C20 CDC 1983, San Antonio CDC 1983 C.20

ESTIMATION IN RANDOM FIELDS WITH SCATTERED DATA

G.BASTIN AND M.GEVERS*

Laboratoire d'Automatique et d'Analyse des Systèmes
Louvain University, Bâtiment Maxwell
B-1348 LOUVAIN-LA-NEUVE (BELGIUM)

ABSTRACT.

The design of linear minimum variance unbiased estimates in 2-D random fields (RF) is a standard problem when the mean and the covariance function of the field are known. Here, we investigate the case where the data are so scarce and so scattered in space that reliable estimates of the covariance function are impossible to obtain by classical procedures. Using the variogram of the RF rather than the covariance function, we develop a procedure for the estimation of a variogram model, which then leads to meaningful estimates of various functionals of the RF.

I. INTRODUCTION.

We present a methodology for the interpolation in 2-dimensional random fields (RF) with scattered data. We are concerned with real-life problems, where the mean and the covariance function of the RF are not known a priori, and where the available data are so scattered (and often so scarce) that these functions are hard to estimate by usual ways. The major difficulty, then, is the preliminary identification of a mathematical model of the RF. The paper is therefore mainly concerned with this identification problem.

We consider a 2-dimensional RF $Z(x,y)$ over a domain Ω : $(x,y) \in \Omega \subseteq R^2$, and we assume that a realization of this RF is available in the form of a finite-dimensional vector of measurements $Z=(z_1,\ldots,z_N)$, where $z_i=z(x_i,y_i)$ is a realization of $Z(x_i,y_i)$. The N locations are scattered in the domain Ω. We want to construct an optimal linear estimator $\hat{f}(z)$ for various functionals $f(z)$ of the RF :

$$\hat{f}(z) = \lambda_o + \sum_{i=1}^{N} \lambda_i z_i \qquad [1.1]$$

The functional $f(z)$ can, e.g., be the estimate of $Z(x,y)$ at a point (x_o,y_o), the surface integral of $Z(x,y)$ over the domain Ω, or the derivative of the RF at a point (x_o,y_o). A typical problem is the contour mapping of a spatially distributed variable : e.g., from rainfall measurements at a few locations, one may want to estimate the rainfall at all points of a grid, or the average rainfall over a basin.

Some common features in many applications are:

- the measurement locations are scarce and not equispaced;
- the mean of the RF is almost never zero;
- the mean and the spatial covariance function are seldom known and are hard to estimate.

Typical systems engineering applications are in geostatistics and hydrosciences. Our own interest for this problem grew out of a parameter estimation problem in 2-D groundwaterflow models. Over the years we have developed a method for the identification of a model for the RF, and we have applied this technique to a large number of problems.

As every control engineer knows, linear minimum variance unbiased estimation with a random process requires the knowledge of the covariance function of this process. The same is true for estimation in random fields. In practice, however, the covariance function is not known. In this paper we shall suggest that, for interpolation in RF, the covariance function should be replaced by another function, called variogram, which contains the same information, but which has several advantages over the covariance function. In particular, a stationary variogram exists for RF for which no stationary covariance exists, such as Wiener fields and other non stationary fields with stationary increments which are very common in 2-D problems.

In section 2, the expression of the Best Linear Unbiased Estimator (BLUE) is given for the case of point-wise interpolation (i.e. $f(z) = z(x_o,y_o)$). Two methods are presented: one where the mean and the covariance function of the RF are assumed stationary and known, the other where the mean is unknown but the variogram is stationary and known. In section 3, we discuss these two methods and explain why the second is preferable. We also show that the experimental variogram or covariance function (i.e. estimated from the data) is very chaotic and must be replaced by an analytic model. Section 4 presents and compares two identification methods for the variogram model; this is the main contribution of this paper. Finally, the usefulness of our 2-D estimation technique is illustrated in section 5, where we present results obtained on a real-life application.

II. OPTIMAL INTERPOLATION IN A RANDOM FIELD.

Notations and Definitions.

