Spatial Interpolation Methods: A Review

Nina Siu-Ngan Lam

ABSTRACT. Two forms of spatial interpolation, the interpolation of point and areal data, are distinguished. Traditionally, point interpolation is applied to isarithmic, that is, contour mapping and areal interpolation to isopleth mapping. Recently, areal interpolation techniques have been used to obtain data for a set of administrative or political districts from another set of districts whose boundaries do not coincide. For point interpolation, the numerous methods may further be classified into exact and approximate. Exact methods include most distance-weighting methods, Kriging, spline interpolation, interpolating polynomials, and finite-difference methods. Approximate methods include power-series trend models, Fourier models, distance-weighted least-squares, and least-squares fitting with splines. Areal interpolation methods, on the other hand, are classified according to whether they preserve volume. Traditional areal interpolation methods which utilize point interpolation procedures are not volume-preserving, whereas the map overlay and pycnophylactic methods are. It is shown that methods possessing the volume-preserving property generally outperform those that do not.

KEY WORDS: two-dimensional interpolation, contouring, Kriging, spline, trend surface, volume-preserving, map overlay, pycnophylactic, areal interpolation

The spatial interpolation problem can be simply stated as follows. Given a set of spatial data either in the form of discrete points or for subareas, find the function that will best represent the whole surface and that will predict values at other points or for other subareas. This general problem has long been a major concern in many disciplines. In geography and cartography, the main applications of different spatial interpolation methods are in isoline mapping. With the advance of computing technology, and with increased use of multivariate analysis of data collected for varying units, spatial interpolation methods have been applied to other problems in geographic research as well. For example, the study of the effects of socio-economic characteristics on voting behavior requires the comparison of ward data with census tract data, and boundary segments of these two sets of units rarely coincide. It would be useful to examine the nature and characteristics of various interpolation methods so that appropriate selections can be made for various applications.

In giving a systematic review of interpolation methods, a classification will be used that divides them into point methods and areal methods. Point interpolation deals with data collectable at a point, such as temperature readings or elevation, whereas areal interpolation deals with data, such as population counts by census tracts, that are aggregated over a whole area. Maps of the former type of data are often referred to as isometric, maps of the latter as isopleth (Hsu and Robinson 1970). Point or isometric methods will be further subdivided into "exact" and "approximate" methods according to whether they preserve the original sample point values, whereas areal or isopleth methods will be subdivided according to whether they preserve volume. The nature of each class of interpolation methods and its relative merits will be examined. Worked examples of selected interpolation methods are also given in the appendix. Although reviews of spatial interpolation methods have appeared before, they are either oriented toward disciplines other than cartography (Crain 1970; Leberl 1975; Schumaker 1976; Schut 1976), or they do not include discussion of areal

Nina Siu-Ngan Lam is assistant professor of geography at the Ohio State University, Columbus, OH 43210. The author wishes to thank Dr. M. F. Goodchild and the reviewers for their comments on an earlier draft of this paper.

0094-1689/83$2.50

interpolation (Rhind 1975). It is the intent of this paper to bring together selected methods that are useful in mapping and map-related problems.

POINT INTERPOLATION

Numerous algorithms for point interpolation have been developed in the past. But none of them is superior to all others for all applications, and the selection of an appropriate interpolation model depends largely on the type of data, the degree of accuracy desired, and the amount of computational effort afforded. Even with computers available, some methods are too time-consuming and expensive to be justified for certain applications. In all cases, the fundamental problem underlying all these interpolation models is that each is a sort of hypothesis about the surface, and that hypothesis may or may not be true.

These point interpolation methods may be classified in any of a number of ways. For example, some classify the methods according to the spatial extent of data points involved, that is, as either global methods, in which all sample points are utilized in determining value at a new point, or piecewise methods, in which only nearby points are used. Some classify Kriging as a statistical technique and identify the remainder as analytical methods (Delfiner 1976). In the present paper, the numerous point interpolation methods are classified as either exact or approximate methods because the characteristic of preserving or not preserving the original sample point values on the inferred surface seems fundamental in analyzing accuracy and in examining the nature of interpolation methods (Wren 1975). The methods of the "exact" type include interpolating polynomials, most distance-weighting methods, Kriging, spline interpolation, and finite difference methods. The group of "approximate" methods includes power-series trend models, Fourier models, distance-weighted least-squares, and least-squares fitting with splines (Figure 1).

It should be mentioned briefly that two different approaches may be used for contour mapping, a main application of spatial interpolation methods. Given a set of irregularly-spaced data points, the first approach to contouring first forms a set of triangles from the data points. The contours are then drawn through the triangles using different interpolation methods (Gold and others 1977). The second approach requires interpolation of the data points to a mesh of grids and then traces the contours through the mesh of interpolated values (Walters

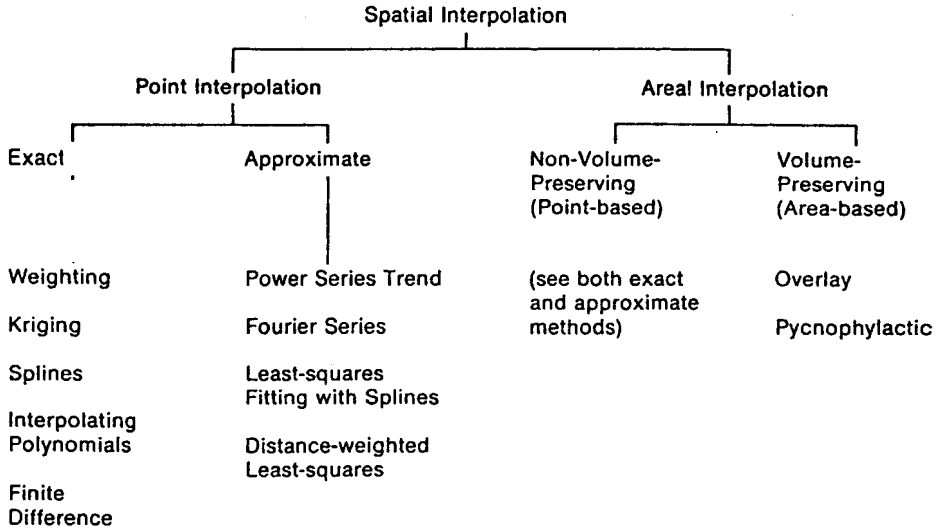


Figure 1. Types of spatial interpolation methods.

1969). In this approach, it is very unlikely that the contoured surface would pass through the data points even if an exact interpolation method were used, unless the data points coincide with the grids. Yet in the first approach the contoured surface constructed by an exact method will pass through each data point since the irregular triangular grids are the data points (McCullagh and Ross 1980).

Exact Methods

Given the set of N data points, one of the simplest mathematical expressions for a continuous surface that intersects these points is the *interpolating polynomial* of the lowest order that passes through all data points. One common form of this polynomial is

$$f(x,y) = \sum_{i,\ j=0}^{N} a_{ij}x^i y^j. \qquad (1)$$

The coefficients a_{ij} are determined by solving the set of equations

$$f(x_i,y_i) = z_i,\ i = 1, \ldots\ldots, N. \qquad (2)$$

The major deficiency of this exact polynomial fit is that since the polynomial is entirely unconstrained, except at the data points, the values attained between the points may be highly unreasonable and may be drastically different from those at nearby data points. This problem may be alleviated to a certain extent by employing lower order piecewise polynomial surfaces to cover the area (Crain and Bhattacharyya 1967). However, piecewise surface fitting might cause such problems as discontinuities at the edges where a certain amount of overlap is necessary, high computation time, and the need to adjust for variations in data density. Other problems include the existence of other solutions for the same set of data (Schumaker 1976) and the inaccuracy of the inverses of large matrices of equation (2) for polynomials of orders greater than 5 (Ralston 1965). As a result, this exact polynomial interpolation method is not generally recommended, and particularly so when the number of data points is large.

The principle of *distance-weighting* methods is to assign more weight to nearby points than to distant points. The usual expression is

$$f(x,y) = \left[\sum_{i=1}^{N} w(d_i)z_i\right] \Big/ \left[\sum_{i=1}^{N} w(d_i)\right], \qquad (3)$$

where $w(d)$ is the weighting function, z_i is the data value at point i, and d_i is the distance from point i to (x,y).

Although weighting methods are often used as exact methods (Sampson 1978), they can also be approximate depending on the weighting functions. For those weighting functions where $w(0) = \infty$, such as $w = d^{-1}$, the weighting method will give the exact value of the original sample points. On the other hand, for a negative exponential weighting function, the method will only approximate the original values at the locations of the sample points. Lancaster and Salkauskas (1975) discuss the relative merits of various weighting functions. An example of this method using $w = d^{-1}$ is shown in the appendix.

