Cahiers de Géostatistique, Fascicule 1
Compte-rendu des Journées de Géostatistique,
6-7 Juin 1991, Fontainebleau, pp. 225-246

RESTRICTED MAXIMUM LIKELIHOOD ESTIMATION OF SEMIVARIOGRAM VALUES IN NESTED DESIGNS.

A. STEIN

Dept. Soil Science & Geology
Agricultural University
P.O. Box 37
6700 AA Wageningen
The Netherlands

L.C.A. CORSTEN

Emeritus Professor of Math. Statistics
Ritzemabosweg 20
6703 AX Wageningen
The Netherlands

ABSTRACT

In spatial studies semivariogram values for a small number of distance classes may be obtained by means of a nested sampling design. Following a previous paper [Corsten and Stein, *subm.*], the present paper addresses the question which estimation procedure one should follow in order to obtain estimations for semivariogram values in nested designs. Particular attention is given to Restricted Maximum Likelihood Estimation (REML). It is shown that REML-estimation performs well as compared to expected mean squares. The existence only of asymptotic variances of the estimates for the variance of semivariogram values, however, is felt to be a real disadvantage. The study is illustrated by two designs with artificial data and by one design with data emerging from soil science. In unbalanced designs, REML estimation may yield semivariogram values with (only asymptotic) variances of about 70% the size of the variances obtained by expected mean squares.

TABLE OF CONTENTS

1. INTRODUCTION
2. NESTED DESIGNS
3. ESTIMATING SEMIVARIOGRAM VALUES IN NESTED DESIGNS
4. EXAMPLES AND RESULTS
5. DISCUSSION
6. REFERENCES
APPENDIX

1. INTRODUCTION

In spatial inventory studies common use is made of nested sampling plans. In particular, when a pilot study is carried out to get a first impression of differences in soil and geologic characteristics nested designs have yielded important results [Burrough, 1986; Oliver and Webster, 1986; Van Dongen and Widianto, *in prep.*]. The main advantage is that observation locations have several differing intermediate distances, ranging from small to large. When no information is available on the scale of variation of a spatially varying characteristic a nested design may yield preliminary information. In order to analyse the relation between observations and the separation distance between their observation locations, common use is made of semivariograms. Although the semivariogram is a function of the separation distance between observation locations only in the absence of a trend, as a first approach it may well serve the purpose.

In [Corsten & Stein, *subm.*], included in [Stein, 1991], it was proven that semivariograms for certain distances are equivalent to the cumulative sum of several variance components each of those belonging to a particular distance and available from a classical ANOVA procedure. Moreover, the mean of pair differences to estimate the semivariogram, the so-called intuitive estimator, is equivalent to values obtained from expected mean squares in balanced designs, but differences are observed for unbalanced designs. It has also been shown that in unbalanced designs neither estimation procedure has uniformly minimum variance: for certain values of the variance components in some unbalanced designs the intuitive semivariogram estimator has smaller variance, in other situations the ANOVA estimator has. However, serious doubts were cast whether nested sampling schemes must be advocated as a general procedure to estimate semivariograms, since values for only a small number of distance classes are obtained.

There are other ways to estimate variance components. Recently maximum likelihood and restricted maximum likelihood procedures have received much attention in literature [Patterson and Thompson, 1974; Engel, 1991; Dietrich and Osborne, 1991; Pettitt and McBratney, *in press*]. In the present study attention will be given to estimation of the variance components and their variances to be used for semivariogram estimation in nested designs.

The main problem we want to deal with in this paper is whether the use of one estimation method is to be preferred above any other procedure. Attention will therefore be given to several data sets, of which one emerges from a practical study.

In Chapter 2 attention will be given to the general description of nested designs and the relation between variance components obtained from a classical analysis of variance and estimated semivariogram values. In Chapter 3 different procedures to estimate variance components in nested designs will be outlined. In Chapter 4 attention is paid to several applications. In an appendix special attention is given to the use of expected mean squares for variance components estimation.

2. NESTED DESIGNS

We will consider observations which are collected in space on a spatially varying characteristic. One may think of soil properties like the content of a pollutant, or the thickness of the plough layer, a geological characteristic like the ore grade or the content of a geochemical mineral, a meteorological characteristic like the mean annual precipitation, or other environmental and sociological variables. In many studies such variables have been treated as regionalized variables (see, for example, [Journel & Huijbregts, 1978]). On such variables observation are collected, sometimes following a predetermined sampling scheme, such as the nodes of a triangular, square or hexagonal grid or the equally spaced points along a transect.

In order to determine the scale of variation which may be helpful to fix the distance between grid nodes, or points on a transect, in advance, a pilot study may be carried out using nested designs, for which relatively little effort is required to obtain observations for several different distances [Webster, 1985]. Use of analysis of variance may yield the necessary information to make an assessment of an adequate grid mesh.

As a nested design we consider a design where observations are collected at several levels of variation with systematically decreasing distances between them [Gower, 1962]. The prior choice for the levels is usually given by relevant sources of variation. When, for example, observations are collected for different areas, within every area for

different soil types, within different soil types for different land use types, for different land use types for different parcels, for different parcels both on the ridge and in the furrow, and each observations is replicated twice, one obtains a six-level nested design. Distances between the observations decrease in passing from the most global level to the most detailed level. A nested design results in a nested classification of the observations.

