Mathematical Geology, Vol. 30, No. 7, 1998

Spatial Breakdown Point of Variogram Estimators[1]

Marc G. Genton[2]

In the context of robust statistics, the breakdown point of an estimator is an important feature of reliability. It measures the highest fraction of contamination in the data that an estimator can support before being destroyed. In geostatistics, variogram estimators are based on measurements taken at various spatial locations. The classical notion of breakdown point needs to be extended to a spatial one, depending on the construction of most unfavorable configurations of perturbation. Explicit upper and lower bounds are available for the spatial breakdown point in the regular unidimensional case. The difficulties arising in the multidimensional case are presented on an easy example in $\mathbb{R}^2$, as well as some simulations on irregular grids. In order to study the global effects of perturbations on variogram estimators, further simulations are carried out on data located on a regular or irregular bidimensional grid. Results show that if variogram estimation is performed with a 50% classical breakdown point scale estimator, the number of initial data likely to be contaminated before destruction of the estimator is roughly 30% on average. Theoretical results confirm the previous statement on data in $\mathbb{R}^d$, $d \geq 1$.

KEY WORDS: spatial statistics, scale estimation, robustness, classical breakdown point.

INTRODUCTION

Variogram estimation is a crucial stage of spatial prediction, because it determines the kriging weights. It is important to have a variogram estimator which remains close to the true underlying variogram, even if outliers (faulty observations) are present in the data. Otherwise kriging can produce noninformative maps. As a consequence, many robust variogram estimators have been proposed in the literature, in order to remedy the lack of robustness of Matheron's classical variogram estimator (1962). An important robustness feature of such an estimator is its breakdown point, which indicates how many data points need to be replaced by arbitrary values to destroy the estimator. In this paper, this notion is extended to variogram estimators, because it is useful in practice to know, loosely speaking, "how far" the robustness of a variogram estimator extends.

Consider a spatial stochastic process $\{Z(\mathbf{x}) : \mathbf{x} \in D\}$, where D is a fixed

[1]Received 6 February 1997; accepted 15 December 1997.

[2]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307. e-mail: genton@math.mit.edu

0882-8121/98/1000-0853$15.00/1

subset of $\mathbb{R}^d$, $d \geq 1$. Assume that this process is ergodic and satisfies the hypothesis of intrinsic stationarity given by

$$\text{(a) } E(Z(\mathbf{x})) = \mu = \text{constant}, \qquad \forall \mathbf{x} \in D$$

$$\text{(b) } \text{Var}(Z(\mathbf{x} + \mathbf{h}) - Z(\mathbf{x})) = 2\gamma(\mathbf{h}), \qquad \forall \mathbf{x}, \mathbf{x} + \mathbf{h} \in D$$

where $2\gamma(\mathbf{h})$ is the variogram. Let $2\hat{\gamma}(\mathbf{h})$ be a variogram estimator based on a sample $\mathcal{Z} = \{Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n)\}$ of the spatial process.

The most natural approach to variogram estimation is via scale estimation (Cressie and Hawkins, 1980; Cressie, 1993; Genton, 1998). The stochastic process of differences at lag $\mathbf{h}$, $V(\mathbf{h}) = Z(\mathbf{x} + \mathbf{h}) - Z(\mathbf{x})$, has zero expectation and a variance of $2\gamma(\mathbf{h})$. Let $\mathcal{V}_\mathbf{h} = \{V_1(\mathbf{h}), \ldots, V_{N_\mathbf{h}}(\mathbf{h})\}$ be the sample of $V(\mathbf{h})$ corresponding to the sample $\mathcal{Z} = \{Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n)\}$ of Z, where $N_\mathbf{h}$ is the cardinality of $N(\mathbf{h}) = \{(\mathbf{x}_i, \mathbf{x}_j) : \mathbf{x}_i - \mathbf{x}_j = \mathbf{h}\}$. If $S_{N_\mathbf{h}}(\mathcal{V}_\mathbf{h})$ is a scale estimator of the process $V(\mathbf{h})$, a natural variogram estimator is given by

$$2\hat{\gamma}(\mathbf{h}) = (S_{N_\mathbf{h}}(\mathcal{V}_\mathbf{h}))^2, \qquad \mathbf{h} \in \mathbb{R}^d \tag{1}$$

The classical notion of breakdown point of a scale estimator is given in the following definition.

Definition 1. The sample breakdown point of a scale estimator $S_{N_\mathbf{h}}(\mathcal{V}_\mathbf{h})$ is defined by

$$\varepsilon^*_{N_\mathbf{h}}(S_{N_\mathbf{h}}, \mathcal{V}_\mathbf{h}) = \max\left\{\frac{m}{N_\mathbf{h}} \;\middle|\; \sup_{\overline{\mathcal{V}}_\mathbf{h}} S_{N_\mathbf{h}}(\overline{\mathcal{V}}_\mathbf{h}) < \infty \text{ and } \inf_{\overline{\mathcal{V}}_\mathbf{h}} S_{N_\mathbf{h}}(\overline{\mathcal{V}}_\mathbf{h}) > 0\right\}$$

where $\mathcal{V}_\mathbf{h}$ is a sample of size $N_\mathbf{h}$ and $\overline{\mathcal{V}}_\mathbf{h}$ is obtained by replacing any m observations of $\mathcal{V}_\mathbf{h}$ by arbitrary values.

Roughly speaking, the classical breakdown point gives the maximum fraction of bad outliers (in our case $m/N_\mathbf{h}$) the scale estimator can cope with. This indicates how many data points can be replaced by arbitrary values before the scale estimator explodes (tends to infinity) or implodes (tends to 0). For instance, in the location framework, it is well known (Huber, 1981) that the mean has a classical breakdown point of 0%, whereas the median attains 50%, the highest possible value. Further discussions of this concept can be found in Hampel (1971, 1974, 1976), Huber (1981, 1984), and Donoho and Huber (1983). The sample breakdown point $\varepsilon^*_{N_\mathbf{h}}$ of most scale estimators is known, or can be computed. Note that this is at the level of the process of differences $V(\mathbf{h})$, on which the scale estimator is applied. However, in geostatistics, one is much more interested in the breakdown point related to the initial data $\mathcal{Z}$, which are spatially located. Therefore, the previous definition loses its meaning because the location of the perturbed data becomes important. In fact, the effect of the perturbation of a point located on the boundary of the spatial domain D, or inside of it, can