There are several disadvantages to weighting methods. First, the choice of a weighting function may introduce ambiguity, especially when the characteristics of the underlying surface are not known. Second, the weighting methods are easily affected by uneven distributions of data points since an equal weight will be assigned to each of the points even if it is in a cluster. This problem has long been recognized (Delfiner and Delhomme 1975), and has been handled either by averaging the points or selecting a single point to represent the cluster (Sampson 1978). How far apart points should be from each other before one can consider some of them redundant remains another question. Morrison (1974) even suggested that interpolation should not be carried out unless the data point distribution has a nearest neighbor statistic of more than 1.0, to indicate randomness. Such a rule is rather too simple as well as controversial since the nearest neighbor statistic itself is subject to a number of problems, not the least of which is that some very

non-random patterns also yield a value of 1.0.

Finally, the interpolated values of any point within the data set are bounded by $\min(z_i) \leq f(x,y) \leq \max(z_i)$ as long as $w(d_i) > 0$ (Crain and Bhattacharyya 1967). In other words, whatever weighting function is used, the weighted average methods are essentially smoothing procedures. This is considered to be an important shortcoming because, in order to be useful, an interpolated surface, such as a contour map, should predict accurately certain important features of the original surface, such as the locations and magnitudes of maxima and minima—even when they are not included as original sample points.

However, the simplicity of the principle, the speed in calculation, the ease of programming, and reasonable results for many types of data have led to a wide application of the weighting methods as well as improvements of various types. A combination of a weighting method with other procedures also has been used, most notably in SYMAP interpolation (Shepard 1970).

Kriging is perhaps the most distinctive of interpolation methods. The term is derived from the name of D. G. Krige, who introduced the use of moving averages to avoid systematic overestimation of reserves in the field of mining (Krige 1976). Matheron (1971) has generalized the theory to the case of nonstationary data, and the resulting method was later termed Universal Kriging. It has become a major tool in the field of geostatistics in the last two decades (Guarascio and others 1976; Mousset-Jones 1980). Recent applications of Kriging to other fields are increasing (McCullagh 1975).

Kriging regards the statistical surface to be interpolated as a regionalized variable that has a certain degree of continuity. In some cases, a regionalized variable may have a minimal degree of continuity in that no matter how short the distance between two samples, their values are simply independent of each other. Such variables will have a "nugget" effect on the estimation procedures (Figure 2b). Regionalized variables may also have a certain degree of anisotropy, whereby the zone of influence of a sample does not have the same extent in all directions. Yet there must be a structure or spatial autocorrelation, that is, a dependence between sample data values, which decreases with their distance apart. These characteristics of regionalized variables are quantified by the sample variances and covariances, that is, the autocovariance matrix, from which the Kriging estimates of unknown points are determined (Rendu 1970).

Because different assumptions about the regionalized variables may be involved, two systems of Kriging procedures, simple Kriging and Universal Kriging, can be distinguished. Within the system of simple Kriging, two different assumptions may further be distinguished and these relate to two approaches for estimating the autocovariance matrix. In the first approach, the covariogram function, expressing the relation between the covariance of the sample points and their distance, is used. It is expected that the covariogram is a decreasing function of distance (Figure 2a); however, in actual applications the covariograms will diverge from this theoretical behavior. This approach to simple Kriging is based on the stationarity assumption, which holds that all the sample points are taken randomly and independently from one simple probability distribution. This assumption, in turn, implies that the probability density function and the autocovariance matrix can be estimated.

However, natural phenomena with this stationarity characteristic seldom exist. Hence, interpolation may be based upon the second but less restrictive assumption, the intrinsic or quasi-stationarity assumption, in which only the increments of the function but not the function itself are required to be stationary (David 1977; Goodchild 1979). Instead of the covariogram, the variogram, which represents the relationship between the mean-square difference between sample values and their inter-

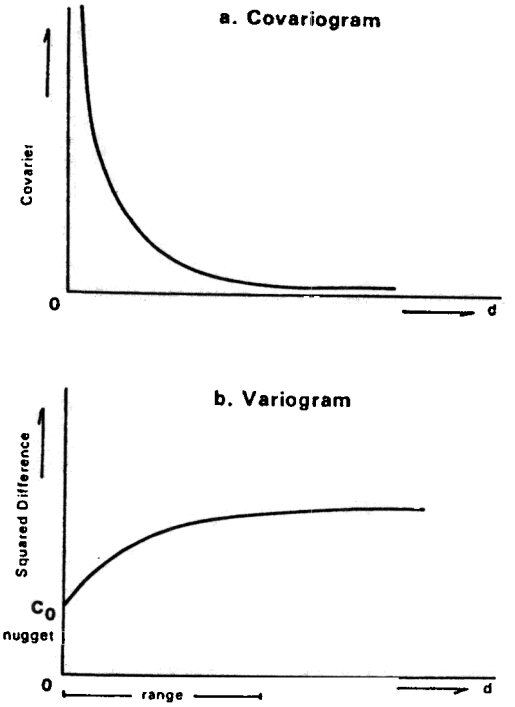


Figure 2. Examples of covariogram and variogram.

vening distance, is now used. Mathematically, the variogram (2r) or semivariogram (r) is defined by

$$r = \tfrac{1}{2}N \sum_{i=1}^{N} [z(x_i + d) - z(x_i)]^2, \quad (4)$$

where d is the distance between two samples. This function is expected to increase with the distance between samples, taking a value close to zero for small distances, and becoming a constant for distances larger than the zone of influence, or range (Figure 2b). Similarly to the covariogram, the experimental variogram will often deviate significantly from this theoretical model. Once the variogram or covariogram has been estimated from the samples, it is possible to calculate the elements of the autocovariance matrix.

Because these assumptions of simple Kriging imply a certain amount of stationarity over space, and because regionalized variables are often nonstationary, an alternative hypothesis is desirable. Universal Kriging assumes that the increments of the regionalized variable have some properties of stationarity only within a neighborhood and that the trend or drift for a neighborhood can be described by a polynomial function. The residuals from the drift are now assumed to have a constant variogram within a neighborhood.

Once the coefficients of the autocovariance matrix for a given set of samples are determined, the estimates for unknown points can be calculated by a linear combination of the weighted sample values

$$z^* = \sum \lambda_i z(x_i), \quad (5)$$

where λ_i are weights to be determined under the following two conditions:

$$E(Z^* - Z) = 0 \quad (6)$$

$$\text{var}(Z^* - Z) = \text{min.} \quad (7)$$

The first is a universality condition which states that Z must be an unbiased estimate. The second condition, the optimality condition, implies that λ_i should have values such that the estimation variance of the difference $(Z^* - Z)$ be minimum (Matheron 1963). The Kriging estimate thus obtained is the best linear unbiased estimate (BLUE). The corresponding estimation variance provided for an unknown point is the Kriging error. For points that belong to the set of samples, Kriging returns the original data values, and so constitutes an exact interpolation procedure.

Calculation of Kriging estimates under the two systems, simple and universal, can be found in a number of sources (David 1976 and 1977; Goodchild 1979). An example of calculating simple Kriging estimates is given in the appendix. Generally, simple Kriging has more restrictive assumptions but fewer computational problems, whereas the assumptions of Universal Kriging are more general but difficulty of calculation is greater. Universal Kriging uses a different set of equations for each point estimate in different neighborhoods. The

variogram represents the residuals instead of the observed values, which would require that local drifts be known first. Since true drifts are not known, they must be estimated from the available sample values; the variogram calculated from them is also an estimate of the true variogram (Olea 1974).

How closely the variogram of the estimated residuals corresponds to the true but unknown variogram depends upon the appropriateness of the function selected to represent the drift, the function selected to represent the variogram, and the size of the chosen neighborhood. These three problems are closely related, and none can be determined independently of the others. The usual procedure is first to assume a simple form of the variogram of residuals and then to select a neighborhood size. Next the drifts within the neighborhoods are estimated and the experimental variogram of residuals calculated. The two variograms are then compared (Huijbregts and Matheron 1971). The result of a search for the drift and variogram is not unique: there are always several combinations of drift and variogram that may be equally satisfactory.

Among other problems associated with Universal Kriging is the selection of an appropriate size of neighborhood. If the neighborhood is large, a regular and slowly varying drift is obtained, but also a more complicated underlying variogram, and vice versa. Choice of neighborhood will also affect the continuity properties of the estimates, which may lead to serious bias in interpretation. If the change of data points from one neighborhood to the next is too abrupt, there may be discontinuities even though the actual phenomenon is continuous (Delfiner 1976). A closely related problem is the determination of drift and variogram under different scales. An area of higher elevation on a topographical surface can be regarded as a "mountain", and hence a drift at one scale, or it may enter the random part (variogram) at another scale (Matheron 1971). There are other criticisms: the method is not reliable unless a very large number of sample values are available (Rendu 1970); the improved accuracy provided by Kriging will not always justify the computational effort required (Matheron 1963; Olea 1974); and the difference in accuracy between local cubic polynomial interpolation and Kriging is marginal (Kubik and Botman 1976).