In order to formulate the statistical model, consider a vector of observations $\underline{y}$ which are collected according to a h stage nested design with fixed distances $0 = r_0 < r_1 < .. < r_h$ among pairs of observations, perhaps proportional to each other with a common factor. In practical studies factors ranging from 5 to 10 have been used [Van Dongen and Widianto, *in prep.*; Miesch, 1975]. In this paper we will follow the common convention to underline random variables, in order to distinguish them from fixed values. For any r_i a classification A_i of the observations is defined, in which any class a_{ij}, $j = 1,\ldots,m_i$, of A_i, possibly of different sizes, consists of observations which have intermediate distance up to r_i. It is noted that the actual distances are only approximately equal to each other. Let the number of elements in a_{ij} be denoted by the generic symbol n_i; in particular, $n_h = n$ while n_0 is always equal to 1. The subscript i decreases as the classification becomes more refined. To be more specific, at each level i of the hierarchy is associated the partition A_i of the samples. The classes a_{ij}, $j = 1,\ldots,m_i$, satisfy two conditions:

i) any two points x_1 and x_2 in the same class of A_i have distance $|x_1-x_2| \leq r_i$;

ii) for $i \geq 1$ any class a_{ij} of A_i is a union of classes in A_{i-1}.

Obviously, A_0 is the finest classification, in which every class contains exactly one observation, whereas A_h consists of one class, containing the n observations.

Next with each classification A_i a subspace of R^n can be associated consisting of the set of vectors which are constant within the classes a_{ij} of A_i. This subspace will be called A_i as well. The space A_{i+1} is a subspace of A_i, which, by noting that the subspace A_0 equals R^n, holds for $i = 0,\ldots,h-1$. We notice that A_h is spanned by 1_n, the vector consisting of n elements all equal to one.

For a classical nested design with independent observations within each stage and with homogeneous variances at the same stage the covariance matrix V of $\underline{y}$ would be $V = \sum_{i=0}^{h-1} \sigma_i^2 V_i = \sum_{i=0}^{h-1} \sigma_i^2 U_i U_i'$, where the j^{th} column of U_i equals the n-vector with elements 1 in the j^{th} class of A_i and 0 elsewhere. The columns of U_i are an orthogonal base of A_i. Each variance component σ_i^2 is associated with a level of nesting in such a way that the variance component σ_0^2 is the variance within A_1, the class of points with the same distance r_1, the variance component σ_i^2 is the variance within the class of points with the same the distance r_{i+1}, and σ_{h-1}^2 is the variance within A_h, the class of points with distances up to r_h, being the largest distance considered.

In short, we will use the following linear model:

$$E\underline{y} = \mu 1_n \qquad (1)$$

$$\mathrm{Var}(\underline{y}) = \sum_{i=0}^{h-1} \sigma_i^2 U_i U_i'.$$

Values for the parameters μ, σ_0^2, σ_1^2 ,..., σ_{h-1}^2 have to be estimated from the available data.

A feature commonly encountered in spatial studies is that observations close to each other are more similar to each other than observations which are separated by a larger distance. Use can be made of the semivariogram to describe the (spatial) dependence between observations, which, as a function of the distance r between observations is, defined as

$$\gamma(r) = \tfrac{1}{2}E[(\underline{y} - \underline{y}_r)^2] \qquad (2)$$

where $\underline{y}$ and $\underline{y}_r$ form a pair of points separated by a distance r. The factor ½ allows one to establish a straightforward analogy between the semivariogarm and the covariance function c(r) in the case that $\mathrm{Var}(\underline{y}) = c(0)$ exists: $\gamma(r) = c(0) - c(r)$, and hence the limiting value for $r \to \infty$ equals $\mathrm{Var}(\underline{y})$.

Based on the fact that to the matrix of semivariogram values between observations any constant may be added without changing the covariance

structure between the observations it has been shown [Corsten and Stein, *subm.*] that the semivariogram value for the i^{th} distance equals the cumulative sum of the i variance components for the levels up to the $i\text{-}1^{th}$ one:

$$\gamma(r_i) = \sum_{j=0}^{i-1} \sigma_j^2. \qquad (3)$$

Eq. (3) holds generally and exactly for balanced as well as unbalanced designs. Summation of estimated variance components therefore yields an estimation of h semivariogram values in nested designs.

3. ESTIMATING SEMIVARIOGRAM VALUES IN NESTED DESIGNS

Different estimation procedures are distinguished to estimate semivariogram values by means of variance components in nested designs. In this paper we will compare mean squared pair difference (MSPD) estimation with expected mean squares (EMS) and with restricted maximum likelihood (REML) estimation.

3.1. Expected mean squares.

Expected mean squares are commonly applied to estimate variance components in ANOVA procedures. We will here present only the main results of expected mean square estimation of variance components; further details are elaborated on in the Appendix. Define P_i to be the orthogonal projection on the orthogonal complement of A_{i+1} in A_i $(i = 0,\ldots,h\text{-}1)$. In matrix form P_i may be written as $P_i = U_i(U_i'U_i)^{-1}U_i' - U_{i+1}(U_{i+1}'U_{i+1})^{-1}U_{i+1}'$, from which we notice that $P_i\underline{y}$ is the difference vector between the projection of $\underline{y}$ on A_i and its projection on A_{i+1}, representing variation between the classes of A_i within those of A_{i+1}. We notice that $E(P_i\underline{y}) = 0$ and hence that $E[(P_i\underline{y})'(P_i\underline{y})] = E[\underline{y}'P_i\underline{y}] = tr(P_iV)$. With

$$k_{ij} = \Sigma n_j^2\left(\frac{1}{n_i} - \frac{1}{n_{i+1}}\right) \qquad (4)$$

for $i \geq j \geq 0$, summation extending to all classes of A_j, we obtain the following expected sums of squares equivalent to (4):

$$E(\underline{y}'P_i\underline{y}) = \sum_{j=0}^{i} k_{ij}\sigma_j^2 \qquad (5)$$

which holds for $i = 0,\ldots,h-1$.

Equating $E(\underline{y}'P_i\underline{y})/k_{i0}$ to $MS_i = y'P_iy/k_{i0}$, one can solve the linear equations emerging from (5) to obtain unbiased estimators $\hat{\sigma}_i^2$ of σ_i^2, which by summing according to (3) yield unbiased estimators $\tilde{\gamma}_i$ of h semivariogram values. After some algebraic manipulations, one can deduce with the following successive operators

$$Q_0 = \frac{1}{k_{00}} P_0$$

$$Q_i = \frac{1}{k_{ii}} P_i + \sum_{j=0}^{i-1} \frac{k_{i,j+1} - k_{ij}}{k_{ii}} Q_j \qquad (6)$$

that

$$\tilde{\gamma}_{i+1} = y'Q_iy \qquad (7)$$

for $i=0,\ldots,h-1$.

