Stoch Environ Res Ris Assess (2007) 21:391–403
DOI 10.1007/s00477-006-0072-3

ORIGINAL PAPER

Reducing fluctuations in the sample variogram

Xavier Emery

Published online: 5 September 2006

Abstract In the analysis of regionalized data, irregular sampling patterns are often responsible for large deviations (fluctuations) between the theoretical and sample semi-variograms. This article proposes a new semi-variogram estimator that is unbiased irrespective of the actual multivariate distribution of the data (provided an assumption of stationarity) and has the minimal variance under a given multivariate distribution model. Such an estimator considerably reduces fluctuations in the sample semi-variogram when the data are strongly correlated and clustered in space, and proves to be robust to a misspecification of the multivariate distribution model. The traditional and proposed semi-variogram estimators are compared through an application to a pollution dataset.

Keywords Spatial statistics · Variogram inference · Weighted variogram · Noncentered covariance · Variogram declustering

1 Introduction and scope of the work

Variogram analysis is a key step for modeling the spatial distribution of regionalized data and is a requirement in most applications concerned with spatial interpolation or with uncertainty characterization at unsampled locations, through either kriging or simulation techniques (Chilès and Delfiner 1999). The use of semi-variograms for modeling spatially or temporally correlated data is popular in many disciplines in the physical and engineering sciences, including hydrology, mining and petroleum engineering, meteorology, forestry, agricultural land management, soil and environmental sciences. However, practitioners are often confronted to difficulties in the calculation, interpretation and posterior fitting of the sample semi-variogram, especially when dealing with small datasets or when the sampling pattern is highly irregular in space.

The motivation of this work is to propose an alternative approach to the traditional semi-variogram estimator, in order to reduce fluctuations in the estimator and to ease variogram analysis. Henceforth, the term *fluctuation* refers to the deviation between a parameter calculated from a dataset (in the present case, the sample semi-variogram) and its expected value (the theoretical semi-variogram) (Matheron 1989). This definition makes sense if the attribute under study is regarded as a realization of a random field $\{Z(\mathbf{x}), \mathbf{x} \in \mathrm{R}^d\}$ (in general, $d = 1$, 2 or 3).

In the following, we focus on the case of stationary random fields, for which the finite-dimensional distributions are shift-invariant (Matheron 1971). Under this assumption, the semi-variogram between two variables $(Z(\mathbf{x}), Z(\mathbf{x} + \mathbf{h}))$ representing the values of the attribute at locations $\mathbf{x}$ and $\mathbf{x} + \mathbf{h}$ only depends on the separation vector $\mathbf{h}$:

$$\gamma(\mathbf{h}) = \frac{1}{2} E\Big\{[Z(\mathbf{x}) - Z(\mathbf{x}+\mathbf{h})]^2\Big\}. \quad (1)$$

In practice, this semi-variogram has to be estimated from a finite set of data, corresponding to the values of

X. Emery (✉)
Department of Mining Engineering, University of Chile,
Avenida Tupper 2069, Santiago, Chile
e-mail: xemery@cec.uchile.cl

the attribute monitored at given locations $\{\mathbf{x}_1,..., \mathbf{x}_n\}$. In this respect, a common assumption is that the sampling pattern is not preferential, i.e. the sampling density does not depend on the data values.

The next sections give an overview of the traditional semi-variogram estimator and the current alternatives to it, and propose a new estimator consisting of a weighted average of the squared data and products of paired data values. The weights are determined so as to account for the spatial clustering of the data and the redundancy in their values. The traditional and weighted semi-variograms are then compared through the analysis of a few sampling patterns and finally applied to a case study in environmental science.

2 Tools for variogram analysis

2.1 Traditional sample semi-variogram

For $\mathbf{h} \in \mathrm{R}^d$, let $N(\mathbf{h})$ be the subset of $\{1,..., n\}$ such that $\forall\ i \in N(\mathbf{h})$, $\{\mathbf{x}_i, \mathbf{x}_i + \mathbf{h}\} \subset \{\mathbf{x}_1,..., \mathbf{x}_n\}$ and let $n(\mathbf{h})$ be the cardinal of $N(\mathbf{h})$; in particular, $n(\mathbf{0}) = n$. The traditional semi-variogram estimator is defined by substituting an arithmetic average for the expected value in Eq. 1 (Matheron 1971):

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2n(\mathbf{h})} \sum_{i \in N(\mathbf{h})} [Z(\mathbf{x}_i) - Z(\mathbf{x}_i + \mathbf{h})]^2. \quad (2)$$

The estimator in Eq. 2 is unbiased, i.e. its expected value is equal to the theoretical semi-variogram (Eq. 1), and is currently the most widespread in practice. Since most applications of variogram analysis deal with uneven sampling patterns, lag vectors are usually binned to form lag classes, i.e. tolerances are introduced on the norm and on the orientation of $\mathbf{h}$. The use of tolerances deteriorates the quality of the semi-variogram estimator and may introduce biases. We will momentarily pass over this problem, which will be tackled again when examining the weighted sample semi-variogram in the next section.

Despite its unbiasedness, the traditional sample semi-variogram (Eq. 2) is sensitive to the occurrence of extreme data values, to data sparsity and to irregular sampling patterns, in particular in the presence of clusters of data (Armstrong 1984; Srivastava and Parker 1989; Rivoirard 2001; Kovitz and Christakos 2004). Alternative tools are therefore sometimes preferred in geostatistical studies. The main ones are presented and discussed in the following subsections.

2.2 Robust and resistant semi-variogram estimators

An estimator is said to be 'robust' if it is efficient under a given statistical model and still performs well when the available data do not conform to this model. In contrast, the concept of 'resistance' is model-free and means that the estimator is not affected by a (even large) change in a few data values.

Robust and/or resistant semi-variogram estimators are generally related to one of the following approaches or a combination of them: (1) to replace the average operator in Eq. 2 by a quantile operator (Armstrong and Delfiner 1980; Dowd 1984; Genton 1998); (2) to replace the squared increments by increments of lower order, such as the absolute increments or their square roots (Cressie and Hawkins 1980; Genton 1998); (3) to clip large increment values, or (4) to calculate the sample semi-variogram of a nonlinear transform of the original data, e.g. of their logarithms (Armstrong 1984; Rivoirard 1987). The resulting semi-variogram estimators are less sensitive to the presence of extreme data values than the traditional estimator and are helpful to 'clean up' a sample semi-variogram affected by outliers.

A model of the bivariate distributions of the random field $\{Z(\mathbf{x}), \mathbf{x} \in \mathrm{R}^d\}$ is needed to rescale the aforementioned estimators and ensure that their expected values match the theoretical semi-variogram. In general, bivariate normality is assumed, although other types of distributions may be considered (Emery 2005a). However, should the bivariate distribution model be incorrect, the semi-variogram estimators would be biased (inaccurate).

2.3 Covariance and correlogram

Alternatively, one may be interested in other measures of spatial continuity, such as the covariance function and the correlogram, from which the theoretical semi-variogram can be derived. Several estimators of the covariance and correlogram have been proposed in the geostatistical literature (Journel and Huijbregts 1978; Cressie and Glonek 1984; Isaaks and Srivastava 1988, 1989). All of them make use of an estimate of the prior mean of the random field $\{Z(\mathbf{x}), \mathbf{x} \in \mathrm{R}^d\}$ and, concerning the correlogram, of its prior variance. However, the uncertainty in the true mean and variance is ignored, which introduces a bias (Cressie 1993, p. 71). Corrections to such a bias may be considered if the data are mutually independent, a usually unrealistic assumption when dealing with regionalized data.

