Cokriging Point Data on Moisture Deficit

A. STEIN,* W. VAN DOOREMOLEN, J. BOUMA, AND A. K. BREGT

ABSTRACT

Existing computer calculations by simulation of the moisture deficit (MD) were used for 500 point observations in an area of 404 ha with sandy soils in the Netherlands. The statistical prediction techniques kriging and cokriging were used to predict MD-values in 100 points selected at random from the 500 available points. The MD-data at the remaining 400 points were used for the predictions. The mean variance of prediction error (MVPE) and the mean squared error of prediction (MSEP) decreased only slightly when kriging is compared with cokriging using the mean highest water-table level as a covariable. The number of MD-values used in cokriging could, however, be reduced from 400 to 160 with only a small loss of accuracy in using 400 observations on the mean highest water-table (MHW) as a covariable. As the MD-variables being considered are four times as expensive to determine as the covariable, which is routinely estimated during soil survey, this also represents a considerable reduction of costs. Cokriging can thus be a useful technique to more effectively utilize available soil survey information.

COMPUTER SIMULATION techniques for predicting the soil moisture deficit are being used in the Netherlands to estimate yield reductions due to lowering of the water table by water extraction for municipal water supply (e.g. Bouma et al., 1980a,b). Soil maps are used to estimate basic hydrological properties such as hydraulic conductivity, moisture retention and water table fluctuations (e.g. Wösten et al., 1985). Simulations are carried out for so-called representative soil profiles for each mapping unit. This is a clear disadvantage, because heterogeneity within units is thus ignored. An alternative, more attractive procedure is to produce simulations for each separate soil boring, followed by predictions for unvisited spots by means of the prediction technique of kriging and averaging to obtain a predictive map (e.g. Bregt et al., 1987; De Wit and Van Keulen, 1987). Costs of borings, data gathering and simulation calculations are, however, relatively high. Procedures have to be developed, therefore, to cut costs to the point that maximum information is provided at minimum cost.

The objective of this study was to explore use of available soil survey information for development of such procedures. A potentially attractive procedure is the statistical prediction technique of cokriging (McBratney and Webster, 1983; Yates and Warrick, 1987), which not only uses observations of a particular variable, such as the moisture deficit, but also observations of covariables such as soil or water table characteristics that are routinely assembled during soil survey. Cokriging could conceivably result in a considerable reduction of costs in achieving a comparable degree of accuracy by using fewer relatively expensive variables and more relatively inexpensive co-variables.

A. Stein, Dep. of Soil Statistics, W. van Dooremolen and J. Bouma, Dep. of Soil Science, Agricultural Univ. Box 37, 6700 AA Wageningen, Netherlands; A.K. Bregt, Dep. of Applied Information Science and Statistics, the Netherlands Soil Survey Inst., Wageningen, Netherlands. Received 21 Dec. 1987. *Corresponding author.

Published in Soil Sci. Soc. Am. J. 52:1418–1423 (1988).

MATERIALS AND METHODS

Study Area and Procedures

In 1985 a detailed soil survey was carried out in the Mander area in the eastern part of the Netherlands by the Dutch Soil Survey Institute (Stiboka) to study the effect of groundwater extraction on the production of grassland (Stoffelsen and Van Holst, 1985; Wösten et al., 1987). The Mander area consists of sandy soils, which are classified as Haplaquods, Humaquepts and Plaggepts (Wösten et al., 1987). Fluctuations of the water table are routinely characterized by Dutch soil survey in terms of mean highest (MHW) and mean lowest (MLW) levels (Van der Sluijs and De Gruijter, 1985; Fig. 1a). They are classified as indicated in Table 1. Lowering of water tables results in change in productions, e.g. of grassland due to lower fluxes of water from the water table to the root zone. Farmers have to be financially compensated for these production losses by municipal water companies. Computer simulation techniques by the simulation model LAMOS are used to obtain the necessary quantitative expressions for the moisture deficit and the associated yields, as has been discussed elsewhere (Bouma et al., 1980a,b). Simulation calculations were made for a 30-yr period. For this study attention was arbitrarily focused on two results, viz. the 30-yr avg. value for the moisture deficit for grassland (MD30) and the value for the dry year 1976 (MD76) (Fig. 1b, c). In the area of 404 ha being studied, 500 soil borings were made. This observation density is generally assumed to be representative for surveys of scale 1:10 000. Hydraulic conductivity and moisture retention data for major soil horizons were used as they were obtained from a previous study (Wösten et al., 1985, 1987). Simulations were carried out for each boring location. To evaluate the predictions by kriging and cokriging, 100 borings were selected at random from the data set to serve as a test set. Predictions were made for the test set with the calculated values for the moisture deficit of the remaining 400 borings (see statistical procedures). Attention was focused on the average precision obtained for the 100 test locations. Concerning the scale of maps a rule of thumb for soil survey recommends that approximately four observations per cm^2 of map area are to be taken. This rule was questioned, however, by investigating the possibility of taking fewer observations. More observations would not be feasible for economic reasons. Data sets with reduced numbers of observations were generated by random numbers corresponding to four map scales (Table 2). Finally an analysis of costs was made to allow a financial comparison between kriging and cokriging.