By means of (7) an expression for the variance of $\tilde{\gamma}_i$ is obtained. As is well known, $var(\underline{y}'A\underline{y}) = 2tr(AVAV)$ for any symmetric A, if $\underline{y}$ follows a Gaussian distribution [Searle, 1987]. Hence, under normality, $var(\tilde{\gamma}_i) = 2tr(Q_{i-1}VQ_{i-1}V)$.

3.2. Mean squared pair differences.

As mentioned before the semivariogram value for any distance is most commonly estimated by half the mean of squared pair differences of all pairs of points with (approximately) that distance between the points. The intuitive unbiased semivariogram estimator $\hat{\gamma}_i$ is defined as half the mean of squared differences among all pairs of observations at distance r_i. In order

to allow a comparison of this MSPD estimator with the EMS estimator, it is written as a quadratic form:

$$\hat{\gamma}_i = \frac{1}{2N_i} \underline{y}' D_i \underline{y}, \qquad (8)$$

where N_i is the number of pairs with distance r_i and the symmetric matrix D_i equals $U'_{i-1}U_{i-1} - U'_iU_i + H_i$, where H_i is a diagonal matrix with $(H_i)_{jj} = n_i - n_{i-1}$, n_i and n_{i-1} being the size of the class of A_i and A_{i-1}, respectively, containing the j^{th} element. Again under normality, the variance of estimator (10) is obtained as $\text{var}(\hat{\gamma}_i) = (1/2N_i^2)\text{tr}(D_iVD_iV)$.

3.3. Restricted Maximum likelihood.

For the Restricted Maximum Likelihood (REML) estimation procedure the method of scoring as outlined in [Patterson and Thompson, 1974] is adequate. In fact, REML is an extension of ordinary maximum likelihood, taking a fixed effect into account. It is shown by several authors that the ordinary maximum likelihood estimator in a mixed model leads to extremely biased estimates and are therefore undesirable. In the model formulation (1), the general mean μ is the one fixed effect, to which the other, random, effects are orthogonal. By means of the REML procedure estimates for the random effect parameters are obtained. An estimation of μ is obtained afterwards as $\hat{\mu} = (1'_nV^{-1}1_n)^{-1}1'_nV^{-1}y$.

In the present paper restricted maximum likelihood is formulated in terms of increments of the observations. Vectors of increment coefficients of the observations are orthogonal to the (fixed) expectation vector and use of increments avoids the search for a generalized inverse of a singular matrix. When there are n observations there are n-1 independent increments. To obtain them, define $\underline{z}$ by means of $\underline{z} = C\underline{y}$, where C, of size n-1 * n contains a basis for the vector of increment coefficients. For example C may be obtained by deleting any arbitrary row of $(I - 1_n(1_n1'_n)^{-1}1'_n)$. The row choice is irrelevant, since the space spanned by the remaining rows remains the same and any basis of that space can be tarnsformed into another one, e.g. the one arising by deleting another row. We notice that $C1_n = 0$ indeed,

and hence that $\underline{z}$ has coefficients orthogonal to 1_n. The variance of $\underline{z}$ equals

$$W = \mathrm{Var}(\underline{z}) = CVC' = \sum_{j=0}^{h-1} \sigma_j^2 C U_j U_j' C' = \sum_{j=0}^{h-1} \sigma_j^2 W_j \qquad (9)$$

The likelihood equations are obtained by putting $\underline{z}'W^{-1}W_iW^{-1}\underline{z}$ equal to its expectation $\mathrm{tr}(W^{-1}W_i)$. As can be shown by standard methods [Kitanidis, 1983], this yields the following system of linear equations:

$$\mathrm{tr}(W^{-1}W_iW^{-1}\underline{z}.\underline{z}') = \sum_{j=0}^{h-1} \sigma_j^2 \mathrm{tr}(W^{-1}W_iW^{-1}W_j),\ i = 0,\ldots,h-1. \qquad (10)$$

In order to solve (10) one starts with a preliminary vector of estimates $(\tilde{\sigma}_0^2,\ldots,\tilde{\sigma}_{h-1}^2)$. These are inserted for $(\sigma_0^2,\ldots,\sigma_{h-1}^2)$ yielding a new matrix $\tilde{W} = W_1\tilde{\sigma}_0^2 + \ldots + W_h\tilde{\sigma}_{h-1}^2$, which may be used in turn to improve the estimate of σ_i^2. Semivariogram values are obtained by adding the estimated variance components according to (3). The matrix F, defined as $\{f_{ij}\}$ with $f_{ij} = \mathrm{tr}(W^{-1}W_iW^{-1}W_j)$, is the inverse of the Fisher information matrix [Patterson and Thompson, 1974], which is obtained during the solution of (10). Asymptotic estimates for the variances of the variance components are obtained as the coefficients of the matrix F^{-1}. As an estimate of asymptotic variance of the semivariogram estimator the sum of the variances of the variance components plus twice the sum of their covariances will serve. Hence

$$\mathrm{Var}(\gamma_i) = f_i' F^{-1} f_i, \qquad (12)$$

where f_i is the vector of length h, with the first i components equal to 1 and extended with h-i+1 components equal to zero.