2.4 Noncentered covariance

Rivoirard et al. (2000) suggest the use of the noncentered covariance to estimate the semi-variogram, by putting:

$$\tilde{\gamma}(\mathbf{h}) = \frac{1}{n}\sum_{i\in N(\mathbf{0})} Z(\mathbf{x}_i)^2 - \frac{1}{n(\mathbf{h})}\sum_{j\in N(\mathbf{h})} Z(\mathbf{x}_j)\, Z(\mathbf{x}_j+\mathbf{h}). \quad (3)$$

This estimator avoids estimating the prior mean value and is unbiased. It constitutes an alternative to the traditional sample semi-variogram and may sometimes outperform it. For instance, suppose that the data are mutually independent (pure nugget effect) with mean zero. Let us introduce their successive moments:

$$\forall k \in \mathrm{N}, \quad \forall \mathbf{x} \in \mathrm{R}^d, \quad m_k = E[Z(\mathbf{x})^k]. \quad (4)$$

Assume that the two sets $\{Z(\mathbf{x}_j),\ j \in N(\mathbf{h})\}$ and $\{Z(\mathbf{x}_j+\mathbf{h}),\ j \in N(\mathbf{h})\}$ are disjoint (i.e. the tail datum of a pair is never the head datum of another pair), which implies that $n(\mathbf{h})$ is less than or equal to $n/2$. Under this condition, the fluctuation variances of the traditional sample semi-variogram (Eq. 2) and of the semi-variogram estimated via noncentered covariance (Eq. 3) are:

$$\begin{aligned}\mathrm{var}[\hat{\gamma}(\mathbf{h})] &= \frac{(2n(\mathbf{h})+1)m_2^2+m_4}{2n(\mathbf{h})} - \gamma^2(\mathbf{h}),\\ \mathrm{var}[\tilde{\gamma}(\mathbf{h})] &= \frac{(n-1)m_2^2+m_4}{n} + \frac{m_2^2}{n(\mathbf{h})} - \gamma^2(\mathbf{h}).\end{aligned} \quad (5)$$

The difference between both variances simplifies into:

$$\mathrm{var}[\hat{\gamma}(\mathbf{h})] - \mathrm{var}[\tilde{\gamma}(\mathbf{h})] = \frac{[n-2n(\mathbf{h})](m_4-m_2^2)}{2n\,n(\mathbf{h})}, \quad (6)$$

which is nonnegative since $n \geq 2\, n(\mathbf{h})$. This result indicates that, if the correlations between data are low, the traditional sample semi-variogram is likely to present greater fluctuations around the theoretical model than the estimator based on the noncentered covariance, which should therefore be preferred in practice. A comparative study of these two semi-variogram estimators (Eqs. 2, 3) in the context of spatially correlated data will be made further on.

2.5 Weighted sample semi-variogram

The idea of weighting data pairs in semi-variogram calculation is first due to Omre (1984), who introduced an estimator of the form:

$$\gamma_{\mathrm{wt}}(\mathbf{h}) = \frac{1}{2}\sum_{i\in N(\mathbf{h})} \omega_i [Z(\mathbf{x}_i) - Z(\mathbf{x}_i+\mathbf{h})]^2. \quad (7)$$

Several methods for determining the weights $\{\omega_i,\ i \in N(\mathbf{h})\}$ have been proposed by Omre (1984), Rivoirard (2001), Richmond (2002) and Kovitz and Christakos (2004), based on the geometrical configuration of the available data. The weighted semi-variogram (Eq. 7) is unbiased as soon as the weights assigned to the data pairs add to one. Distributional assumptions are required if one wishes to determine the weights minimizing the fluctuation variance, i.e. the variance of the difference between the estimator $\gamma_{\mathrm{wt}}(\mathbf{h})$ and its expected value $\gamma(\mathbf{h})$. For instance, Emery and Ortiz (2005) examined the case of a Gaussian random field $\{Z(\mathbf{x}),\ \mathbf{x} \in \mathrm{R}^d\}$ and defined a system of linear equations to derive the "optimal" weights.

3 On a more general form of weighted semi-variogram

3.1 Principle

Let us expand the traditional semi-variogram estimator in Eq. 2 as follows:

$$\begin{aligned}\hat{\gamma}(\mathbf{h}) = &\frac{1}{n(\mathbf{h})}\sum_{i\in N(\mathbf{h})} \frac{Z(\mathbf{x}_i)^2 + Z(\mathbf{x}_i+\mathbf{h})^2}{2}\\ &- \frac{1}{n(\mathbf{h})}\sum_{j\in N(\mathbf{h})} Z(\mathbf{x}_j)Z(\mathbf{x}_j+\mathbf{h}).\end{aligned} \quad (8)$$

This expression is comparable to the estimator based on the noncentered covariance defined in Eq. 3, except that the first term (corresponding to the noncentered covariance at lag zero) is not estimated from all the data, but only from a subset corresponding to the paired data at lag $\mathbf{h}$. The basic idea of this work is to seek a more general semi-variogram estimator of the form

$$\gamma^*(\mathbf{h}) = \sum_{i\in N(\mathbf{0})} \omega_i Z(\mathbf{x}_i)^2 + \sum_{j\in N(\mathbf{h})} \lambda_j Z(\mathbf{x}_j)Z(\mathbf{x}_j+\mathbf{h}). \quad (9)$$

Henceforth, this estimator will be referred to under the generic name "weighted sample semi-variogram" or "declustered sample semi-variogram". Note that the estimator in Eq. 7 is a particular case of Eq. 9. To make $\gamma^*(\mathbf{h})$ as accurate and precise as possible, the weights $\{\omega_i,\ i \in N(\mathbf{0})\}$ and $\{\lambda_j,\ j \in N(\mathbf{h})\}$ will be determined in order to minimize the mean squared fluctuation:

$$E\left\{[\gamma^*(\mathbf{h}) - \gamma(\mathbf{h})]^2\right\}. \quad (10)$$

3.2 Derivation of the optimal weights

The previous criterion amounts to looking for an unbiased estimator and to minimizing the fluctuation variance. Unbiasedness implies the following constraints on the sum of weights:

$$\begin{aligned} &\sum_{i\in N(\mathbf{0})} \omega_i = 1, \\ &\sum_{j\in N(\mathbf{h})} \lambda_j = -1. \end{aligned} \quad (11)$$

Under this condition, the fluctuation variance is

$$\mathrm{var}\{\gamma^*(\mathbf{h}) - \gamma(\mathbf{h})\} = E\{\gamma^*(\mathbf{h})^2\} - \gamma(\mathbf{h})^2. \quad (12)$$

To express the expected value of the squared estimator, let us introduce the vectors $\Omega = (\omega_1, \ldots, \omega_n)^{\mathrm{T}}$ and $\Lambda = (\lambda_1, \ldots, \lambda_{n(\mathbf{h})})^{\mathrm{T}}$, as well as the matrices $\boldsymbol{M}_1$, $\boldsymbol{M}_2$ and $\boldsymbol{M}_{12}$ defined for any quadruple of indices $(i, i', j, j') \in N(\mathbf{0})^2 \times N(\mathbf{h})^2$ by:

$$\begin{aligned} \boldsymbol{M}_1(i,i') &= E[Z(\mathbf{x}_i)^2 Z(\mathbf{x}_{i'})^2] \\ \boldsymbol{M}_{12}(i,j) &= E[Z(\mathbf{x}_i)^2 Z(\mathbf{x}_j) Z(\mathbf{x}_j + \mathbf{h})] \\ \boldsymbol{M}_2(j,j') &= E[Z(\mathbf{x}_j) Z(\mathbf{x}_j + \mathbf{h}) Z(\mathbf{x}_{j'}) Z(\mathbf{x}_{j'} + \mathbf{h})]. \end{aligned} \quad (13)$$