Nevertheless, the attractions of Kriging are several. First of all, from a theoretical point of view, Kriging utilizes the theory of regionalized variables which allows the drawing of statistical inference. The model itself represents an improvement over other interpolation techniques, especially polynomial interpolation, since the degree of interdependence of the sample points has been taken into account. The Kriging estimate is based on the structural characteristics of the samples which are summarized in the covariogram or variogram function and thus result in an optimal unbiased estimate. Kriging also provides an estimate of the error and confidence interval for each of the unknown points, an asset not provided by other interpolation procedures. This error information reflects the density and distribution of control points and the degree of spatial autocorrelation within the surface, and therefore is very useful in analyzing the reliability of each feature in the Kriged map. The error map can also be used to determine where more information is needed so that future sampling can be planned.

Spline functions are widely discussed topics in mathematics, but applications in geography and cartography are relatively few. They have only recently been applied to isopleth mapping (Tobler and Lau 1978 and 1979).

First consider the two-dimensional case. Given a set of n points along a profile $x_0 < x_1 < \ldots . < x_n$, a spline function $s(x)$ of degree m with the knots $x_0, x_1, \ldots . , x_n$ is a function defined on the entire line such that in each interval (x_i, x_{i+1}) for $i = 0, \ldots , n$, $s(x)$ is given by some polynomial of degree m or less, and

that $s(x)$ and its derivatives of order 1, 2, ... , $m - 1$ are continuous everywhere (Giloi 1978). For $m = 1$, 2, or 3 a spline is called linear, quadratic, or cubic, respectively. Thus, a quadratic spline must have one continuous derivative at each knot, the cubic two. The cubic splines are the most widely used and they are called bicubic splines in the three-dimensional case. In some cases, the knots need not be the data points at which the values are given, and the splines in these cases are only an approximation of the data. However, the case of coincident knots and data points seems to be the most widely used, and most spline interpolations are exact.

Extending splines to the three-dimensional case is not easy since a three-dimensional spline is not a simple cross product of univariate splines. Furthermore, there is an ambiguity in dividing the surface into patches such that the spline functions can be applied. Hessing and his co-workers (1972) first extended bicubic spline interpolation to irregularly spaced data by drawing lines through the data points to form quadrangles. Extra points were needed at some intersections of these lines in order to complete the quadrangles, and the values for these extra points had to be determined before beginning the interpolation. An example of bicubic spline interpolation using the algorithm in Späth (1974) is given in the appendix.

Another approach is to divide the surface into triangles by connecting the data points at vertices of these triangles. The fact that there are many ways of making the triangulation of the same set of data points complicates the interpolation problem. However, selection of triangles has long been a concern in digital terrain modeling (Peucker and others 1976) and algorithms for dividing the surface into acceptable or optimal triangles according to some criteria have been designed (Cavendish 1974).

One generalization of spline functions has led to the use of spline blending. This method is useful for the construction of a surface which interpolates not only function values at isolated points but also at points along grid lines. If data are dense along lines, there may be a real advantage in using this method (Gordon 1969 and 1971). In addition, the B splines, which search for the least number of non-zero subintervals—for a linear spline the number is two—also have been suggested for handling large numbers of data points since computations are more reliable and efficient (Ahlberg 1970; De Boor 1976).

The use of spline functions in spatial interpolation offers several advantages. They are piecewise, and hence involve relatively few points at a time and should be closely related to the value being interpolated; they are analytic; and they are flexible. Splines of low degree, such as the bicubic splines, are always sufficient to interpolate the surface quite accurately. Bhattacharyya and Holroyd (1971) illustrated that when compared with other interpolation methods, specifically the inverse square distance-weighting method and the Gram-Schmidt orthogonalization procedure, spline interpolation is highly accurate since all important small-scale features are retained. However, there are difficulties associated with this technique. In addition to the problem of defining patches over a surface, all of the spline interpolation and blending methods introduce anomalies that are not in the original surface (Lancaster and Salkauskas 1975).

The principle behind *finite difference* methods is the assumption that the desired surface obeys some differential equations, both ordinary and partial. These equations are then approximated by finite differences and solved iteratively. For example, the problem may be to find a function z such that

$$\frac{\delta^2 z}{\delta x^2} + \frac{\delta^2 z}{\delta y^2} = 0 \qquad (8)$$

inside the region, and $z(x_i,y_i) = 0$ on the boundary. This is the LaPlace equation; and a finite difference approximation of this equation is

$$z_{ij} = (z_{i-1,j} + z_{i+1,j} + z_{ij-1} + z_{ij+1})/4, \quad (9)$$

where z_{ij} is the value in cell ij. This equation in effect requires that the value at a point is the average of its four neighbors, resulting in a smooth surface. For a smoother surface, other differential equations may be used. Also, the "boundary conditions" may be applied not only to the boundary but also within the region of interest (Briggs 1974; Swain 1976). This point interpolation technique has a striking similarity with the pycnophylactic areal interpolation method, which will be discussed later.

The principle involved in these finite difference methods is generally simple though the solution of the set of difference equations is time-consuming. Yet, the surface generated from these equations has no absolute or relative maxima or minima except at data points or on the boundary. In addition, interpolation beyond the neighborhood of the data points is poor, and unnatural contouring can occur for certain types of data, such as broad flat areas (Crain 1970). Moreover, in some cases there might be no value assigned to certain points.

Approximation Methods

The methods to be discussed in this section are concerned with determining a function, $f(x,y)$, which assumes values at the data points approximately but not generally equal to the observed values. Thus, there will be an "error" or residual at every data point. In order to obtain a good approximation, the errors must be kept within certain bounds by some error criterion. Two commonly used criteria are the minimax, which minimizes the maximum value of e over all i and the least-squares, which minimizes the sum of squares of residuals

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (f(x_i,y_i) - z_i)^2 = \min. \quad (10)$$

Since the determination of $f(x,y)$ according to the minimax criterion is rather complicated even in two dimensions, the least-squares criterion is frequently used (Crain and Bhattacharyya 1967).

Ordinary least-squares polynomials are of the same general form as equation (1), but in this case the number of terms in the polynomial, m, is less than the total number of data points, N, with the addition of an error term:

$$f(x,y) = \sum_{i,j=0}^{m} a_{ij}x^i y^j + e. \quad (11)$$

These methods are also called trend-surface models since they are often used to simplify the surface into a major trend and associated residuals. Since interpolation means prediction of function values at unknown points, and trend in this case is regarded as a simplified function able to describe the general behavior of the surface, predictions of values thus follow the trend (Torelli 1975). An example of fitting a first-degree trend is given in the appendix.

Although often criticized (Norcliffe 1969; Unwin 1975), applications of this trend model to both physical and socio-economic phenomena has been very extensive (Chorley and Haggett 1965; Wren 1973). Problems associated with these trend models for interpolation are apparent. In the first place, the trend model assumes a distinction between a deterministic trend and a stochastic random surface (noise) for each phenomenon, which may be arbitrary in most cases. Such distinction requires a serious theoretical background which is often missing in geography. Actually, in most of the geosciences, the so-called trend may present the same stochastic character as the noise itself. Hence, a distinction between them is only a matter of scale, which is similar to the problem of distinguishing drift and variogram in Universal Kriging.

Miesch and Connor (1968) have compared fitted surfaces constructed from polynomial terms and arbitrary terms, with approximately the same number of terms being used in each case. Although both fitted surfaces could explain roughly the same proportion of total variance, they led to markedly different patterns of residuals. Since both models have approximately the same proportion of variance explained, the choice between

them therefore depends largely on the *a priori* knowledge of surface form.

The estimation of values using trend models is highly affected by the extreme values and uneven distribution of data points (Krumbein 1959). The problem is further complicated by the fact that some of the data points are actually more informative than others. For example, in topographic maps, the data points taken from the peaks, pits, passes, and pales (Warntz 1966; Peucker 1972) are more significant than the points taken from the slope or plain. Hence, the answer to how many data points are required for a reliable result is not known.

Compared with Kriging, the variance given by least-squares polynomials is the variance between the actual and the estimated values at sample points, which is generally less than the variance at points not belonging to the set of sample points (Matheron 1967). The mean-square error from the polynomial fit is not related to the estimation error as illustrated clearly by Delfiner and Delhomme (1975). The experiment in Lam (1981) further indicates that polynomial trend surfaces having the same amount of variance explained, represented by r^2, may have a drastic difference in the mean error between the estimated and the actual values for all points.

Another interesting problem neglected in most of the literature about trend models relates to the accuracy across the map. Zurflueh (1967) showed that a polynomial surface fit becomes unreliable at the edge of the map, causing severe problems when two adjacent areas have to be fitted with polynomials.

