G. H. Schut
National Research Council
Ottawa, Canada K1A 0R6

Two Interpolation Methods

Evaluation experiments indicate that if the prescribed procedure is followed, the method of prediction and filtering is superior to the method of pointwise interpolation and smoothing.

Introduction

The theory of stationary random functions has become the mathematical basis for a method of interpolation and smoothing or, in other words and with a somewhat wider interpretation, prediction and filtering. This method is used in geophysics where it was introduced by Moritz (1963) and it was introduced in photogrammetry by Kraus (Kraus, 1972; Kraus and Mikhail, 1972) with an assumption that is not based on the theory.

Recently also, two other interpolation methods have been published (Arthur, 1972; Hardy, 1971, 1972) which make use of the formalism of the theory. However, their authors do not refer to this theory but base their formulation on other considerations.

The present paper describes first the conditions under which the theory is applicable and the extent to which the three methods apply the theory. For this purpose, the following sections contain a summary of the so-called *correlation theory* of those functions and an analysis of the three methods.

Very different from these methods is a method of pointwise interpolation and smoothing which is a generalization of the *method of moving averages*. It has recently been developed for specific cases of interpolation (Schut, 1970) and smoothing (Schut, 1972). For the purpose of comparison, a more general description of this method is given here.

Finally, the prediction method and the method of pointwise interpolation are applied in various ways to the interpolation of heights of points on a surface in three-dimensional space. For this purpose, a known analytical surface has been selected. This serves to obtain information on the degree to which in practical applications the requirements of correlation theory must be satisfied and it serves to compare the results of the two methods.

Abstract: *A recently published interpolation method is based on the correlation theory of stationary random functions. This method is here discussed together with two variants which are not overtly based on correlation theory but employ its formalism and differ only in the degree to which they deviate from the specifications of this theory. A method of a different type which is a generalization of the method of moving averages is presented in a more general form than in earlier publications. Finally, results obtained with the two methods are compared. The results show that the first method can be expected to give better results but only if the requirements and specifications of correlation theory are carefully satisfied and obeyed.*

Correlation Theory

A random function of one parameter, say x, is a function whose values are random variables. With it are associated distribution functions which specify the probabilities that the values of the function, individually and in

sets of two or more, lie within specified ranges. The random function is called *stationary* of these probability distributions and are independent of the value of x.

The correlation theory of these functions is based on the first and second moments of the distribution functions. The first moment is the mean value of the function. The second moment is a function, $B(\mu)$, of the separation μ in x of two values of the random function. $B(\mu)$ is called the *correlation* function or *covariance* function. The theory has been developed for the case where the first moment is zero.

To apply the theory to a given case where a number of function values at equally spaced values of x are given, the first moment is made zero by subtracting the mean from each value. The function $B(\mu)$ is then computed for each multiple μ of the x-spacing by taking the mean of the products of all pairs of function values spaced by the distance μ. $B(0)$ is the mean square of all values of the random function. The correlation function can be normalized by dividing all its values by $B(0)$.

Because the correlation function is a statistical concept, it cannot be accurately determined if only a limited number of values of the random function is available. In such a case, if the correlation function is not known in advance, a correlation function must be assumed that agrees reasonably well with the values calculated from the available values of the random function. It must have the following properties based upon its statistical origin: it has its maximum as well as its maximum absolute value at $\mu = 0$ and its graph is symmetric with respect to $\mu = 0$. Further, if i and j are two values of x and $\mu = i-j$, the matrix $\mathbf{B}$ whose element b_{ij} equals $B(\mu)$ must be nonnegative definite.

According to (Yaglom, 1962) a correlation function which is often used in cases where precise information is not available is

$$B(\mu) = exp\,(-a|\mu|);\ a>0. \qquad (1)$$

This function is always positive and with increasing values of μ it rapidly approaches zero. The correlation is not necessarily always positive; it can, e.g., have the character of a damped oscillation. In that case, a more suitable correlation function is

$$B(\mu) = (exp\,(-a|\mu|))\ cos(b\mu);\ a>0,\ b>0. \qquad (2)$$

Let now the correlation function of a stationary random function be known and let the values ξ_i of the latter function at $x_i = i\,(i = -11, -2, \ldots, -n)$ have been measured. According to correlation theory, the best linear estimate of this function at $x = m$ is:

$$\xi_m = \mathbf{a}'\mathbf{b}\ (m = 0,1, \ldots) \qquad (3a)$$

Here, the vector $\mathbf{a}$ is solved from

$$\mathbf{Ba} = \xi \qquad (3b)$$

ξ is the column vector whose components are the values ξ_i of the stationary random function, $\mathbf{B}$ is the above-mentioned matrix whose elements have the values $B(0)$ to $B(n-1)$ and $\mathbf{b}$ is the vector whose components are the values $B(\mu)$ for $\mu = m-1, m-2, \ldots, m-n$.

If the values ξ_1 are affected by errors η_i, the actual observations are $\zeta_i = \xi_i + \eta_i$. Let it be assumed that the errors η_i are also elements of a stationary random function, that their mean value is zero and that the two random functions are not correlated. The correlation function of the actual observations is then the sum of the correlation functions of the two random functions:

$$B(\mu) = B_\xi(\mu) + B_\eta(\mu) \qquad (4)$$

The best linear estimate of the value ξ_m of the random function at $x = m$ is now obtained from

$$\xi_m = \mathbf{a}'\mathbf{b} \qquad (5a)$$

in which, as before, $\boldsymbol{b}$ is the vector whose components are the values of the correlation function $B_\xi(\mu)$ for $\mu = m-1, m-2, \ldots, m-n$. The vector $\boldsymbol{a}$ is now solved from

$$\boldsymbol{Ba} = \zeta\ . \qquad (5b)$$

ζ is the vector whose components are the actual observations of the stationary random function and the elements of the matrix $\boldsymbol{B}$ are now derived from Equation 4.

Obviously, the formulas (3a,b) and (5a,b) can be extended to the case where m is not a non-negative integer. If m is any rational non-integer number, the formulas serve for interpolation between the measured values or for extrapolation beyond them. Both these cases have been covered by the name prediction. If $m = -1, \ldots, -n$, Formulas 3a,b reproduce the actual observations. Formulas 5a,b give here the best linear estimate of the values ξ_i of the random function. This has been called *filtering*.

Interpolation in Two Dimensions

Yaglom (1962) also mentions the concepts of a stationary random function in multidimensional space and the correlation theory

of such a function. This random function is said to be homogeneous and isotropic if its correlation function is a function of only the distance between points in the field, not of location or direction.

Moritz (Moritz, 1963; Heiskanen and Moritz, 1967) has applied these concepts to interpolation in two-dimensional space. His ξ_i are the gravity anomalies in a given area. Assuming that the average product of two gravity anomalies is a function of the separation of their locations only, this average product again defines the correlation function. Moritz's interpolation formulas are formally identical with Equations. 3a,b but these now have the following interpretation. The reference points, that is, the points at which the gravity anomalies are determined, are ordered in an arbitrary sequence. The components of the vector ξ are the gravity anomalies, ordered in this sequence. The element b_{ij} of the matrix $\boldsymbol{B}$ is the value of the correlation function for the distance between the two points with sequence numbers i and j. The component b_i of the vector $\boldsymbol{b}$ is the value of the correlation function for the distance between the point at which the gravity anomaly is to be predicted and the point with sequence number i.

A more recent publication (Moritz, 1973) extends the interpolation and filtering to the case where the observations are the sum of a linear function of unknown parameters, a stationary random function and measuring errors. The linear function contains the systematic part of the observations and may be called the trend function. If the trend function were known, its values could be subtracted from the observations and subsequently the correlation theory could be applied. In Moritz's formulation, however, the parameters in the trend function are computed simultaneously with the interpolation and filtering. This implies that the correlation function must either be known in advance or be determined otherwise.

Kraus's Linear Least-Squares Interpolation

Kraus has adopted Moritz's (1963) formulation for his least-squares interpolation (Kraus, 1972; Kraus and Mikhail, 1972). He makes use of Equations. 5a,b with the additional and reasonable assumption that the random measuring errors are uncorrelated. This makes $B_\eta(\mu)$ non-zero only if μ equals zero.