It comes:

$$E\{\gamma^*(\mathbf{h})^2\} = \Omega^{\mathrm{T}} \boldsymbol{M}_1 \Omega + 2\Omega^{\mathrm{T}} \boldsymbol{M}_{12} \Lambda + \Lambda^{\mathrm{T}} \boldsymbol{M}_2 \Lambda. \quad (14)$$

Minimization subject to the unbiasedness constraints (Eq. 11) leads to the following system of linear equations, in which μ_1 and μ_2 are Lagrange multipliers, and $\mathbf{0}_v$ (resp. $\mathbf{1}_v$) is a column vector with v entries equal to 0 (resp. 1):

$$\begin{aligned} \boldsymbol{M}_1 \Omega + \boldsymbol{M}_{12} \Lambda + \mu_1 &= \mathbf{0}_n \\ \boldsymbol{M}_{12}^{\mathrm{T}} \Omega + \boldsymbol{M}_2 \Lambda + \mu_2 &= \mathbf{0}_{n(\mathbf{h})} \\ \mathbf{1}_n^{\mathrm{T}} \Omega &= 1 \\ \mathbf{1}_{n(\mathbf{h})}^{\mathrm{T}} \Lambda &= -1. \end{aligned} \quad (15)$$

To solve this system of equations, one needs to specify a multivariate distribution model so as to express the matrices of fourth-order moments (Eq. 13). A first example, which is widely used in the geostatistical simulation of continuous attributes, is that of a stationary multivariate Gaussian distribution with mean zero and covariance function $C(\mathbf{h})$. In such a case, for any set of locations $\{\mathbf{x}_\alpha, \mathbf{x}_\beta, \mathbf{x}_\delta, \mathbf{x}_\varepsilon\}$ (not necessarily distinct), one has (Isserlis 1918; Triantafyllopoulos 2003):

$$\begin{aligned} E[Z(\mathbf{x}_\alpha)Z(\mathbf{x}_\beta)Z(\mathbf{x}_\delta)Z(\mathbf{x}_\varepsilon)] = {} & C(\mathbf{x}_\alpha - \mathbf{x}_\beta)C(\mathbf{x}_\delta - \mathbf{x}_\varepsilon) \\ & + C(\mathbf{x}_\alpha - \mathbf{x}_\delta)C(\mathbf{x}_\beta - \mathbf{x}_\varepsilon) \\ & + C(\mathbf{x}_\alpha - \mathbf{x}_\varepsilon)C(\mathbf{x}_\beta - \mathbf{x}_\delta). \end{aligned} \quad (16)$$

A second example of interest is that of a multivariate gamma random field $\{Z(\mathbf{x}), \mathbf{x} \in \mathrm{R}^d\}$ with shape parameter 0.5, obtained by squaring a stationary standard Gaussian random field $\{Y(\mathbf{x}), \mathbf{x} \in \mathrm{R}^d\}$. The fourth-order moments of the former are eighth-order moments of the latter and can be expressed in the following fashion (Isserlis 1918; Triantafyllopoulos 2003):

$$\begin{aligned} &E[Z(\mathbf{x}_\alpha)Z(\mathbf{x}_\beta)Z(\mathbf{x}_\delta)Z(\mathbf{x}_\varepsilon)] \\ &\quad = E[Y^2(\mathbf{x}_\alpha)Y^2(\mathbf{x}_\beta)Y^2(\mathbf{x}_\delta)Y^2(\mathbf{x}_\varepsilon)] \\ &\quad = 2\sum_{(I,J)} \prod_{k=1}^{4} \rho_Y(\mathbf{x}_{i_k} - \mathbf{x}_{j_k}), \end{aligned} \quad (17)$$

in which the sum is calculated over all the possible pairs of subsets $I = \{i_1, i_2, i_3, i_4\}$ and $J = \{j_1, j_2, j_3, j_4\}$ such that $I \cap J = \varnothing$ and $I \cup J = \{\alpha,\alpha,\beta,\beta,\delta,\delta,\varepsilon,\varepsilon\}$. In Eq. 17, $\rho_Y(\mathbf{h})$ stands for the covariance function of the Gaussian random field $\{Y(\mathbf{x}), \mathbf{x} \in \mathrm{R}^d\}$; the covariance of the gamma random field $\{Z(\mathbf{x}), \mathbf{x} \in \mathrm{R}^d\}$ is twice the square of this function: $C(\mathbf{h}) = 2\rho_Y(\mathbf{h})^2$.

3.3 Comments

1. The problem of the ''optimal'' estimation of semi-variogram weights is similar to that of estimating the expected value of a random field by ordinary kriging, using a weighted average of the data values (Matheron 1971, p. 127). Here, instead of a first-order moment, one estimates a second-order moment of the random field from two types of information (the squared data values and the products of paired data values), leading to an ordinary cokriging-type system of linear equations.
2. The proposed semi-variogram estimator (Eq. 9) should not be used in the scope of non-stationary models, in particular intrinsic models (with an infinite variance) or models with drifts. In the former case, the expected value of the semi-variogram estimator (Eq. 9) and of its square (Eq. 14) may not be finite. In the latter case, the expression of the theoretical semi-variogram

(Eq. 1) is no longer valid (the variance of the increment should be used instead of the expected squared increment) and the semi-variogram estimator (Eq. 9) is likely to be biased.

3. From the multivariate Gaussian and gamma examples, it is seen that the determination of the weights (Eqs. 15, 16, 17) requires knowledge of the covariance function $C(\mathbf{h})$, which is precisely the purpose of semi-variogram calculation. To break the deadlock, a prior covariance model must be assumed, for instance the one derived from traditional variogram analysis. To avoid that the result is sensitive to the prior model, a sequential approach could be used, in which the covariance obtained at a current step becomes the prior model of the subsequent step. The rate of convergence of such an approach is still unknown to a great extent, although some experiments made by the author indicate that few iterations (less than 5) may suffice in practice (see last section for a case study).
4. Even if the traditional semi-variogram shows no interpretable structure and one chooses a pure nugget covariance as an initial guess, the proposed approach may lead to a non-uniform weighting of the data pairs. Indeed, data pairs that share one datum (e.g. the tail datum of a pair is also the head datum of another pair) are partially redundant and are likely to be down-weighted with respect to the other pairs.
5. Likewise, one has to choose a type of multivariate distribution in order to express the fourth-order moments (Eq. 13) and to calculate the optimal weights. The choice of such a multivariate distribution may be guided by further considerations, for instance it is reasonable to decide on a multivariate Gaussian distribution if the goal of the geostatistical study is to perform multigaussian kriging or simulation. Otherwise, the type of multivariate distribution is an arbitrary decision of the practitioner and is used as a reference for ''declustering'' the sample semi-variogram: this approach is as valid as using a declustering algorithm based on the geometrical configuration of the data (Rivoirard 2001; Richmond 2002). Note that the declustered semi-variogram may become imprecise if the multivariate distribution model is ill suited to the available data, although it will always remain unbiased because of constraints (Eq. 11).
6. The proposed approach provides an estimate of the semi-variogram and, at the same time, an estimation variance that measures the expected amplitude of the sample semi-variogram fluctuation (Eq. 12). Such a variance accounts for the number of pairs considered in semi-variogram calculation and for the spatial redundancies between these pairs, and can be used for fitting a semi-variogram model by weighted least squares (Chilès and Delfiner 1999, p. 109).
7. The methodology is easily applicable if the number of data and data pairs involved in semi-variogram calculation is relatively small (say, $n + n(\mathbf{h}) < 1{,}000$), for which the system of linear equations (Eq. 15) can be solved by matrix inversion. This situation corresponds to the case of small datasets and is the most critical in practical applications for inferring the semi-variogram. When the number of data increases, the solution to system (Eq. 15) can be approximated by iterative algorithms (Greenbaum 1997), using for instance the weights corresponding to the traditional sample semi-variogram as the initial guess.
8. If tolerances on vector $\mathbf{h}$ are used, the semi-variogram estimator (Eq. 9) becomes:

$$\gamma^*(\mathbf{h}) = \sum_{i \in N(\mathbf{0})} \omega_i Z(\mathbf{x}_i)^2 + \sum_{j \in N(\mathbf{h})} \lambda_j Z(\mathbf{x}_j) Z(\mathbf{x}_j + \mathbf{h}_j), \quad (18)$$

where $\{\mathbf{h}_j,\ j \in N(\mathbf{h})\}$ are vectors falling into the tolerance region attached to $\mathbf{h}$. In this case, the conditions in Eq. 11 do no longer guarantee that the estimator is unbiased, as the weighting may not be uniform over the tolerance region. Table 1