If there is some definite reason for assuming that the surface takes some recurring or cyclical form, then a trigonometric polynomial, or *Fourier series model*, may be most applicable. The Fourier model basically takes the form

$$z = a_{00} + \sum_{i=1}^{M}\sum_{j=1}^{N} F(a_{ij},p_i,q_j) + e, \qquad (12)$$

where $p_i = 2\pi ix/M$ and $q_j = 2\pi jy/N$. M and N are the fundamental wavelengths in the x and y directions. The Fourier series $F(a_{ij},p_i,q_j)$ is usually defined as

$$F(a_{ij},p_i,q_j) = cc_{ij}\cos(p_i)\cos(q_j) + cs_{ij}\cos(p_i)\sin(q_j) + sc_{ij}\sin(p_i)\cos(q_j) + ss_{ij}\sin(p_i)\sin(q_j). \qquad (13)$$

CC_{ij}, CS_{ij}, SC_{ij}, SS_{ij} are the four Fourier coefficients for each a_{ij} (Bassett 1972). With this equation a surface can be decomposed into periodic surfaces with different wavelengths. The Fourier models have been mainly used in describing and comparing physical surfaces (Harbaugh and Preston 1968; Harbaugh and others 1977). It has been suggested by Curry (1966) and Casetti (1966) that the model is particularly useful for studying the effects of areal aggregation on surface variability. It is possible to combine trend and Fourier models so that a polynomial of low order is used to extract any large-scale trend; the residuals from this surface are analyzed by Fourier models (Bassett 1972).

Distance-weighted least-squares may be used to take into account the distance-decay effect (McLain 1974; Lancaster and Salkauskas 1975). In this approach, the influence of a data point on the coefficient values is made to depend on its distance from the interpolated point. The error to be minimized becomes

$$\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} w(d_i)(f(x_i,y_i)-z_i)^2, \qquad (14)$$

where w is a weighting function. Its choice again has a serious effect on the interpolation results. Computation time is increased by the calculation of the weighting function.

Another variation of least-squares is *least-squares fitting with splines*. Although a number of authors have suggested that this method will yield adequate solutions for most problems (Hayes and Halliday 1974; Schumaker 1976; McLain 1980), it involves a number of technical difficulties such as the problem of rank-deficiency in matrix manipulations, the choice of knots for spline approximation, and problems

associated with an uneven distribution of data points.

AREAL INTERPOLATION

The areal interpolation problem is more common to geography than to other fields. The literature concerning this type of interpolation, however, is very scanty. Applications of areal interpolation procedures in the past, as mentioned above, have mainly been in isopleth mapping, which seems to be regarded as a fundamental problem in this field (Mackay 1951 and 1953). An extended application of areal interpolation methods is the transformation of data from one set of boundaries to another. As indicated before, this type of application has increased rapidly in importance and has become a major focus in the study of the areal interpolation problem. It is in this sense that the term "areal interpolation" is used in the remainder of this paper. Although the nature of the data is different, the study of the areal interpolation problem is closely related to point interpolation since the traditional approach to areal interpolation requires the use of point interpolation procedures. Therefore, the problems associated with point interpolation models should be understood first before examining the underlying structure of areal interpolation.

For convenience, following Ford (1976), the geographic areas for which data are available will be called source zones and those for which data are needed will be called target zones. Two approaches, volume-preserving and non-volume-preserving, can be used to deal with the areal interpolation problem (Figure 1).

Non-Volume-preserving Methods

This approach generally proceeds by overlaying a grid on the map and assigning a control point to represent each source zone. Point interpolation schemes are then applied to interpolate the values at each grid node. Finally, the estimates of the grid points are averaged together within each target zone, yielding the final target-zone estimate. In this approach, the major variations between the numerous methods are the different point interpolation models used in assigning values to grid points. It is therefore also termed the "point-based areal interpolation approach". The specific point interpolation methods are identical to those already discussed.

There is evidence that this approach is a poor practice (Porter 1958; Morrison 1971). First of all, the choice of a control point to represent the zone may involve errors. If the distribution of the phenomenon is symmetrical and relatively uniform, the center-of-area would be a convenient control point, and the estimated value for each grid would be reliable. Unfortunately in reality, zones such as census tracts and counties for which the data are aggregated are seldom symmetrical, and the patterns of distributions of most socio-economic phenomena are uneven. Secondly, ambiguity occurs in assigning values at the grid points in some conditions, particularly when opposite pairs of unconnected centers have similar values which contrast with other opposite pairs—a classic problem of locating isopleths mentioned by many authors (Mackay 1951 and 1953).

Thirdly, this approach utilizes point interpolation methods and hence cannot avoid the fundamental problems associated with them. As mentioned above, the most important problem underlying the interpolation process is that an *a priori* assumption about the surface is involved. Very often, this assumption is rather arbitrary and most geographical phenomena are, in fact, very complex in nature and it is difficult to reduce the data in such a fashion that it can be analyzed simply. Ironically, the abundant use of interpolation procedures found in the field of cartography is associated with scanty research on the reliability of the specific interpolation method used (Hsu and Robinson 1970; Jenks and Caspall 1969; Jenks and others 1969; Morrison 1971; Stearns 1968).

Still other factors including, for example, the spatial arrangement and the density of data points suggested by a number of authors (Hsu and Robinson 1970; Morrison 1971) may seriously affect the validity of the interpolation result. In applying point interpolation methods to areal data, the problem is further complicated by the fact that the accuracy of the result is subject to sources of error implicit in the original aggregation procedure. The size and shape of the source and target zones (Coulson 1978) and the distribution of the values of the variable for interpolation (Ford 1976) are major factors affecting the validity of the results.

The most important problem of this approach, however, is that it does not conserve the total value within each zone. This problem has long been neglected in most of the pertinent literature, although sometimes it is indirectly implied (Schmid and MacCannell 1955). Tobler (1979) addressed this property explicitly and applied it to both point and areal interpolation problems. The idea of volume-preserving can be simply expressed as follows. First consider the two-dimensional case, as shown in Figure 3. A smooth curve can always be constructed so that the area under the curve in each category is retained. The same general procedure is necessary in the three-dimensional case; the interpolated surface is required to be smooth while preserving volume in each source zone. Because of the volume-preserving property, the isoline map drawn from the interpolated values can be converted into a bivariate histogram simply by computing the volume under the isoline surface, and so the original value of each source zone can be constructed. This is what Tobler called an inversion property, and is closely related to the volume-preserving property. Volume-preserving is a very useful property because it gives greater fidelity to the approximation of grid values in each source zone so that subsequent estimation of a value for each target zone is less subject to error.

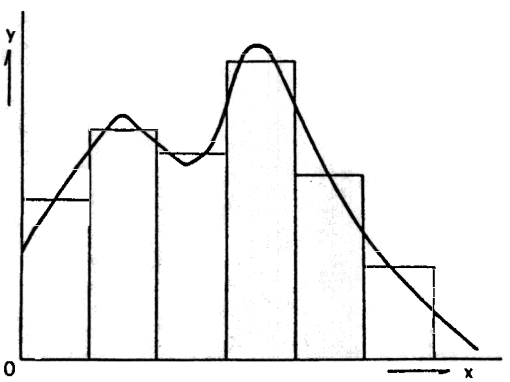


Figure 3. Volume-preserving property in the two-dimensional case.

Volume-preserving Methods

The second approach to areal interpolation, called the "area-based areal interpolation approach" in this paper, preserves volume as an essential requirement for accurate interpolation. Furthermore, the zone itself is now used as the unit of operation instead of the arbitrarily assigned control point. Hence, no point interpolation process is required. So far, two different methods utilizing this approach can be distinguished.

The *overlay* method of areal interpolation superimposes the target zones on the source zones. The values of the target zones are then estimated from weights which are determined from the size of the overlapping areas. Similar procedures have been described in a number of disciplines in a widely scattered literature (Markoff and Shapiro 1973; Crackel 1975). Recently, the overlay method itself has become a major function of many geographic information systems.

Areal interpolation using map overlay is intuitively simple. Once the overlay product of the source zones and the target zones is obtained, the area of each individual polygon can be measured. One can construct a matrix A consisting of the area of each of the m target zones (rows) in common with each of the n source zones (columns), with elements denoted by a_{ts}. Also, let the column vec-

tors U (of length n), V (of length m) represent the source zone values and the target zone estimates (notation follows Goodchild and Lam 1980). The next step of the estimation procedure will differ slightly depending on the type of the aggregate data describing the source zones. First of all, for data in the form of absolute figures or counts, such as total population and income, an estimate of target zone t is obtained by:

$$V_t = \sum_s U_s a_{ts}/\sigma_s, \qquad (15)$$

where σ_s denotes the area of source zone s. In matrix representation, $V = WU$, where W is a weight matrix containing elements of a_{ts}/σ_s. A small example of using the overlay method is given in the appendix.