In addition, he assumes that the most appropriate correlation function is a Gaussian curve. After normalization, this gives

$$B_\xi(\mu) = exp\,(-a\mu^2);\ a>0 \qquad (6)$$

For this, there is no justification in the theory of random functions. If the first moment has been reduced to zero, for large values of μ the distribution function $B_\xi(\mu)$ is equally likely to have negative as positive values. This can be illustrated with practical examples. An experimentally derived correlation function in (Moritz, 1963) attains negative values. Also, in (Kraus, 1972) the experimental data in Figure 2 can be fitted much better with a function that becomes negative for large values of μ. Finally, in a present investigation of film deformation Dr. H. Ziemann of the Photogrammetric Research Section of NRC is having the same experience. In all these cases, the experimental data can be fitted much better with a correlation function of the type

$$B_\xi(\mu) = (exp(-a\mu^2))\ cos\,(b\mu);\ a>0,\ b>0. \qquad (7)$$

Kraus (1972) states that this method of interpolation, which he applies here to film deformation correction, is independent of the type and structure of the systematic deformation. Although this is true of the computation, it is not true of the result. Best results, in the statistical sense, are obtained only if the data comprise a realization of a random function with all the restrictions discussed in the preceding section. This cannot be achieved by simply subtracting a constant from all observations to make the first moment equal to zero.

Recognizing this, in (Kraus and Mikhail, 1972) the concept of the trend function is utilized. As it is put here: relative to the trend function, the data must have positive and negative regions in a more or less random fashion. The trend function is determined first and the interpolation and filtering are performed only after reducing the data by the trend function.

Arthur's Interpolation of a Function of Many Variables

Arthur (1965) devised an interpolation method which made use of Equations 3a,b. The correlation function used here is

$$B(\mu) = 1-\mu^2;\ \mu = d/a, \qquad (8)$$

d being the distance between two points and a being a constant distance. By making a larger than the largest distance between two points, $B(\mu)$ remains positive for all values of μ. For criticism of this method, the reader should refer to (Schut, 1970).

Arthur (1972), after having given a strangely incorrect listing of this criticism, has changed his correlation function to a Gaussian curve. Based on a desired closest possible approach to a linear interpolation in the case of two reference points only, he now chooses the correlation function

$$B(\mu) = exp\ (-2.5\mu^2);\ \mu = d/a. \qquad (9)$$

Now, a is the average distance between the points with known values.

As Arthur (1972) states that no interpolation method has a theoretical basis, he is apparently unaware that his method differs from the one derived in correlation theory only in the criterion for selection of the correlation function.

Hardy's Interpolation with Multiquadric Equations

Hardy (1971, 1972) developed an interpolation method which, also, makes use of Equations 3a,b. The role of the correlation function is here assumed by one of the functions

$$B(d) = (d^2 + C)^{1/2} \qquad (10)$$

and

$$B(d) = d^2 + C \qquad (11)$$

Here, d is again the distance between two points whereas C is a constant to which a value is assigned in advance.

Hardy (1971) has given the interpolation its geometrical interpretation: it is simply a summation of n functions. Each term in Equation 3a contributes one function to the summation. Each function contains only one adjustable parameter which is the component a_i of the vector $\boldsymbol{a}$. In the two-dimensional case, this function becomes a surface of revolution with its extreme at the i-th reference point and with $a_iB(d)$ as a vertical section.

At first sight, Equations 10 and 11 may seem to be very unsuitable as correlation functions because $B(\mu)$ increases with increasing values of μ. However, it must be borne in mind that both here and in the case of the rigorous application of correlation theory the a_i are not restricted to positive values and therefore in both cases both maxima and minima can occur at the reference points. Hardy reports reasonable-looking results in the use of his formulation for contouring. These results are obtained by carefully choosing the reference points at significant terrain points such as highs and lows.

This seems to indicate that the form of the correlation function may not be very critical especially if a rather large number of reference points is used. Nevertheless, to avoid sharp peaks or dips at the reference points, one should avoid functions such as Equations 1 and 2, and Equation 10 with $C = 0$.

In a more sophisticated version of the method (Hardy, 1971), a polynomial of low degree is added to the first parts of Equations 3a,b. The coefficients of this polynomial are determined simultaneously with the components of the vector $\boldsymbol{a}$ by adding suitable condition equations to Equations 3b. This means that here a trend surface is determined simultaneously with the vector $\boldsymbol{a}$.