Table 1 Optimal weights $\{\lambda_j, j \in N(\mathbf{h})\}$ assigned to the products of paired data

Data pair		Optimal weight of the data pair without additional constraint (Eq. 19)	Optimal weight of the data pair with additional constraint (Eq. 19)
Tail data abscissa	Head data abscissa		
0	9	– 0.8424	– 0.4419
0	10	– 0.0243	– 0.0698
0	11	– 0.1334	– 0.4883

One-dimensional configuration with four locations at coordinates 0, 9, 10 and 11, a lag distance $\|\mathbf{h}\|$ equal to 10 and a lag tolerance of five. The reference model is a multivariate Gaussian distribution with an exponential covariance function with practical range 30

shows a simple example for which the pairs $(Z(\mathbf{x}_j), Z(\mathbf{x}_j + \mathbf{h}_j))$ with $|\mathbf{h}_j| > |\mathbf{h}|$ are down-weighted with respect to the other pairs. To ensure unbiasedness, an additional constraint must be introduced:

$$\sum_{j \in N(\mathbf{h})} \lambda_j C(\mathbf{h}_j) = -C(\mathbf{h}), \tag{19}$$

and a third Lagrange multiplier has to be included in the system of equations (Eq. 15). Note that the additional constraint (Eq. 19) is useless and would entail a singularity in the system if $C(\mathbf{h}_j)$ is the same for all $j \in N(\mathbf{h})$. This happens if the sampling pattern is regular, if no tolerance on vector $\mathbf{h}$ is used, or if all the data pair separations $\{||\mathbf{h}_j||, j \in N(\mathbf{h})\}$ are greater than the range of the assumed covariance model.

4 Comparison of the traditional, covariance-based and weighted semi-variogram estimators

In this section, the performances of four semi-variogram estimators are compared for two configurations in R^2: a regular sampling over a square domain with size L and a highly clustered sampling over the same domain (Fig. 1). Each configuration contains exactly 100 data. The estimators under study are the traditional sample semi-variogram (Eq. 2), the estimator based on the noncentered covariance (Eq. 3), the optimally weighted sample semi-variogram (Eqs. 9,18), and the ''classical'' weighted sample semi-variogram (Eq. 7) based on univariate declustering weights obtained with the cell method (Isaaks and Srivastava 1989). Concerning the latter, the weight assigned to each data pair has been chosen proportional to the product of declustering weights of the two data, as suggested by Rivoirard (2001) and Kovitz and Christakos (2004). For the regular sampling case, this estimator matches the traditional semi-variogram estimator and therefore will not be examined. Instead of the fluctuation variance, the estimators will be compared on the basis of the fluctuation relative standard deviations (square root of the fluctuation variances divided by the semi-variogram value), which are dimensionless.

4.1 Multivariate Gaussian and gamma distribution models

For each configuration shown in Fig. 1, two models are examined: the standard multivariate Gaussian and gamma distributions (Eqs. 16, 17). Both models are associated with an isotropic exponential covariance function, therefore only omni-directional sample semi-variograms are calculated (i.e. with a 90° tolerance on the azimuth). Sensitivity to the practical range of the covariance function is performed by determining the fluctuation relative standard deviations for three range values: $L/10$, $L/2$ and L.

The results (Figs. 2, 3) call for the following comments.

1. The optimally weighted sample semi-variogram substantially improves the traditional sample semi-variogram when the spatial correlations are

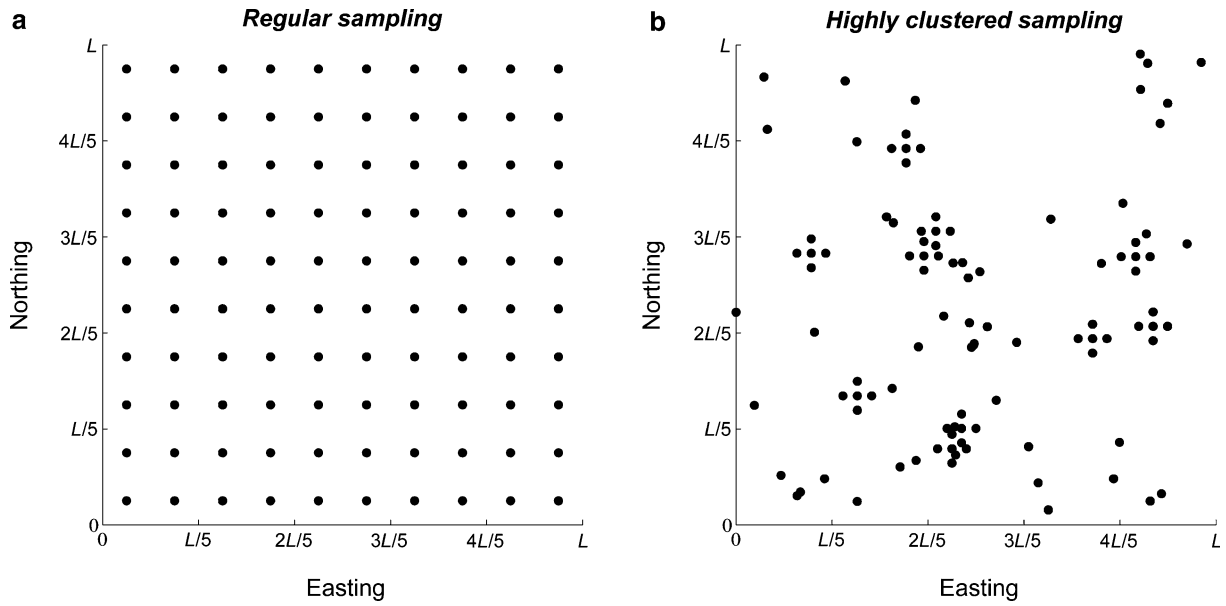

Fig. 1 Two-dimensional sampling patterns: **a** regular sampling, **b** highly clustered sampling

