WORKING PAPER

KRIGING AND OTHER ESTIMATORS OF SPATIAL FIELD CHARACTERISTICS (WITH SPECIAL REFERENCE TO ENVIRONMENTAL STUDIES).

V. Fedorov

October 1987
WP-87-99

International Institute
for Applied Systems Analysis
A-2361 Laxenburg/Austria

KRIGING AND OTHER ESTIMATORS OF SPATIAL FIELD CHARACTERISTICS (WITH SPECIAL REFERENCE TO ENVIRONMENTAL STUDIES).

V. Fedorov

October 1987
WP-87-99

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
A-2361 Laxenburg, Austria

Foreword

During the last two years, Valeri Fedorov has been bringing his very considerable statistical talents to bear on the design of environmental monitoring systems and on the analysis of experimental data in the environmental sciences. In this particular Working Paper, Valeri has examined the concept of ***kriging***, a method used to recover spatial patterns from point measurements, first applied in the geological sciences. Unfortunately, environmental fields usually vary in time as well as space. This leads to severe difficulties which have not been fully realized by environmental scientists. The nature of these difficulties is elaborated in this paper.

R.E. Munn
Leader
Environmental Program

KRIGING AND OTHER ESTIMATORS OF SPATIAL FIELD CHARACTERISTICS (WITH SPECIAL REFERENCE TO ENVIRONMENTAL STUDIES).

V. Fedorov

1. Introduction

During the last decade use of the kriging method (or simply "kriging") increased greatly in research related to the analysis of spatial variability of environmental parameters; see for instance, Barnes (1980), Dennis and Seilkop (1986), Clark et al. (1986), Endlich et al. (1986), Finkelstein and Seilkop (1981), Der Megreditchan (1979) and McBratney and Webster (1981). It is difficult to understand the reason for this increasing popularity (but with some occasional dark spots; see for instance, Akima (1974), Armstrong (1984)). Is it its comparative simplicity, statistical effectiveness, mathematical elegancy or the mesmerizing impact of the word "kriging" (sometimes "universal kriging")? In this paper an attempt to clarify the situation and to understand its place in the statistical theory of spatial pattern analysis and the admissibility of the kriging method in environmental studies is undertaken.

Technical details and computing aspects of the kriging method are avoided but the reader can find them for instance, in Journal and Huijbregts (1978), Ripley (1983) and Thiebaux and Peddler (1987). The bibliographic overview by Bell and Reeves (1979) provides a guide to the original theoretical literature (mainly by Krige, Matheron and Journel) and the numerous applied papers.

It is worthwhile to note that the method is the analogue for spatial processes of Wiener-Kolmogorov prediction theory. It "has been developed and used mainly by Matheron and his school in the mining industry. He christened the method kriging after D.G. Krige" (Ripley (1983) section 4.4).

II. Linear Estimators, Kriging

Let some value $y(x)$ be observed at points $x_1, x_2, \ldots, x_n \in X$, where X is two dimensional in the majority of applications. Points x_i, $i = \overline{1,n}$, could be spaced at the intersection points of some regular grid but it is not essential for the theory and does not cause serious difficulties for modern software if they are not.

The estimation of value y at the prescribed point x_0 is of interest in the analysis of spatial patterns. For the purpose of this paper, it is sufficient to consider only the linear estimator of $y_0 = y(x_0)$:

$$\hat{y}_0 = \lambda^T y \ , \tag{1}$$

where "T" stand for transposition, $\lambda \in R^n$, $y^T = (y_1, y_2, \ldots, y_n)$, $y_i = y(x_i)$.

A practitioner wishes to have $\hat{y}_0$ as close to its true value y_0 as possible. In other words, one has to minimize some measure of discrepancy between $\hat{y}_0$ and y_0 :

$$\hat{y}_0 = Arg \min_{\lambda}\{\text{discrepancy } (\hat{y}_0, y_0)\}. \tag{2}$$

The choice of the "distance" is defined by the structures of $y(x)$. Frequently some constraints can be imposed on λ in optimization problem (2).

Two main approaches can be easily traced in the corresponding literature: the first is deterministic and closely related to classical function approximation theory, while the second is based on the assumption that $y(x)$ is generated by a random field (see for instance Katkovnik (1985), Miccheli and Wahba (1981) and Ripley (1983)).

In both cases, the most crucial assumptions are related to the "smoothness" of surface $y(x)$. In the deterministic case, these assumptions usually concern the derivatives of $y(x)$. When $y(x)$ is generated by a random field, assumptions concerning its correlation structure are mainly used.

This paper concerns the latter case. The square risk (discrepancy measure)

$$v^2(\lambda) = E[(\hat{y}_0 - y_0)^2] = \int (\lambda^T y - y_0)^2 \, p(y_0, y) dy_0 dy, \tag{3}$$

where $p(y_0, y)$ is the density (assuming its existence) of the random vector $(y_0, y) \in R^{n+1}$, will be used as a criterion of the optimality of $\hat{y}_0$. It is known that the linear estimator (1)-(3) is the best one only if $y(x)$ is a Gaussian random field. In other cases, it is referred to as *the best linear estimator*.

Observations with known mean-covariance structure.

Let us assume that:

(a) for any x and x' the mean $m(x) = E[y(x)]$ and the covariance $C(x, x') = E[(y(x) - m(x))(y(x') - m(x')]$ are known. In matrix notation, this means particularly that

$$\begin{gathered} m_0 = E(y_0), \quad m = E(y), \\ C_{00} = E[(y_0 - m_0)^2], \; C_{01}^T = C_{10} = E[(y - m)(y_0 - m_0)], \\ C_{11} = E[(y - m)(y - m)^T]. \end{gathered} \tag{4}$$

are given.

From definition (3) and (4), it follows that:

$$v^2(\lambda) = C_{00} + m_0^2 - 2\lambda^T (C_{10} + m_0 m) + \lambda^T (C_{11} + m m^T)\lambda. \tag{5}$$

The minimization of (5) leads to *the best linear (unconstrained) estimator:*

$$\hat{\lambda}_1 = (C_{11} + m m^T)^{-1} (C_{10} + m m_0), \tag{6}$$

$$v^2(\hat{\lambda}_1) = \min_{\lambda} v^2(\lambda) = C_{00} + m_0^2 - (C_{01} + m^T m_0)(C_{11} + m m^T)^{-1} (C_{10} + m^T m_0). \tag{7}$$

When linear constraints are imposed:

$$F\lambda = L, \tag{8}$$

where F is a $(k \times n)$-matrix and L is a $(k \times 1)$-vector, then the solution of λ_2 of (5) is defined by the system:

$$\begin{bmatrix} C_{11} + m m^T & F^T \\ F & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} C_{01} + m m_0 \\ L \end{bmatrix}, \tag{9}$$

where μ corresponds to Lagrangian multiples.

The variance of the prognosis is equal to:

$$v^2(\hat{\lambda}_2) = v^2(\hat{\lambda}_1) + \Psi^T[F^T(C_{11}+mm^T)^{-1}F]^{-1}\Psi\,, \tag{10}$$

where $\Psi = F^T(C_{11}+mm^T)^{-1}(C_{01}+m_0 m)-L$. The last term in (10) is positive due to positive definiteness of the matrix $F^T(C_{11}+mm^T)^{-1}F$.