Secondly, density data, such as population densities, are converted first to absolute figures by multiplying by σ_s. The result is then converted back to densities by dividing by the target zone area Γ_t:

$$V_t = \left(\sum_s U_s \sigma_s a_{ts}/\sigma_s\right)/\Gamma_t = \sum_s U_s a_{ts}/\Gamma_t. \qquad (16)$$

Finally, for data which are in the form of ratios or proportions, such as percent of males in the population, additional interpolation procedures have to be included. Since ratios simply compare two absolute figures or two densities, it is necessary to perform separate areal interpolation procedures for both the numerator U_{s1} and the denominator U_{s2}, where $U_s = U_{s1}/U_{s2}$:

$$V_t = \left(\sum_s U_{s1} a_{ts}/\sigma_s\right) \Big/ \left(\sum_s U_{s2} a_{ts}/\sigma_s\right). \qquad (17)$$

If the data for the denominator or the numerator are densities, then the procedures will be similar to those discussed earlier but will use equation (16) instead.

The major problem with this method is that it assumes homogeneity within each source zone (McAlpine and Cook 1971). In other words, if the value of each source zone is the same everywhere, subsequent reaggregation into target zones will yield exact estimates. Yet if the value of each source zone is unevenly distributed within its domain, estimation of target zone values from the amount of overlapping areas may not be reliable. In this latter case, the reliability of target zone estimates will be governed mainly by the nature and degree of the inhomogeneity of the source zone description and by the size of the target zone in relation to the corresponding source zone. Ford (1976), in his study of a contour reaggregation problem using point-based areal interpolation methods, concluded that several conditions should be considered in order to achieve acceptable results. These conditions are indeed extensions of the notion of spatial homogeneity.

Unfortunately, source zones having homogeneous distributions seldom occur in the real world. Non-homogeneity arises mainly from the fact that most thematic maps are only generalizations of very detailed investigations made on individual samples. The size of the samples and the method of sampling become important in determining the quality and accuracy of the thematic map. Very often the source zones were originally delineated for other purposes and may not represent the most important information for the target zones. Moreover, imperfect knowledge of the spatial distribution of the phenomenon and the assignment of values or identifiers to zones may produce imprecise zone definitions. Finally, other technical problems involved in the process of transferring the map from graphic to digital format, such as digitization errors and generalization errors, should not be neglected.

The *pycnophylactic* interpolation method was originally suggested by Tobler (1979) for isopleth mapping. The method assumes the existence of a smooth density function which takes into account the effect of adjacent source zones. The density function to be found must have the pycnophylactic, or volume-preserving, property, which can be defined in the following discrete way. Let p_k be the population of zone k, A_k the area of zone k, z_{ij} the density in cell ij, and α the area of a cell. Set q^k_{ij} equal to 1

if ij is in zone k; otherwise set it at 0. Then, the pycnophylactic conditions include:

$$\sum_{ij} \alpha\, z_{ij} q^k_{ij} = p_k,$$

$$\sum_{ij} \alpha\, q^k_{ij} = A_k, \qquad (18)$$

and

$$\sum_{k} q^k_{ij} = 1.$$

The smooth density function can be obtained by minimizing the sum of the squares of the partial derivatives,

$$\iint \left(\left(\frac{\delta z}{\delta x} \right)^2 + \left(\frac{\delta z}{\delta y} \right)^2 \right) \delta x \delta y, \qquad (19)$$

which is also called the Dirichlet's integral. Without the pycnophylactic and the non-negativity constraints, the minimum is given by the LaPlace equation and the finite approximation of this equation is the same as equation (9), which requires the values of any grid point to approach the averages of its four neighbors. Other smoothing conditions may be used depending on the type of application.

The interpolation procedure begins by assigning the mean density to each grid cell superimposed on the source zones, and then modifies this by a slight amount to bring the density closer to the value required by the governing partial differential equation. The volume-preserving condition is then enforced by either incrementing or decrementing all of the densities within individual zones after each computation. Since the assignment of values outside the study area will affect the measure of smoothness near the edge and consequently inward, the selection of a boundary condition should be based on the specific type of application and on the information available for the areas outside. In general, two types of boundary conditions can be specified. The first, known as the Dirichlet condition, specifies a numerical value for cells along the boundary. For example, those cells which fall outside of the study area can be assigned a density of zero when dealing with a study area bounded by water. The other, called the Neuman condition, requires the specification of the rate of change of the densities across the boundary. These two constraints can be used simultaneously if desired. The interpolation procedures are illustrated with a small example in the appendix.

Compared with the polygon overlay method, the pycnophylactic method represents a conceptual improvement since the effects of neighboring source zones have been taken into account. Moreover, homogeneity within zones is not required. However, the smooth density function imposed is again only a hypothesis about the surface and does not necessarily apply to many real cases. The choice of an appropriate smooth density function and of a boundary condition thus depends heavily on the characteristics of the surface and individual applications. In some cases, the smooth density function may not apply globally to the whole surface and some side conditions may be included. Tobler (1979) has suggested a number of possibilities that could be used as target equations other than the smoothing criterion. They are also related to some long-established geographical theories, such as Christaller's central place theory, and may have more realistic significance. Another possibility is to utilize the existing information on the source zones in the estimation procedures, such as the variogram function used in Kriging or the spatial autocorrelation characteristics; these may lead to a better result.

The quality of the pycnophylactic estimates also depends on two factors. First of all, the cell size should be sufficiently small to warrant both the smoothness and the volume-preserving conditions. Second, the real-world example used by Lam (1980) has shown that some of the source zones cannot maintain the two properties simultaneously even with a fine lattice. These source zones were found to have large variations of values within zones. Although the pycnophylactic method does

not require homogeneity within zones, rapid variations of values within zones seem to influence the quality of the estimates.

Comparisons between the point-based and the area-based approaches have been made by using a real example (Lam 1980). In general, judging from the theoretical and the limited empirical bases, the latter is far more desirable than the former because of the volume-preserving property of the latter. Within the group of area-based methods, the overlay method does not consider the smoothness of the changes of values between zones while assuming homogeneity within, whereas the pycnophylactic method imposes smoothness on the interpolated grid values without requiring within-zone homogeneity. These two methods can be linked together as the two ends of a continuum between a discontinuous and a maximally smooth density surface. There should be some real-world cases where reliable interpolation occurs somewhere along the continuum, such as by imposing only a certain degree of smoothness of the density surface but not as much as the pycnophylactic method does, or by including some side conditions. In choosing between these two methods, one must consider the underlying structure of the surface as well as the methods by which the zones are delineated.

CONCLUSIONS

The problem of spatial interpolation has long been recognized by a variety of disciplines. Although the interpolation of point data has been studied extensively, areal interpolation has seldom been examined. The review of point interpolation has shown that various methods have individual advantages and disadvantages, and the choice of an interpolation model depends largely on the type of data, the degree of accuracy desired, and the amount of computational effort afforded. In general, exact or piecewise methods are more reliable than approximate or global methods because of the former's simplicity, flexibility, and reliability. The former are represented by most weighting methods, Kriging, and spline interpolation, and the latter are represented by trend-surface models. In all cases, point interpolation models are seriously affected by the quality of the original data, especially the density and the spatial arrangement of data points, and the complexity of the surface.

Areal interpolation is subject to other sources of error because of areal aggregation. The quality of the areal interpolation estimates depends largely on how the source and target zones are defined, the method of data collection, the degree of generalization or method of aggregation, and the characteristics of the partitioned surface. It is shown from both theoretical and limited empirical evidence that the area-based, or volume-preserving, approach is more reliable than the traditional point-based, or non-volume-preserving, approach. Overlay and pycnophylactic methods represent different models for a statistical surface, and it is expected that the overlay method will yield better estimates if the surface is discontinuous, whereas the pycnophylactic method gives better results when smoothness is a real property of the surface. In cases where the surface is intermediate between discontinuous and maximally smooth, different target equations and side conditions should be imposed for reliable results, but such methods have not yet been developed.

REFERENCES

Ahlberg, J. H. 1970. Spline approximation and computer-aided design. *Advances in Computers* 10:275–89.

Armstrong, M., and Jabin, R. 1981. Variogram models must be positive-definite. *Mathematical Geology* 13, no. 5:455–59.

Bassett, K. 1972. Numerical methods for map analysis. *Progress in Geography* 4:217–254.

Bhattacharyya, B. K. 1969. Bicubic spline interpolation as a method for treatment of potential field data. *Geophysics* 34, no. 3:402–23.