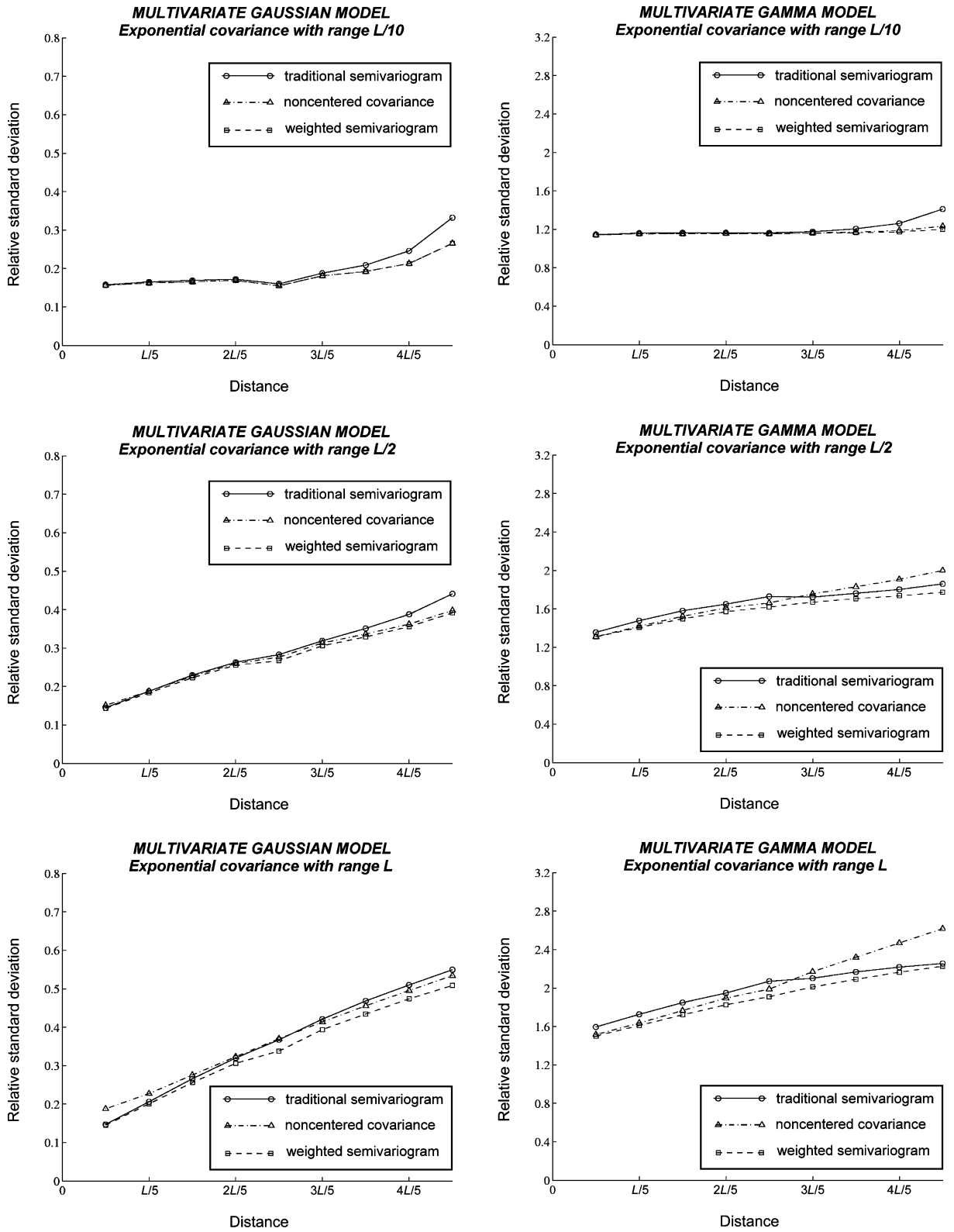

Fig. 2 Fluctuation relative standard deviations for the traditional semi-variogram (*solid lines*), semi-variogram based on noncentered covariance (*dash dots*) and optimally weighted semi-variogram (*dashed lines*), associated with a regular sampling in R^2. Semi-variograms are calculated for lags multiple of the sampling mesh ($L/10$), with no tolerance on the lag distance and a 90° tolerance on the azimuth

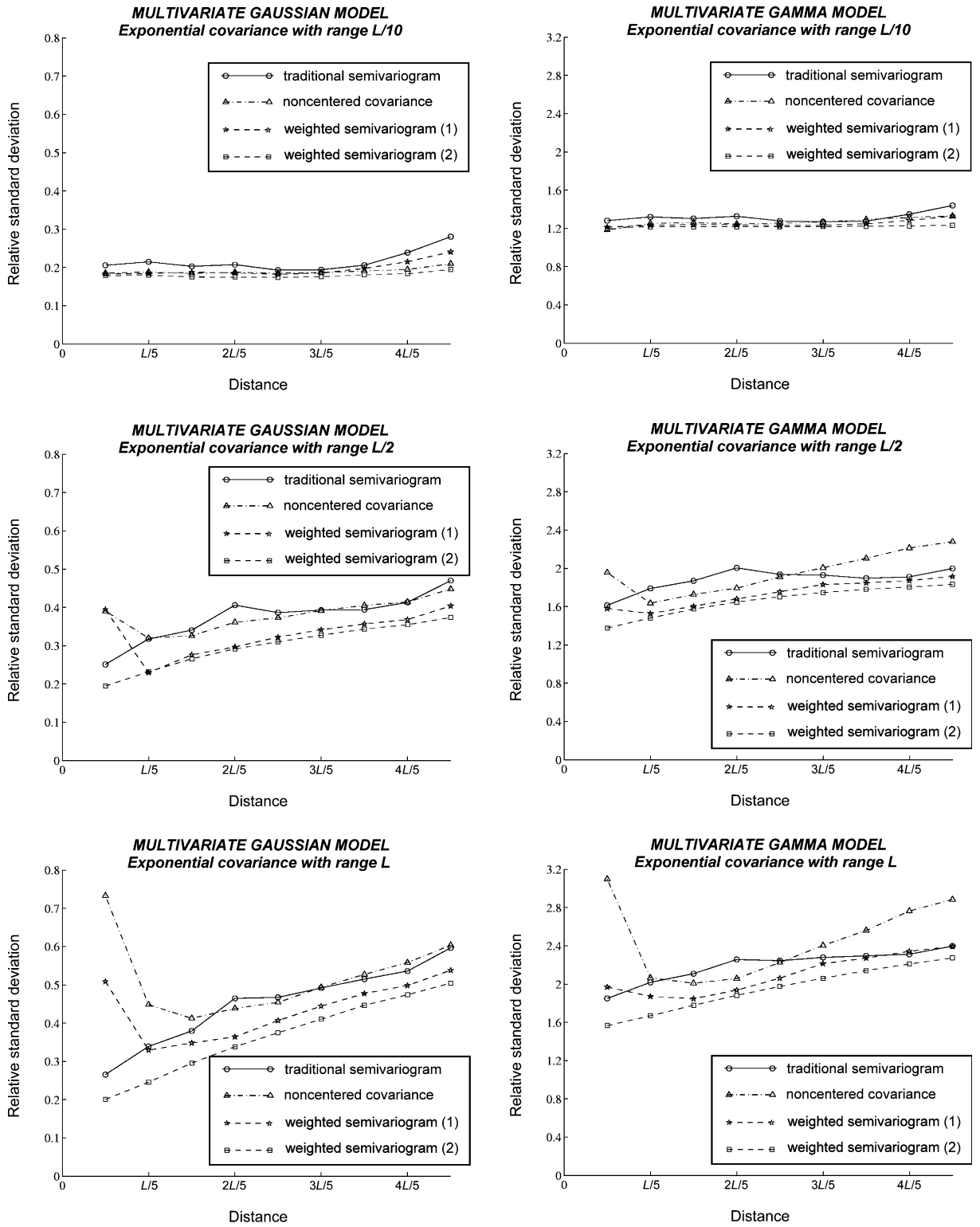

Fig. 3 Fluctuation relative standard deviations for the traditional semi-variogram (*solid lines*), semi-variogram based on noncentered covariance (*dash dots*) and weighted semi-variograms (*dashed lines*), associated with a clustered sampling in R^2. Weighted semi-variogram (*1*) uses univariate declustering weights obtained with the cell method, while weighted semi-variogram (*2*) uses the optimal weights derived from Eq. 15. Semi-variograms are calculated for lags multiple of $L/10$, with a tolerance on the lag distance equal to half the lag and a 90° tolerance on the azimuth

important (large ranges) and when the sampling pattern is highly irregular (clustered sampling). In the other cases (weak correlations between data or even sampling pattern), improvements are not so significant, so that in practice there is no need for declustering the sample semi-variogram.