We consider a RF $Z(x,y)$, $(x,y) \in \Omega \subseteq R^2$, for which the following functions are defined :

- the mean (assumed stationary) :
$$m = E\{Z(x,y)\} \qquad [2.1]$$
- the spatial covariance kernel :
$$R(i,j) = E\{[Z(x_i,y_i)-m][Z(x_j,y_j)-m]\} \qquad [2.2]$$
where (x_i,y_i) and (x_j,y_j) are two arbitrary points in Ω.
- the spatial variogram
$$\gamma(i,j) = \frac{1}{2} E\{[Z(x_i,y_i)- Z(x_j,y_j)]^2\} \qquad [2.3]$$

We consider two special classes of random fields:

a) (Weak sense) stationary random fields:

In addition to the stationary mean assumption, we assume that the covariance is stationary :

$$R(i,j) = R(d_{ij}) \qquad [2.4]$$

where d_{ij} is the Euclidean distance between the points (x_i,y_i) and (x_j,y_j). In this case the RF variance is finite and stationary : $\sigma^2 = R(o)$; the variogram is, by definition, also stationary, $\gamma(i,j)=\gamma(d_{ij})$, and is related to the covariance function as follows :

$$\gamma(d) = \sigma^2 - R(d) \qquad [2.5]$$

b) Intrinsic random fields :
In addition to the stationary mean assumption, we assume that the variogram (but not necessarily the covariance) is stationary :

$$\gamma(i,j) = \gamma(d_{ij}) \qquad [2.6]$$

This is a wider class than class a), since it involves not only stationary RF but also non stationary RF with stationary increments (like e.g. 2-D Wiener fields). In this last case, the relationship [2.5] is no longer valid : the RF variance can be infinite if $\lim_{d\to\infty} \gamma(d)=\infty$.

Optimal Interpolation

We consider now the following situation :
. given a finite realization $Z = (z_1,\ldots,z_N)$ of the RF measured at N scattered points in Ω, find an optimal linear minimum variance unbiased estimate of $Z(x_o,y_o)$ at an arbitrary point $(x_o,y_o) \in \Omega$.

We solve this problem both for stationary and intrinsic random fields.

METHOD 1 : Stationary fields with known mean and covariance.

The linear minimum variance estimator of $z(x_o,y_o)$ is given by (see e.g. [1]) :

$$\hat{z}_o \equiv \hat{z}(x_o,y_o) = \nu_o + \sum_{i=1}^{N} \nu_i z_i = m + \sum_{i=1}^{N} \nu_i (z_i - m) \qquad [2.7]$$

where $\{\nu_i,\ i=1,\ldots,N\}$ is the solution of the system :

$$\sum_{j=1}^{N} \nu_j R(d_{ij}) = R(d_{oi}) \qquad i=1,\ldots,N \qquad [2.8]$$

The interpolation error variance is given by :

$$\sigma_o^2 = R(o) - \sum_{i=1}^{N} \nu_i R(d_{oi}) \qquad [2.9]$$

This estimator is unbiased and is a straightforward extension of the well known Levinson predictor for stochastic processes.

METHOD 2 : Intrinsic fields with unknown mean and known variogram.

We look for a linear minimum variance unbiased estimator of the form:

$$\hat{z}_o = \lambda_o + \sum_{i=1}^{N} \lambda_i z_i \qquad [2.10]$$

The λ_i are the solution of the following system :

$$\sum_{j=1}^{N} \lambda_j \gamma(d_{ij}) + \mu = \gamma(d_{oi}) \qquad i=1,\ldots,N \qquad [2.11a]$$

$$\sum_{j=1}^{N} \lambda_j = 1 \quad , \quad \lambda_o = 0 \qquad [2.11b]$$

where μ is a Lagrange parameter. The interpolation error variance is given by:

$$s_o^2 = \mu + \sum_{i=1}^{N} \lambda_i \gamma(d_{oi}) \qquad [2.12]$$

III. DISCUSSION.

In most practical applications, one does not know a priori if the RF is stationary or not, and the values of m, R(d) or $\gamma(d)$ are not given. The only information on the RF is the set of data. From these, one must:
a) make a stationarity assumption
b) estimate the mean m and the covariance R(d), or the variogram $\gamma(d)$. The first step is to draw the "experimental variogram".

Computation of the experimental variogram.

Since in most practical cases the points are not equispaced, the experimental variogram is estimated as follows. The interval of useful distances is divided into m subintervals $[d_i,d_{i+1}]$, i=1,...,m , and for each subinterval the following estimator is used :

$$\hat{\gamma}(\bar{d}_i) = \frac{1}{2N_i} \sum_{k,j} (z_k - z_j)^2 \qquad [3.1]$$

where the sum is over all couples of points (k,j) separated by a distance d such that $d_i \leq d < d_{i+1}$, N_i is the number of such couples, and $\bar{d}_i = \frac{1}{2}(d_i+d_{i+1})$.