IV. EXAMPLES AND RESULTS

4.1. Four-stage balanced design

The estimation procedures were applied to several data sets. The first data set contains 16 observations collected according to a 4-stage balanced

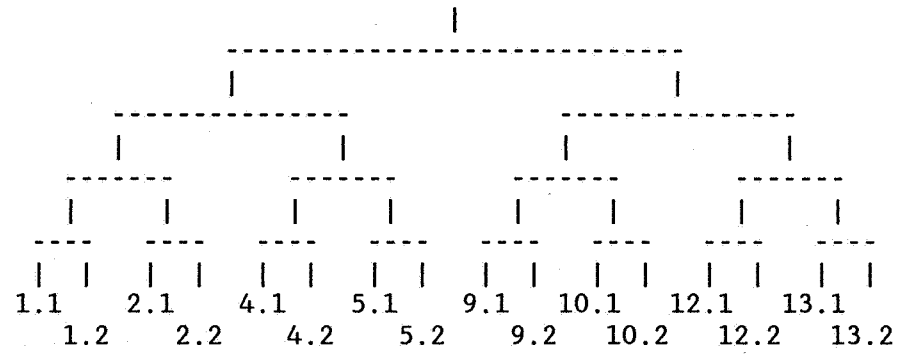

Fig. 1. Observations collected according to a four-stage nested design with two branches at each level.

		MSPD	EMS	REML
γ_1	Est.	0.005	0.005	0.005
	Var.	*0.000*	*0.000*	*0.000*
γ_2	Est.	0.5025	0.5025	0.5025
	Var.	*0.125*	*0.125*	*0.125*
γ_3	Est.	4.752	4.752	4.752
	Var.	*20.28*	*20.28*	*20.28*
γ_4	Est.	34.503	34.503	34.503
	Var.	*2053.1*	*2053.1*	*2053.1*

TABLE 1. Estimated semivariogram values and their estimated variances as obtained with different estimation procedures for the four-stage balanced design.

design, with two branches at each level, yielding a total of 16 observations (fig. 1). Values are assigned to the observations in such a way that small differences occur among observations which are close to each other, and larger differences for observations which are further apart.

Estimated semivariogram values and their estimated variances, which for

the EMS and MSPD procedure are obtained by inserting estimated values for the variance components in the matrix V, are summarized in table 1. For this balanced design we notice that the three procedures yield identical results for γ_i (i = 1, 2, 3 and 4) both for the estimated semivariogram values and for their variances. In particular the variances of the semivariogram values for both EMS and for MSDP were obtained as a the quadratic form in σ_0^2, σ_1^2, σ_2^2 and σ_3^2 with the symmetric matrices

$$\begin{bmatrix} A_1 = 0.25 \end{bmatrix} \begin{bmatrix} A_2 = 0.188 & 0.25 \\ 0.25 & 0.5 \end{bmatrix} \begin{bmatrix} A_3 = 0.156 & 0.188 & 0.188 \\ 0.188 & 0.25 & 0.5 \\ 0.188 & 0.5 & 1.0 \end{bmatrix}$$

$$\begin{bmatrix} A_4 = 0.141 & 0.156 & 0.188 & 0.25 \\ 0.156 & 0.313 & 0.375 & 0.5 \\ 0.188 & 0.375 & 0.75 & 1 \\ 0.25 & 0.5 & 1 & 2 \end{bmatrix}$$

for var(γ_1), var(γ_2), var(γ_3) and var(γ_4), respectively. Uniform minimum variance unbiased estimators of the variance components for balanced designs in the sense of Lehmann [Lehmann, 1983], are therefore obtained by REML estimation as well as by EMS and by MSPD.

4.2. Three-stage unbalanced design.

The second data set contains 8 observations collected according to a three-stage unbalanced design (fig. 2). Although the data are artificial one may think in a practical soil study that two soil units are sampled, separated by a distance of approximately 1 km, in each of which two parcels

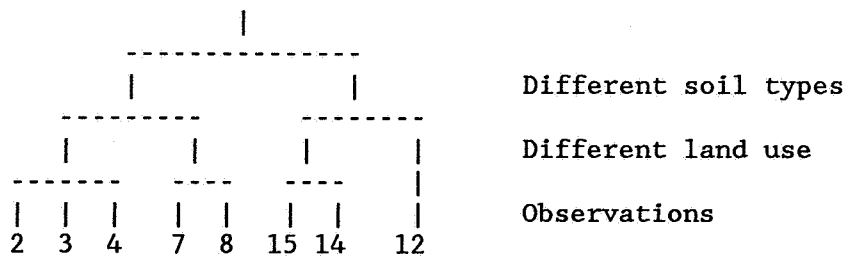

Fig. 2. Observations collected according to a three-stage nested design with one, two or three branches at each level.

		MSPD	EMS	REML
γ_1	Est.	0.800	0.750	0.744
	Var.	*0.333*	*0.281*	*0.277*
γ_2	Est.	8.75	7.973	7.113
	Var.	*86.42*	*62.47*	*46.279*
γ_3	Est.	42.77	43.198	36.153
	Var.	*2983*	*3104*	*2116*

TABLE 2. Estimated semivariogram values and their estimated variances as obtained with different estimation procedures for the three-stage unbalanced design

are sampled, separated by a distance approximately equal to 100m, with different land use types, in each of which 1, 2 or 3 observations are taken, separated by 10m.

Estimated semivariogram values and their variances are summarized in table 2. In contrast to the four-stage design we notice apparent differences. In particular, variances obtained with the REML estimation procedure are substantially lower than those obtained with the MSDP and the EMS estimation procedure. We conclude that REML estimation may be preferred to EMS and MSDP estimation in this example.

For both MSDP and EMS var(γ_1), var(γ_2) and var(γ_3) were again calculated as a quadratic form in terms of variance components σ_0^2, σ_1^2 and σ_2^2 with the matrices as follows:

MSDP:

$$A_1 = \begin{bmatrix} 0.52 \end{bmatrix} \quad A_2 = \begin{bmatrix} 0.406 & 0.567 \\ 0.567 & 1.25 \end{bmatrix} \quad A_3 = \begin{bmatrix} 0.333 & 0.409 & 0.533 \\ 0.409 & 0.827 & 1.075 \\ 0.533 & 1.075 & 2 \end{bmatrix}$$

EMS:

$$A_1 = \begin{bmatrix} 0.5 \end{bmatrix} \quad A_2 = \begin{bmatrix} 0.395 & 0.535 \\ 0.535 & 1.082 \end{bmatrix} \quad A_3 = \begin{bmatrix} 0.321 & 0.401 & 0.533 \\ 0.401 & 0.810 & 1.075 \\ 0.533 & 1.075 & 2 \end{bmatrix}$$

From this we notice that for this particular design the variances obtained with EMS are uniformly lower than the variances obtained with MSDP. This does not hold generally, however.