Pointwise Interpolation

An entirely different method of interpolation and smoothing is the method of *moving averages*. Here, the interpolated value of a function at any point is computed as a weighted average of the values at the reference points. The weight attached to a reference value is a function of the distance from the interpolated point to the reference point. Therefore, the interpolated value is computed independently for each point.

Taking the case of the height interpolation over a given area, the interpolated value may be interpreted as the height of a horizontal plane. Therefore, in the method of moving averages, for each interpolated point a horizontal plane is determined and its height is taken as the interpolated value at the point.

The method of interpolation and smoothing developed by Schut (1970, 1972) is a generalization of the method of moving averages. Taking again the case of the height interpolation, the horizontal plane is replaced by a tilted plane or even by a curved surface. Again, for each interpolated point such a surface is computed and its height at the point is accepted as the interpolated value.

One important consideration in this method is the choice of the surface which is computed for each point. To save computation time, it should have only a few parameters. A simple formulation is a polynomial with respect to the planimetric coordinates. It need not have higher than first- or second-degree terms.

A second important consideration is the choice of weights. The weight should be a monotonically decreasing function of distance. As a result, the totality of interpolated points defines a continuous surface which cannot be given an analytical formulation. A rather sharp drop-off of the weight at small values of the distance produces a surface which fits well at the reference points. A slower drop-off produces a smoothing effect.

In experiments with interpolation in planimetric block adjustment (Schut 1970), the best results were obtained with the weight function.

$$w = (1-r)^3 (1-r^2)^3/r. \qquad (12)$$

Here, r was the ratio between the distance to a reference point and a maximum distance beyond which reference points were not used. These results were obtained with a fixed distance which was somewhat larger than the largest distance in the block. For ratios smaller than 0.01, r was made equal to this value. This prevents the weight from rising to infinity as r approaches zero.

The use of a maximum distance by which all distances are divided serves two purposes. Firstly, it makes it possible to specify a weight function which can be used independent of the size of the block and of the unit of measurement. Secondly, by making the weight function approach to zero as the distance to a reference point approaches the maximum distance, reference points beyond this distance are not used in the computation. In the case of a dense net of reference points, the proper choice of a maximum distance can save much computation time.

A weight function which proved to be suitable in a case where a strong smoothing was required (Schut 1972) is:

$$w = 1 - 2r^2 \quad (r \leq 0.5)$$

$$w = 2(1 - r)^2 \quad (r \geq 0.5) \qquad (13)$$

Various other weight functions could be used. Arthur (1972) has suggested replacing r in the denominator of the weight function for interpolation by r^2. This causes a considerably sharper dropoff in weight with increasing values of r than does an increase in the powers of $1 - r$ and $1 - r^2$.

A function which can be adapted to both interpolation and smoothing is

$$w = exp\,(-ar^2). \qquad (14)$$

If the constant a is equal to 14 or 20, this function varies from unity at $r = o$ to less than 10^{-6} or 10^{-8}, respectively, at $r = 1$. The smoothing effect could be varied by varying either one or both of the constant a and the maximum distance. However, both these measures can affect the number of points that effectively participates in the interpolation. Rather, the degree of smoothing should be controlled by using the weight function

$$w = exp\,(-ax^2);\ x = r/(b + (1 - b)r) \quad . \qquad (15)$$

Here, a is a fixed value which may be taken to be 14 and b is a variable parameter. In the case of a fairly regular distribution of reference points and the use of a maximum distance which is about four times the average distance between those points, the value $b = 0.2$ gives very little smoothing, $b = 1$ gives a fair amount of smoothing, and $b = 2$ gives a very considerable amount of smoothing.

Experiments

To evaluate and compare the results that can be obtained with these methods, they were applied to the interpolation of points on a known analytical surface in three-dimensional space. In this application, and by reason of the geometric interpretation, a method based on the formalism of correlation theory may be called the *multisurface method* and the method of pointwise interpolation may be called the *moving surface method.*

The analytical surface was constructed by Dr. V. Kratky of the Photogrammetric Research Section for a non-topographic application of photogrammetry. Profiles in x-direction through this surface form a wavelike pattern while profiles in y-direction vary from a straight line to a convex curve. A set of 17×17 points was used covering about 125 mm in x-direction (about one wavelength) and 160 mm in y-direction. The maximum variation in height is about 26 mm and the mean height is zero.