Statistical Procedures

In carrying out the statistical analyses the statistical prediction techniques kriging and cokriging have been used. Cokriging, closely linked to kriging (Matheron, 1973), allows the use of a second variable (the covariable) in predicting values of the variable of interest (the predictand). The new approach being used here, forms a special version of a more general approach dealing with polynomial trends up to any order. The method is here considered to be a statistical prediction technique (Kendall and Stuart, 1973) rather than a numerical interpolation technique (Vauclin et al., 1983). The

Table 1. Groundwater classes and mean highest and mean lowest groundwater levels.

Class	Groundwater level, cm	
	Mean highest	Mean lowest
3	<40	80-120
5	<40	>120
6	40-80	>120
7	>80	>120

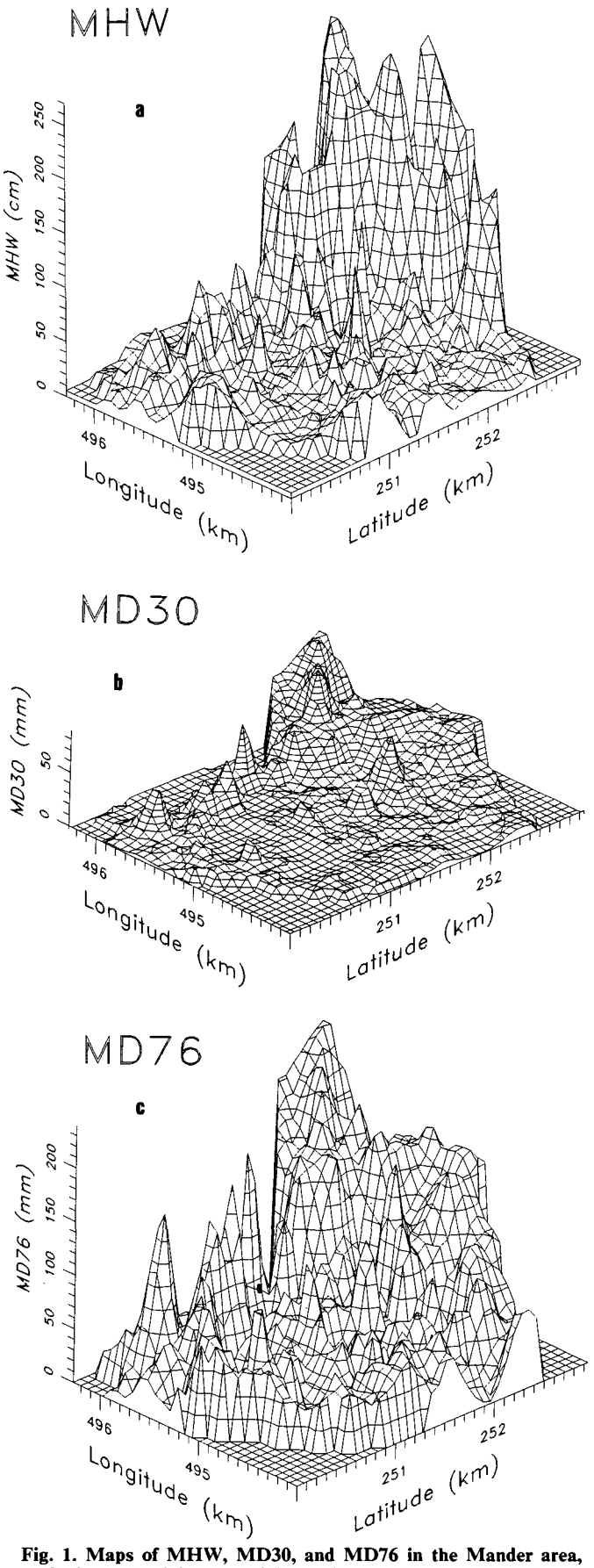

Fig. 1. Maps of MHW, MD30, and MD76 in the Mander area, obtained by kriging the original soil survey data.