In other words, when demanding the fulfillment of some properties for $\hat{\lambda}$ (the fulfillment of (8)), one sacrifices its precision in the sense of (3):

$$v^2(\hat{\lambda}_1) \leq v^2(\hat{\lambda}_2)\,. \tag{11}$$

Unknown mean structure, kriging.

Formulae (6), (7) are of theoretical interest but for a practitioner their usefulness is very restricted. Knowledge of both $m(x)$ and $C(x,x')$ is more often an exception than a frequently met situation. Therefore, a number of attempts to construct estimators that do not use $m(x)$ and $C(x,x')$ or at least part of this information have repeatedly been undertaken. One of them is ***kriging***, where knowledge of $m(x)$ is not necessary. Unfortunately one has to pay for this by the additional assumption:

(b) the mean of $y(x)$ does not depend upon x, e.g., $m(x) \equiv m_0$, and by the following constraint imposed on λ:

$$\sum_{i=1}^{k} \lambda_i = 1\,. \tag{12}$$

In terms of (8), this means that $F = l^T$ is the vector $(1\times k)$ with all elements equal to 1 and $L = m_0$. Constraint (12) provides an unbiasness of the corresponding estimator $\lambda^T y$:

$$E[\lambda^T y] = m_0\,,\ \sum_{i=1}^{n} \lambda = m_0\,.$$

From (9) and (12) it follows that the ***kriging estimator*** $\hat{\lambda}_k$ is the solution to the following linear system

$$\begin{bmatrix} C_{11} & l \\ l^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} C_{01} \\ m_0 \end{bmatrix}, \tag{13}$$

which does not include vector m.

The restricted nature of assumption (b) was evident at the very beginning of kriging history and the so-called ***universal kriging*** estimator $\hat{\lambda}_{uk}$ was proposed (Huijbregts and Matheron (1971)), which was based on the assumption:

(b') the mean of $y(x)$ can be presented in the form $m(x) = \vartheta^T f(x)$, where $f(x)$ is a vector $(k\times 1)$, $k < n$, of ***a priori*** known functions. To obtain the unbiased estimator, one has to impose the following constraint:

$$E[\lambda^T y] = \lambda^T F^T \vartheta = f^T(x)\vartheta\,, \quad \text{for any } x \in X\,,$$

or

$$F\lambda = f(x)\,, \tag{14}$$

where

$$F = [f(x_1), f(x_2), \ldots, f(x_n)] .$$

From (9) and (14) is follows that the ***universal kriging estimator*** $\hat{\lambda}_{uk}$ is the solution of the system:

$$\begin{bmatrix} C_{11} & F^T \\ F & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} C_{01} \\ f(x) \end{bmatrix}, \tag{15}$$

which similarly to (13) does not include vector m.

The estimator $\hat{\lambda}_{uk}$ is unbiased if (b ′) is fulfilled but $v^2(\hat{\lambda}) \leq v^2(\hat{\lambda}_{uk})$.

For the universal kriging estimator, (10) can be transformed to

$$v^2(\hat{\lambda}_{uk}) = C_{00} - C_{01}C_{11}^{-1}C_{10} + \\ + [f(x_0) - FC_{11}^{-1}C_{10}]^T (FC_{11}^{-1}F^T)^{-1} [f(x_0) - FC_{11}^{-1}C_{10}] \tag{16}$$

Kriging is more attractive theoretically when (a) is changed into the following assumption (Huijbuegts and Matheron (1971)):

(a') $E[y(x) - y(x')] = \Delta(x,x'), E[(y(x) - y(x'))^2] = 2\gamma(x,x')$,

where it is assumed that only the ***increments*** of the random field admit the first and second order moments. From theory, it is known that there are cases when these moments do not exist for $y(x)$ itself while they exist for the corresponding increments. Function $\gamma(x,x')$ is called a ***semivariogram*** and

$$2\gamma(x,x') = [m(x) - m(x')]^2 + C(x,x) + C(x',x') - 2C(x',x') ,$$

if mean $m(x)$, variance $C(x,x)$ and covariance $C(x,x')$ exist. Usually in practice one ***believes*** in their existence.

In applied studies using kriging, authors prefer to work with the semivariogram $\gamma(x,x')$ instead of the familiar covariance $C(x,x')$ although it does not make acceptance of the results any easier (see the next section).

To summarize this subsection, one can say that kriging is ***a particular case of linear estimation theory*** or, to be more specific, some generalization of the Wiener-Kolmogorov filter.

Probably it is also worthwhile to notice that a universal kriging estimator can be constructed in the framework of the generalized least square method (see section IV) when the model

$$y(x) = \vartheta^T f(x) + \varepsilon(x)$$

is considered under the assumption that the covariance structure of the random error $\varepsilon(x)$ is known.

The simple kriging estimator can be considered also as a particular case of the ***moving average*** (see, for instance, Katkovnik (1985)):

$$\hat{y}_a = \sum_{i=1}^{n} \lambda_i y_i \ , \ \sum_{i=1}^{n} \lambda_i = 1 \quad ,$$

where weights λ_i are defined by (3).

Observations with unknown mean-covariance structure

Assumption (a) (or the theoretically slightly milder assumption (a')) is drastically restrictive in practice for environmental analysis. In the publications related to kriging, the author could not find any approach other than using the empirical estimates $\hat{m}(x)$ and $\hat{C}(x,x')$ in place of $m(x)$ and $C(x,x')$.

Probably this recommendation is worthwhile in engineering sciences where there exists the possibility of repeating similar experiments and where one can use so-called learning samples to restore $m(x)$ and $C(x,x')$ and then to use kriging. But in this case the simpler and well established technique can be used. For instance, instead of (1), (8) one can use the estimator:

$$\hat{y} = \lambda_0 + \lambda^T y$$

that in case of (3) is defined by the following formulae:

$$\hat{\lambda}_{03} = m_0 - m^T C_{11}^{-1} C_{10}\ ,\ \hat{\lambda}_3 = C_{11} C_{10}\ ,$$
$$\hat{y}_0 = m_0 + C_{01} C_{11}^{-1} (y - m)\ , \qquad (17)$$

and

$$v^2(\hat{\lambda}_{03}, \hat{\lambda}_3) = C_{00} - C_{01} C_{11}^{-1} C_{10} \le v^2(\hat{\lambda}_1) \le v^2(\hat{\lambda}_2)\ .$$

If the learning sample is sufficiently large and one can forget the uncertainties of $\hat{m}(x)$ and $\hat{C}(x,x')$ then (17) is better than kriging.

In environmental applications and by the way, in geostatistics (the homeland of the kriging method), it is more realistic to use the term *observations* than *experiments*. Therefore, the problem of a good learning sample becomes here especially acute and usually there is no hope that the errors of estimates $\hat{m}(x)$ and $\hat{C}(x,x')$ can be neglected.

In this case, the average in formulae (7) and (10) (where all *a priori* unknown values are substituted by their estimates) has the conditional character:

$$E[(\hat{y} - y)^2 / \text{ learning sample}] = v^2(\hat{\lambda})$$

Of course the total variance, taking into account the randomness of learning sample, will be greater than $v^2(\hat{\lambda})$ defined either by (7) or (10). To some extent they can be considered as *estimates of the low bounds for the corresponding total variances*.