This surface is clearly not a representative example of an isotropic stationary random function. Not only does it show systematic patterns, but the patterns in the x- and y-directions are different. Results obtained with this surface should be instructive because it corresponds to the situation often encountered in practical applications. For instance, heights in a digital terrain model can hardly be regarded as such a function.

The multisurface method was used with correlation functions of the type proposed by Kraus and by Arthur:

$$B(d) = exp\,(-ad^2/b^2)$$

in which a and b are parameters whose values are specified in advance and d is the distance. This function actually has only one parameter, the ratio a/b^2.

When first using this method with correlation functions of this type and with various values of the ratio, very poor interpolation results were obtained. This can be readily explained by the markedly anisotropic behavior of the analytic surface and the geomet-

TABLE 1. RESULTS OF MULTISURFACE METHOD OF INTERPOLATION WITH CORRELATION FUNCTION exp $(-ad^2/b^2)$.

Values of a/b²	y-scale reduction 0.62		y-scale reduction 0.25	
	Max. error	RMS error	Max. error	RMS error
$2.5/27^2 = 34.3/100^2$	2.47 mm	0.56 mm		
$20/100^2$	2.48	0.69	1.06 mm	0.24 mm
$14/100^2$	2.24	0.63	0.41	0.10
$2.5/85^2 = 3.5/100^2$	0.42	0.13		

ric interpretation of the interpolation as a summation of surfaces of revolution.

Arthur (1972) remarked that the use of only distances in the correlation function is not very satisfactory if the distribution of reference points is markedly anisotropic and that the anisotropy should be eliminated by a preliminary affine transformation. In the present situation, this requires the reduction of the y-coordinates by a factor of about 0.62. In Arthur's correlation function, a equals 2.5 and b is the average distance between reference points. After the scale reduction of the y-coordinates and taking into account all directions, the average distance lies between 25 and 30 mm.

Accordingly, further experiments were performed with various scale factors applied to the y-coordinates. Table 1 displays some of the results obtained with the factors 0.62 and 0.25, and varying the ratio a/b^2. With the factor 0.62, a shallow minimum of both the maximum error and the RMS error is reached with $a = 2.5, b = 27$ mm. However, if b is increased beyond 35 mm, the results improve very significantly until the value $b = 85$ mm is reached. Beyond this the interpolation fails, even with double precision arithmetic, because the matrix $\boldsymbol{B}$ becomes singular. At each of several values of a/b^2, the factor 0.25 consistently gave better results than other scale factors.

These results show that it is not sufficient to decide upon a correlation function of this type and to derive the ratio a/b^2 from theoretical desiderata. Also, the introduction of a likely correlation function and of isotropy in the distribution of the reference points are together not sufficient to obtain optimum results.

An additional measure that should be taken is to bring the data into a form that is more representative of an isotropic stationary random function. This can be done by means of the concept of referring the data to a trend surface, and not simply to the horizontal plane at mean height. Simple formulations for such a surface are polynomials and harmonic functions of low degree.

In the present instance, the wavelike form of the analytical surface does not make it possible to fit a polynomial of low degree. A harmonic function could fit rather well but it would have required additional programming. As an interesting alternative, a trend surface was computed by the *moving-surface* method. Moving surfaces of the first and second degree were used with a maximum distance of 100 mm and the weight functions of Equations 14 and 15. Results are shown in Table 2. Comparison with the relevant values in Table 1 shows that the introduction of a trend surface greatly improves the result of the multisurface method.

The interpolation was performed also by the moving surface method. Here, a second-degree moving surface was used in combination with the reduction factor 0.62. The best results that were obtained with various weight functions are shown in Table 3. The maximum distance of 75 mm which was here used in the weight function proved to be the smallest value that was suitable. The use of the weight function of Equation 14 together with a larger value of the maximum distance produces a very strong smoothing effect.