Bhattacharyya, B. K., and Holroyd, M. R. 1971. Numerical treatment and automatic mapping of two-dimensional data in digital form. *Proceedings, 9th International Symposium of Techniques for Decision-making in the Mineral Industry,*

The Canadian Institute of Mining and Metallurgy, Special Volume 12, pp. 148–58.
Briggs, I. C. 1974. Machine contouring using minimum curvature. *Geophysics* 39:39–48.
Casetti, E. 1966. Analysis of spatial association by trigonometric polynomials. *Canadian Geographer* 10:199–204.
Cavendish, J. C. 1974. Automatic triangulation of arbitrary planar domains for the finite element method. *International Journal for Numerical Methods in Engineering* 8:679–96.
Chorley, R. J., and Haggett, P. 1965. Trend surface mapping in geographical research. *Institute for British Geographers, Transactions* 37:47–67.
Coulson, M. R. C. 1978. Potential for variation: A concept for measuring the significance of variations in size and shape of areal units. *Geografiska Annaler* 60B, no. 1:48–64.
Crackel, T. 1975. The linkage of data describing overlapping geographical units—A second iteration. *Historical Methods Newsletter* 8, no. 3: 146–150.
Crain, I. K. 1970. Computer interpolation and contouring of two-dimensional data: A review. *Geoexploration* 8:71–86.
Crain, I. K., and Bhattacharyya, B. K. 1967. Treatment of non-equispaced two-dimensional data with a digital computer. *Geoexploration* 5:173–94.
Curry, L. 1966. A note on spatial association. *Professional Geographer* 18:97–99.
David, M. 1976. The practice of Kriging. In *Advanced geostatistics in the mining industry,* ed. M. Guarascio, M. David, and C. Huijbregts, pp. 33–48. Dordrecht, Holland: Reidel.
———. 1977. *Geostatistical ore reserve estimation.* New York: Elsevier.
De Boor, C. 1976. Splines as linear combinations of B-splines, a survey. In *Approximation theory II,* ed. G. Lorentz, and others, pp. 1–47. New York: Academic Press.
Delfiner, P. 1976. Linear estimation of non-stationary spatial phenomena. In *Advanced geostatistics in the mining industry,* ed. M. Guarascio, M. David, and C. Huijbregts, pp. 49–68. Dordrecht, Holland: Reidel.
Delfiner, P., and Delhomme, J. P. 1975. Optimum interpolation by Kriging. In *Display and analysis of spatial data,* ed. J. C. Davis and M. J. McCullagh, pp. 97–114. Toronto: Wiley.
Ford, L. 1976. Contour reaggregation: another way to integrate data. *Papers, Thirteenth Annual URISA Conference* 11:528–75.
Giloi, W. K. 1978. *Interactive computer graphics.* Englewood Cliffs: Prentice-Hall.
Gold, C. M., Charters, T. D., and Ramsden, J. 1977. Automated contour mapping using triangular element data structures and an interpolant over each triangular domain. *Computer Graphics* 11, no. 2:170–75.
Goodchild, M. F. 1979. The theory of Kriging: Review and interpretation. Unpublished manuscript, Department of Geography, University of Western Ontario, 21 pp.
Goodchild, M. F., and Lam, N. S. 1980. Areal interpolation: A variant of the traditional spatial problem. *Geo-Processing* 1:297–312.
Gordon, W. J. 1969. Spline-blended surface interpolation through curve networks. *Journal of Mathematics and Mechanics* 18, no. 10:931–52.
———. 1971. Blending-function methods of bivariate and multivariate interpolation and approximation. *SIAM Journal for Numerical Analysis* 8:158–77.
Guarascio, M., David, M., and Huijbregts, C., eds. 1976. *Advanced geostatistics in the mining industry.* Dordrecht, Holland: Reidel.
Harbaugh, J. W., and Preston, F. W. 1968. Fourier analysis in geology. In *Spatial analysis: A reader in statistical geography,* ed. B. J. Berry and D. F. Marble, pp. 218–38. Englewood Cliffs: Prentice-Hall.
Harbaugh, J. W., Doveton, J. H., and Davis, J. C. 1977. *Probability methods in oil exploration.* New York: Wiley-Interscience.
Hayes, J. G., and Halliday, J. 1974. The least-squares fitting of cubic spline surfaces to general data sets. *Journal of the Institute of Mathematics and its Application* 14:89–103.
Hessing, R. C., Lee, H. K., Pierce, A., and Powers, E. 1972. Automatic contouring using bicubic functions. *Geophysics* 37:668–74.
Hsu M.-L., and Robinson, A. H. 1970. *The fidelity of isopleth maps—An experimental study.* Minneapolis: University of Minnesota Press.
Huijbregts, C., and Matheron, G. 1971. Universal Kriging. *Proceedings, 9th International Symposium on Techniques for Decision-making in the Mineral Industry,* The Canadian Institute of Mining and Metallurgy, special volume 12, pp. 159–69.
Jenks, G. F., and Caspall, F. C. 1969. The construction of choroplethic maps for use as analytical tools. Unpublished manuscript, University of Kansas.
Jenks, G. F., Caspall, F. C., and Williams, D. L. 1969. The error factor in statistical mapping. *Annals, Association of American Geographers* 59:186–87.
Krige, D. G. 1976. A review of the development of geostatistics in South Africa. In *Advanced geostatistics in the mining industry,* eds. M. Guarascio, M. David, and C. Huijbregts, pp. 279–93. Dordrecht, Holland: Reidel.
Krumbein, W. C. 1959. Trend surface analysis of contour-type maps with irregular control-point spacing. *Journal of Geophysical Research* 64:823–34.
Kubik, K., and Botman, A. G. 1976. Interpolation accuracy for topographic and geological surfaces. *ITC Journal* 4:236–73.
Lam, N. S. 1980. Methods and problems of areal interpolation. Ph.D. dissertation, University of Western Ontario.
———. 1981. The reliability problem of spatial interpolation models. *Modeling and Simulation* 12:869–76.
Lancaster, P., and Salkauskas, K. 1975. An introduction to curve and surface fitting. Unpub-

lished manuscript, Division of Applied Mathematics, University of Calgary, 114 pp.
Leberl, F. 1975. Photogrammetric interpolation. *Photogrammetric Engineering and Remote Sensing* 41, no. 5:603–12.
Mackay, J. R. 1951. Some problems and techniques in isopleth mapping. *Economic Geography* 27:1–9.
———. 1953. The alternative choice in isopleth interpolation. *The Professional Geographer* 5:2–4.
Markoff, J., and Shapiro, G. 1973. The linkage of data describing overlapping geographical units. *Historical Methods Newsletter* 7, no. 1:34–36.
Matheron, G. 1963. Principles of geostatistics. *Economic Geology* 58:1246–66.
———. 1967. Kriging, or polynomial interpolation procedures? *Canadian Mining and Metallurgy Bulletin* 60, no. 665:1041–45.
———. 1971. *The theory of regionalized variables and its application.* Les Cahiers du Centre de Morphologie Mathematique de Fontainebleau, vol. 5, 221 pp.
McAlpine, J. R., and Cook, B. G. 1971. Data reliability from map overlay. Unpublished manuscript, Division of Land Research, CSIRO, Canberra, Australia.
McCullagh, M. J. 1975. Estimation by Kriging of the reliability of the proposed Trent telemetry network. *Computer Applications* 2:357–374.
McCullagh, M. J., and Ross, C. G. 1980. The Delauney triangulation of a random data set for isarithmic mapping. *Cartographic Journal* 17:93–99.
McLain, D. H. 1974. Drawing contours from arbitrary data points. *The Computer Journal* 17:318–24.
———. 1980. Interpolation methods for erroneous data. In *Mathematical Methods in Computer Graphics and Design,* ed. K. W. Brodlie, pp. 87–104. New York: Academic Press.
Miesch, A. T., and Connor, J. J. 1968. *Stepwise regression and non-polynomial models in trend analysis.* Kansas Geological Survey Computer Contribution, no. 27, 40 pp.
Morrison, J. 1971. *Method-produced error in isarithmic mapping.* Technical Monograph, no. CA-5, first edition, Washington, D.C.: American Congress on Surveying and Mapping.
———. 1974. *The accuracy of isometric maps.* Final Report, U.S. Army Research Office Contract DA-ARO-D-31-124-71-G45.
Mousset-Jones, P. F., ed. 1980. *Geostatistics.* New York: McGraw-Hill.
Norcliffe, G. B. 1969. On the use and limitations of trend surface models. *Canadian Geographer* 13:338–48.
Olea, R. A. 1974. Optimal contour mapping using Universal Kriging. *Journal of Geophysical Research* 79, no. 5:695–702.
———. 1975. *Optimum mapping techniques using regionalized variable theory.* Kansas Geological Survey, Series on Spatial Analysis, no. 2.
Peucker, T. K. 1972. *Computer cartography.* Commission on College Geography Resource Paper no. 17, Washington, D.C.: Association of American Geographers.
Peucker, T. K., Fowler, R. J., Little, J. J., and Mark, D. M. 1976. *Digital representation of three-dimensional surfaces by triangulated irregular network.* Technical report no. 10, Office of Naval Research.
Porter, P. 1958. Putting the isopleth in its place. *Proceedings, Minnesota Academy of Science* 25/26:372–84.
Ralston, A. 1965. *A first course in numerical analysis.* New York: McGraw-Hill.
Rendu, J. M. 1970. Geostatistical approach to ore reserve calculation. *Engineering and Mining Journal,* June 1970:114–18.
Rhind, D. W. 1975. A skeletal overview of spatial interpolation. *Computer Applications* 2, no. 3/4: 293–309.
Sampson, R. J. 1978. *Surface II graphics system.* Lawrence: Kansas Geological Survey.
Schmid, C. F., and MacCannell, E. J. 1955. Basic problems, techniques, and theory of isopleth mapping. *Journal of the American Statistical Association* 50:220–39.
Schumaker, L. L. 1976. Fitting surfaces to scattered data. In *Approximation theory II,* ed. G. Lorentz, and others, pp. 203–67. New York: Academic Press.
Schut, G. 1976. Review of interpolation methods for digital terrain models. *The Canadian Surveyer* 30:389–412.
Shepard, D. S. 1970. *A two-dimensional interpolation function for computer mapping of irregularly spaced data.* Papers in Theoretical Geography, Technical report 15, Laboratory for Computer Graphics and Spatial Analysis. Cambridge, Mass., Harvard University.
Späth, J. 1974. *Spline algorithms for curves and surfaces,* transl. W. D. Hoskins and H. W. Sager. Winnipeg: Utilitas Mathematica.
Stearns, R. 1968. A method for estimating the quantitative reliability of isoline maps. *Annals, Association of American Geographers* 58: 590–600.
Swain, C. 1976. A Fortran IV program for interpolating irregularly spaced data using difference equations for minimum curvature. *Computers and Geosciences* 1:231–40.
Tobler, W. R. 1979. Smooth pycnophylactic interpolation for geographic regions. *Journal of the American Statistical Association* 74, no. 367: 519–536.
Tobler, W. R., and Lau, J. 1978. Isopleth mapping using histoplines. *Geographical Analysis* 10, no. 3:272–79.
———. 1979. Interpolation of images via histoplines. *Computer Graphics and Image Processing* 9:77–87.
Torelli, L. 1975. Modern techniques of trend analysis and interpolation. *Annali di Geofisica* 28:271–77.
Unwin, D. 1975. *An introduction to trend surface analysis.* CATMOG, no. 5. Norwich: GeoAbstracts, University of East Anglia.