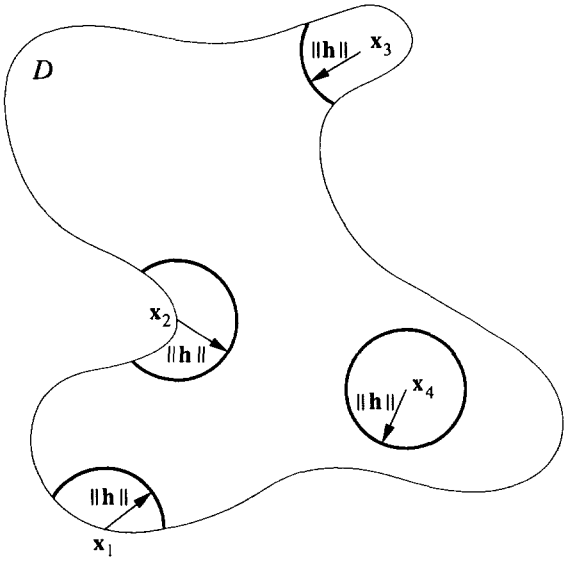

Figure 1. The effects of perturbations of points located on the boundary of the spatial domain D or inside of it. For example, a perturbation located at $\mathbf{x}_4$ has much more impact on the variogram estimator than one located at $\mathbf{x}_1$.

be quite different and depends notably on the lag vector $\mathbf{h}$. Figure 1 depicts this behavior for an isotropic variogram and a fixed lag vector $\mathbf{h}$.

A perturbation located at $\mathbf{x}_i$, $i = 1, \ldots, 4$, generates a perturbation of all differences with the points on the circumference or arc of circumference of the circles centered at $\mathbf{x}_i$, $i = 1, \ldots, 4$, with radius $\|\mathbf{h}\|$. This means, for example, that a perturbation located at $\mathbf{x}_4$ has much more impact on the variogram estimator than one located at $\mathbf{x}_1$. For that reason, the concept of a spatial breakdown point, which is more suitable for variogram estimators, is introduced in the next section. Its link with the classical breakdown point is studied by mean of theoretical results and by simulations. Some examples on typical variogram estimators are presented. Because the spatial breakdown point is only a local notion, for a fixed lag vector $\mathbf{h}$, the last section is devoted to the study of more global effects of perturbations.

SPATIAL BREAKDOWN POINT

Denote by I_m a subset of size m of $\{1, \ldots, n\}$. Recalling the link between the sample $\nabla_{\mathbf{h}} = \{V_1(\mathbf{h}), \ldots, V_{N_{\mathbf{h}}}(\mathbf{h})\}$ of the process of differences $V(\mathbf{h})$ and

the sample $\mathcal{Z} = \{Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n)\}$ of the initial process Z, it is now possible to write the following definition.

Definition 2. The spatial sample breakdown point of a variogram estimator $2\hat{\gamma}(\mathbf{h}) = (S_{N_\mathbf{h}})^2$ is defined by

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

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.

Note that in opposition to Definition 1, the configuration (i.e., the spatial location) of the perturbation is now taken into account, by adding the supremum and infimum on I_m. This definition is justified by the fact that a variogram estimator can be destroyed by a single configuration of perturbation, indexed in I_m. Therefore, it is quite possible to find other configurations, with more than $\varepsilon_n^{S_p}(S_n, \mathcal{Z})\%$ of perturbations, which do not demolish the estimator. Notice furthermore that this definition is local, in the sense that it is valid for a fixed $\mathbf{h}$.

The study of the spatial breakdown point of variogram estimators is now analyzed for data in $\mathbb{R}^1$, on a regular spatial support. In this case, it is possible to compute the maximal number of perturbed differences by using most unfavorable configurations of perturbation. Moreover, lower and upper bounds for the spatial breakdown point are available. Next, the situation of data on regular grids in $\mathbb{R}^d$, $d \geq 2$, is investigated. In this case, it is no longer possible to compute the maximal number of perturbed differences. This question turns into a complex problem of number theory, as shown by a simple example in $\mathbb{R}^2$. The case of irregular grids in $\mathbb{R}^d$, $d \geq 1$, is also investigated by simulations.

Regular Configurations in $\mathbb{R}^1$

Consider the simple situation when the spatial stochastic process is located on a regular unidimensional support and note $\mathcal{Z} = \{Z_1, \ldots, Z_n\} = \{Z(x_1), \ldots, Z(x_n)\}$ a realization of it. In this case, the variogram is automatically isotropic. Fix a positive lag distance $h \in \mathbb{R}$ and note $\mathcal{V}_h = \{V_1(h), \ldots, V_{n-h}(h)\} = \{Z_i - Z_j : i < j; j - i = h\}$. For $m = 1$ perturbed data point, it follows that, if $h < n/2$, one perturbation located at x_i, with $h < i \leq n - h$, generates the perturbation of two differences, whereas for $0 < i \leq h$ or $n - h < i \leq n$, a single difference is perturbed. Finally, if $h \geq n/2$, one perturbation located at x_i, with $0 < i \leq n - h$ or $h < i \leq n$, affects one difference, and none in the other cases. Therefore, to one perturbed observation corresponds at

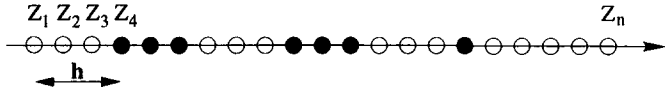

Figure 2. The most unfavorable configuration of perturbation in $\mathbb{R}^1$, for the case $h = 3$, $m = 7$, and $n = 21$.

most two perturbed differences. If $m \geq 1$, we are interested in finding the most unfavorable configuration of perturbed data for a fixed h. Such a configuration is shown in Figure 2 for the case $h = 3$, $m = 7$, and $n = 21$. Open circles represent unperturbed observations, whereas filled circles represent perturbed observations. There are m filled circles. Construction of this configuration consists in placing h unperturbed observations, followed by h perturbed observations, followed by h unperturbed observations, and so on until exhaustion of the m filled circles.