The plight becomes worse when the same set of data $y^T = (y_1, y_2, \ldots, y_n)$ is used both for the estimation of C_{01}, C_{11} and kriging. In that case, the author has failed to find in the literature any serious theoretical results on the evaluation of the bias and the variance of the kriging estimators.

A sensitivity analysis examining how the estimate $\hat{y}_{uk} = \hat{\lambda}_{uk}^T y$ will change for given small perturbations in C_{11} and C_{10} was done by Warnes (1986). But it is a numerical analysis of the stability of system (15), rather than a statistical analysis of the corresponding estimator. which, for instance, must determine how much the variance of $\hat{y}_{uk}$ will change under small random perturbations of the aforementioned matrices.

To conclude this section, one can say that in the case of *a priori* unknown C_{01} and C_{11}, *the kriging approach gives some hints on how to choose weights for the linear estimator* $\lambda^T y$ *but it gives no serious guarantee of their optimality*.

III. Examples

In all three examples presented here, attention will be focussed on the most sensitive aspect of kriging : fitting of $\gamma(x,x')$ or $C(x,x')$, leaving aside other aspects of the problem.

1. *Distribution of residual contamination from atmospheric nuclear tests* (Barnes (1980)).

This example illustrates how the semivariogram was estimated using data on isotope ^{241}Am for one ground zero site, Smallboy. The data was extracted from a more extensive study related to the analysis of residual contamination after atmospheric nuclear tests. The final objective was to evaluate the contours of ^{241}Am activity. All experimental and geographical details can be found in (Barnes (1980)) and in the references given there. In the cited report it was assumed that the observed value $y(x)$ is a random spatially weak stationary field, e.g., $E[y(x)] \equiv m$ and $\gamma(x,x') = \gamma(h)$, where $h = \sqrt{(x-x')^T(x-x')}$

This assumption is essential because firstly it allows one to estimate $\gamma(x,x')$ and secondly it makes it possible to use simple kriging ($m(x) \equiv m$, see assumption (b) from the previous section). The consistency of the assumption with reality can be evaluated, at least partly, with the help of Figure 1. Here, the "observed" values $\tilde{\gamma}(h)$ are plotted by numbers which correspond to the directions presented at the top-right corner of Figure 1.

The "observations" were calculated by the following procedure (Barnes (1980)): "All points within a small angle of true east-west of a given point are put in the "east-west" class and points that are "approximately" h away of a given point are put in the distance h class. The size of small angle and the closeness of the distance approximation can be controlled by the user to reduce the error introduced by these approximations".

The dashed line in Figure 1 corresponds to the final approximation of $\gamma(h)$ and it is clear that both anisotropy and possible drift were ignored. Barnes did not mention how $\hat{\gamma}(h)$ was fitted to the data, but what is evident is that it does not follow the points at all, except maybe in the interval $0 \leq h \leq 1000ft$.

From the definition of the semivariogram, it follows that in the case when the drift of $m(x)$ can be ignored and $y(x)$ is weakly spatially stationary, then $\gamma(h) = C' - C(h)$, where $C \equiv C(x,x)$ and $C(h) = C(x,x')$.

It seems that for the physical phenomenon (distribution of contamination) considered by Barnes, $C(h) \geq 0$ for all h and $\lim_{h \to \infty} C(h) = 0$ (implicitly this is assumed, see p.6 of cited paper, where various types of $\gamma(h)$ are considered). Therefore, for the upper bound of $\gamma(h)$, one has

$$\sup_h \gamma(h) = C,$$

e.g., the plateau of $\gamma(h)$ (or "sill") has to be equal to the variance C of observed values. From Figure 1 it follows that $C \simeq 3400$. Unfortunately, the scale of the vertical axis was not accurately defined. Does it correspond to $\gamma(h)$ or to $2\gamma(h)$? So it could be that $C \simeq 1700$, but this would not change the situation, i.e., where one cannot be sure that the inaccuracy of $\hat{\gamma}(h)$ defined by the dashed line in Figure 1 is not less than 30%. How can it violate the results of the kriging procedure? Is there any hope for its optimality? Or maybe it is a very indirect way to construct a mediocre moving average estimator with very restrictive intermediate assumptions?

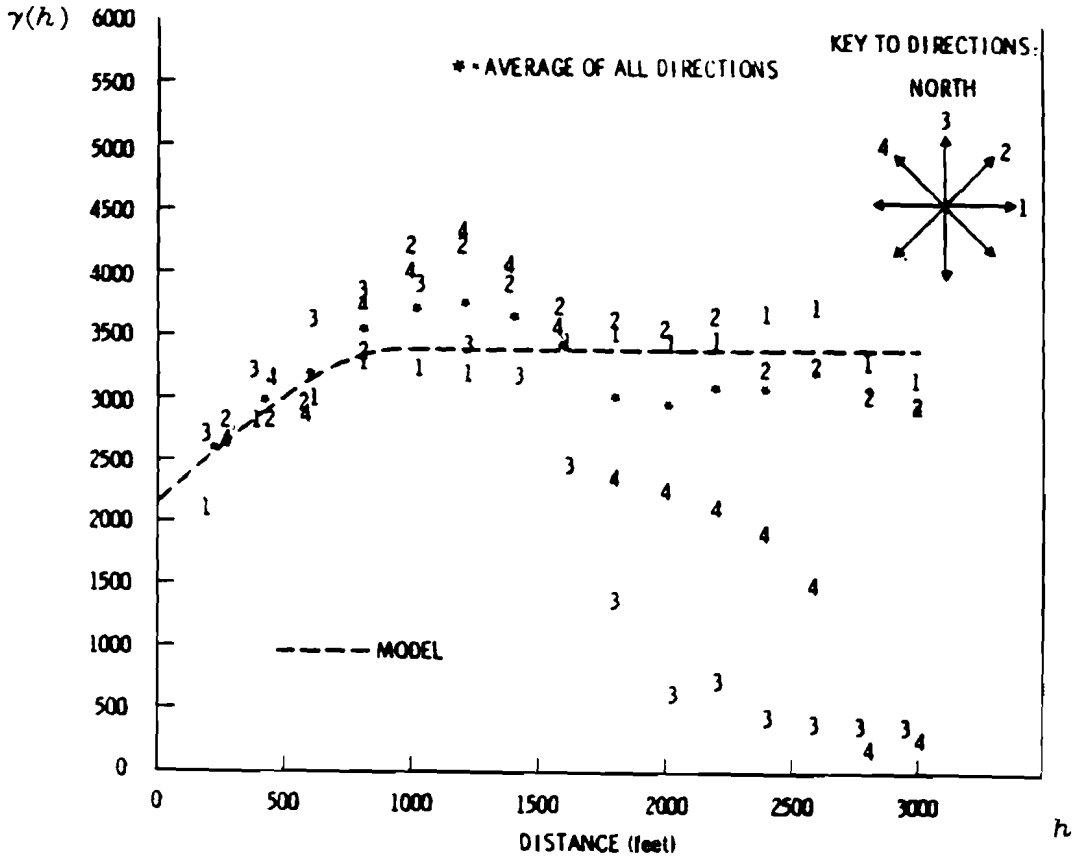

Figure 1: Pointwise estimates for the semivariogram $\gamma(h)$, Barnes (1980).

Probably it is better to use the moving average directly because it needs simpler computing and more straightforward and explicit assumptions about the nature of $y(x)$.