Table 2. Correspondence between observation density, the approximate mapping scale, and the number of observations in the data sets.

Observ. density, obs./km^2	99	44	25	16	11
Approx. map scale	10 000	15 000	20 000	25 000	30 000
No. of observ.	399	177	100	64	44

technique has been developed in the context of the Generalized Linear Model with dependent observations (Rao, 1973), and can be described as follows. Throughout the text we use boldface type to denote that a variable is stochastic, in order to distinguish it from nonstochastic variables. A predictor t is used in a number of points, e.g. the nodes of a grid. Let one such point be denoted by z_0 and let a prediction t of the value of y_0 in z_0 be based on n observations of the predictand in the observation points $z_1, \ldots, z_n$ and m observations of the covariable in the observation points $z_{n+1}, \ldots, z_{n+m}$. Let X be a matrix, $m + n$ by 2, x_0 a two-vector and y a stochastic $m + n$-vector, composed of an n-vector $\mathbf{y}^1$ and an m-vector $\mathbf{y}^2$ pointing to the predictand and the co-variable, respectively, such that

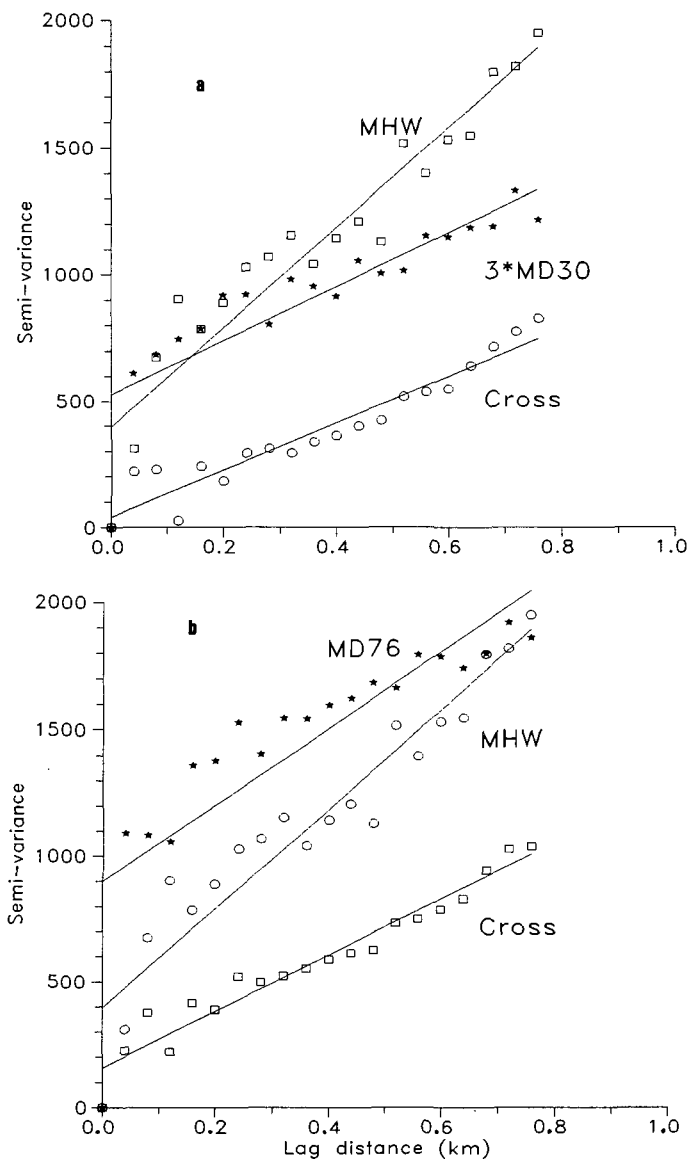

Fig. 2. Semivariograms for MD30, MD76, and MHW, and their cross variograms with MHW. The MD30 values are multiplied by 3 to obtain comparable pictures.

$$X = \begin{bmatrix} 1_n 0 \\ 0\ 1_m \end{bmatrix}; \; x_0' = (1\ 0); \; \mathbf{y} = \begin{bmatrix} \mathbf{y}^1 \\ \mathbf{y}^2 \end{bmatrix};$$