This configuration ensures that the most possible differences are perturbed (i.e., each filled circle perturbes two differences). Moreover, perturbations do not overlap for a given lag distance h, which means that no difference between two perturbed observations is ever taken. Let $v_{\max}(h, m, n)$ be the maximal number of perturbed differences for given h, m, and n. This function depends on the relation between m and h. Let p and q be the two nonnegative integers such that $m = ph + q$ and $q < h$. By disjunction of cases, it is then possible to compute the function $v_{\max}(h, m, n)$ explicitly:

$$v_{\max}(h, m, n) =$$

$$\begin{cases} n - h & \text{if } m = n/2 \\ 2m & \text{if } n/2 > m \geq h,\ q = 0,\ n - 2m \geq h \\ n - h & \text{if } n/2 > m \geq h,\ q = 0,\ n - 2m < h \\ 2m & \text{if } n/2 > m \geq h,\ q \geq 1,\ n - 2ph \geq 2h + q \\ n - 2h + q & \text{if } n/2 > m \geq h,\ q > 1,\ 2h + q > n - 2ph \geq 2h \\ 2ph + q & \text{if } n/2 > m \geq h,\ q \geq 1,\ 2h > n - 2ph \geq h + q \\ n - h & \text{if } n/2 > m \geq h,\ q \geq 1,\ h + q > n - 2ph \geq 0 \\ 2m & \text{if } m < h,\ m + 2h \leq n,\ n - 2m < h \\ m + n - 2h & \text{if } m < h,\ m + 2h > n,\ m < n - h,\ h < n/2 \\ m & \text{if } m < h,\ m + 2h > n,\ m < n - h,\ h \geq n/2 \\ n - h & \text{if } m < h,\ m + 2h > n,\ m \geq n - h \end{cases}$$

Notice that the case $m > n/2$ makes no sense because it implies that more than half of the differences are perturbed. No equivariant scale estimator can be that resistant (Huber, 1981). Figure 3 shows the function $v_{\max}(h, m, 100)$. The hollows appearing in $v_{\max}$ are highly related to the relation betwen m and h, and vary also with n.

The following theorem and corollary examine the relation between the classical breakdown point (usually known) and the spatial one. Afterward, some applications on variogram estimators are presented.

Theorem 1. For each $h \in \{1, \ldots, n-1\}$ and for each integer $M = (n-h)\varepsilon^*_{n-h}(S_{n-h}, \mathcal{V}_h) \leq n/2$, the sample breakdown point and the spatial sample breakdown point of a variogram estimator $2\hat{\gamma}(h) = (S_{n-h})^2$ verify the double inequality

$$2\varepsilon_n^{S_P}(2\hat{\gamma}(h), \mathcal{Z}) \leq \varepsilon^*_{n-h}(S_{n-h}, \mathcal{V}_h) \leq \frac{2n}{n-h}\varepsilon_n^{S_P}(2\hat{\gamma}(h), \mathcal{Z})$$

with the first equality if and only if $h = n/2$ or $M = n/2$, and with the second equality if and only if h and M are such that $v_{\max}(h, M, n) = 2M$.

Proof. In order to prove the first inequality, consider the function

$$\delta(h, m, n) = \frac{v_{\max}(h, m, n)}{n-h} - 2\frac{m}{n}$$

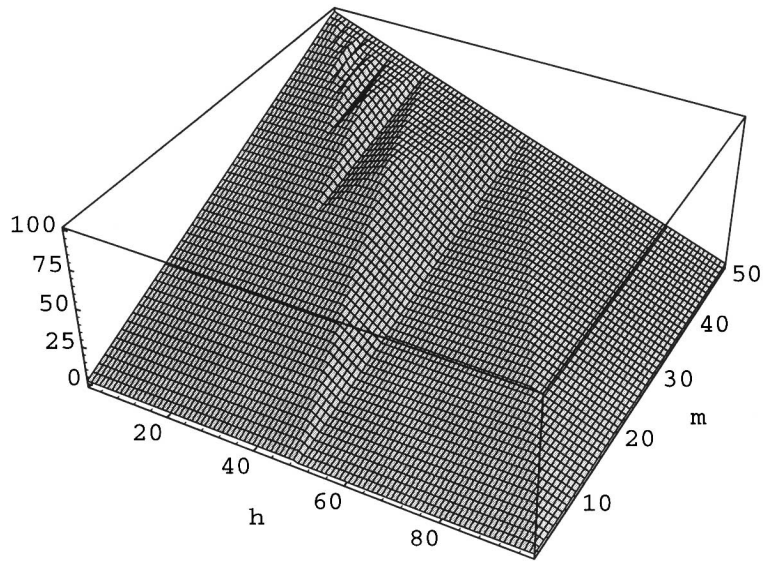

Figure 3. The function $v_{\max}(h, m, 100)$. The hollows appearing in $v_{\max}$ are highly related to the relation between m and h, and vary also with n.

We have to show that the function δ is nonnegative for all possible integers m. If $v_{\max}(h, m, n) = n - h$, then $\delta(h, m, n) = 1 - 2m/n \geq 0$ because $n/2 > m$. If $v_{\max}(h, m, n) = 2m$, then $\delta(h, m, n) = 2m/(n - h) - 2m/n \geq 0$ because $n - h < n$. If $v_{\max}(h, m, n) = n - 2h + q$, then

$$\delta(h, m, n) = \frac{n - 2h + q}{n - h} - \frac{2m}{n} = \frac{n^2 - (p + 2)hn + m(2h - n)}{n(n - h)}$$

$$\geq \frac{n^2 - (p + 2)hn + h(p + 1)(2h - n)}{n(n - h)}$$

because $m < h(p + 1)$ and $2h - n < 0$

$$= \frac{n - 2(p + 1)h}{n} \geq 0 \quad \text{because} \quad n - 2ph \geq 2h$$

If $v_{\max}(h, m, n) = 2ph + q$, then

$$\delta(h, m, n) = \frac{2ph + q}{n - h} - \frac{2m}{n} = \frac{2mh - nq}{n(n - h)}$$

$$\geq \frac{2mh - 2hq(p + 1)}{n(n - h)} \quad \text{because} \quad n - 2h(p + 1) < 0$$

$$= \frac{2hp(h - q)}{n(n - h)} \geq 0 \quad \text{because} \quad h > q$$

If $v_{\max}(h, m, n) = m + n - 2h$, then

$$\delta(h, m, n) = \frac{m + n - 2h}{n - h} - \frac{2m}{n} = \frac{m(2h - n) + n(n - 2h)}{n(n - h)}$$

$$\geq \frac{-h(n - 2h) + n(n - 2h)}{n(n - h)} \quad \text{because} \quad m < h$$

$$= \frac{n - 2h}{n} \geq 0 \quad \textit{because} \quad h < \frac{n}{2}$$

If $v_{\max}(h, m, n) = m$, then

$$\delta(h, m, n) = \frac{m}{n - h} - \frac{2m}{n} = \frac{m(2h - n)}{n(n - h)} \geq 0 \quad \text{because} \quad h < \frac{n}{2}$$