2. Kriging analysis of the regional patterns of the chemical constituents of precipitation.

In the report on "Statistical analysis of precipitation chemistry measurements" by Endlich et al, (1986), Ch. 4, the kriging approach was applied to interpolate the yearly medians of the daily laboratory pH and analyte concentrations at each site and for the yearly total deposition of H^+ and other analytes. To perform kriging the authors separated "long-term spatial and temporal trends from year-to-year fluctuations". The quadratic spatial polynomial plus the linear time trend were used to approximate the logarithms of these trends. The fitting technique was the ordinary least square method. The residuals (observed values minus the least square estimates) were taken as inputs for kriging. Their variations from year-to-year were *assumed to be independent,* and the covariation structure was *"assumed to be both stationary (independent of location) and isotropic (independent of direction)"*

The semivariograms were evaluated by a linear model:

$$\gamma(h) = \vartheta_1 + \vartheta_2 h \ .$$

Fitting was done by the weighted least square method with weights proportional to the number of site-pairs and inversely proportional to distance h. Typical examples of this fitting are presented in Figure 2. It is difficult to evaluate the goodness of fitting because the inverse values of weights were not plotted. However, it seems that the fit is not very good. Besides the linear approximation does not reflect the fact that the observations from the mutually remote sites have not to be correlated and $\lim_{h \to \infty} \gamma(h) = const$, i.e., $\gamma(h)$, has to approach saturation (compare with the previous example and with the definition of a semivariogram).

There are some other points that can be criticized:

(a) Use of the least square method to remove trends and subsequent application to kriging is not consistent with theory: one has to use at least universal kriging instead of these two steps. Only in this case are the estimates optimal in the sense of (3) or at least approximately optimal if the estimates for $C(h)$ and subsequently for C_{1x} and C_{11} are sufficiently precise. It seems that the use of covariances (if they exist) is more convenient both from the theoretical and the computational points of view. In what follows, $C(h)$ or $C(x,x')$, or C_{1x}, C_{11} will be used without comments.

With respect to residuals u_i, the authors of the report did not notice that their variance-covariance matrix is singular. Therefore all "kriging" formulae (see, for instance, (9), (15), (16)) ***cannot be used directly*** (the solution will not be unique). Probably, the use of some estimate $\hat{C}_{11}$ instead of C_{11} will regularize computations. But this "regularization" simultaneously means the loss of optimality of the prognoses $\hat{y}(x)$ due to the poor estimation of C_{11}.

The singularity of C_{11} can be easily verified.

Assume that $y = F^T\vartheta + \varepsilon$ (or $y_i = f^T(x_i)\vartheta + \varepsilon_i$), where $F = (f(x_1), \ldots, f(x_n))$. Then the least square estimator $\bar{\vartheta}$ is defined by the formula (see, for instance, Anderson, 1971):

$$\bar{\vartheta} = (FF^T)^{-1}Fy \quad ,$$

the vector of residuals equals

$$u = y - F^T\bar{\vartheta} = [I - F^T(FF^T)^{-1}F]\varepsilon$$

and the variance-covariance matrix of the residuals

$$\bar{c}_{11} = [I - F^T(FF^T)^{-1}F]C_{11}[I - F^T(FF^T)^{-1}F] \ .$$

The projection matrix $I - F^T(FF^T)^{-1}F$ has rank $(n - m)$, where m is the dimension of ϑ. Therefore rank $C_{11} = n - m$, i.e. C_{11} is singular ($|C_{11}| = 0$).

(b) The method of semivariogram estimation can be improved (from the statistical point of view) almost without an increase of computations.

For the sake of simplicity let us assume that observed values (i.e., ordinates in Figure 2) are normally distributed. Then (see, for instance, Seber, 1977, Ch. 14):

$$\text{Var}\,[(y - y')^2] = 2\gamma^2(h) \ ,$$

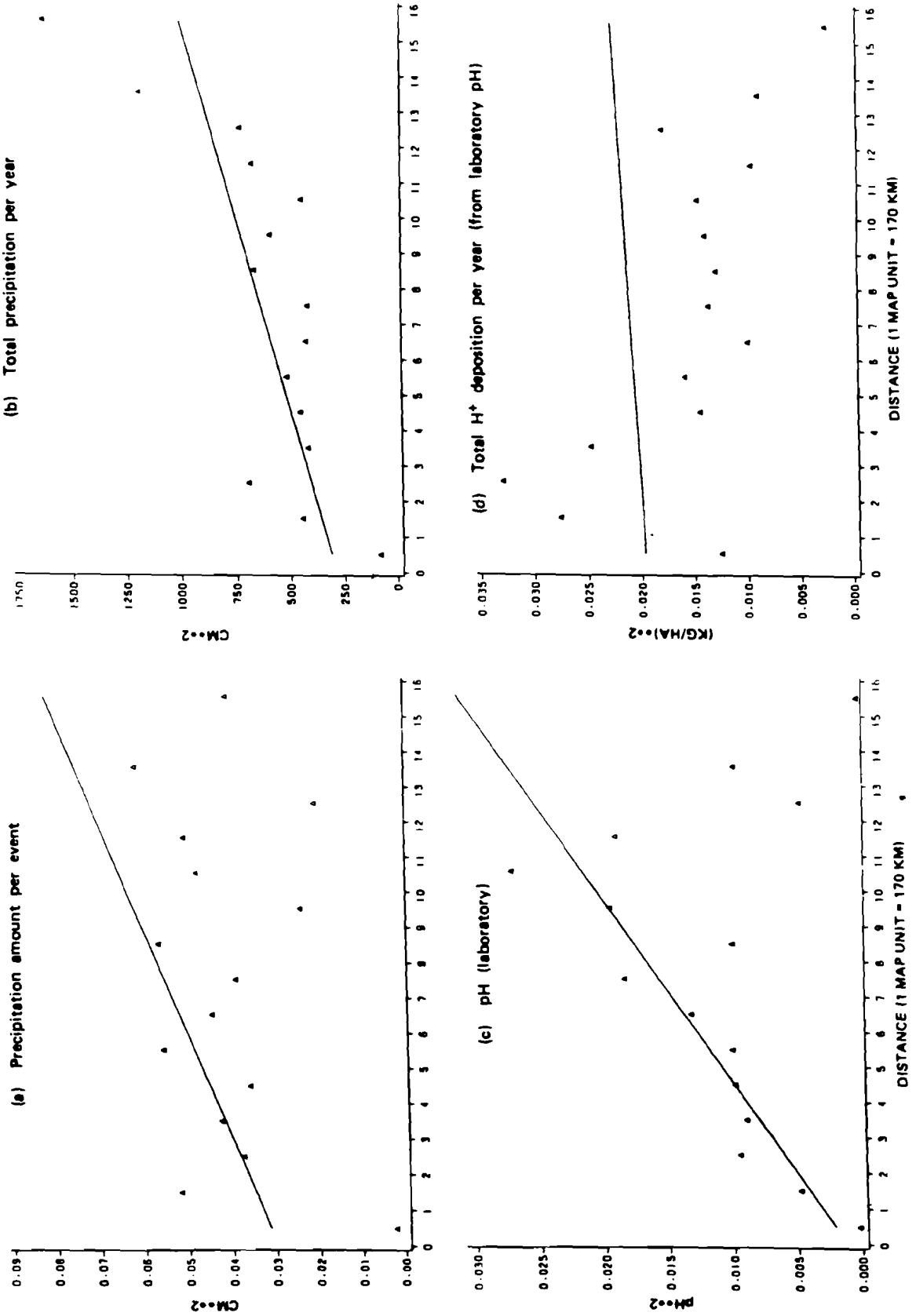

Figure 2: Empirical and fitted semivariogram functions, Endlich et al (1986).

where h is the distance between observation points. In the case of other distributions, this formula is more complicated.