1_n is a vector consisting of n ones and 1_m is a vector consisting of m ones. The model can be formulated as

$$\mathbf{y} = x\beta + \mathbf{e},$$

with $E(\mathbf{e}) = 0$, and $\text{var}(\mathbf{e}) = C$ and unknown parameter vector β. We have to predict $\mathbf{y}_0 = x_0'\beta + \mathbf{e}_0$ with $\text{var}(\mathbf{e}_0) = c_{00}$. Under the assumption of stationarity and finite variance the elements of C are given by the covariance functions $c_1(h)$ and $c_2(h)$ for the predictand and the covariable, respectively, and the cross-covariance function $c_{12}(h)$ between both of them, all three depending on the distance h between observation points. The value for c_{00} is given by $c_1(0)$. A generalized least squares estimator for β is $\hat{\beta}$, defined as $\hat{\beta} = (X'C^{-1}X)^{-1}X'C^{-1}\mathbf{y}$.

In general, as in this study, the demand of finite variance is too strong. Use can be made of semivariograms $g_1(h)$ and $g_2(h)$ for the predictand and the covariable and of the cross-variogram $g_{12}(h)$ for the spatial interaction between predictand and covariable, respectively, depending on the distance h between observation points. These variograms can be estimated by

$$g_k(h) = \Sigma(Y_i^k - Y_{i+h}^k)^2/2N(h) \text{ for } k = 1,2; \quad [1]$$

$$g_{12}(h) = \Sigma(Y_i^1 - Y_{i+h}^1)(Y_i^2 - Y_{i+h}^2)/2N'(h); \quad [2]$$

Summation in Eq. [1] is for every h taken over all $N(h)$ pairs of observation points of the predictand and the covariable, respectively, being a distance h apart, and in Eq. [2] for every h over all $N'(h)$ pairs of points for which observations of the two variables are available being a distance h apart. Let G be the $n + m$ by $n + m$ matrix with element $g_{ij} = -g(h_{ij})$, being the opposite of the value of the covariance between observation points z_i and z_j and let g_0 be the $(n + m)$ vector with element g_{0i} being the opposite of the value of the covariance between z_0 and the ith observation point. The nugget effect is denoted by g_{00}. We then obtain as a stochastic predictor

$$\mathbf{t} = x_0'\hat{\beta} + g_0'G^{-1}(\mathbf{y} - X\hat{\beta}). \quad [3]$$

As a measure for the precision of a prediction the variance of the prediction error is used, being equal to

$$\text{var}(\mathbf{t} - \mathbf{y}_0) = g_{00} - g_0'G^{-1}g_0 + x_a'(X'G^{-1}X)^{-1}x_a \quad [4]$$

where x_a is defined as $x_a = x_0 - g_0'G^{-1}X$.

Since the value of Eq. [4] is zero in an observation point (both kriging and cokriging are exact predictors), a test set of 100 points was randomly selected from the original set of 500 calculated data. For the test set predictions are carried out on the basis of MD-data in a neighborhood of the eight nearest points of the remaining 400 points. The variances of the prediction errors, Eq. [4], were averaged over the test set, yielding the mean variance of prediction error (MVPE). Also, predictions in the test locations were afterwards compared with the corresponding observations, giving a second measure of the performance of the predictor, the mean of squared errors of prediction (MSEP). The two measures are equal to

$$\text{MSEP} = (1/n) * \sum_{i=1} (t_i - y_i)^2 \quad [5]$$

and

$$\text{MVPE} = (1/n) * \sum_{i=1}^{n} \text{var}(\mathbf{t}_i - \mathbf{y}_{0i}), \quad [6]$$

respectively. In both Eq. [5] and [6] summation is carried out over the test set, in Eq. [5] expression [4] is used, whereas

in Eq. [6] the prediction is compared with the observed value. Although these measures are fundamentally different in that the MVPE is stochastic and the MSEP is not, they can both be used to obtain an indication of the quality of predictions. The relation between the variance of the prediction error and the squared error of prediction is given by

$$E(\mathbf{t} - \mathbf{y}_0)^2 = \text{var}(\mathbf{t} - \mathbf{y}_0) + \{E(\mathbf{t} - \mathbf{y}_0)\}^2. \quad [7]$$

The term $E(\mathbf{t} - \mathbf{y}_0)^2$ is estimated by Eq. [5], in other words, the MSEP is an estimator of the mean value of the second moment. Because $\{E(\mathbf{t} - \mathbf{y}_0)\}^2$ is equal to zero, the prediction model is considered appropriate, if the estimator for $E(\mathbf{t} - \mathbf{y}_0)^2$ is about equal to the estimator for $\text{var}(\mathbf{t} - \mathbf{y}_0)$, that is if the MSEP and the MVPE are about equal. In practice, of course, differences may exist between the MVPE and the MSEP, due to deficiencies in the prediction model, i.e. the bias of the model. The MVPE is always smaller than the MSEP.