Finally, if $h = n/2$ or $m = n/2$, then $\delta(h, m, n) = 0$ and therefore equality is reached. The second inequality follows from the fact that a perturbation on a single observation generates the perturbation of at most two differences. Thus, the perturbation of m observations generates the perturbation of at most $2m$

differences, and $v_{\max}(h, m, n) \leq 2m$. Consequently, we have the inequality

$$\varepsilon^*_{n-h}(S_{n-h}, \nabla_h) = \frac{v_{\max}(h, M, n)}{n-h} \leq \frac{2M}{n-h} = \frac{2n}{n-h}\frac{M}{n} = \frac{2n}{n-h}\varepsilon_n^{S_p}(2\hat{\gamma}(h), \mathbb{Z})$$

with equality if and only if $v_{\max}(h, M, n) = 2M$. □

By transforming the double inequality for $\varepsilon^*_{n-h}(S_{n-h}, \nabla_h))$ of the previous theorem in a double inequality for $\varepsilon_n^{S_p}(2\hat{\gamma}(h), \mathbb{Z})$, we obtain the following corollary.

Corollary 1.1. For each $h \in \{1, \ldots, n-1\}$ and for each integer $M = (n-h)\varepsilon^*_{n-h}(S_{n-h}, \nabla_h) \leq n/2$, the sample breakdown point and the spatial sample breakdown point of a variogram estimator $2\hat{\gamma}(h) = (S_{n-h})^2$ verify the double inequality

$$\frac{n-h}{2n}\varepsilon^*_{n-h}(S_{n-h}, \nabla_h) \leq \varepsilon_n^{S_p}(2\hat{\gamma}(h), \mathbb{Z}) \leq \frac{1}{2}\varepsilon^*_{n-h}(S_{n-h}, \nabla_h)$$

with the first equality if and only if h and m are such that $v_{\max}(h, m, n) = 2m$, and with the second equality if and only if $h = n/2$ or $m = n/2$.

Let us look into some examples. Matheron's classical variogram estimator (1962), as well as the one of Cressie and Hawkins (1980), are based on scale estimators whose sample breakdown point is 0 (Genton and Rousseeuw, 1995; Genton, 1998). Therefore, by Corollary 1.1, the spatial sample breakdown point of these variogram estimators is also 0, for every lag h. This means that a single outlier in the data can destroy them.

Genton (1998) proposes a highly robust variogram estimator, based on an equivariant scale estimator $S_{n-h} = Q_{n-h}$, whose sample breakdown point is $\varepsilon^*_{n-h} = \lfloor (n-h)/2) - 1 \rfloor / (n-h)$, the highest possible value (Rousseeuw and Croux, 1993), where $\lfloor \cdot \rfloor$ stands for the integer part. Figure 4 shows the spatial sample breakdown point $\varepsilon_{100}^{S_p}(S_{100}, \mathbb{Z})$ of this highly robust variogram estimator, for each lag distance h, represented by the black curve. The upper and lower bounds given in the previous corollary are shown by the light grey curves. As it was stated, the spatial sample breakdown point equals its lower bound as long as $v_{\max}(h, M, n) = 2M$, and equals its upper bound if $h = n/2$. The interpretation of this figure is as follows. For a fixed h, if the percentage of perturbed observations is below the black curve, the estimator 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 4.

Regular and Irregular Configurations in $\mathbb{R}^d$, $d \geq 1$

The previous results have been obtained for data located on a unidimensional and regular support. We would like to extend our results to d-dimensional

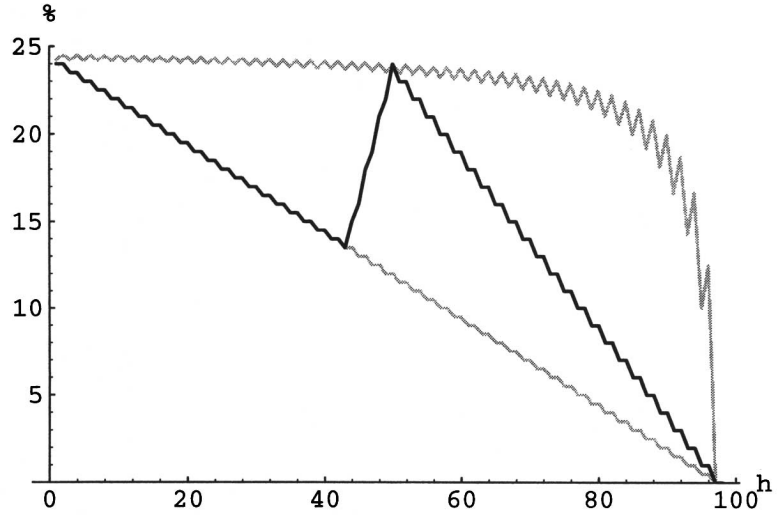

Figure 4. The spatial sample breakdown point (in black) as a function of the lag distance h, for a scale equivariant estimator, with maximal sample breakdown point. The upper and lower bounds are drawn in light grey.

supports, $d \geq 1$, either regular or irregular. In the regular case, we suppose that there are $n = \Pi_{i=1}^{d} n_i$ observations, where n_i is the number of locations along the ith axis. In the irregular case, we assume that only N locations among all the n points of a regular grid are being observed. Notice that a totally irregular configuration, that is to say when points are not located on a grid, may be reduced to an irregular grid by using tolerance neighborhoods, as described in Cressie (1993). Our two notions of regular and irregular grid are therefore quite general.