The experimental variogram has the graphical appearance of a broken line: a typical example is shown in fig.1 for a piezometric field, which will be further studied in section 5.

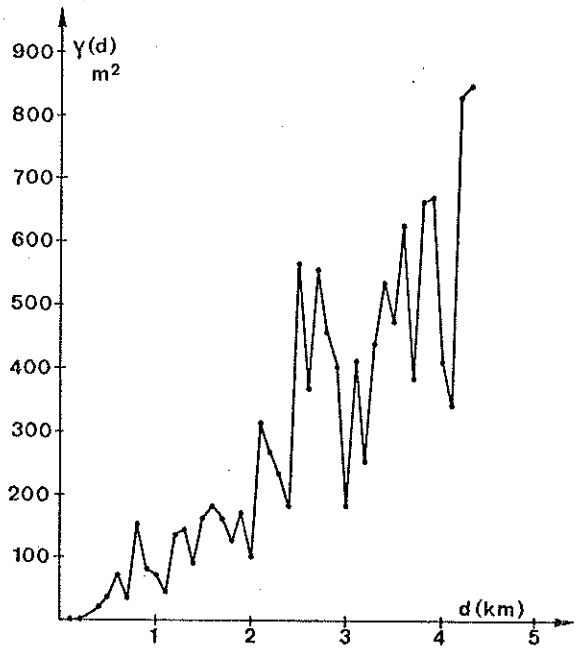

Fig.1: Experimental variogram of the water level (piezometry) in a 6km x km aquifer.

Comments

1) From fig.1, a stationarity assumption on the covariance R(d) is difficult to validate. Indeed, a necessary condition for a stationary R(d) is:

$\lim_{d\to\infty} R(d) = 0$, or equivalently (see [2.7]) $\lim_{d\to\infty} \gamma(d) = \sigma^2$.

From fig. 1 it is clearly impossible to decide whether $\gamma(d)$ will asymptotically reach a finite value σ^2 or not, since no data are available for large distances.

2) A stationary variogram exists for a wider class of random field models than a stationary covariance (see section 2). In case of doubt on the stationarity, it is therefore safer to opt for method 2.

3) We could have developed the arguments of this section using a representation of the experimental covariance (rather than the experimental variogram). Note that this requires the preliminary estimation of an estimate $\hat{m}$ of the mean of the RF. This shows an additional advantage for method 2 : not only is m not required for the interpolation, but it does not even have to be estimated to obtain an unbiased estimate of the variogram.

Conclusion

For the reasons that we have just indicated, we conclude that method 2 is preferable for the optimal interpolation of real-life random fields.

IV. IDENTIFICATION OF A VARIOGRAM MODEL.

Introduction

The formulas [2.10]-[2-11] give a straightforward solution to the interpolation problem in random fields. However, the use of the "experimental variogram" in

[2.11] would lead to completely absurd values for the estimates $\hat{z}$; this will be illustrated in section 5. Therefore it is necessary to use analytical variogram models, which obey some structural admissibility constraints, and to infer their parameters from the data. This is the most difficult part of the estimation problem in RF.

In this section, we shall present some common variogram models, and describe two basically different parameter estimation approaches :
- either the parameters are adjusted so that the variogram fits the experimental variogram(by a least-squares or generalized least-squares method) ;
- or they are adjusted so as to minimize the interpolation errors computed with this variogram model (by a minimum interpolation error or a maximum likelihood method).

Variogram models

The shape of the experimental variogram obtained inmost practical applications indicates that fairly simple parametric models can be used to describe the variograms [2] :

$$\gamma(d) = \alpha\, d^{\beta} \qquad [4.1a]$$

$$\gamma(d) = \alpha\, [1 - \exp(-\beta d)] \qquad [4.1b]$$

$$\gamma(d) = \alpha\, [1 - \exp(-\beta d^2)] \qquad [4.1c]$$

$$\gamma(d) = \alpha \log(1+\beta d) \qquad [4.1d]$$

Note that all these models have the form

$$\gamma(d;\alpha,\beta) = \alpha\gamma^*(d;\beta) \qquad [4.2]$$

With this form, α is called the "scale factor" while $\gamma^*(d;\beta)$is called the shaping factor or the spatial autocorrelation factor. Then it is important to observe that

a) the optimal interpolator $\hat{z}_o = \sum_{i=1}^{N} \lambda_i z_i$ is independent of the scale factor(see [2.11a]).