2. To a lesser extent, the optimally weighted sample semi-variogram (Eq. 9) also outperforms the weighted semi-variogram estimator based on univariate declustering weights (Eq. 7). However, the latter still performs relatively well in the presence of clustered sampling and strong spatial correlations (Fig. 3) and is far much simpler to calculate than the former. It therefore constitutes a shortcut approach to variogram declustering, although its efficiency is not guaranteed: for instance, one observes in Fig. 3 that the relative error for the first lag is sometimes higher than that obtained with the traditional semi-variogram estimator.
3. The comparison between the Gaussian and gamma models proves that fluctuations in the sample semi-variogram are much more important in the latter than in the former model, which reflects that semi-variogram inference is more arduous when dealing with skewed and long-tailed distributions.
4. As mentioned earlier, the semi-variogram estimator based on the noncentered covariance (Eq. 3) may outperform the traditional sample semi-variogram (Eq. 2), in particular when the correlations between data are small (short ranges). However, it turns out to be less precise at small lag distances when the data are significantly correlated and highly clustered in space.

4.2 Robustness to model misspecification

The weighting of the data and data pairs in the proposed semi-variogram estimator (Eqs. 9,18) depends on the choice of a prior multivariate distribution model. In this subsection, a simple exercise is carried out to analyze the implications of a model misspecification in the precision of the weighted sample semi-variogram derived from Eq. 15.

Figure 4 displays the relative standard deviations of the semi-variogram fluctuation when a standard multivariate Gaussian model is chosen by default, whereas the distribution of the available data actually corresponds to a multivariate gamma model. Both models are assumed with the same correlogram (isotropic exponential) and the two previous sampling patterns are examined. Overall, little difference is observed in the results in comparison to the ones displayed in Figs. 2 and 3 for the gamma model. In general, despite the model misspecification, the weighted semi-variogram remains more precise than the traditional and covariance-based semi-variograms, which suggests that the former is still a worthy alternative when the type of multivariate distribution is uncertain.

This observation may be explained by the fact that the true (gamma) and assumed (Gaussian) multivariate distributions have similar features: both of them correspond to diffusion-type random fields with gradational transitions in space (Chilès and Delfiner 1999, p. 402) and with the same correlogram model, which turns out to characterize well enough the spatial continuity of the available data. Of course, the conclusion may not hold if a flagrant error in the multivariate distribution model is made, i.e. if the spatial continuity of the data is completely mistaken. However it should be stressed that, even in such a case, the weighted sample semi-variogram remains an unbiased estimator of the theoretical semi-variogram and can therefore complement the traditional sample semi-variogram.

5 A case study in environmental science

An application of the previous concepts to a real dataset is now presented. This application deals with soil pollution at a smelter site located near Dallas, Texas. To assess the extent of the pollution and its impact on human health, a sampling campaign has been performed under guidance of the U.S. Environmental Protection Agency. Specifically, a set of 180 soil samples have been collected over a circular area with a radius of about 1.7 km (Fig. 5a), in each of which the lead concentration has been measured. Although the spacing between sample locations is quasi-regular, one notices that several areas are not well recognized, in particular flooded areas to the northeast of the pollution plume and in the eastern part of the smelter site. Further details on how the samples have been collected as well as a listing of the data coordinates and values are reported in Isaaks (1984).

The histogram of the lead concentrations is positively skewed and long-tailed, with more than 70% of the data below 300 mg/kg and less than 6% above 2,000 mg/kg (Fig. 5b; Table 2). The maximum measured concentration is 10,400 mg/kg and is located in a junkyard on the extreme east part of the sampled area.

Assume that the goal of the geostatistical analysis is to simulate the lead concentrations over the smelter site, in order to assess the probability that these concentrations (upscaled to remediation units) exceed a

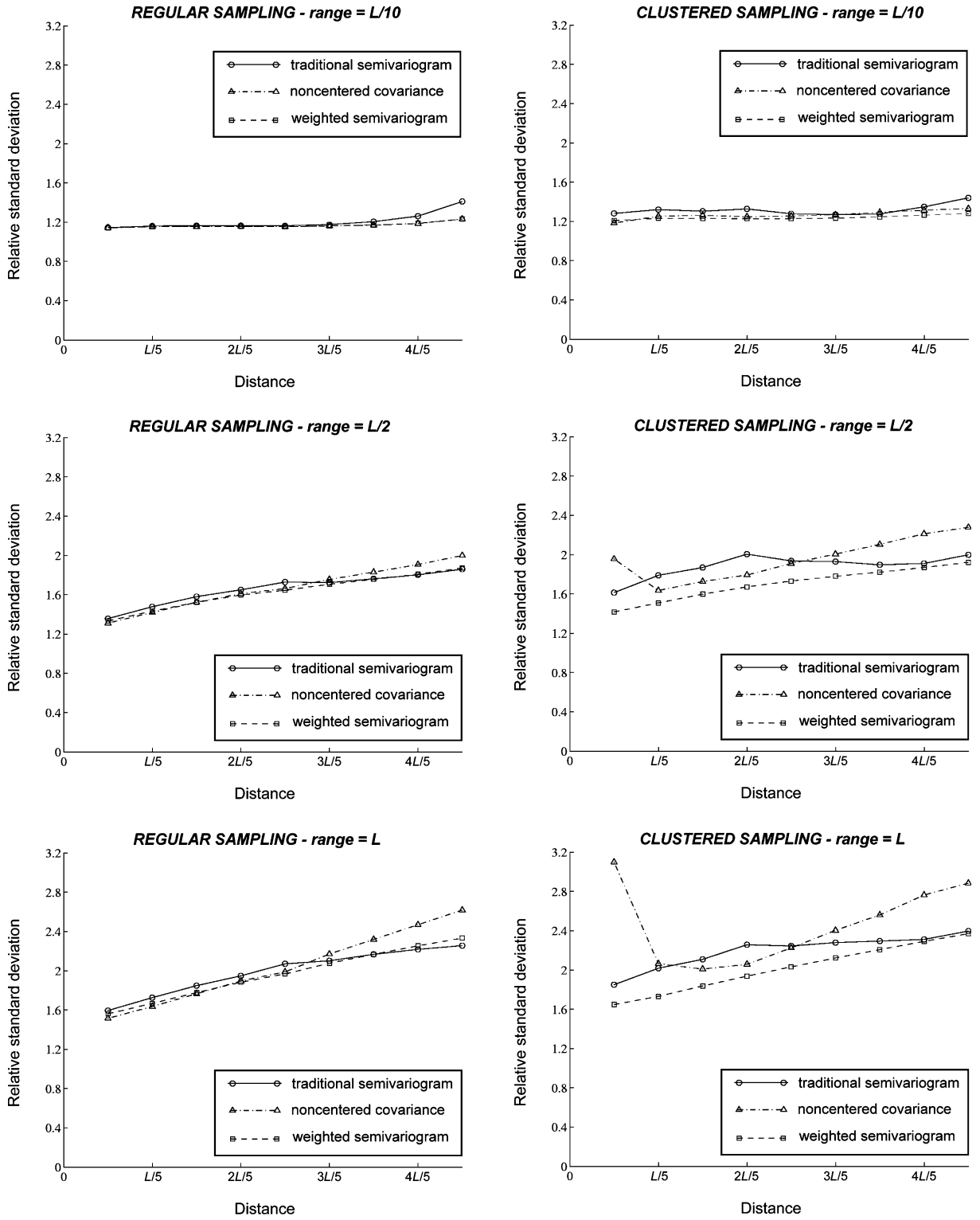

Fig. 4 Fluctuation relative standard deviations for the traditional semi-variogram (*solid lines*), semi-variogram based on noncentered covariance (*dash dots*) and optimally weighted semi-variogram (*dashed lines*). Parameters for semi-variogram calculation are the same as in Figs. 2 and 3. The weights are calculated by assuming a multivariate Gaussian distribution, while the true model corresponds to a multivariate gamma distribution with shape parameter 0.5