If the spatial domain $D \subset \mathbb{R}^d$ of the spatial stochastic process is a regular or irregular grid, and if the variogram is isotropic, then it becomes intractable to compute explicitly the function $v_{\max}(h, m, n)$ which counts the number of perturbed differences for given h, m, and n. Nevertheless, we conjecture that the behavior of the spatial sample breakdown point is quite similar to the one in the unidimensional case, as is shown by the following example. Let us consider a regular grid of size $n = 10 \times 10 = 100$. We are interested in the most unfavorable configuration of perturbation for a fixed h. Figure 5 depicts this configuration for each h. Perturbed observations are represented by black points, on which the corresponding number of perturbed differences (between one and four) is indicated. As in the unidimensional case, we construct blocks of observations of size $h \times h$, which are placed as a checkerboard on the grid, beginning with a block of perturbed observations in the left upper corner. In

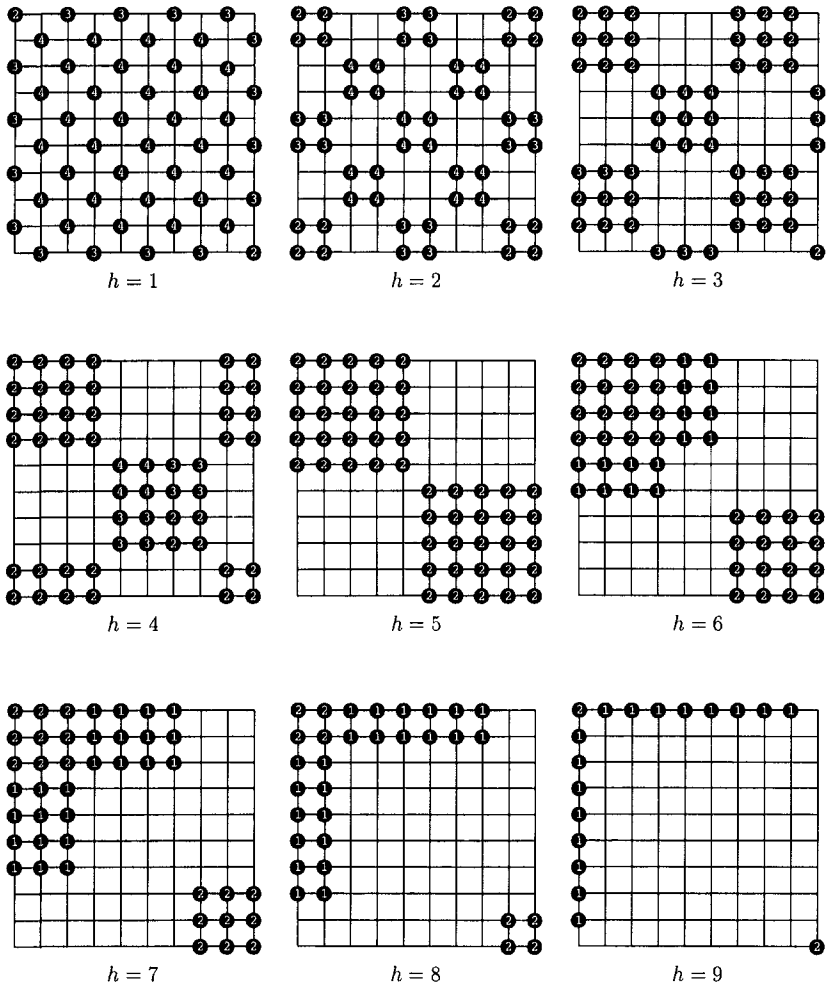

Figure 5. The nine most unfavorable configurations of perturbation for a regular grid of size $n = 10 \times 10 = 100$. Perturbed observations are represented by black points, on which the corresponding number of perturbed differences (between one and four) is indicated.

order to ensure that a perturbation affects four differences, the maximum possible value in $\mathbb{R}^2$, it has to be located at least at a distance h from the border of the grid. If $h > 5$, one has to eliminate the overlapping perturbations. Therefore, this configuration ensures that the most possible differences are perturbed. Moreover, perturbations do not overlap for a given lag distance h, which means that no difference between two perturbed observations is ever taken.

Table 1 shows the behavior of the spatial sample breakdown point for the

Table 1. The Spatial Sample Breakdown Point for a Regular Grid of Size $n = 10 \times 10 = 100$: Its Behavior Is Close to the One of the Unidimensional Case

h	N_h	$\left\lfloor \frac{N_h}{2} - 1 \right\rfloor$	D_4	D_3	D_2	D_1	$\varepsilon_n^{S_p}$
1	180	89	32	16	2	0	22.5
2	160	78	20	16	16	0	20.0
3	140	69	10	16	26	0	20.0
4	120	59	4	8	40	0	22.0
5	100	49	0	0	50	0	25.0
6	80	39	0	0	32	16	20.0
7	60	29	0	0	18	24	15.0
8	40	19	0	0	8	24	12.0
9	20	9	0	0	2	16	8.0

previous grid of size $n = 10 \times 10 = 100$. The first column contains the lag h, to which corresponds a number $N_h = n_1(n_2 - h) + n_2(n_1 - h) = 2n_1(n_1 - h)$ of differences along the axes-directions of the grid. If we choose a scale equivariant estimator for the variogram, with a sample breakdown point of $\varepsilon_{N_h}^* = \lfloor N_h/2 - 1 \rfloor / N_h$ (for example, take the highly robust variogram estimator proposed by Genton (1998)), then the third column shows the maximal number of differences tolerated by the estimator. The columns D_i, $i = 1, \ldots, 4$ contain the number of initial perturbations which can destroy i differences, for each h. Finally, the last column shows the spatial sample breakdown point. We note that its behavior is quite close to the one shown in Figure 4.

GLOBAL EFFECTS OF PERTURBATIONS

The spatial sample breakdown point is a theoretical tool, indicating the worst-case behavior of the variogram estimator for each h. It allows to judge the resistance of the variogram estimates, and consequently their respective reliability. As already mentioned, this concept is of local nature. 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, we start by carrying out some simulations, which are afterward supported by theoretical results.

Simulations

On a regular grid, of size $n = 10 \times 10 = 100$, randomly perturb $m = 25$ observations and compute the percentage p of perturbed differences for each h.