b) the variance of the interpolation error [2.12] can be written :

$$s_o^2 = \alpha\, V^*(\beta) \qquad [4.3]$$

Admissible parametrizations

The models [4.1] do not define admissible variogram models for all values of (α,β). The choice of the parametrization is constrained by the following conditions :

a) The coefficient matrix of the system [2.11a] must be non singular for all possible locations of the measure points.
b) The interpolation variance s_o^2 must be positive for all possible locations of the measure points.

From these constraints, the following necessary conditions for an admissible parametrization can be derived (see [3]) :

$$1)\quad \gamma(0;\alpha,\beta) = 0 \qquad [4.4a]$$

$$2)\quad \gamma(d;\alpha,\beta) > 0 \quad \text{for } d>0 \qquad [4.4b]$$

$$3)\quad \gamma(2d;\alpha,\beta) < 4\,\gamma(d;\alpha,\beta) \text{ for } d>0 \qquad [4.4c]$$

This implies in particular that $\alpha > 0$ and $\beta > 0$ for all models [4.1],and furthermore that $\beta < 2$ for the model [4.1a].

Estimation of α and β by the least squares method.

Given the measurements $z_1, \ldots, z_N$, one can compute experimental squared increments :

$$q_{ij} = \tfrac{1}{2}(z_i - z_j)^2 \qquad i=1,\ldots N \;;\; j=i+1,\ldots,N \qquad [4.5]$$

q_{ij} is an unbiased estimate of $\gamma(i,j)$and we can write :

$$q_{ij} = \gamma(d_{ij}) + v_{ij} \text{ with } E[v_{ij}] = 0 \qquad [4.6]$$

Having chosen a theoretical model $\gamma(d;\alpha,\beta)$, the parameters α and β are then obtained by minimizing the cost function :

$$J(\alpha,\beta) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \{q_{ij} - \alpha\gamma^*(d_{ij};\beta)\}^2 \qquad [4.7]$$

$\partial J/\partial\alpha = 0$ yields

$$\hat{\alpha}_{LS}(\,) = \frac{\sum\sum q_{ij}\,\gamma^*(d_{ij},\beta)}{\sum\sum [\gamma^*(d_{ij},\beta)]^2} \qquad [4.8]$$

β is then obtained by minimizing

$$J^*(\beta) = J(\hat{\alpha}_{LS}(\beta),\beta\,) = \sum_i \sum_j \{q_{ij} - \hat{\alpha}_{LS}(\beta)\,\gamma^*(d_{ij},\beta)\}^2 \qquad [4.9]$$

Although very simple in principle, the least-squares method has a serious drawback :the observations q_{ij} are correlated, and this can lead to the very undesirable situation where the addition of new observations can deteriorate the quality of the estimated parameters (see[5]).
Therefore, it is advisable to replace the LS method by a generalized LS method, which does not have this drawback (see[3] for details).

Estimation of β by an interpolation error (IE)method.

In this second approach, we use the minimization of the interpolation errors as a criterion for the estimation of the parameters of the variogram model. This is analogous to the idea of using a prediction error identification method for the estimation of the parameters of a dynamical model when this model is to be used for prediction purposes. Recall that, when the variogram has the form $\gamma(d) = \alpha\,\gamma^*(d,\beta)$, the optimal interpolation is independent of α. The scale factor influences only the interpolation error variance.Therefore, by minimizing some measure of the interpolation error, we shall be able to estimate β only. The method proceeds as follows.
1°) At each measure point (i = 1,...,N) an optimal estimate $\hat{z}_i$ is computed based on the N-1 other measure points, using the interpolation formulas of section 2. In matrix notations, one can write

$$\hat{Z} = \Lambda(\beta)\, Z \qquad [4.10]$$

$\Lambda(\beta)$ is a NxN matrix with zeroes on the diagonal; it depends only on β and on the location of the measure points.

2°) A vector of interpolation errors is defined

$$E \triangleq Z - \hat{Z} = [I_N - \Lambda(\beta)]\; Z \qquad [4.11]$$

The mean square interpolation error can then be defined as

$$E_q \triangleq (E^T E)^{1/2} = \{Z^T [I - \Lambda(\beta)]^T\, [I - \Lambda(\beta)] Z\}^{1/2} \qquad [4.12]$$

3°) The estimate $\hat{\beta}$ is obtained by minimizing E_q with respect to β.