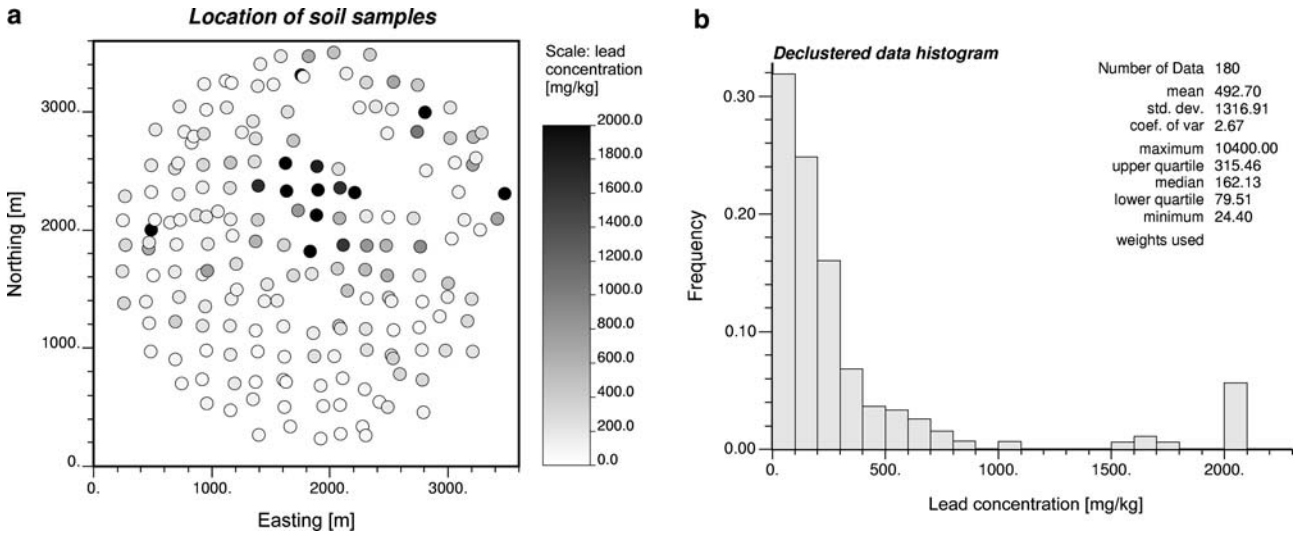

Fig. 5 Posting of the available data (**a**) and histogram of lead concentrations (**b**). The histogram has been declustered using the cell method, with a reference cell size of 600 m × 600 m. The *last bar* of the histogram corresponds to lead concentrations greater than 2,000 mg/kg

regulatory threshold. This requires defining a prior model that represents the multivariate distribution of lead concentrations. In this respect, the aforementioned multivariate gamma model is deemed adequate, as it is associated with skewed univariate distributions and with an asymmetry in the spatial correlation of the indicators around the median threshold (Emery 2005b, p. 428): high values are spatially more continuous than low values and tend to cluster in space, as in the map displayed in Fig. 5a.

The first step of the study consists in transforming the original lead values into a set of data with a standard gamma univariate distribution with shape parameter 0.5. The reader is referred to Emery (2005b, 2006) for details and implementation aspects on the gamma scores transformation and on the conditional simulation of multivariate gamma random fields.

A preliminary covariance or semi-variogram model is required as an initial guess for describing the spatial correlation of the gamma scores data. Since no clear anisotropy can be detected, an omni-directional sample semi-variogram is calculated for lags multiple of the average sampling mesh (230 m), with a tolerance on the distances of half the lag. This sample semi-variogram is fitted by a nugget effect with a sill equal to 0.05 plus an isotropic exponential model with a sill equal to 0.55 and practical range of 1,850 m (Fig. 6a).

Having specified the multivariate distribution (multivariate gamma) and its parameters (prior semi-variogram model), one can determine the weighted semi-variogram estimator for the same lag distances and tolerances as in Fig. 6a. The idea is then to update the semi-variogram model, to use it as an initial guess for re-calculating the weighted sample semi-variogram, and to loop until convergence is reached. In the present case study, three iterations suffice to obtain the final model (Fig. 6d). Although its shape is similar to that of the preliminary model shown in Fig. 6a, it has a lower sill, lower nugget effect and a significantly lower practical range (1,000 m only).

6 Conclusions

The optimally weighted sample semi-variogram complements the traditional sample semi-variogram and may improve the determination of the spatial structure of a set of regionalized data. It is unbiased and produces the smallest fluctuations under a given multivariate distribution model. In practice, the

Table 2 Basic univariate statistics of lead concentrations before and after cell declustering

	Mean (mg/kg Pb)	Variance $(\text{mg/kg Pb})^2$	Lower quartile (mg/kg Pb)	Median (mg/kg Pb)	Upper quartile (mg/kg Pb)
Unweighted	430.4	9.90×10^5	78.75	159.0	314.5
Declustered	492.7	1.73×10^6	79.51	162.1	315.5