By means of the relative precision (rp) different prediction techniques are compared with kriging based on 400 observation points. It is defined as

$$\text{rp} = \text{MVPE}^0/\text{MVPE}^1 \quad [8]$$

where MVPE^0 is the MVPE of kriging with 400 observation points and MVPE^1 is the MVPE of the compared prediction technique. A value of the r.p. >1, for instance, points to a prediction technique that is more precise than kriging with 400 observation points.

Table 3. Coefficients of weighted linear regression models for variograms, $\gamma(h) = a + b*h$ (see also Fig. 2).

	a	b
MD30, mm	74	89
MD76, mm	1030	1560
MHW, cm	495	1780
MD30-MHW	12	295
MD76-MHW	171	1090

Table 4. Correlation coefficients.

		MD30	MD76	MHW	MHL
Moisture deficit					
30-yr avg.	MD30	1.000			
Dry year 1976	MD76	0.934	1.000		
Groundwater level					
Mean highest	MHW	0.587	0.636	1.000	
Mean lowest	MLW	0.551	0.600	0.915	1.000

Table 5.1a. MVPE of MD76 for different observation densities obtained by means of reference variograms.

	Predictand				
Covariable	99	44	25	16	11
99	171	215	269	360	521
44		225	288	384	556
25			294	397	578
16				413	603
11					615

Table 5.2a. MVPE of MD30 for different observation densities obtained by means of reference variograms.

	Predictand				
Covariable	99	44	25	16	11
99	12.4	15.6	19.6	26.1	38.1
44		16.4	21.0	28.0	40.5
25			21.4	28.9	42.1
16				30.0	43.8
11					44.7

Two approaches are followed for the production of the test set. First, it can be assumed that no data are available in the test-set points, except their coordinates, which is the case when a map has to be produced by means of predictions in grid nodes. Second, it can be assumed that in each of the test-set points an observation of the covariable is available. This distinction will be denoted by cokriging 1 and cokriging 2, respectively. To select the most promising covariables, correlation coefficients and crossvariograms were determined between calculated moisture deficit and different soil survey data, such as actual mean highest and mean lowest water table, rooting depth, loam fraction from top- and subsoil, amount of organic matter and clay fraction of the topsoil.

RESULTS

Comparing Kriging with Cokriging

Calculations for the MD30 and the MD76, as made for the 400 boring locations in the study area, were on the average 10.3 and 57.7 mm, while the individual standard deviations were 14.5 and 50.7 mm, respectively. These values could be used for estimation in any unvisited point in the area, if spatial dependency would be absent. Better estimates are, however, obtained when the spatial dependence is taken into account as is inherent to the kriging and the cokriging procedures.

Semivariograms were calculated for the variables being considered. A linear model without sill fits sufficiently well (Fig. 2a, 2b). Fitting was carried out by means of weighted linear regression, weights being based on the number of pairs of points in the distance classes (Table 3). For cokriging the mean highest watertable before extraction (MHW) was used as a covariable as it showed the highest correlation with MD30 and MD76 (Table 4) and a clear spatial dependence. The variables MD30 and MD76 and the covariable MHW are known in every sample location. To simulate differences in scale, the observation density for MD30 and MD76 of the original data set was randomly reduced, such that the number of observations per unit map sheet be equal on every scale. The MSEP and the MVPE values for kriging and cokriging 1 and cokriging 2 were compared for observation densities 11 obs/km^2 for MD30 and MD76 and 99 obs./km^2 for MHW. Semivariograms were calculated for variables and covariables for all data points, as well as for the

Table 5.1b. MVPE of MD76 for different observation densities obtained by means of scale dependent variograms.

	Predictand				
Covariable	99	44	25	16	11
99	171	242	299	455	539
44		254	325	502	600
25			335	533	654
16				560	699
11					713

Table 5.2b. MVPE of MD30 for different observation densities obtained by means of scale dependent variograms.