The method of the iterative least squares:

$$\vartheta_{s+1} = Arg \min_{\vartheta} \sum_{i=1}^{n} r_i [\gamma_i - \gamma(h_i, \vartheta)]^2 / \gamma^2(h_i, \vartheta_s) , \; \hat{\vartheta} = \lim_{s \to \infty} \vartheta_s , \qquad (18)$$

gives asymptotically ($n \to \infty$) optimal estimates for ϑ and $\gamma(h, \vartheta)$, i.e., the method minimizes var $(\hat{\vartheta})$ and var $\gamma[(h, \hat{\vartheta})]$, see Malyutov (1982). In (18) r_i is the number of observations for every h_i , ϑ stands for unknown parameters. In practice usually the iterative procedure (18) is terminated after 3,4 steps.

(c) For all four fitted lines in Figure 2, the intercept (i.e., ϑ_1) is significantly greater 0. It means that there exists a so-called ***nugget*** effect (see, for instance, Gilbert and Simpson (1984)), i.e., a discontinuity of the covariance function $C(x, x')$. This could probably occur in geostatistics when one analyses the deposition of some ore minerals, but in the analysis of pollutants in fluids or atmospheric contamination, it seems unreal. This is also confirmed by the existence in each part of Figure 2 of observation points located close to the origin. Presumably, only observations with $h \leq 3 \div 4$ still satisfy the kriging assumptions and more distant observations are either violated by trends and anisotropy, or the covariance function is negative for $h \geq 4$. The results presented in Figure 3 (the model $\gamma(h) = (\vartheta_1 z + \vartheta_2 z^{2)} / (1 + \vartheta_3 z^2)$, $z = e^{-h} - 1$, was fitted to the data with the help of iterative use of BMDP AR program, 1983) are a good confirmation of this assertion.

The above critique leads to the same conclusion as in the previous example.

3. Sulfur deposition model evaluation.

One of the main goals of the report by Clark et al (1986), was the comparison of the several models currently used for computing of sulfur deposition in North America. Roughly, the strategy consists of the following steps:

- observations from irregularly located observation stations are used to estimate values at regular grid points with the help of kriging,
- model predictions over some grid are used to obtain gridpoint values with the help of kriging,
- both sets of results are compared by procedures mainly based on methods of mathematical statistics.

In this approach, for instance, "a confidence interval" for the difference in observed and predicted values was constructed as follows:

$$y_{pred} - y_{obs} \pm 1.96 \, [Var(y_{pred}) + Var(y_{obs})]^{1/2} , \qquad (19)$$

where

y_{pred} = kriging estimate for model prediction at the grid point,

y_{obs} = kriging estimate for the observed data at the grid point,

Var $(\cdot)$ = kriging variance estimate.

If the deviation of kriged values from true values that they estimate have normal (Gaussian) distribution, then (19) defines 95% confidence interval. See Clark et al (1986), p. 5.12.

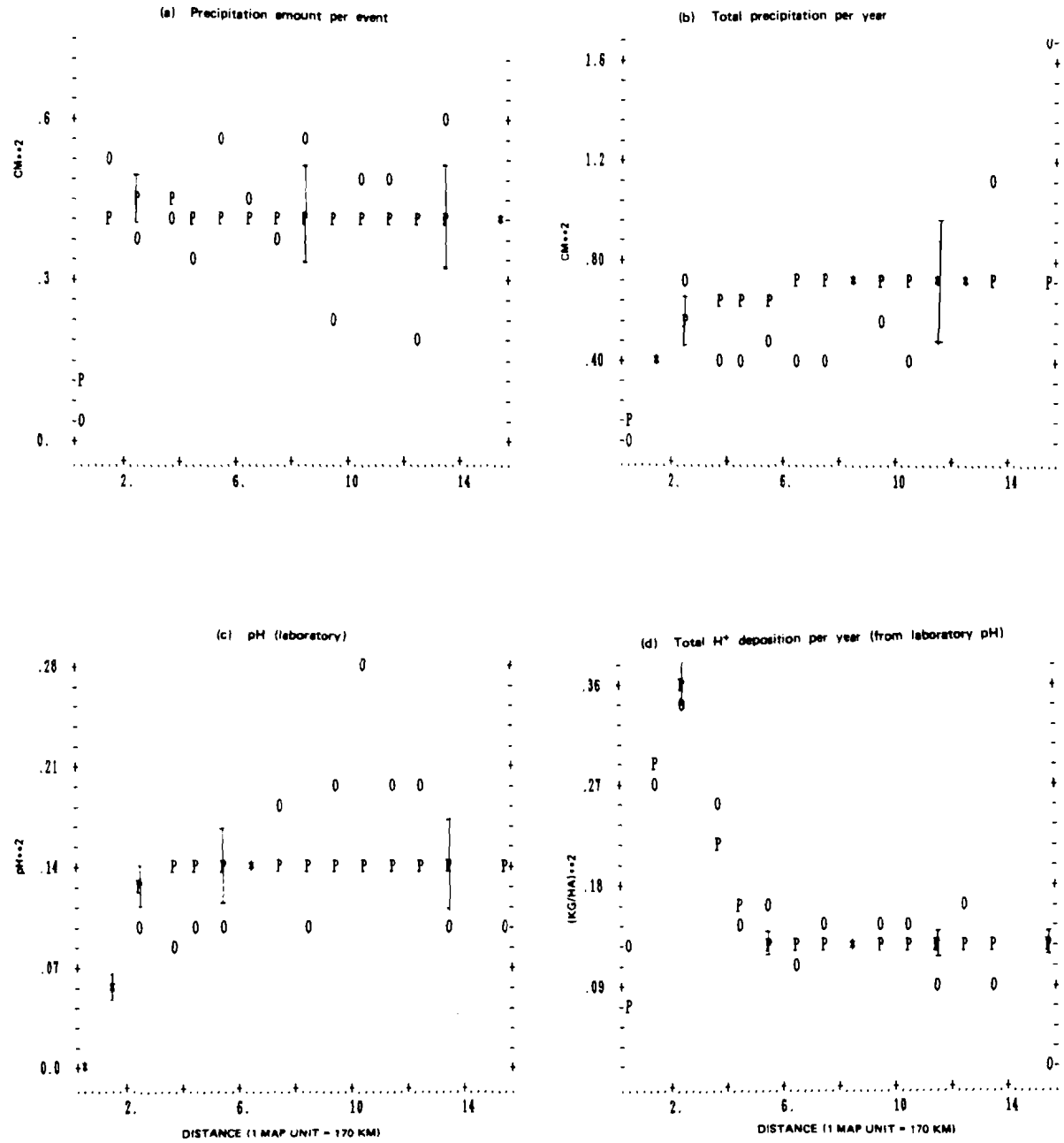

Figure 3: Fitted semivariogram $\gamma(h) = (\vartheta_1 z + \vartheta_2 z^2)/(1+\vartheta_3 z^2), z = e^h - 1$, vertical lines stand for the standard error of prognoses.