TABLE 2. RESULTS OF MULTISURFACE METHOD AFTER REDUCING DATA TO A TREND SURFACE; CORRELATION FUNCTION, exp $(-14d^2/100^2)$; SCALE FACTOR, 0.62

Construction of trend surface	Max. error	RMS error
moving surface of degree 1, using Eq. (14) with $a=14$	0.53 mm	0.18 mm
moving surface of degree 2, using Eq. (14) with $a=14$	0.12	0.04
moving surface of degree 2, using Eq. (15) with $a=14$, $b=0.5$	0.24	0.09

TABLE 3. RESULTS OF MOVING SURFACE METHOD, USING A SECOND-DEGREE MOVING SURFACE AND A MAXIMUM DISTANCE OF 75 MM.

Weight function	*Max. error*	*RMS error*	*Max. error at reference points*
$(1 - r)^3 (1 - r^2)^3/r$	0.35 mm	0.12 mm	0.01 mm
$(1 - r)^3 (1 - r^2)^3/r^2$	0.36	0.12	0.00
Eq. 14; $a=20$	0.30	0.11	0.11
Eq. 15; $a = 20, b = 0.2$	0.35	0.12	0.00
Eq. 14; $a=14$	0.30	0.11	0.30
Eq. 15; $a=14, b=0.2$	0.41	0.14	0.00

Weight functions that result in little or no smoothing together with a smaller value of the maximum distance tend to produce gross errors in some points with relatively few reference points nearby.

CONCLUSIONS

The experiments show that the method of prediction and filtering that is based on correlation theory can give excellent results even if the data have a very systematic character.

They show also that it is not sufficient, as done in Arthur's method, to use the formalism of the method although deriving the correlation function from other considerations.

The choice of the Gaussian curve as correlation function, adopted by Kraus and by Arthur, could give good results but only because the known analytical surface which was interpolated made it possible to determine the most advantageous value of this parameter experimentally. By far the best results were obtained by first reducing the data not by their arithmetic mean but with respect to a trend surface.

Accordingly, to obtain best results in practical applications where the data will often have a systematic character, first a trend function must be determined which represents the overall systematic trend in the data, and the data must be reduced by subtracting for each point the value of the trend function. Next, the correlation function should be determined from the reduced data. There is no good reason why this function should be restricted to a Gaussian curve. A good alternative would be the function given by Equation 7. Subsequently only, the interpolation or filtering should be performed.

One may conclude from the experiments that, if this procedure is followed, the method of prediction and filtering is superior to the method pointwise interpolation and smoothing. On the other hand, if this procedure is not followed, the method can easily give inferior results. This conclusion reduces the importance of the method of pointwise interpolation and smoothing but it leaves it not entirely without interest.

REFERENCES

1. Arthur, D.W.G., 1965, "Interpolation of a function of many variables," *Photogrammetric Engineering*, 31:2, pp. 348-349.
2. Arthur, D.W.G., 1973, "Interpolation of a function of many variables, II," *Photogrammetric Engineering*, 39:3, pp. 261-266.
3. Hardy, R. L., 1971, "Multiquadric equations of topography and other irregular surfaces," *Journal of Geophysical Research*, 76:8, pp. 1905-1915.
4. Hardy, R. L., 1972, "Analytical topographic surfaces by spatial intersection," *Photogrammetric Engineering*, 38:5, pp. 452-458.
5. Kraus, K., 1972, "Film deformation correction with least squares interpolation," *Photogrammetric Engineering*, 38:5, pp. 287-493.
6. Kraus, K., and Mikhail, E. M., 1972, "Linear least-squares interpolation, *Photogrammetric Engineering*," 28:10, pp. 1016-1029.
7. Moritz, H., 1963, "Statistische Methoden in der gravimetrischen Geodäsie," *Zeitschrift für Vermessungswesen*, 88:10, pp. 409-416.
8. Moritz, H., 1973, "Determination of the gravity field by collocation," *Bollettino di Geodesia e Scienze Affini*, 32:1, pp. 1-13.
9. Schut, G. H., 1970, "External block adjustment of planimetry," *Photogrammetric Engineering*, 36:9, pp. 974-982.
10. Schut, G. H., 1972, "A computerized adjustment of APR strips," *The Canadian Surveyor*, 26:1, pp. 45-48.
11. Yaglom, A. M., 1962. *An Introduction to the Theory of Stationary Random Functions*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.