Comments

a) The drawback of the method is that it does not take into account the geometry of the measure points, because the criterion [4.11]gives the same weight to all interpolation errors. Therefore large interpolation errors at the border of the domain will tend to have an unduly large effect on the choice of β. The answer to this difficulty is to replace the criterion [4.12]by a maximum likelihood (ML) criterion, assuming a Gaussian distribution for the RF. The likelihood function incorporates the covariance matrix of the errors,$R(\beta)$, which is a function of the geometric location of the

measure points. The ML method provides an estimate of both α and β, in a decoupled way. It is presented in detail in [3].

b) The interpolation error described above can also yield an estimate of the scale factor α as follows. Denote by $e_i(\beta)$ the interpolation error at the i-th measure point and by s_i^2 the variance of the corresponding interpolation error. Then, by [4.3],

$$s_i^2 = \alpha \, V_i^*(\beta) \qquad [4.13]$$

This suggests the following estimator for α :

$$\hat{\alpha}_{AML}(\beta) = \frac{1}{N} \sum_{i=1}^{N} \frac{e_i^2}{V_i^*} \qquad [4.14]$$

This estimator is called $\hat{\alpha}_{AML}$ (for "approximate maximum likelihood") because it was in fact obtained as an approximation to the ML estimator for α derived in [3].

c) The IE method and the ML method have been applied to a large number of estimation problems in random fields. They have shown to perform consistently better than the LS or generalized LS methods, particularly when few data are available. The application presented in section 5 will illustrate this point.

V. APPLICATION : CONTOUR MAPPING OF THE WATER LEVEL IN A GROUNDWATER RESERVOIR.

A typical application is the contour mapping of the piezometric level (i.e. the level of the top of the water table) in a groundwater reservoir. Such a contour mapping requires the estimation of the piezometric level at all nodes of a grid covering the domain (in order to establish a chart of the piezometry) from measurements made at a few piezometers scattered within the reservoir (Fig.2).

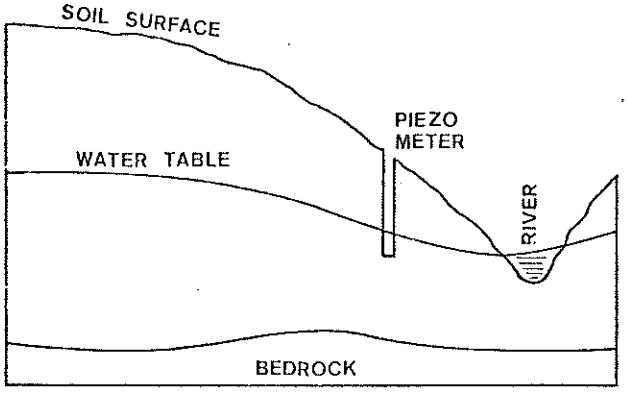

Fig.2: Groundwater reservoir with piezometer.

Here we shall use this application to illustrate some of the features of the theory presented before. The studied domain consists of a 6 km x 6 km area around Louvain-la-Neuve (Belgium). 28 piezometers were available and observed during October and November 1977: their locations will be indicated by dots on figs. 4 and 5. A distance d_{max} = 5 km is considered. This distance has been divided into 50 segments of 100 meters each, and the experimental variogram has been drawn as explained in section 3. The result is presented in fig. 1.

This experimental variogram is fairly chaotic, and if it were used as such to compute interpolated values of the piezometry, it would lead to a mean square interpolation error at the data points of E_q = 171 m ! This is totally unacceptable, since the data points are all between 56 and 117 m. This example shows that the use of analytic variograms is absolutely essential. We shall see that with such models E_q will be of the order of 3 meters.

The following table shows the result of the estimation of the parameters α and β for the models [4.1] and for some combinations of the estimation methods proposed in section 4.

Variogram model	Estimation method	$\hat{\alpha}$	$\hat{\beta}$	Eq	J_{LS}
αd^β	LS ①	91.49	1.29	3.27	$3.51\ 10^7$
	IE - AML ③	31.20	1.44	3.01	
$\alpha[1-\exp(-\beta d^2)]$	LS ②	792.	0.081	13.82	$3.49\ 10^7$
	IE - AML ④	277.	0.986	5.34	
$\alpha[1-\exp(-\beta d)]$	LS and IE		$\beta \to 0$		
$\alpha \log(1+\beta d)$	LS and IE		$\beta \to 0$		

(note: d in km, Eq in meters)

Table 1.