4.3. Actual data collected by means of a six-stage unbalanced design

The third data set is the one found in [Pettitt and McBratney, *in press*]. This papers considers actual data on the A horizon thickness of about 100 sites in a field near Forbes in New South Wales, Australia. An overview of the sampling locations is given in figure 3, whereas a kriged map of the thickness of the A-horizon is presented in figure 4. In every block observations are located according to the design presented in figure 5.

In fact, the four observations with intermediate distances 125m, 25m, 5m and 1m in every block occur on transects, whereas the two observations with intermediate distance 0.2m are located 0.1m above and below the transect. The transects systematically follow a prescribed direction (0^o, 60^o or 120^o with the horizontal axes) in the 18 blocks throughout the area in order to study anisotropy. The design is termed by the authors a 5-stage staggered nested design with blocks. In our terminology this design will be called an unbalanced 6-stage design.

Blocks have been used to distribute the observations evenly over the area [Pettitt and McBratney, *in press*], but were not imposed by considerations like differences in land use or soil type or different management conditions, etc. Use of this design has several attractive properties: the presence of many observations at short distances from each other allows a precise estimation of the nugget effect, the combination of blocks and transects yields an operationally efficient sampling scheme and the use of proportionally increasing distances allows detection of multiple sources of variation.

Estimated semivariogram values by means of variance components as well as for their variances are shown in table 3. MSPD estimates and MSE estimates are identical (up to calculation precision) due to the fact that at every level one of the branches has exactly one observation. However, these estimates differ slightly from the REML estimates. Variance components estimated by means of the REML estimation procedure, on which the estimated semivariogram are (linearly) based have lower variances as compared to the other two procedures, except for that for γ_3.

Since imposing the blocks in this study is rather artificial, we have

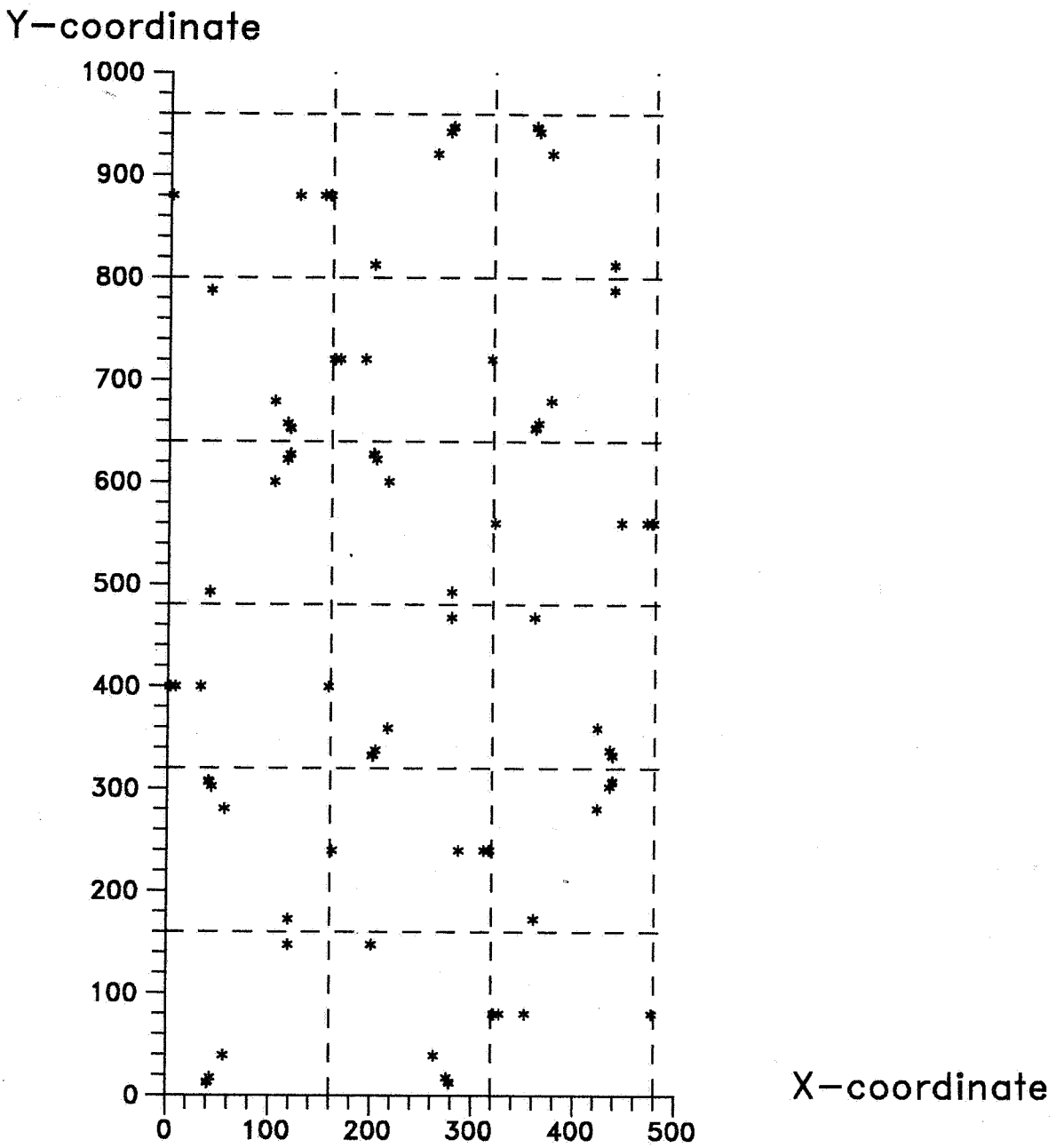

Fig. 3. Map of the sampling locations in the Forbes site in New South Wales, Australia. Within each of the 18 blocks 6 observations are located on transects as described in the text.