	Predictand				
Covariable	99	44	25	16	11
99	12.4	18.0	19.4	33.9	37.9
44		18.9	21.2	36.8	41.8
25			22.0	38.7	45.3
16				40.4	48.1
11					49.1

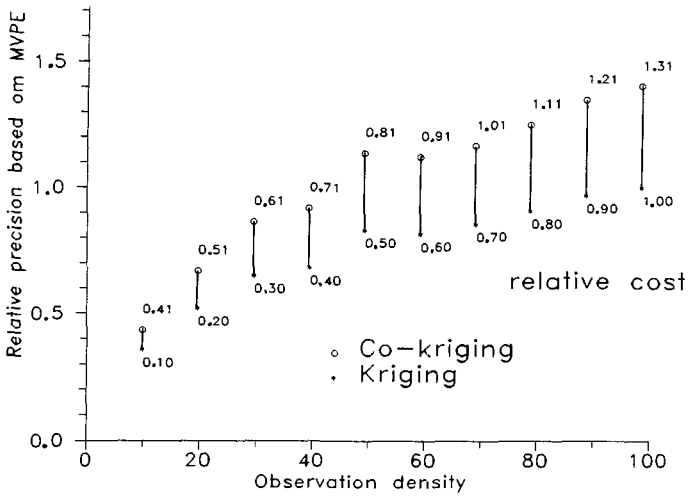

Fig. 3. The relative precision based on the MVPE obtained by kriging and cokriging 1 for different observation densities.

four sets of reduced numbers of observations. Of particular interest for soil survey is the case of having few observations of the predictand and many of the covariable. The results for the different observation densities are summarized in Tables 5.1a to 5.2b and in Fig. 3. In calculating semivariogram values a minimum of 30 pairs of observations in every distance class is needed. This restricted the number of possible combinations to be tested (Fig. 4a,b). Only semivariograms for MD30 (Fig. 4a) and cross variograms for MD30 and MHW (Fig. 4b) are shown; comparable figures were obtained for MD76.

Results can be summarized as follows:

1. According to the MVPE and the MSEP values obtained, cokriging 2, i.e. cokriging with an observation of the covariable in the prediction point, leads to more precise predictions than cokriging 1. A prediction obtained by cokriging is more precise than a prediction obtained by kriging with the same observations of the predictand. MVPE values indicate that the standard deviations for MD30 and MD76 are 4- to 20-mm lower for cokriging than for kriging (Table 6).
2. There is a small loss in precision by reducing the number of observations in the predictand up to 40% of the original number. Further reductions result in a larger increase (Fig. 5).
3. A change in the observation density of the covariable had less severe implications for the precision than a change in the observation density of the predictand. Reducing the number of observations by 89% from 399 to 44, for instance, resulted for the predictand in an increase of the MVPE of 400%, and reducing the number of observations of the covariable in an increase of the MVPE of 130%. Incorporation of a covariable reduced the increase of the MVPE from 400 to 300%.
4. Use can be made of semivariograms based on the reduced numbers of observations without seriously affecting the quality of the predictions.

Table 6. Comparison of cokriging and kriging for 11 obs. per km^2 for the predictand and 99 obs. per km^2 for the covariable.

		Kriging	Cokriging 1	Cokriging 2
MSEP	MD30	126	111	102
	MD76	2460	2230	1970
MVPE	MD30	44.7	38.1	31.7
	MD76	615	521	433