Table 2. Simulation of the Average Percentage $\bar{p}$ of Perturbed Differences for Each h When $m = 25$ Observations of a Regular Grid of Size $n = 10 \times 10 = 100$ Are Perturbed

h	sim 1	sim 2	sim 3	$\bar{p}$	$\hat{\sigma}_{\bar{p}}$
1	45.0	46.1	42.2	44.0	0.1
2	43.8	46.3	41.3	43.9	0.1
3	46.4	44.3	47.9	43.8	0.1
4	45.0	48.3	44.2	43.8	0.1
5	44.0	44.0	45.0	44.0	0.1
6	40.0	40.0	40.0	44.0	0.1
7	40.0	40.0	41.7	44.0	0.1
8	40.0	37.5	42.5	44.5	0.2
9	30.0	20.0	50.0	44.4	0.3

Table 2 contains the percentages p for three replications of this simulation, as well as an average $\bar{p}$ and its standard deviation $\hat{\sigma}_{\bar{p}}$ over 1000 replications.

The results are quite surprising. First, the average of the percentages seems to be stable, independent of h, as shown in the fifth column. Second, the average of the percentages is close to 44% and not to 50%, as one would expect at first sight. If we vary the number m of perturbed observations, we obtain in average, and for each h, a percentage $\bar{p}$ of perturbed differences presented in Table 3.

Note that if m/n is small, $\bar{p}$ equals approximately $2m/n$, whereas it is slightly smaller if m/n is large. This decrease comes from differences taken between two perturbed observations. Similar simulations are carried out on an

Table 3. Simulation 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

Table 4. Simulation of the Average Percentage $\bar{p}$ of Perturbed Differences for Each h, When $m = 25$ of the $N = 100$ Observations of an Irregular Grid Are Perturbed

h	sim 1	sim 2	sim 3	$\bar{p}$	$\hat{\sigma}_{\bar{p}}$
1	45.7	44.2	40.7	43.8	0.1
2	49.4	40.0	38.9	43.9	0.1
3	44.4	32.4	40.0	43.7	0.1
4	44.3	48.3	41.7	44.1	0.2
5	42.9	36.5	43.9	43.8	0.2
6	40.7	37.5	40.4	44.0	0.2
7	39.1	43.1	48.9	43.8	0.2
8	48.7	55.8	44.4	44.0	0.2
9	37.9	51.5	43.6	43.8	0.2
10	44.0	46.7	48.5	44.1	0.3
11	42.1	42.9	46.4	43.8	0.3
12	45.5	60.0	50.0	44.2	0.4
13	33.3	50.0	66.7	44.6	0.6

irregular grid. From a regular grid of size $n = 14 \times 14 = 196$, $N = 100$ observations are selected randomly, among which $m = 25$ are perturbed randomly. The results, presented in Table 4, are similar to those obtained in the case of a regular grid. However, the standard deviation $\hat{\sigma}_{\bar{p}}$ is slightly higher.

Theoretical Results

Some theoretical results are now established, allowing one to understand the results of the previous simulations. The configurations of regular or irregular grids are treated separately. Afterwards, some examples are discussed.

First consider a d-dimensional regular grid of $n = \Pi_{i=1}^{d} n_i$ points. Select randomly m points among the n points of the grid and perturb them. Let M_h be the number of perturbed differences for each h and N_h be the number of distances equal to h among the n points of the grid. Then, M_h is a random variable and the expectation $E(M_h/N_h)$ corresponds to the average $\bar{p}$ given in the previous simulations. Let $H_{j,h}$ be the random variable which counts the number of perturbed differences by a single perturbed point, the jth one, $j = 1, \ldots, m$, for each h and K_h be the random variable which counts the number of perturbed differences between two perturbed points, separated by a lag distance h. The random variable M_h may be decomposed in

$$M_h = \sum_{j=1}^{m} H_{j,h} - K_h$$

This relation expresses the fact that $\Sigma_{j=1}^{m} H_{j,h}$ counts differences between two perturbed points twice. For the next lemma, we need the notation

$$(n_i - h)_+ = \begin{cases} n_i - h & \text{if } h < n_i \\ 0 & \text{otherwise} \end{cases}$$

Lemma 1. Let $D \subset \mathbb{R}^d$, $d \geq 1$, be a d-dimensional regular grid of $n = \Pi_{i=1}^{d} n_i$ points. The expectation of $H_{j,h}$ equals $E(H_{j,h}) = 2 \Sigma_{i=1}^{d} (n_i - h)_+ / n_i$, and the expectation of K_h equals $E(K_h) = m(m-1)N_h/(n(n-1))$.

Proof. First consider the case $d = 1$. Perturb a single point, randomly selected on a segment of length n_1. Then, the random variable $H_{j,h}$ equals the random variable X_1 defined by

$$X_1 = \begin{cases} 1 + B_1 & \text{if } 0 < h \leq \dfrac{n_1}{2} \\ B_2 & \text{if } \dfrac{n_1}{2} < h < n_1 \\ 0 & \text{if } n_1 \leq h \end{cases}$$

where $B_1 \sim \mathcal{B}((n_1 - 2h)/n_1)$ and $B_2 \sim \mathcal{B}(2(n_1 - h)/n_1)$ are Bernoulli random variables, as is shown in Figure 6. If $d \geq 1$, the random variable $H_{j,h}$ behaves like $H_{j,h} = \Sigma_{i=1}^{d} X_i$, with

$$X_i = \begin{cases} 1 + B_{i,1} & \text{if } 0 < h \leq \dfrac{n_i}{2} \\ B_{i,2} & \text{if } \dfrac{n_i}{2} < h < n_i \\ 0 & \text{if } n_i \leq h \end{cases}$$

where $B_{i,1} \sim \beta((n_i - 2h)/n_i)$ and $\mathcal{B}_{i,2} \sim \mathcal{B}(2(n_i - h)/n_i)$. Thus, the expectation of $H_{j,h}$ is

$$E(H_{j,h}) = E\left(\sum_{i=1}^{d} X_i\right) = \sum_{i=1}^{d} E(X_i) = 2 \sum_{i=1}^{d} \frac{(n_i - h)_+}{n_i}$$

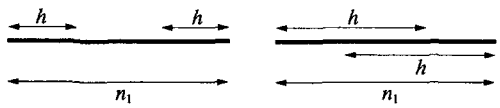

Figure 6. A unidimensional configuration with $0 < h \leq n_1/2$ and $n_1/2 < h < n_1$.

There are a total of $n(n-1)/2$ distances between the n points and $m(m-1)/2$ distances between the m perturbed points. Hence, the random selection of m points among n corresponds to the random selection of $m(m-1)/2$ distances among $n(n-1)/2$ distances, from which N_h are equal to h. It follows that K_h is an hypergeometric random variable, with expectation given in the lemma's statement. □

The expectation of the percentage M_h/N_h can now be computed, and leads to the following result.