From the previous considerations (see the concluding part of section II and the first two examples) it is clear that the variances computed by substitution of true values of $C(x)$ and $C((x,x'))$ (or $\gamma(x,x')$) by their estimates (which are usually very poor) have a very remote relation with reality. Therefore, one can use (19), keeping in mind a number of "ifs" due to the violation of assumptions about "pure kriging" and due to the assumption on the normality of distributions of y_{pred} and y_{obs}.

Together with this technical remark there is one more general comment. The strategy of model comparison with data can be described by the scheme presented in Figure 4. It is clear from this scheme that ***one compares distorted images of two objects***. In principle the distortion due to projection I can be very small because a user can continue computing until his models give good results on any given grid. Then the whole procedure of comparison can be described as comparison of the given set of models with one very simple model which is defined by the kriging method assumption (see Section II). Implicitly the authors believe that this model is better than any other considered in their report.

Probably the approach schematically presented in Figure 4 would be more fruitful.

IV. Some Alternatives to Kriging

Generalized least squares (g.l.s.) estimator. In this subsection the links between the kriging approach and the old-fashion least squares estimators will be illuminated.

Let similarly to (b') from section II

$$y_i = f^T(x_i)\vartheta + \varepsilon_i \ , \ i = 1,n \quad ,$$

where $f(x)$ is the vector of known basic functions,

$$E[\varepsilon_i] = 0 \ , \quad E[\varepsilon_i \varepsilon_{i'}] = C_{11,ii'} \ , \quad \text{or in the matrix presentation}$$

$$y = F^T\vartheta + \varepsilon \ , \ E[\varepsilon\varepsilon^T] = C_{11} \quad . \tag{21}$$

To estimate the value of the response $y(x)$ at a given point x, one can follow the two steps procedure:

- compute the parameters estimates $\hat{\vartheta}$,
- predict $y(x)$ using the linear estimator

$$\hat{y}(x) = f^T(x)\hat{\vartheta} + \lambda_0^T u \quad ,$$

where u is the vector of residuals, i.e. $u = y - F^T\vartheta$.

Vector λ_0 is defined as a solution of the following optimization problem

$$\lambda_0 = Arg \min_{\lambda} E[y(x) - f^T(x)\hat{\vartheta} - \lambda^T(y - F^T\hat{\vartheta})]^2 \quad ,$$

or

$$\lambda_0 = Arg \min_{\lambda} E[u(x) - \lambda^T u]^2 \quad . \tag{22}$$

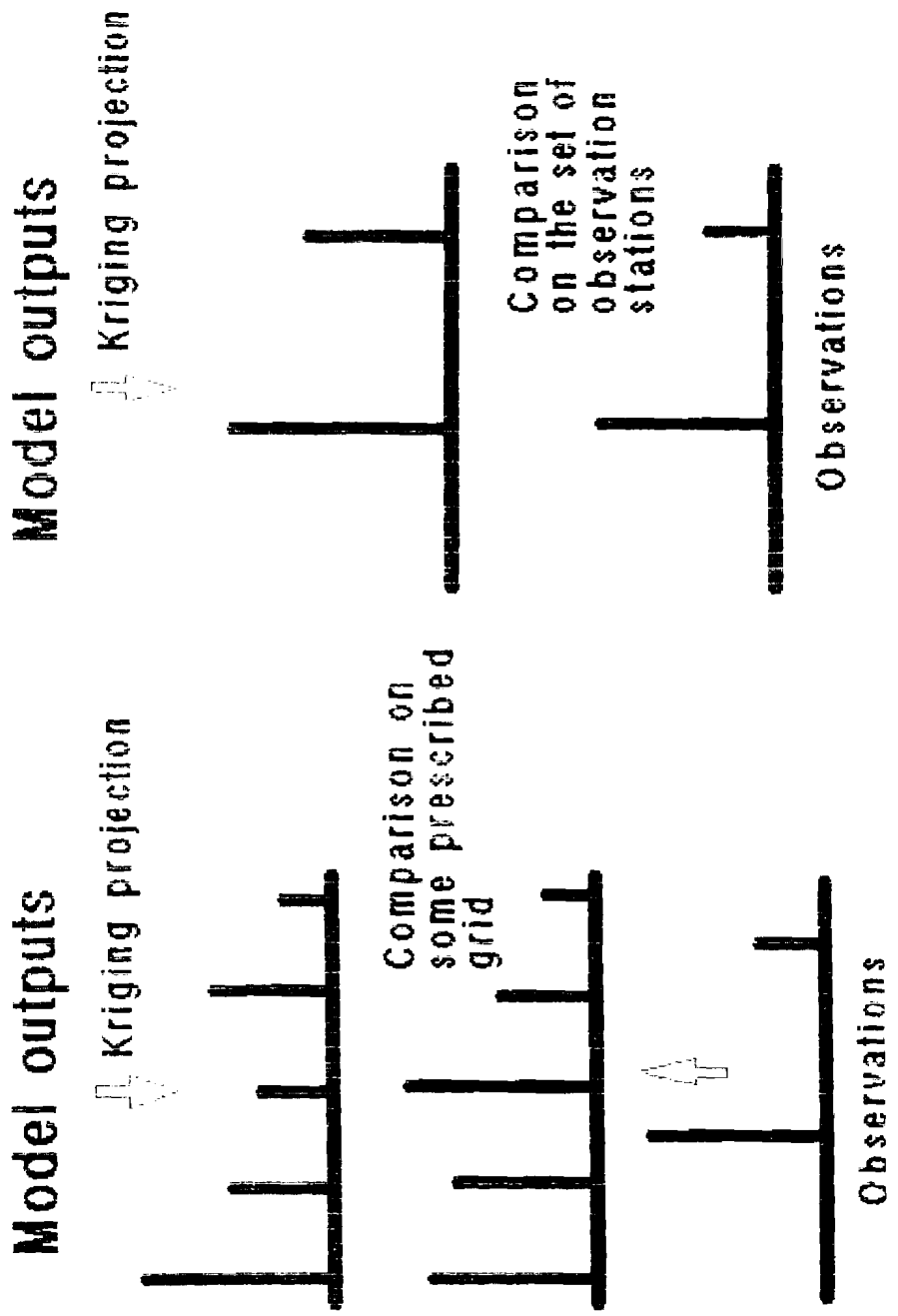

Figure 4: Comparison of model outputs with data.