The numbers ①, ②, ③, ④ refer to Fig.3.

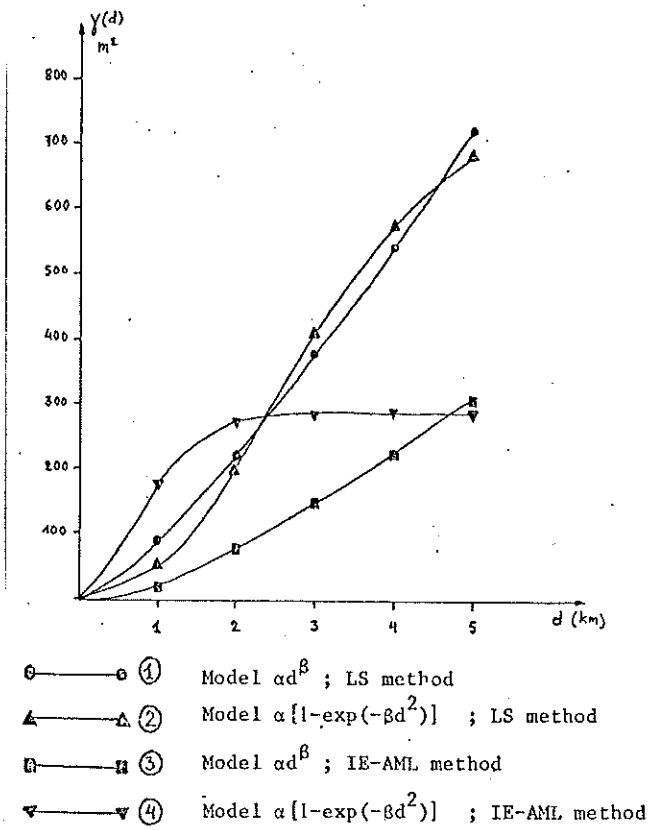

Fig.3: Estimated variogram models for the piezometric random field.

In this table, LS means that $\hat{\alpha}$ and $\hat{\beta}$ are computed by the Least-squares Method, IE - AML means that $\hat{\beta}$ is computed

by the Interpolation Method and $\hat{\alpha}$ by the approximate maximum likelihood formula [4.14]. Fig.3 illustrates graphically some of the identified models.

COMMENTS

1) As can be expected from fig.1, only variograms with positive curvature (i.e. αd^{β} and $\alpha [1-\exp(-\beta d^2)]$) can be reasonably fitted to the experimental variogram. Indeed both LS and IE methods converge to an estimate $\beta = 0$ for the models with negative curvature.
2) With the mean-square interpolation error Eq as criterion, table 1 shows that αd^{β} is better than $\alpha[1-\exp(-\beta d^2)]$, even though the latter gives in fact a better least-squares fit to the experimental variogram.
3) For the model αd^{β}, the mean square errors Eq obtained with the LS and the IE methods are quite close (3.27 and 3.01). This does not mean that both models are equivalent : although the interpolated values will be very close, the estimation error variance s^2 (which is proportional to $\hat{\alpha}$: $s^2 = \hat{\alpha} V^*(\beta)$, see [4.3]) will be twice as large with the LS model than with the IE model.

To conclude, the model $\gamma(d;\alpha,\beta) = \alpha d^{\beta}$ is selected with $\alpha = 31.2$ and $\beta = 1.44$. A grid with square elements of size $\Delta x = \Delta y = 0.5$ km is superimposed on Ω, and $z(x,y)$ is estimated at each node of the grid using the interpolation formulas [2.10] and [2.11]. A contour map can then be obtained from this grid and is represented in fig. 4.