Theorem 2. Let $D \subset \mathbb{R}^d$, $d \geq 1$, be a d-dimensional regular grid of $n = \Pi_{i=1}^d n_i$ points. The expectation of M_h/N_h equals

$$E(M_h/N_h) = \left(2 - \frac{m-1}{n-1}\right)\frac{m}{n}$$

Proof. On the regular grid, N_h is defined by $N_h = \Sigma_{i=1}^d [(n_i - h)_+ \Pi_{j \neq i} n_j]$. Therefore, by Lemma 1, we have

$$\begin{aligned} E(M_h/N_h) &= \sum_{j=1}^{m} E(H_{j,h}/N_h) - E(K_h/N_h) \\ &= \frac{2m \Sigma_{i=1}^d \dfrac{(n_i - h)_+}{n_i}}{\Sigma_{i=1}^d [(n_i - h)_+ \Pi_{j \neq i} n_j]} - \frac{m(m-1)}{n(n-1)} \\ &= \frac{2m \Sigma_{i=1}^d [(n_i - h)_+ \Pi_{j \neq i} n_j]}{n \Sigma_{i=1}^d [(n_i - h)_+ \Pi_{j \neq i} n_j]} - \frac{m(m-1)}{n(n-1)} \\ &= \left(2 - \frac{m-1}{n-1}\right)\frac{m}{n} \end{aligned}$$

□

We now study the case of a d-dimensional irregular grid $D \subset \mathbb{R}^d$, $d \geq 1$, of N points. It is obtained by randomly selecting N points from a regular grid of $n = \Pi_{i=1}^d n_i$ points. Next, randomly select m points among the N points of the irregular grid. The random variables M_h, $H_{j,h}$ and K_h are defined as in the regular case. However, N_h is no longer constant, but random. There is a similar lemma to the one of the regular case.

Lemma 2. Let $D \subset \mathbb{R}^d$, $d \geq 1$, be a d-dimensional irregular spatial grid of N points, selected randomly among the $n = \Pi_{i=1}^d n_i$ points of a regular grid. The expectation of $H_{j,h}$ equals $E(H_{j,h}) = 2(\Sigma_{i=1}^d (n_i - h)_+/n_i)(N-1)/(n-1)$, and the expectation of K_h equals $E(K_h) = m(m-1)E(N_h)/(N(N-1))$.

Proof. Demonstration of this lemma is similar to the one of Lemma 1. First consider the case $d = 1$. Perturb a point x, randomly selected on a segment

of length n_1. Then, the random variable $H_{j,h}$ is equal to the random variable X_1 defined by

$$X_1 = \begin{cases} B_{1g} + B_{1d} & \text{if } 0 < h \leq \dfrac{n_1}{2} \\ B_2 & \text{if } \dfrac{n_1}{2} < h < n_1 \\ 0 & \text{if } n_1 \leq h \end{cases}$$

where B_{1g}, $B_{1d} \sim \mathcal{B}((n_1 - h)/n_1 \cdot (N - 1)/(n - 1))$ and $B_2 \sim \mathcal{B}(2(n_1 - h)/n_1 \cdot (N - 1)/(n - 1))$ are Bernoulli random variables, as shown in Figure 6. The probability for the point x to have a neighbor belonging to the N points of the irregular grid and being at distance h on the left of x is $(n_1 - h)/n_1 \cdot (N - 1)/(n - 1)$. This probability is obtained by conditioning. Effectively, $(n_1 - h)/n_1$ is the probability for the point x to have a neighbor at distance h on its left and $(N - 1)/(n - 1)$ is the probability for the point x to belong to the N points of the irregular grid, given that it has a neighbor at distance h on its left. The same is true for a neighbor at distance h on his right. The probability for B_2 is obtained by the same way. If $d \geq 1$, the random variable $H_{j,h}$ behaves like $H_{j,h} = \Sigma_{i=1}^{d} X_i$, with

$$X_i = \begin{cases} B_{i,1g} + B_{i,1d} & \text{if } 0 < h \leq \dfrac{n_i}{2} \\ B_{i,2} & \text{if } \dfrac{n_i}{2} < h < n_i \\ 0 & \text{if } n_i \leq h \end{cases}$$

where $B_{i,1g}$, $B_{i,1d} \sim \mathcal{B}((n_i - h)/n_i \cdot (N - 1)/(n - 1))$ and $B_{i,2} \sim \mathcal{B}(2(n_i - h)/n_i \cdot (N - 1)/(n - 1))$. Thus, the expectation of $H_{j,h}$ equals

$$E(H_{j,h}) = E\left(\sum_{i=1}^{d} X_i\right) = \sum_{i=1}^{d} E(X_i) = 2\left(\sum_{i=1}^{d} \frac{(n_i - h)_+}{n_i}\right)\frac{N - 1}{n - 1}$$

There are $N(N - 1)/2$ distances between the N points and $m(m - 1)/2$ distances between the m perturbed points. Hence, the random selection of m points among N corresponds to the random selection of $m(m - 1)/2$ distances among $N(N - 1)/2$ distances, from which $E(N_h)$ are equal to h exactly. It follows that K_h is an hypergeometric random variable, with expectation given in the lemma's statement. □

An approximation of the expectation of the percentage M_h/N_h can now be computed:

Theorem 3. Let $D \subset \mathbb{R}^d$, $d \geq 1$, be a d-dimensional irregular grid of N points, randomly selected among the $n = \Pi_{i=1}^{d} n_i$ points of a regular grid. Then, the expectation $E(M_h/N_h)$ equals approximately

$$\left(2 - \frac{m-1}{N-1}\right)\frac{m}{N}$$

for N large enough.

