}} = (2.124, 2.074, \mathbf{27.261})^T$

Figure 6. An example of estimation with Matheron's classical variogram estimator of the perturbed spherical variogram and fit with WLS or GLSE (the underlying variogram is the solid line and the fitted variogram is the dashed line). Reprinted from *Mathematical Geology 30*(4), 323–345, with permission of the International Association for Mathematical Geology.



(a) WLS on sph(1,2,15) with $\epsilon = 0\%$
$\widehat{\boldsymbol{\theta}} = (0.845, 2.746, \mathbf{16.183})^T$

(b) GLSE on sph(1,2,15) with $\epsilon = 0\%$
$\widehat{\boldsymbol{\theta}} = (0.771, 2.472, \mathbf{14.975})^T$

(c) WLS on sph(1,2,15) with $\epsilon = 5\%$
$\widehat{\boldsymbol{\theta}} = (1.179, 3.038, \mathbf{16.827})^T$

(d) GLSE on sph(1,2,15) with $\epsilon = 5\%$
$\widehat{\boldsymbol{\theta}} = (1.074, 2.761, \mathbf{14.842})^T$

(e) WLS on sph(1,2,15) with $\epsilon = 10\%$
$\widehat{\boldsymbol{\theta}} = (1.481, 2.970, \mathbf{20.285})^T$

(f) GLSE on sph(1,2,15) with $\epsilon = 10\%$
$\widehat{\boldsymbol{\theta}} = (1.247, 2.632, \mathbf{16.581})^T$

Figure 7. An example of estimation with the highly robust variogram estimator of a perturbed spherical variogram and fit with WLS or GLSE (the underlying variogram is the solid line and the fitted variogram is the dashed line). Reprinted from *Mathematical Geology 30*(4), 323–345, with permission of the International Association for Mathematical Geology.

(Cressie, 1985) WLS or generalized least squares GLSE. The data is perturbed by $\epsilon = 5\%$ and $\epsilon = 10\%$ of observations from a Gaussian $N(0, 25)$ distribution, with mean zero and variance 25. On each graph, the underlying variogram is represented by a solid line and the fitted variogram by a dashed line. The effect of perturbations is noticeable by a greater vertical variability of the variogram estimations. For Matheron's estimator, a horizontal deformation is added, which leads to an increase of the range, expressed through the parameter c. This phenomena occurs to a much lesser extent for the highly robust variogram estimator. When fitting, the method GLSE tends to reduce this effect. Therefore, the combination of the highly robust variogram estimator and of GLSE fitting gives the best estimation of the parameter c, which is the most important one for kriging.

Some examples of application of the robustness methodology discussed in this paper can be found in the literature. Eyer and Genton (1999) present an application of highly robust variogram estimation in astronomy, where outlying values can sometimes be present in data sets. The variogram is used to determine a pseudo-period in the pulsation of variable stars. Simulations show that one single outlying value can completely mask the determination of the pseudo-period when using Matheron's classical variogram estimator, whereas the highly robust estimator remains unaffected. Furrer and Genton (1999) analyze a data set of sediments from Lake Geneva, in Switzerland, using the module S+SPATIALSTATS of the software SPLUS. They apply the methodology of highly robust variogram estimation, as well as variogram fitting by generalized least squares (GLSE). A similar type of analysis can be found in Genton and Furrer (1998b), in the context of rainfall measurements in Switzerland.

5 Conclusion

In this paper, we addressed some robustness problems occurring in the analysis of spatial data. First, the use of a highly robust variogram estimator has been suggested, and we studied some of its properties, such as spatial breakdown point, effect of linear trend in the data, as well as location outliers. Second, we described the selection of variogram models by means of nonparametric derivative estimation. Finally, the fit of the variogram model by generalized least squares has been discussed. Two new SPLUS functions for highly robust variogram estimation and variogram fitting by generalized least squares are made available on the Web at `http://www-math.mit.edu/~genton/`, as well as a MATLAB code for variogram model selection via nonparametric derivative estimation. Note that the kriging step has not been robustifyied yet, and further research is needed in this direction.

6 References

Bochner, S. (1955). *Harmonic Analysis and the Theory of Probability.* Berkeley: University of California Press.