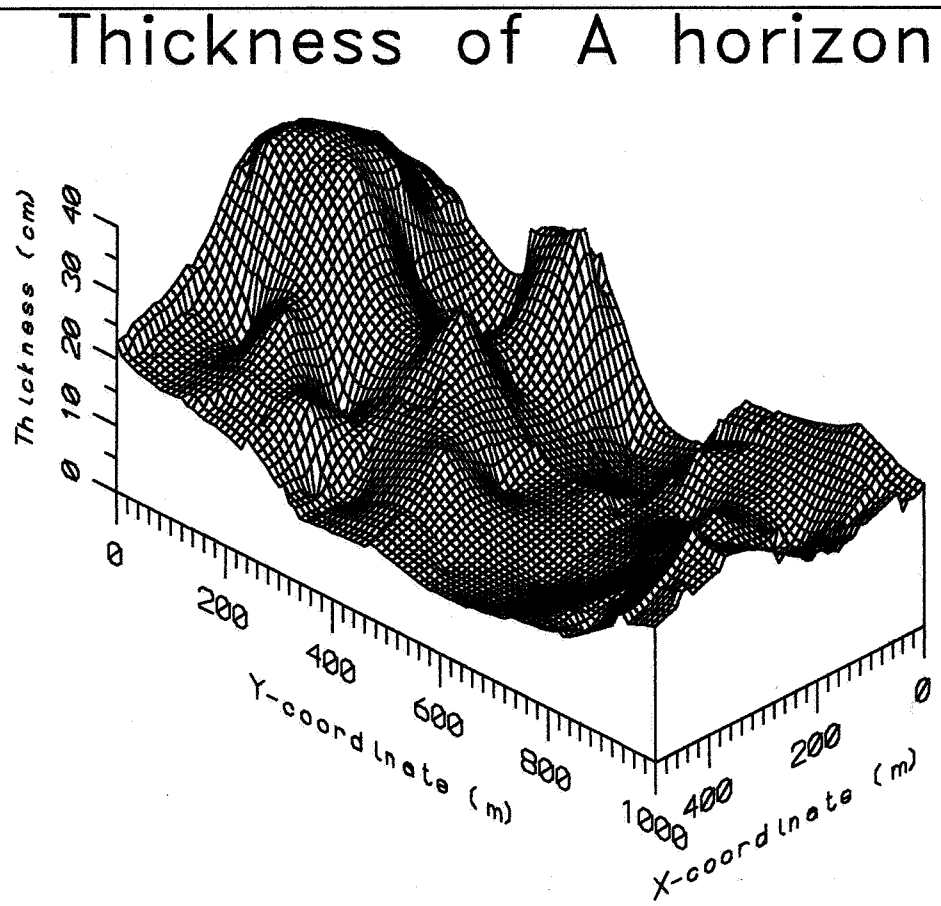

Fig. 4. Kriged map of the thickness of the A horizon in the Forbes site.

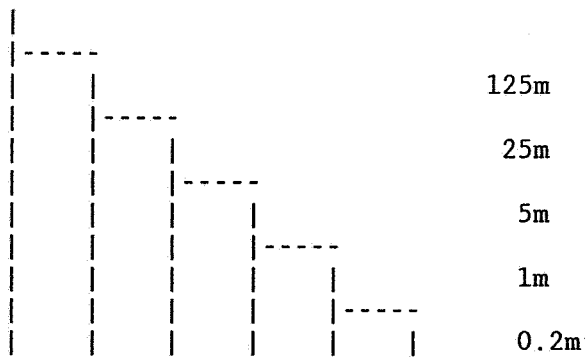

Fig. 5. Observations collected according to a (highly) unbalanced six-stage nested design.

		MSPD	EMS	REML
γ_1	Est.	1.132	1.132	1.143
	Var.	*0.142*	*0.142*	*0.145*
γ_2	Est.	9.767	9.767	9.150
	Var.	*10.003*	*10.003*	*8.633*
γ_3	Est.	12.264	12.264	14.593
	Var.	*11.639*	*11.639*	*16.885*
γ_4	Est.	32.597	32.597	30.290
	Var.	*94.284*	*94.284*	*73.852*
γ_5	Est.	40.846	40.846	40.602
	Var.	*128.048*	*128.048*	*114.561*
γ_6	Est.	103.314	103.314	104.509
	Var.	*811.035*	*811.033*	*757.410*

TABLE 3. Estimated semivariogram values and their estimated variances as obtained with different estimation procedures for the six-stage unbalanced design.

estimated the semivariogram also by ignoring the blocks (Fig. 6). A Gaussian model fitted the experimental semivariogram best, as judged by the weighted error sum of squares, yielding a sill value equal to 109.2 cm^2 and a nugget effect equal to 19.9 cm^2, whereas 'eye-fitting' would have yielded an exponential model with a nugget effect equal to about 9 cm^2. It appears that the nugget effect obtained by means of the experimental semivariogram neglects the semivariogram value at the smallest distance, obtainable by means of the nested ANOVA: the observed value equals 9 cm^2, which is in agreement with the estimate for γ_2 obtained with the nested ANOVA, but not with that for γ_1. We further notice that the estimate for the sill is well in agreement with the variance components obtained by means of the nested ANOVA, but the distance at which this value is reached (259m), a value of crucial importance for many soil studies [Webster, 1985] is not obtainable from the nested ANOVA, for which the largest distance at which a variance component is obtained equals 125m.