It is known (see, for instance, Seber, 1977) that the generalized least square estimator (which is the best linear unbiased estimator)

$$\hat{\vartheta} = Arg \min_{\vartheta} (y - F^T\vartheta)^T C_{11}^{-1} (y - F^T\vartheta)$$

can be calculated implicitly:

$$\hat{\vartheta} = M^{-1} F C_{11}^{-1} y \ , \ M = F C_{11}^{-1} F^T \quad . \tag{23}$$

Formally (22) coincides with (3) and at first glance all considerations from section II can be applied to (23). Nevertheless, there exist some specific features of (22):

- It is not necessary to impose any constraints on λ to get an unbiased estimator, because $E[u(x)] = 0$ and $E[u] = 0$, due to the unbiased nature of $\hat{\vartheta}$, $E[\hat{\vartheta}] = \vartheta_t$.
- The covariance structure of vector u is singular, i.e. $|D_{11}| = 0$, where $D_{11} = E[uu^T]$. Therefore the majority of formulae from section II cannot be used. The straightforward calculations give:

$$\begin{aligned} D_{11} &= E[uu^T] = C_{11} - F^T M^{-1} F \quad , \\ D_{1x}^T &= D_{x1} = E[u(x)u^T] = C_{x1} - C_{x1} C_{11} sp - 1 F^T M^{-1} F \quad , \\ D_{xx} &= E[u^2(x)] = C_x \cdot - QC_{x1} C_{11}^{-1} F^T M^{-1} f(x) + f^T(x) M^{-1} f(x) \quad . \end{aligned} \tag{24}$$

Unfortunately one cannot use (17), where $\lambda = D_{11}^{-1} D_{1x}$, because D_{11} is singular (see also comments in Example 2) and one must find a way to solve the singular linear system:

$$(C_{11} - F^T M^{-1} F)\lambda = C_{1x} - F^T M^{-1} F C_{11}^{-1} C_{1x} \tag{25}$$

One of the simplest solutions of (25) is:

$$\lambda_0 = C_{11}^{-1} C_{1x} \quad , \tag{26}$$

(C_{11} assumed to be regular).

It has to be stressed that C_{11} is the ***variance-covariance matrix of the vector ε but not the vector of residuals u!*** Combining (24) and (26) one finds that the variance of the estimator $\hat{y}(x) = f^T(x)\hat{\vartheta} + C_{x1} C_{11}^{-1} u$ is equal to

$$\begin{aligned} E[y(\hat{x}) - y(x)]^2 &= E[u(x) - \lambda_0^T u]^2 = C_x - C_{x1} C_{11}^{-1} C_{1x} + \\ &+ [f(x) - F C_{11}^{-1} C_{1x}]^T M^{-1} [f(x) - F C^{-1} C_{1x}] \quad . \end{aligned} \tag{27}$$

If the estimator $\bar{y}(x) = f^T(x)\hat{\vartheta}$ is used, then

$$E[y(x) - \bar{y}(x)] = C_x + f^T(x) M^{-1} f(x) \quad .$$

The same result will be obtained if $C_{1x} = 0$ (observation at point x is uncorrelated with other observations).

Expression (27) coincides with (16), i.e. ***the universal kriging can be considered as a particular application of the generalized least square method.***

Application to the least square method allows one to trace a common take in a number of applied studies (see examples) viz., the estimates of D_{11} and D_{x1} are used to calculate λ_0 (or $\hat{\lambda}_2$); see also (24).

In spite of the theoretical clarity of the generalized least square method, its applicability to real empirical situations is very problematic because of the necessity to know matrices C_{11} and C_{1x} and one can repeat here all considerations from section II related to the case with an unknown mean-covariance structure.

Moving least squares estimator. In the majority of applications (see section III) the correlated observational errors are used to simulate the deviations of real processes from a comparatively simple trend approximation. Example 2 is a good example. Unfortunately, the model (i.e., deviations which are random values with covariance structure C_x, C_{1x}, C_{11}) contains too many unknown parameters to provide good prognoses. The use of kriging-related approaches where C_x, C_{1x}, C_{11} are substituted by their rough estimates (and sometimes wrongly constructed theoretically) misleads readers (and probably authors also) in the evaluation of the precision of prognoses.

It seems that in the examples considered, the moving average or its slightly modified version – ***moving least squares estimator*** can give better results and clearer and more direct understanding of the bounds of admissibility of the assumptions used.

Let x be a point where the prognosis has to be made, and $x_i = x + u_i$ be locations of observation points. Assume that:

(a) In the vicinity of point x the following approximation is valid

$$y_i = \vartheta_0(x) + \vartheta^T(x) f(u_i, x) + \varepsilon_i(x) \ , \ i = \overline{1,n} \quad , \tag{28}$$

where y_i is the result of observation at point $x_i = x + u_i$, $\vartheta_0(x)$ and $\vartheta(x)$ are parameters to be estimated, $\varepsilon_i(x)$ is the observation (and approximation) error, $f(u,x)$ is a vector of given functions vanishing when $u \to 0$.

The estimator can be defined as

$$\{\hat{\vartheta}_0(x), \hat{\vartheta}(x)\} = Arg \min_{\vartheta_0, \vartheta} \sum_{i=1}^{n} \lambda(u_i, x)[y_i - \vartheta_0 - \vartheta^T f(u_i, x)]^2 \ , \ \hat{y}(x) = \hat{\vartheta}_0(x) \tag{29}$$

Function $\lambda(u,x)$ has to reflect the confidence in using approximation (28) at point u. Normally $\lambda(u,x)$ is a unimodal function and

$$\lambda(0,x) = \max_u \lambda(u,x) \ , \ \lim_{u \to \infty} \lambda(u,x) = 0 \quad . \tag{30}$$

Using different $\lambda(u,x)$ one can easily vary the smoothness $\hat{y}(x)$. Due to (30), the approaches similar to (29) are frequently addressed by the ***distance-weighted least squares method*** (see Ripley, 1981, Ch. 37). In the case when there is no prior "physical" information about $y(x)$ (or $f(u,x)$), one can consider (28) as the Taylor approximation of the response $y(x)$. For the second order Taylor approximation, one will have in the two-dimensional case:

$$\vartheta_0(x) = y(x) \ , \ \vartheta^T(x) = \left(\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \frac{\partial^2 y}{\partial x_1^2}, \frac{\partial^2 y}{\partial x_2^2}, \frac{\partial^2 y}{\partial x_1 \partial x_2}\right) \quad ,$$

$$f^T(u_i, x) = (u_{1i}, u_{2i}, u_{1i}^2, u_{2i}^2, u_{1i} u_{2i}) \tag{31}$$

where $\varepsilon_i(x)$ is the remainder term at point $x - u_i$. For a sufficiently dense observation network, approximation (31) usually serves reliably. In more general case one can use any reasonable (better if supported by some physical considerations) $f(u,x)$ vanishing when $u \to 0$.

The standard least squares technique provides simple algorithms and formulae for the calculation of $\hat{y}(x) = \hat{\vartheta}_0(x)$ (see, for instance, Golub and Van Loan (1983), Seber (1977)). For theoretical analysis, it is convenient to use the following formulae:

$$\hat{y}(x) = \hat{\vartheta}_0(x) + \hat{\vartheta}^T(x) f(0,x), \tag{32}$$

where

$$\begin{pmatrix} \hat{\vartheta}_0(x) \\ \hat{\vartheta}(x) \end{pmatrix} = M^{-1}(x) Y(x),$$

$$M(x) = \begin{pmatrix} M_{00}(x) & M_{10}(x) \\ M_{01}(x) & M_{11}(x) \end{pmatrix} = \begin{pmatrix} 1 & \sum_{i=1}^{n} w_i f^T(u_i,x) \\ \sum_{-1}^{n} w_i f(u_i,x) & \sum_{i=1}^{n} w_i f(u_i,x) f^1(u_i,x) \end{pmatrix},$$

$$Y(x) = \begin{pmatrix} \sum_{i=1}^{n} w_i y_i \\ \sum_{i=1}^{n} w_i y_i f(u_i,x) \end{pmatrix}, \quad w_i = \frac{(\lambda(u_i,x)}{\sum_{i=1}^{n} \lambda(u_i,x)}.$$

The estimator (29)–(32) can be considered as some approximation scheme which can be used in both deterministic and stochastic approaches. Usually in the stochastic case, it is easier to evaluate the discrepancy between $\hat{y}(x)$ and the true value $y(x)$, of course paying for that by the additional and practically nonverified assumption:

(b) The observation errors $\varepsilon_i(x)$ are random values with $E[\varepsilon_i(x)] = 0$, $E[\varepsilon_i(x)\varepsilon_i(x)], = \lambda^{-1}(u_i,x)\delta_{ii}$.