Walters, R. F. 1969. Contouring by machine: A user's guide. *Bulletin, American Association of Petroleum Geologists* 53:2324–40.

Warntz, W. 1966. The topology of a socio-economic terrain and spatial flows. *Papers of the Regional Science Association* 17:47–61.

Wren, A. E. 1973. Trend surface analysis—A review. *Canadian Society of Exploration Geophysics Journal* 9:39–44.

———. 1975. Contouring and the contour map, a new perspective. *Geographical Prospecting* 23:1–17.

Zurflueh, E. G. 1967. Applications of two-dimensional linear wavelength filtering. *Geophysics* 32:1015–35.

APPENDIX

This appendix illustrates with a brief worked example the general procedures involved in interpolation using some of the methods discussed in the paper. Since it is impossible to illustrate all of them here, only those methods which seem to have a great potential for cartographic applications are shown. They include 1) distance-weighting, kriging, bicubic spline, and trend surface for point interpolation, and 2) overlay and pycnophylactic for areal interpolation.

Point Interpolation

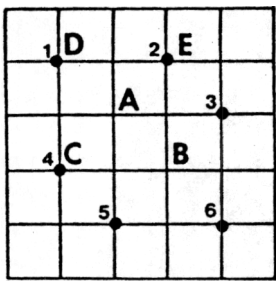
Consider a simple surface designed by a 5 × 5 matrix, with the values for six sample points known (Figure 4a). The point interpolation problem is to determine the values for those grid points whose values are not given.

For the *distance-weighting* method, if the inverse distance function $w = d^{-1}$ is used, then using equation (3), the value for point A at (2, 3) becomes

$$z(2,3) = \frac{(1.4)^{-1}40 + (1.4)^{-1}24 + \ldots + (2.8)^{-1}25}{(1.4)^{-1} + (1.4)^{-1} + \ldots + (2.8)^{-1}} = 29.4 \quad (20)$$

To calculate the *simple Kriging* estimate of point A, the semivariogram function of the surface has to be determined by using equation (4). For example, since there are two pairs of sample points having a distance of 1.4 units, the semivariogram function value for this distance is

(a)



	x	y	z
1		4	40
2	3	4	24
3	4	3	30
4	1	2	24
5	2		32
6	4		25

(b)

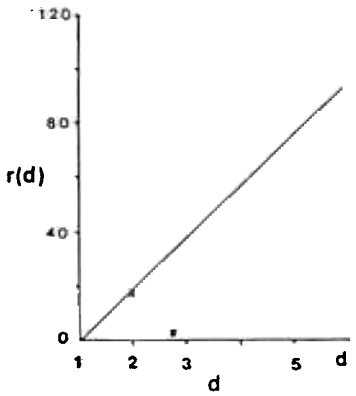


Figure 4. Hypothetical surface and its variogram.

$$r(1.4) = \tfrac{1}{2}\left[\frac{(24-30)^2 + (24-32)^2}{2}\right] = 25. \quad (21)$$

The semivariogram function values for other distances are calculated in a similar way; they are plotted in Figure 4b. One may derive from the points in the variogram a theoretical distribution of the whole surface. Notice that the final variogram used must be a monotonically increasing continuous function of distance, otherwise it may result in the calculation of negative estimation variances for certain points on the surface (Armstrong and Jabin 1981). The most widely used models for variogram are linear, spherical, and exponential (Mousset-Jones 1980; Olea 1975). For the sake of brevity, a linear model, such

as $r(d) = bd$, where b is the slope, is used here. A regression line is fitted through the origin and the slope is found to equal 14.83.

The next step is to apply the following system of linear equations to solve for the weights λ_j:

$$\sum_{j=1}^{n} \lambda_j \, r(d_{ij}) + u = r(d_{iA}) \qquad (22)$$

for all $i = 1, n$, where $r(d_{ij})$ is the semivariogram function value for the distance between sample point i and j, and u is the Lagrange multiplier. In matrix form:

$$\begin{Bmatrix} \lambda_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \lambda_n \\ u \end{Bmatrix} = \begin{bmatrix} r(d_{11}) & \cdots\cdots & r(d_{1n}) & 1 \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ r(d_{n1}) & \cdots\cdots & r(d_{nn}) & 1 \\ & \cdots\cdots & & \\ 1 & & 1 & 0 \end{bmatrix}^{-1} \times \begin{Bmatrix} r(d_{1A}) \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ r(d_{nA}) \\ 1 \end{Bmatrix} \qquad (23)$$

In this example,

$$\begin{Bmatrix} \lambda_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \lambda_6 \\ u \end{Bmatrix} = \begin{bmatrix} 0 & 29.7 & 47.5 & 29.7 & 47.5 & 62.3 & 1 \\ 29.7 & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ \cdot & & & & & & \cdot \\ 62.3 & 47.5 & 29.7 & 47.5 & 29.7 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}^{-1} \times \begin{Bmatrix} 20.8 \\ 20.8 \\ 29.7 \\ 20.8 \\ 29.7 \\ 41.5 \\ 1 \end{Bmatrix} = \begin{Bmatrix} 0.26 \\ 0.27 \\ 0.09 \\ 0.27 \\ 0.09 \\ 0.03 \\ -5.35 \end{Bmatrix}. \qquad (24)$$

Once the λ's are found, the kriging estimate of A can be calculated by simply applying equation (5); A is found to be equal to 29.69. In addition, the estimate of variance at point A is given by

$$u + \sum_{i=1}^{n} \lambda_i r(d_{iA}) = -5.35 + (0.26)20.8 + \ldots + (0.03)41.5 = 17.88. \qquad (25)$$