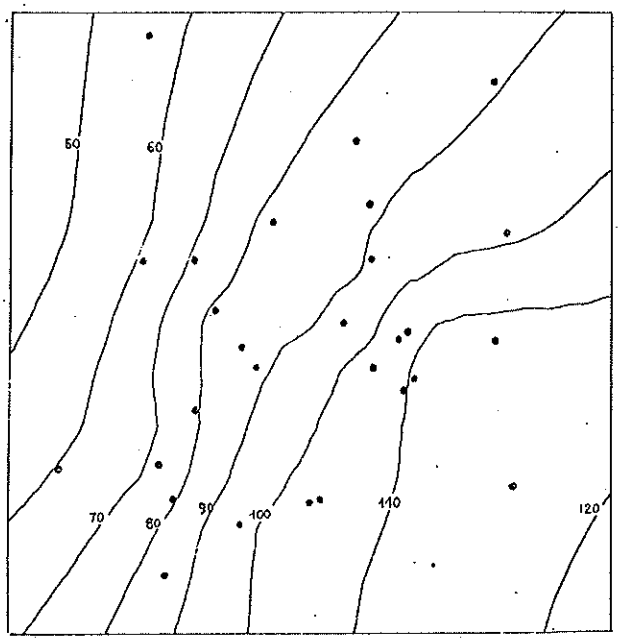

Fig.4: Contour map of the piezometry (in meters above sea level). The dots indicate the locations of the piezometers.

Since the optimal interpolation technique provides not only piezometric estimates, but also estimated standard deviations, a contour map can also be drawn showing the areas of equal standard deviations. This is shown in fig. 5. As could be expected, the standard deviation increases in areas where there are few measure points. In practice, the computation of the experimental variogram, the fitting of various standard variogram models, the optimal interpolation and the contour mapping can be done automatically by a software package KRIGEA/CARTO developed at Louvain University by the Automatic Control Group.

VI. CONCLUSION.

We have shown that the standard techniques for interpolation in random fields cannot be applied to many real-life problems, because the spatial covariance

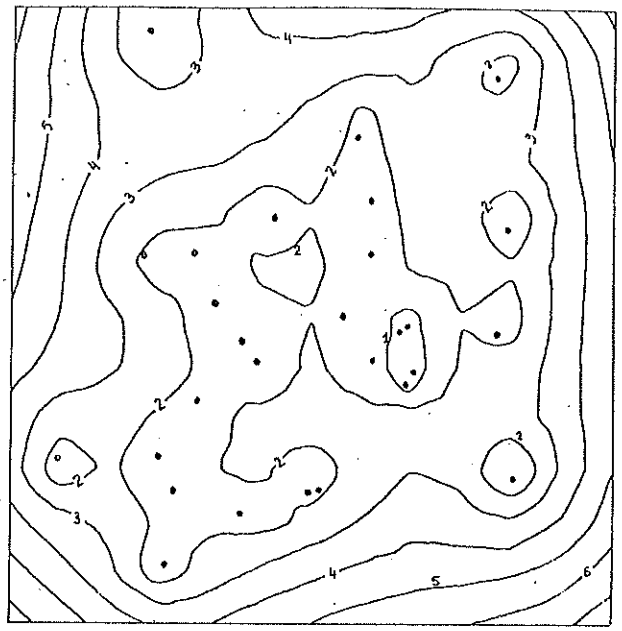

Fig.5: Contour maps of equal standard deviation for the estimated piezometric levels of Fig.4.

function is unknown and its estimate will lead to absurd answers when the data points are scarce. The solution we propose is twofold : first, replace the covariance function by a variogram, because the latter covers a much wider class of RF's and is easier to estimate; next, use an analytical model for the variogram. The major problem then becomes the identification of a variogram model. We have presented a class of admissible models and given two methods for the estimation of their parameters. We have illustrated their performances through a typical application and shown that the interpolation error method of parameter estimation gives much better results. Other real-life applications are presented in [4]. Hopefully, the reader should by now be convinced that the methodology presented in this paper is applicable to a wide range of 2-D problems.

REFERENCES.

1. PAPOULIS A., "Probability, random variables and stochastic processes", MacGraw Hill, 1965.
2. MATHERON G., "The intrinsic random functions and their applications", Adv. Appl. Prob. ,vol.5,1973.
3. BASTIN G. and M.GEVERS, "Identification and optimal estimation of random fields from scattered point-wise data - Part I .Theory", submitted for publication.
4. BASTIN G. and M.GEVERS, "idem - Part II : Applications", submitted for publication.
5. GEVERS M. and G.BASTIN, "On the estimation of the variogram in spatial interpolation methods used in groundwaterflow modeling", in "Applications of Information and Control Systems", Reidel Publ. Cy, 1980, pp. 60-68.