If (b) holds and $f(0,x) = 0$ (it is quite a usual case, compare with (31)) then:

$$Var\ \hat{y}(x) = \{(\sum_{i=1}^{n} \lambda(u_i,x)(1 - M_{01}(x) M_{11}^{-1}(x) M_{10}(x)\}^{-1} \tag{33}$$

One has to notice that the observed value $y_i = y(x_i) + \varepsilon_i(x)$. In many applications it is reasonable to choose $\lambda^{-1}(u,x) = \sigma^2 + \delta(u,x)$, where σ^2 is the variance of an observation error, $\delta(u,x)$ comprises local stochastic fluctuations and $\delta(0,x) = 0$, $\lim_{u \to \infty} \delta(u,x) = \infty$. The estimation scheme (29)–(33) can be generalized in the case of correlated observations. Application to this case seems not to be useful because due to the local character of (28), the model already takes into account local tendencies and changes while in the kriging (or similar) approaches they are handled via the correlation structure of an observed field.

The cautious reader will notice that proposed estimator demands a rather tedious calculation necessitating inversion of matrix $M(x)$ for every x taken into consideration.

At first glance, one can easily avoid this by using to models describing the observed field in the vicinity of *fixed point* x_0:

$$y_i = \vartheta_0(x_0) + \vartheta^T(x_0) f(u_i, x_0) + \varepsilon_i(x_0) \quad , \tag{34}$$

and using subsequently the simplified version ($C_{11,ii'} = \delta_{ii'} \cdot \lambda^{-1}(u_i)$) of the technique discussed in the previous subsection, when

$$\hat{y}(x) = \hat{\vartheta}_0(x_0) + \hat{\vartheta}^T(x_0) f(u, x_0) \, , \; u = x - x_0 \quad . \tag{35}$$

It is clear that for all x, where $\hat{y}(x)$ has to be calculated, one has to solve the least squares problem only once. Unfortunately, it is necessary to pay for this simplification (which does not seem to be very crucial in our computerized age) by:

(a) Estimator (32) is smooth (continuous, differentiable) if functions $\lambda(u, x)$ and $f(u, x)$ are smooth. Estimator (35) will "jump" when x_0 will be changed (it is changed discretely). To avoid discontinuities, one needs to address to the least square method merging together prognoses based on different x_0. But the merging procedure removes the simplicity of (35).

(b) Model (28) provides the best approximation at point x, which is of interest, while (34) is oriented to some fixed point x_0 which can be quite remote from the moving x.

In conclusion, it is worthwhile to note that in the majority of applied studies in fluid flows, in meteorology or the atmospheric pollution studies, in contrast to geological applications, one has temporal as well as spatial information: see Example 2. The methods discussed in this section can easily incorporate temporal data explicitly expanding matrix F by adding a time dimension in generalized least squares case or in (29)–(32) the summation has to be taken over space and time. In the kriging approach, temporal information can be implicitly used through improvement of the estimates for C_{11} and C_{1x}.

References

Akima, H. (1975) Comments on "Optimal contour mapping using universal kriging" by Ricardo A. Olea *Journal of Geophysical Research*, **80**, pp. 832–836.

Armstrong, M. (1984) Problems with universal kriging. *Mathematical Geology*, **16**, pp. 101–108.

Barnes, M.G. (1980) The use of kriging for estimating the spatial distribution of tadionuclides and other spatial phenomena. Battelle Memorial Institute, Pacific Northwest Laboratory, Richland, Washington, PNL–SA–9051, pp. 20.

Bell, G.D. and Reeves, M. (1979) Kriging and geostatistics: a review of the literature available in English, Proc. Australas. Ins. Min. Metal. No. 269, pp. 17–27.

BMDP (1983) Biomedical Computer Programs, University of California Press.

Clark, T.L., Voldner, E.C., Olson, M.P., Seilkop, S.K. and Alvo, M. (1986) International sulfur deposition model evaluation (ISDME), Report.

Dennis, R.L. and Seilkop, S.K. (1986) The use of spatial patterns and their uncertainty estimates in the models evaluation process. AMS/APCA Conf., pp. xxx.

Der Megreditchan, G. (1979) Optimization des réseaux d'observation des champs météurologiques, *La Meteorologie*, **17**, pp. 51–66.

Endlich, R.M., Eynon, B.P., Ferek, R.J., Valdes, A.D. and Maxwell, C. (1986) Statistical Analysis of Precipitation Chemistry Measurements Over the Ecosystem United States, UAPSP-112, EPRI, Palo Alto, California.

Finkelstein, P.L. and Seilkop, S.K. (1981) Interpolation error and the spatial variability of acid precipitation. Proc. of the 7th Conference on Probability and Statistics in Atmospheric Sciences of AMS, AMS, Boston, pp. 206–212.

Gilbert, R.O. and Simpson, J.C. (1984) Kriging for estimating spatial pattern of contaminants: potential and problems. *Environment Monitoring and Assessment*, **9**, pp. 113–135.

Golub, G.H. and Van Loan, Ch. F. (1983) Matrix computations. The Johns Hopkins University Press, Baltimore, pp. 476.

Huijbregts, C. and Matheron, G. (1971) Universal kriging (an optimal method for estimating and contouring in trend surface analysis). Decision making in the mineral industry. *Can. Inst. Min. Met. Spec.*, Vol. **12**, pp. 159–169.

Journel, A.G. and Huijbregts, Ch.J. (1978) Mining geostatistics. Academic Press, NY, pp. xxx.

Katkovnik, V.Y. (1985) On parametric identification and data smoothing. Nauka, Moscow, pp. 336.

Malyutov, M.B. (1982) Asymptotical properties and applications of the IRGINA-estimator of parameters of generalized regression model. In "Stochastic processes and applications", Moscow, pp. 144–158.

McBratney, A.B. and Webster, R. (1981) The design of optimal sampling schemes for local estimation and mapping of regionalized variables-II. *Computers and Geosciences*, **7**, pp. 335–365.

Miccheli, C.A. and Wahba, G. (1981) Design problems for optimal surface interpolation, Approximation Theory and Applications, Academic Press, NY, pp. 329–349.

Ripley, B.D. (1981) Spatial Statistics, Wiley, NY, pp. 252.

Seber, G.A.F. (1977) Linear regression analysis, Wiley, NY, pp. 456.

Thiebaux, H.J. and Peddler, M.A. (1987) Spatial objective analysis with applications in atmospheric science. Academic Press, NY, pp.

Warner, J.J. (1986) Sensitivity analysis for universal kriging, *Mathematical Geology*, **18**, pp. 653–676.