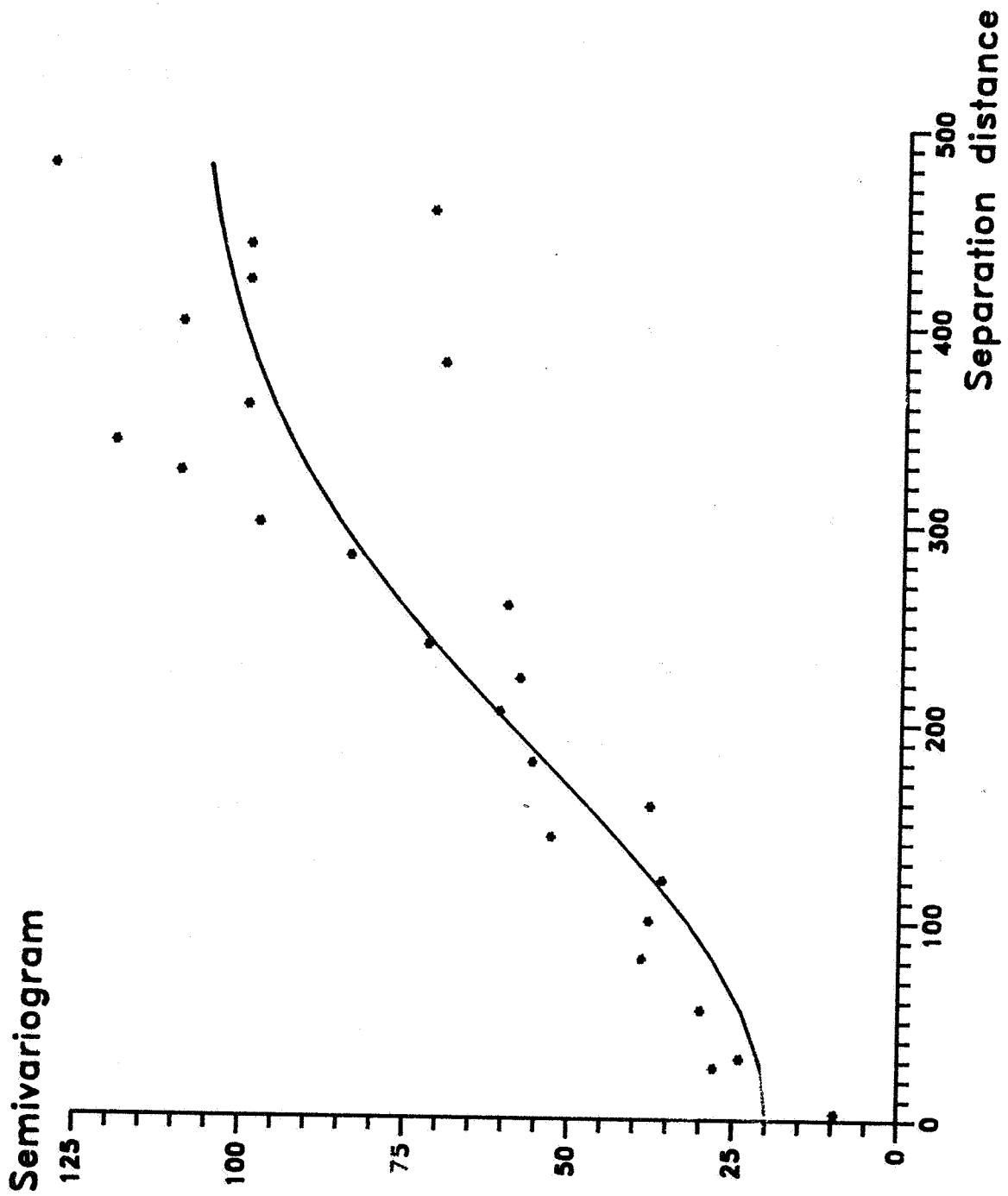

Fig. 6. Experimental semivariogram of the thickness of the A horizon in the Forbes site in New South Wales, Australia. A fitted Gaussian model is shown.

DISCUSSION

The use of nested designs to calculate variance components for spatial semivariograms has certain attractive properties. For example, REML procedures may be used to estimate their values as well as the associated variances. As indicated by this study, REML estimation procedure is attractive as compared to EMS and MSPD estimation, especially if the sampling design is (highly) unbalanced. There appears to be little gain in using MSPD estimation, which, however, yields similar results to the EMS and the REML procedure for balanced designs. Nested designs are helpful in revealing the different sources of variation within a region.

Comparing semivariogram estimates obtained by means of model fitting with those obtained by nested analysis of variance, indicates that imposing a block structure on an area is not very fruitful, although estimating the nugget effect is apparently more precise by the latter procedure. There remains a major drawback when using nested designs, since only a limited number of experimental semivariogram values will be available. No insight can therefore be obtained concerning the range of the semivariogram, nor whether any model is most likely to suit the experimental semivariogram. This prohibits the analysis of spatial data in all details.

ACKNOWLEDGEMENT

The authors are grateful to A.B. McBratney, Derpartment of Crop & Soil Science, University of Sydney, Australia, for giving permission to use the Forbes data.

REFERENCES

Burrough, P.A. 1986. *Principles of geographical information systems for land resources assessment*. Oxford, Clarendon press, 193p.

Corsten, L.C.A., and A. Stein. Are nested sampling schemes recommendable for spatial semivariogram estimation? *Submitted to J. Amer. Stat. Ass.*

Dietrich, C.R., and M.R. Osborne. 1991. Estimation of covariance parameters in kriging via restricted maximum likelihood. *Math. Geol.* 23, 119-135.

Engel, B. 1991. The analysis of unbalanced linear models with variance components. *Statistica Neerlandica*, 44, 195-221.

Gower, J.C. 1962. Variance component estimation for unbalanced hierarchical classifications. *Biometrics*, 18, 537-542.

Journel, A.G. and C.J. Huijbregts. 1978. *Mining geostatistics*. Academic Press, London, 600p.