Cherry, S. (1997). Non-parametric estimation of the sill in geostatistics. *Environmetrics 8*, 13–27.

Cherry, S., J. Banfield, and W.F. Quimby (1996). An evaluation of a non-parametric method of estimating semi-variogram of isotropic spatial processes. *Journal of Applied Statistics 23*, 435–449.

Cressie, N.A.C. (1985). Fitting variogram models by weighted least squares. *Mathematical Geology 17*, 563–586.

Cressie, N.A.C. (1993). *Statistics for Spatial Data* (revised ed.). Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. New York: Wiley.

Cressie, N.A.C. and D. M. Hawkins (1980). Robust estimation of the variogram. I. *Mathematical Geology 12*, 115–125.

Croux, C. and P. J. Rousseeuw (1992). Time-efficient algorithms for two highly robust estimators of scale. In Y. Dodge and J. Whittaker (Eds.), *Computational Statistics*, Volume 1, pp. 411–428. Heidelberg: Physica-Verlag.

Eyer, L. and M.G. Genton (1999). Characterization of variable stars by robust wave variograms: an application to Hipparcos mission. *Astronomy and Astrophysics, Supplement Series 136*, 421–428.

Fang, K.T. and T.W. Anderson (Eds.) (1990). *Statistical Inference in Elliptically Contoured and Related Distributions.* New York: Allerton Press.

Fang, K.T., S. Kotz, and K.W. Ng (1989). *Symmetric Multivariate and Related Distributions*, Volume 36 of *Monographs on Statistics and Applied Probability.* London: Chapman and Hall.

Fang, K.T. and Y.T. Zhang (1990). *Generalized Multivariate Analysis.* Berlin: Springer.

Furrer, R. and M.G. Genton (1999). Robust spatial data analysis of Lake Geneva sediments with S+SpatialStats. *Systems Research and Information Systems 8*, 257–272. special issue on Spatial Data Analysis and Modeling.

Genton, M.G. (1998a). Highly robust variogram estimation. *Mathematical Geology 30*, 213–221.

Genton, M.G. (1998b). Variogram fitting by generalized least squares using an explicit formula for the covariance structure. *Mathematical Geology 30*, 323–345.

Genton, M.G. (1998c). Spatial breakdown point of variogram estimators. *Mathematical Geology 30*, 853–971.

Genton, M.G. (1999). The correlation structure of Matheron's classical variogram estimator under elliptically contoured distributions. *Mathematical Geology 32*, 127–137.

Genton, M.G. and R. Furrer (1998a). Analysis of rainfall data by simple good sense: is spatial statistics worth the trouble? *Journal of Geographic Information and Decision Analysis 2*, 11–17.

Genton, M.G. and R. Furrer (1998b). Analysis of rainfall data by robust spatial statistics using S+SpatialStats. *Journal of Geographic Information and Decision Analysis 2*, 126–136.

Gorsich, D.J. and M.G. Genton (1999). Variogram model selection via nonparametric derivative estimations. *Mathematical Geology 32*, 249–270.

Hampel, F.R. (1973). Robust estimation, a condensed partial survey. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 27*, 87–104.

Hampel, F.R., E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel (1986). *Robust Statistics, the Approach Based on Influence Functions.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: Wiley.

Huber, P.J. (1977). *Robust Statistical Procedures*, Volume 27 of *CBMS-NSF Regional Conference Series in Applied Mathematics.* Philadelphia: Society for Industrial and Applied Mathematics.

Journel, A.G. and Ch.J. Huijbregts (1978). *Mining Geostatistic.* London: Academic Press.

Matheron, G. (1962). *Traité de géostatistique appliquée. I*, Volume 14 of *Mémoires du Bureau de Recherches Géologiques et Minières.* Paris: Éditions Technip.

Rousseeuw, P.J. and C. Croux (1992). Explicit scale estimators with high breakdown point. In Y. Dodge (Ed.), *L_1-Statistical Analyses and Related Methods*, pp. 77–92. Amsterdam: North-Holland.

Rousseeuw, P.J. and C. Croux (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association 88*, 1273–1283.

Shapiro, A. and J.D. Botha (1991). Variogram fitting with a general class of conditionally nonnegative definite functions. *Computational Statistics & Data Analysis 11*, 87–96.