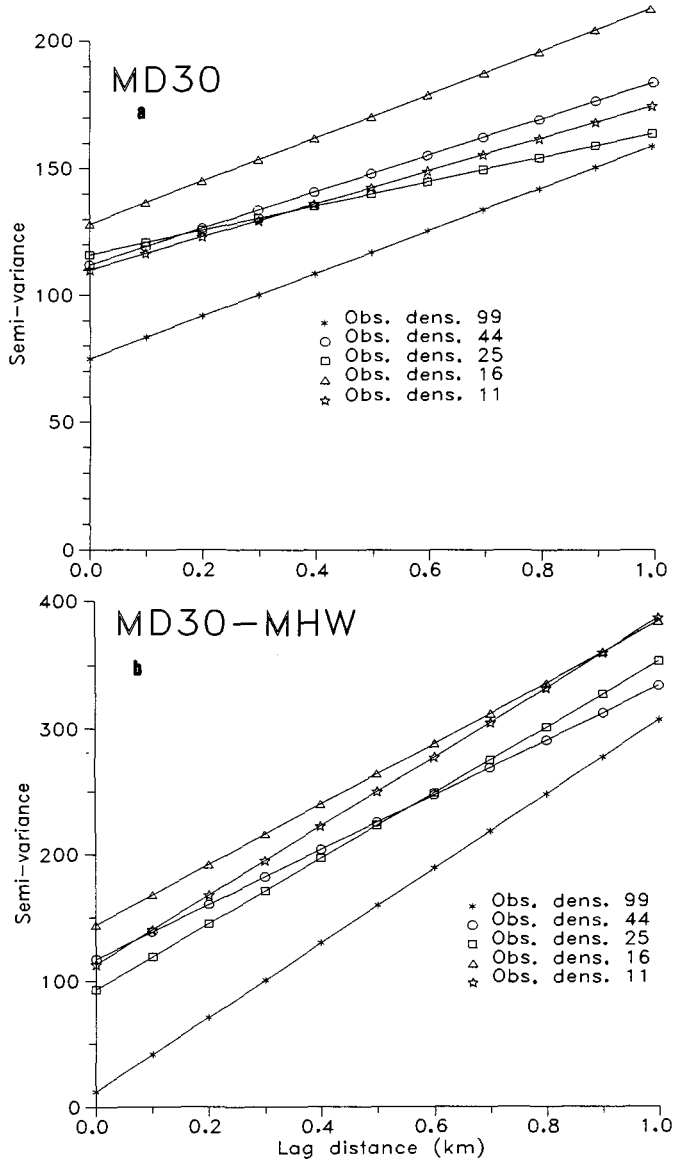

Fig. 4. Fitted models for semivariograms for MD30 and cross variograms for MD30 with MHW for different observation densities.

The Associated Costs

The cost for observations and calculations of the predictand is approximately four times as high as the cost for observations of the covariable. On the basis of the relation between relative cost of a survey and relative precision based on MVPE, is graphically presented in Fig. 5. As can clearly be seen from, for example, the 99 observations per km^2 standard situation, the introduction of observations of a covariable enhances the relative precision (by approximately 40%) and the relative cost (by approximately 30%). Both reduction of costs (by only 10%) and increase of pre-

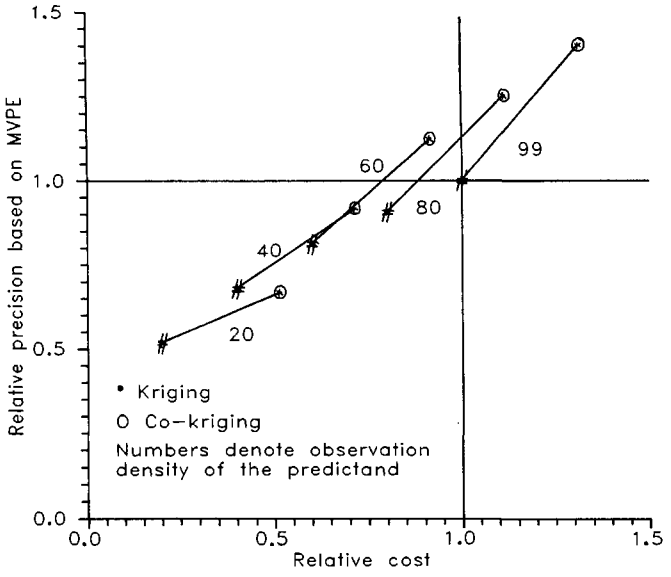

Fig. 5. Relations between kriging and cokriging 1 in terms of the relative precision (based on the MVPE) and the relative cost.

cision (also by 10%) can, for example, be achieved in the situation of a density of 60 observations per km^2 for the predictand and the original density of 99 observations per km^2 for the covariable. Reduction of observation density from 100 to 20 observations per km^2, reduces relative costs to 0.2 and relative precision to 0.5. Inclusion of cokriging reduces relative costs to 0.5 and relative precision to 0.65. Even larger reductions in costs with increase of precision are obtained when the cost ratio between observations of the predictand and the covariable would be > 4:1.

DISCUSSION

This study was focused on one important land quality, the moisture deficit, and used only one covariable. The principle being illustrated is, however, valid for other land qualities and covariables still to be selected. Studies are needed to further explore the practical feasibility of using cokriging.

This study was based on the use of variograms. Some considerations should therefore be kept in mind when carrying out cokriging. The necessary number of data to estimate variograms should be more than 30 to 50 per distance class (Journel and Huijbregts, 1978, p. 194). In order to obtain reliable estimates, the number of paired observations (i.e. observations of the predictand and the covariable in one location) must be relatively high (at least some 40 to 60).