Kitanidis, P.K. 1983. Statistical estimation of polynomial generalized covariance functions and hydrologic applications. *Water Res. Res.* 19, 909-921.

Lehmann, E.L. 1983. *Theory of point estimation*. Wiley, New York, 506p.

Miesch, A.T. 1975. Variograms and variance components in geochemistry and ore evaluation. *Geological Society of America, Memoir* 142, 333 - 340.

Oliver, M.A., and R. Webster. 1986. Combining nested and linear sampling for determining the scale and form of spatial variation of regionalized variables. *Geographical analysis*, 18 (3), 227-242.

Patterson, H.D., and Thompson, R. 1974. Maximum likelihood estimation of components of variance, in Corsten, L.C.A., and Postelnicu, T. (Eds.) *Proceedings of the 8th International Biometric Conference, Constanta*, 197-207.

Pettitt, A.N., and A.B. McBratney. *In press*. Surveys for estimating spatial variance components. *Applied Statistics*.

Stein, A. 1991. *Spatial interpolation*. PhD-thesis, Wageningen, 236p.

Van Dongen, R.P.M., and Widianto. *In prep*. The application of nested analysis of variance for quantified farming systems analysis in East Java.

Webster, R. 1985. Quantitative spatial analysis of soil in the field. In: B.A. Stewart (ed.), *Advances in soil science 3*. Springer Verlag, New York, pp. 1-70.

APPENDIX - Expected mean squares

Expected mean squares are commonly applied to estimate variance components in ANOVA procedures. Turning to relevant expected sums of squares define P_i to be the orthogonal projection on the orthogonal complement of A_{i+1} in A_i ($i = 0,\ldots,h-1$). Then $P_i\underline{y}$ is the difference vector between the projection of $\underline{y}$ on A_i and its projection on A_{i+1}, representing variation between the classes of A_i within those of A_{i+1}.

In order to calculate expected sums of squares we notice that $E(P_i\underline{y}) = 0$ and hence that the required expected sum of squares is $E[(P_i\underline{y})'(P_i\underline{y})] = E[\underline{y}'P_i\underline{y}] = \mathrm{tr}(P_iV)$.

To evaluate $\mathrm{tr}(P_iV) = \sum_{j=0}^{h-1} \mathrm{tr}(P_iU_jU_j')\sigma_j^2$, it is noted first that $P_iU_j = 0$ for $j > i$, since each column of U_j does not alter by projection on A_i or A_{i+1}. Next, observing that P_i may be written as $U_i(U_i'U_i)^{-1}U_i' - U_{i+1}(U_{i+1}'U_{i+1})^{-1}U_{i+1}'$ we find by standard methods that

$$\mathrm{tr}(P_iV) = \sum_{j=0}^{i} \mathrm{tr}(P_iU_jU_j')\sigma_j^2$$

$$= \sum_{j=0}^{i} \{\sigma_j^2 \sum n_j^2 \left(\frac{1}{n_i} - \frac{1}{n_{i+1}}\right)\}, \qquad \text{(A1)}$$

where the second summation extends to all classes of A_j. For example, the coefficient of σ_0^2 in $\mathrm{tr}(P_iV)$ is $\Sigma n_0^2(1/n_i - 1/n_{i+1})$, which equals the number of classes of A_i minus that of A_{i+1}, that is the dimension of P_iR^n, the divisor for obtaining mean squares from sums of squares.

With $k_{ij} = \Sigma n_j^2(1/n_i - 1/n_{i+1})$ for $i \geq j \geq 0$, summation extending to all classes of A_j, we have the following expected sums of squares equivalent to equation (3) before:

$$\begin{aligned} &\text{within } A_1 && E(\underline{y}'P_0\underline{y}) = k_{00}\sigma_0^2 \\ &\text{between } A_i \text{ within } A_{i+1} && E(\underline{y}'P_i\underline{y}) = \sum_{j=0}^{i} k_{ij}\sigma_j^2 \\ &\text{between } A_{h-1} && E(\underline{y}'P_{h-1}\underline{y}) = \sum_{j=0}^{h-1} k_{h-1,j}\sigma_j^2. \end{aligned} \qquad \text{(A2)}$$

Equating $E(\underline{y}'P_i\underline{y})/k_{i0}$ to $MS_i = y'P_iy/k_{i0}$, one can solve the linear equations emerging from (A2) to obtain unbiased estimators $\hat{\sigma}_i^2$ of σ_i^2, which by summing according to (A2) yield unbiased estimators $\tilde{\gamma}_i$ of h semivariogram values. After some algebraic manipulations, one can deduce with the following successive operators

$$\begin{aligned} Q_0 &= \frac{1}{k_{00}} P_0 \\ Q_i &= \frac{1}{k_{ii}} P_i + \sum_{j=0}^{i-1} \frac{k_{i,j+1} - k_{ij}}{k_{ii}} Q_j \end{aligned} \qquad \text{(A3)}$$

that

$$\tilde{\gamma}_{i+1} = y'Q_iy \qquad \text{(A4)}$$

for i=0,...,h-1.

From (A3) and (A4) it follows that the estimators $\tilde{\gamma}_i$ are non-negative, although the variance components σ_j^2 have a positive probability of attaining negative values. Indeed, we notice that Q_i is a linear combination of orthogonal projections with positive coefficients, due to the inequality $k_{ij} < k_{i,j+1}$ for all i; this is based on the fact that the sum of squares of a set of positive numbers is smaller than the square of the sum of the same set.

Based on (A4) an expression for the variance of $\tilde{\gamma}_i$ is obtained. It is well known that $var(\underline{y}'A\underline{y}) = 2tr(AVAV)$ for any symmetric A if $\underline{y}$ follows a Gaussian distribution [Searle, 1987]. Hence, under normality, $var(\tilde{\gamma}_i) = 2tr(Q_{i-1}VQ_{i-1}V)$.