Proof. On the regular grid, N_h is defined by $N_h = \Sigma_{i=1}^{d} [(n_i - h)_+ \Pi_{j \neq i} n_j]$. If N points of the regular grid are selected randomly, then N_h is an hypergeometric random variable, as we have to select $N(N-1)/2$ distances among $n(n-1)/2$ from which $\Sigma_{i=1}^{d} [(n_i - h)_+ \Pi_{j \neq i} n_j]$ are at distance h. Thus, with $f = N(N-1)/(n(n-1))$, we have

$$E(N_h) = f \sum_{i=1}^{d} \left[(n_i - h)_+ \prod_{j \neq i} n_j\right], \quad \text{and}$$

$$\mathrm{Var}(N_h) = f \sum_{i=1}^{d} \left[(n_i - h)_+ \prod_{j \neq i} n_j\right] \left(1 - \frac{2}{n(n-1)} \sum_{i=1}^{d} \left[(n_i - h)_+ \prod_{j \neq i} n_j\right]\right) \left(\frac{1-f}{1 - 2/(n(n-1))}\right)$$

Since

$$E(M_h/N_h) = E(M_h/E(N_h) \quad E(N_h)/N_h)$$

Taylor expansion of order 1 shows that $E(N_h)/N_h$ behaves approximately as

$$1 - \frac{N_h - E(N_h)}{E(N_h)}$$

whose expectation is 1 and whose variance is

$$\frac{\mathrm{Var}(N_h)}{E^2(N_h)} \simeq \frac{1}{N(N-1)} \left(\frac{n-1-2d}{d}\right) \left(\frac{1-f}{1 - 2/(n(n-1))}\right)$$

where we used the fact that $\Sigma_{i=1}^{d} [(n_i - h)_+ \Pi_{j \neq i} n_j]$ can be approximated by nd. Therefore, for N large enough, $E(N_h)/N_h$ is a constant, equal to 1, and in this case, by Lemma 2, $E(M_h/N_h)$ can be approached by

$$\begin{aligned} E(M_h)/E(N_h) &= \sum_{j=1}^{m} E(H_{j,h})/E(N_h) - E(K_h)/E(N_h) \\ &= \frac{2m \left(\sum_{i=1}^{d} \frac{(n_i - h)_+}{n_i}\right) \frac{N-1}{n-1}}{\frac{N(N-1)}{n(n-1)} \sum_{i=1}^{d} [(n_i - h)_+ \Pi_{j \neq i} n_j]} - \frac{m(m-1)}{N(N-1)} \\ &= \frac{2m \sum_{i=1}^{d} [(n_i - h)_+ \Pi_{j \neq i} n_j]}{N \sum_{i=1}^{d} [(n_i - h)_+ \Pi_{j \neq i} n_j]} - \frac{m(m-1)}{N(N-1)} \\ &= \left(2 - \frac{m-1}{N-1}\right)\frac{m}{N} \end{aligned}$$

□

Note that, as in the simulations, the expectation of M_h/N_h for a regular grid is identical to the one for an irregular grid, according to the approximations. This expectation does not depend on h. The variance of M_h/N_h for a regular or irregular grid, is not easy to compute, because perturbed locations are drawn without replacement. In fact, it depends on h and the irregularity of the grid. If the number of perturbed points m and the number n or N of points of the grid are not small, this expectation is approximately equal to $(2 - m/n)m/n$. This means that for a variogram estimator with classical breakdown point of 50%, the maximal percentage of data which can be perturbed is solution of the equation

$$\left(2 - \frac{m}{n}\right)\frac{m}{n} = \frac{1}{2}$$

which leads to

$$\frac{m}{n} = 1 - \frac{\sqrt{2}}{2} = 29.3\%$$

If a highly robust variogram estimator is used, for example the one proposed by Genton (1998), it will have a global resistance to roughly 30% of outliers in the initial observations. On the contrary, Matheron's classical variogram estimator (1962), as well as Cressie and Hawkins' one (1980), have no global resistance at all to any outlier in the initial observations.

CONCLUSIONS

In this paper, the concept of breakdown point has been extended to variogram estimators. If a variogram estimator is based on an equivariant scale estimator with classical breakdown point of 50%, the highest possible value, then the maximal number of initial data which can be perturbed before destroying the estimator, is roughly 30% on average. This has been confirmed by simulations and theoretical results. However, there exist particular configurations of perturbation, called most unfavorable configurations, for which the maximal number of initial data which can be perturbed before destroying the estimator, is much lower than 30%. Of course, variogram estimators like Matheron's classical one or Cressie and Hawkins' one, are not resistant to any perturbation in the data.

ACKNOWLEDGMENTS

This work contains parts of my PhD dissertation, which was written under the generous guidance of Prof. Stephan Morgenthaler, at the Swiss Institute of

Technology, Lausanne. I wish to thank the Swiss National Science Foundation for its financial support.

REFERENCES

Cressie, N., 1993, Statistics for spatial data, 2nd Ed.: John Wiley & Sons, New York, 900 p.

Cressie, N., and Hawkins, D. M., 1980, Robust estimation of the variogram, I: Math. Geology, v. 12, no. 2, p. 115–125.

Donoho, D. L., and Huber, P. J., 1983, The notion of breakdown point, *in* Bickel, P. J., Doksum, K. A., and Hodges, J. L. Jr., eds., A Festschrift for Erich L. Lehmann: Wadsworth, Belmont, p. 157–184.

Genton, M. G., and Rousseeuw, P. J., 1995, The change-of-variance function of M-estimators of scale under general contamination: Jour. Comp. Appl. Math., v. 64, p. 69–80.

Genton, M. G., 1998, Highly robust variogram estimation: Math. Geology, v. 30, no. 2, p. 213–221.

Hampel, F. R., 1971, A general qualitative definition of robustness: Ann. Math. Stat., v. 42, no. 6, p. 1887–1896.

Hampel, F. R., 1974, The influence curve and its role in robust estimation: Jour. Am. Stat. Assoc., v. 69, no. 346, p. 383–393.

Hampel, F. R., 1976, On the Breakdown Points of Some Rejection Rules with Mean: Research Report, no. 11, Fachgruppe für Statistik, ETHZ.

Huber, P. J., 1981, Robust statistics: John Wiley & Sons, New York, 308 p.

Huber, P. J., 1984, Finite sample breakdown of M- and P-estimators: Ann. Math. Stat., v. 12, no. 1, p. 119–126.

Matheron, G., 1962, Traité de géostatistique appliquée, Tome I: Mémoires du Bureau de Recherches Géologiques et Minières, no. 14, Editions Technip, Paris, 333 p.

Rousseeuw, P. J., and Croux, C., 1993, Alternatives to the median absolute deviation: Jour. Am. Stat. Assoc., v. 88, no. 424, p. 1273–1283.