Availability of observations of the covariable in a point where a prediction is to be carried out leads to a reduction of the variance of the prediction error by 30%, instead of 15% if no such observation is available and only kriging is applied. Therefore, use of an observation of a covariable, which is highly correlated with a land quality being studied, enhances the precision of the prediction.

There are substantial financial benefits when using cokriging. Sixty percent of observations of expensive variables could be replaced in this study by less expensive variables without serious loss of precision. Observations on these less expensive variables are to be made in the very locations where a prediction is needed. Up to now it is not clear how to select the most promising covariables. Selection should not only be based upon the correlation coefficient. The spatial relation between the predictand and potential covariables should be taken into account as well. It could be useful to use multivariate techniques with the purpose to combine two or more variables into one new covariable. Studies on this approach are in progress.

In Geographical Information Systems, use of cokriging appears to be attractive, as the performance of a GIS is governed by quality of input data. Kriging offers the opportunity to produce predictive single value maps while cokriging allows incorporation of different soil parameters that are usually available in soil data bases, so as to enhance the precision of output. Computing time will increase when using cokriging, but the availability of high capacity hardware and skillful programming may overcome these problems. Also, in combining two sources of information in terms of variables and covariables, cokriging explicitly uses their spatial relations. Maps of variances of the prediction error can be provided, showing the location of the more or less reliable parts of maps of the predictands.

ACKNOWLEDGMENT

We kindly acknowledge the assistance of G. Stoffelsen from the Netherlands Soil Survey Institute for collecting the data, L.C.A. Corsten, and J.J. de Gruijter for their helpful comments, and M. Hoogerwerf for providing part of the figures.

REFERENCES

Bouma, J., P.J.M. de Laat, R.H.C.M. Awater, H.C. van Heesen, A.F. van Holst, and Th. J. van de Nes. 1980a. Use of soil survey data in a model for simulating regional soil moisture regimes. Soil Sci. Soc. Am. J. 44:808–814.

Bouma, J., P.J.M. de Laat, A.F. van Holst, and Th. J. van de Nes. 1980b. Predicting the effects of changing water-table levels and associated soil moisture regimes for soil survey interpretations. Soil Sci. Soc. Am. J. 44:797–802.

Bregt, A.K., J. Bouma, and M. Jellinek. 1987. Comparison of thematic maps derived from a soil map and from kriging of point data. Geoderma 39:281–291.

de Wit, C.T., and H. van Keulen. 1987. Modelling production of field crops and its requirements. Geoderma 40:253–265.

Journel, A.G., and Chr. J. Huijbregts. 1978. Mining geostatistics. Academic Press, London.

Kendall, M.G., and A. Stuart. 1973. The advanced theory of statistics. Vol. 2. 3rd ed. Griffin, London.

Matheron, G. 1973. The intrinsic random functions and their applications. Adv. Appl. Prob. 5:439–468.

McBratney, A.B., and R. Webster. 1983. Optimal interpolation and isarithmic mapping of soil properties. V. Coregionalization and multiple sampling strategy. J. Soil Sci. 34:137–162.

Rao, C.R. 1973. Linear statistical inference and its applications. John Wiley & Sons, New York.

Stoffelsen, G.H., and A.F. van Holst. 1985. Soil scientific and hydrologic investigations in the Mander watershed area. Internal report. (In Dutch). Soil Survey Inst. Stiboka, Wageningen, Netherlands.

Van der Sluis, P., and J.J. de Gruijter. 1985. Water table classes: A method used to indicate seasonal fluctuation and duration of water tables on Dutch soil maps. Agric. Water Manage. 10(2):109–125.

Vauclin, M., S.R. Vieira, G. Vachaud, and D.R. Nielsen. 1983. The use of cokriging with limited field soil observations. Soil Sci. Soc. Am. J. 47:175–184.

Wösten, J.H.M., J. Bouma, and G.H. Stoffelsen. 1985. Use of soil survey data for regional soil water simulation models. Soil Sci. Soc. Am. J. 49:1238–1244.

Wösten, J.H.M., M.H. Bannink, and J. Bouma. 1987. Land evaluation at different scales: You pay for what you get. Soil Surv. Land Eval. 7:13–24.

Yates, S.R., and A.W. Warrick. 1987. Estimating soil water content using cokriging. Soil Sci. Soc. Am. J. 51:23–30.