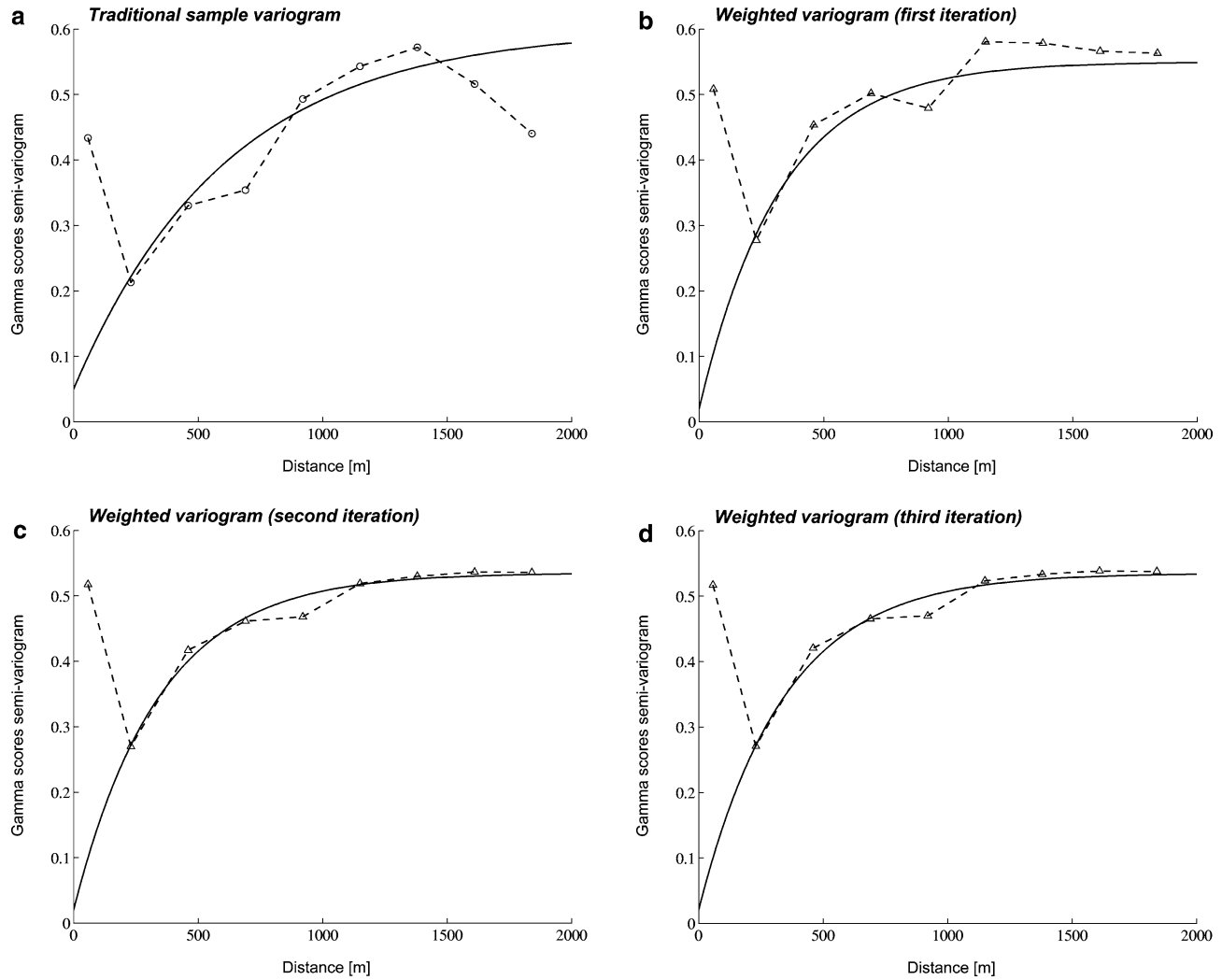

Fig. 6 Omni-directional sample semi-variograms (*dashed lines*) and associated models (*solid lines*) for the gamma score data. The weighted semi-variograms have been calculated by assuming a multivariate gamma distribution with shape parameter 0.5 and by using the semi-variogram fitted at the previous iteration as a prior model

determination of the optimal weighting requires defining the type of multivariate distribution and its parameters, in particular a prior covariance model. The latter can be determined after a traditional variogram analysis, while the former is chosen according to the intended goal of the study (e.g. if multigaussian kriging or simulation are considered, then a multivariate Gaussian distribution is appropriate). Alternatively, the multivariate distribution may constitute only a ''reference'' for representing the expected spatial behavior of the available data.

The proposed methodology is relevant and helpful for variogram analysis when the following conditions are met:

1. one looks for a stationary model, for instance in view of using ordinary kriging;
2. the sampling pattern is irregular;
3. there is a significant spatial correlation, hence there exist redundancies between data values;
4. the number of data and data pairs is less than a few thousands, a situation that often arises in environmental sciences. The determination of the optimal pair weighting is impractical with large datasets, as CPU requirements to solve Eq. 15 become excessive. However, the weighted sample semi-variogram may still be used for inferring the spatial structure at small distances, for which one generally has fewer data pairs.

Although it accounts for spatial redundancies between data values, the weighted sample semi-variogram approach does not necessarily solve problems associated with the occurrence of extreme values (outliers) and with preferential sampling patterns, e.g. those for which high-value areas are over-sampled with

respect to other areas. Such preferential samplings remain a critical issue in structural analysis, as they often entail biases in the sample semi-variogram (Omre 1984, p. 111) and interfere with one's understanding of the spatial continuity.

Acknowledgment The author acknowledges the sponsoring by Codelco-Chile for supporting this research.

References

Armstrong M (1984) Common problems seen in variograms. Math Geol 16(3):305–313. DOI 10.1007/BF01032694

Armstrong M, Delfiner P (1980) Towards a more robust variogram: a case study on coal. Technical report N-671, Ecole Nationale Supérieure des Mines de Paris, Fontainebleau

Chilès JP, Delfiner P (1999) Geostatistics: modeling spatial uncertainty. Wiley, New York, p 695

Cressie NAC (1993) Statistics for spatial data (rev. ed.). Wiley, New York, p 900

Cressie N, Glonek G (1984) Median based covariogram estimators reduce bias. Statist Probab Lett 2(5):299–304. DOI 10.1016/0167-7152(84)90069-5

Cressie N, Hawkins DM (1980) Robust estimation of the variogram. Math Geol 12(2):115–125. DOI 10.1007/BF01035243

Dowd PA (1984) The variogram and kriging: robust and resistant estimators. In: Verly G, David M, Journel AG, Maréchal A (eds) Geostatistics for natural resources characterization. Reidel, Dordrecht, pp 91–106

Emery X (2005a) Variograms of order ω: a tool to validate a bivariate distribution model. Math Geol 37(2):163–181. DOI 10.1007/s11004-005-1307-4

Emery X (2005b) Conditional simulation of random fields with bivariate gamma isofactorial distributions. Math Geol 37(4):419–445. DOI 10.1007/s11004-005-5956-0

Emery X (2006) A disjunctive kriging program for assessing point-support conditional distributions. Comput Geosci 32(7):965–983. DOI 10.1016/j.cageo.2005.10.011

Emery X, Ortiz JM (2005) Histogram and variogram inference in the multigaussian model. Stoch Environ Res Risk Assess 19(1):48–58. DOI 10.007/s00477-004-0205-5

Genton MG (1998) Highly robust variogram estimation. Math Geol 30(2):213–221. DOI 10.1023/A:1021728614555

Greenbaum A (1997) Iterative methods for solving linear systems. Society for Industrial and Applied Mathematics, Philadelphia, p 220

Isaaks EH (1984) Risk qualified mappings for hazardous waste sites: a case study in distribution free geostatistics. unpublished Master's Thesis, Stanford University, Department of Applied Earth Sciences, p 85

Isaaks EH, Srivastava RM (1988) Spatial continuity measures for probabilistic and deterministic geostatistics. Math Geol 20(4):313–341. DOI 10.1007/BF00892982

Isaaks EH, Srivastava RM (1989) An introduction to applied geostatistics. Oxford University Press, New York, p 561

Isserlis L (1918) On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. Biometrika 12:134–139

Journel AG, Huijbregts CJ (1978) Mining geostatistics. Academic, London, p 600

Kovitz JL, Christakos G (2004) Spatial statistics of clustered data. Stoch Environ Res Risk Assess 18(3):147–166. DOI 10.1007/s00477-003-0133-9

Matheron G (1971) The theory of regionalized variables and its applications. Les cahiers du Centre de Morphologie Mathématique de Fontainebleau, Fascicule 5, Ecole Nationale Supérieure des Mines de Paris, Fontainebleau, p 211

Matheron G (1989) Estimating and choosing: an essay on probability in practice. Springer, Berlin Heidelberg New York, p 141

Omre H (1984) The variogram and its estimation. In: Verly G, David M, Journel AG, Maréchal A (eds) Geostatistics for natural resources characterization. Reidel, Dordrecht, pp 107–125

Richmond A (2002) Two-point declustering for weighting data pairs in experimental variogram calculation. Comput Geosci 28(2):231–241. DOI 10.1016/S0098-3004(01)00070-X

Rivoirard J (1987) Computing variograms on uranium data. In: Matheron G, Armstrong M (eds) Geostatistical case studies. Reidel, Dordrecht, pp 1–22

Rivoirard J (2001) Weighted variograms. In: Kleingeld WJ, Krige DG (eds) Proceedings of the 6th international geostatistics congress. Geostatistical Association of Southern Africa, Cape Town, pp 145–155

Rivoirard J, Simmonds J, Foote KG, Fernandes P, Bez N (2000) Geostatistics for estimating fish abundance. Blackwell, Oxford, p 206

Srivastava RM, Parker HM (1989) Robust measures of spatial continuity. In: Armstrong M (ed) Geostatistics. Kluwer, Dordrecht, pp 295–308

Triantafyllopoulos K (2003) On the central moments of the multidimensional Gaussian distribution. Math Sci 28(2):125–128