Several steps are involved in calculating the *bicubic spline* estimate of point A. The procedures for calculating bicubic splines for rectangular surface patches are presented here. For example, to interpolate the value at point A, the bicubic polynomials for the rectangle BCDE enclosing A have to be calculated first. (The value at B is assumed to be 36). These polynomials, which intersect the sample points and also are twice differentiable, that is, smooth across surface patches, are defined by

$$f_{ij}(x,y) = \sum_{k,l=1}^{4} a_{ijkl}(x - x_i)^{k-1}(y - y_i)^{l-1} \qquad (26)$$

for $i = 1, \ldots, n-1, j = 1, \ldots, m-1$, where a_{ijkl} are the coefficients to be determined. In matrix form,

$$a_{ij} = [G(x_i)]^{-1} * S_{ij} * [G(y_j)^T]^{-1} \qquad (27)$$

In bicubic splines, the matrices $G(x_i)$, $G(y_i)$ and their inverses are:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & h^2 & h^2 & h^3 \\ 0 & 1 & 2h & 3h^2 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -3/h^2 & -2/h & -3/h^2 & -1/h \\ 2/h^3 & 1/h^2 & -2/h^3 & 1/h^2 \end{bmatrix} \quad (28)$$

where $h = (x_{i+1} - x_i)$ or $(y_{i+1} - y_i)$. S_{ij} are the matrices consisting of the following elements:

$$S_{ij} = \begin{bmatrix} u_{ij} & q_{ij} & u_{i,j+1} & q_{i,j+1} \\ P_{ij} & r_{ij} & P_{i,j+1} & r_{i,j+1} \\ u_{i+1,j} & q_{i+1,j} & u_{i+1,j+1} & q_{i+1,j+1} \\ P_{i+1,j} & r_{i+1,j} & P_{i+1,j+1} & r_{i+1,j+1} \end{bmatrix}. \quad (29)$$

u_{ij} are the function values given at point ij. p_{ij}, q_{ij}, r_{ij} are the first derivatives along the x, y, and xy dimensions, respectively; they can be calculated by the following sets of equations:

$$\frac{1}{\Delta x_{i+1}} p_{i-1,j} + (2 + p)\left(\frac{1}{\Delta x_{i-1}} + \frac{1}{\Delta x_i}\right) p_{ij} + \frac{1}{\Delta x_i} p_{i+1,j} = (3 + p)\left(\frac{u_{ij} - u_{i+1,j}}{(\Delta x_{i-1})^2} + \frac{u_{i+1,j} - u_{ij}}{(\Delta x_i)^2}\right) \quad (30)$$

for $i = 2, \ldots, n - 1, j = 1, \ldots, m$,

$$\frac{1}{\Delta y_{j-1}} q_{i,j-1} + (2 + p)\left(\frac{1}{\Delta y_{j-1}} + \frac{1}{\Delta y_j}\right) q_{ij} + \frac{1}{\Delta y_j} q_{i,j+1} = (3 + p)\left(\frac{u_{ij} - u_{i,j-1}}{(\Delta y_{j-1})^2} + \frac{u_{i,j+1} - u_{ij}}{(\Delta y_j)^2}\right) \quad (31)$$

for $i = 1, \ldots, n; j = 2, \ldots, m - 1$,

$$\frac{1}{\Delta x_{i-1}} r_{i-1,j} + (2 + p)\left(\frac{1}{\Delta x_{i-1}} + \frac{1}{\Delta x_i}\right) r_{ij} + \frac{1}{\Delta x_i} r_{i+1,j} = (3 + p)\left(\frac{q_{ij} - q_{i-1,j}}{(\Delta x_{i-1})^2} + \frac{q_{i+1,j} - q_{ij}}{(\Delta x_i)^2}\right) \quad (32)$$

for $i = 2, \ldots, n - 1; j = 1, m$, and

$$\frac{1}{\Delta y_{j-1}} r_{i,j-1} + (2 + p)\left(\frac{1}{\Delta y_{j-1}} + \frac{1}{\Delta y_j}\right) r_{ij} + \frac{1}{\Delta y_j} r_{i,j+1} = (3 + p)\left(\frac{p_{ij} - p_{i,j-1}}{(\Delta y_{j-1})^2} + \frac{p_{i,j+1} - p_{ij}}{(\Delta y_j)^2}\right) \quad (33)$$

for $i = 1, \ldots, n; j = 2, \ldots, m - 1$.

Notice that these equations are tridiagonal and differ only on the right-hand sides. Once p_{ij}, q_{ij}, and r_{ij} are solved, equation (27) can be used for calculating the coefficients and equation (26) for interpolating unknown points.

In short, the following values are required for bicubic spline interpolation. The corresponding values in this example are also given below:

	r_{1m}	q_{1m}	····	q_{nm}	r_{nm}
y_n	p_{1m}	u_{1m}	····	u_{nm}	p_{nm}
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
y_1	p_{11}	u_{11}	····	u_{n1}	p_{n1}
	r_{11}	q_{11}	····	q_{n1}	r_{n1}
		x_1	·····	x_n	

	0	8	−6	0
4	−8	40	24	−8
2	6	24	36	6
	0	8	−6	0
		1	3	

(34)

Different boundary conditions can be used. In this example, p_{ij}, and q_{ij}, are found by difference approximation, for example, $p_{11} = (36 - 24)/2 = 6$. The r_{ij} are assumed to be 0. Then, according to equation (27), the coefficients a_{ijkl} are equal to

$$a_{12} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -0.75 & -1 & 0.75 & -0.5 \\ 0.25 & 0.25 & -0.25 & 0.25 \end{bmatrix}$$

$$\times \begin{bmatrix} 24 & 8 & 40 & 8 \\ 6 & 0 & -8 & 0 \\ 36 & -6 & 24 & -6 \\ 6 & 0 & -8 & 0 \end{bmatrix}$$

$$\times \begin{bmatrix} 1 & 0 & -0.75 & 0.25 \\ 0 & 1 & -1 & 0.25 \\ 0 & 0 & 0.75 & -0.25 \\ 0 & 0 & -0.5 & 0.25 \end{bmatrix}$$

$$= \begin{bmatrix} 24 & 8 & 0 & 0 \\ 6 & 0 & -10.5 & 3.5 \\ 0 & -10.5 & 15.75 & -5.25 \\ 0 & 3.5 & -5.25 & 1.75 \end{bmatrix} \quad (35)$$

According to equation (26), the value at A becomes

$$f_{12}(2,3) = a_{11}(2-1)^0(2-1)^0 + a_{12}(2-1)^0(2-1)^1 + \ldots + a_{44}(2-1)^3(2-1)^3 = 31.0 \quad (36)$$

To estimate the value at point A by means of power-series *trend-surface* model, we need to solve for the set of normal equations. Assuming a first-degree trend is used, that is, $Z^* = a_0 + a_1x + a_2y$, the normal equations to be solved are

$$a_0N + a_1\sum x + a_2\sum y = \sum z \quad (37)$$

$$a_0\sum x + a_1\sum x^2 + a_2\sum xy = \sum zx \quad (38)$$

$$a_0\sum y + a_1\sum xy + a_2\sum y^2 = \sum zy. \quad (39)$$

In this example:

$$a_0(6) + a_1(15) + a_2(15) = 175 \quad (40)$$

$$a_0(15) + a_1(47) + a_2(36) = 420 \quad (41)$$

$$a_0(15) + a_1(36) + a_2(47) = 451. \quad (42)$$

After simple algebraic manipulation, a_0, a_1, a_2 are found to be equal to 34.4, −1.8, −0.3. The value at point A therefore is

$$Z_A = 34.4 + (-1.8)2 + (-0.3)3 = 29.9. \quad (43)$$

Areal Interpolation

Consider a hypothetical surface which is partitioned into two different sets of areal units, as shown in Figure 5. Given the boundaries and population values for source zones and the target zone boundaries, the areal interpolation problem is to estimate the population values for each target zone from these source zones.

To obtain an *overlay* estimate for target zone D, for example, simply overlay the two sets of zones, find the area of intersection of each resultant polygon (Figure 5), and then apply equation (15):

$$V_D = \sum_s U_s a_{is}/\sigma_s = 10 \times 4/6 + 40 \times 2/6 = 20.0. \quad (44)$$

The *pycnophylactic* estimates for target zones can be found by the following algorithm. 1) Superimpose a mesh of grids (4 × 4 in this example) on the source zones. 2) Assign the mean population density of the source zone to each grid within the zone (Figure 6a). 3) In each iteration, change each grid cell value into the average of its four neighbors as specified in equation (9). Neighbors outside the boundary of the study area can be assigned to 0 or other values. In this example, they are simply not taken into account for averaging. Hence, the value of cell (1, 2) becomes (1.67 + 1.67 + 5.00)/3 = 2.78. Figure 6b shows the modified grid values after this step. 4) Add all the grid values in

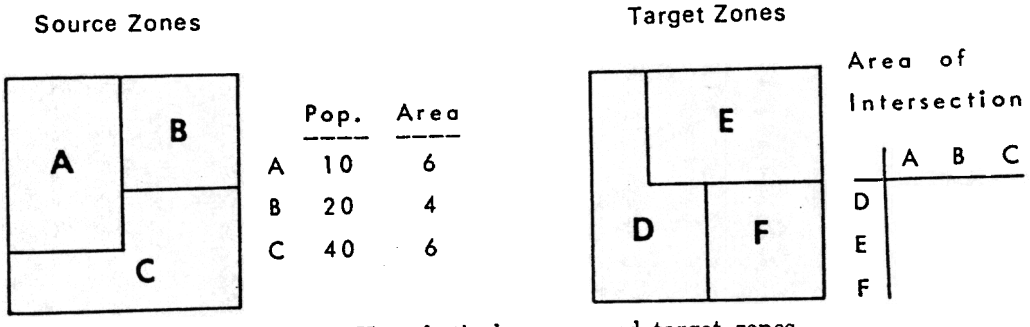


Figure 5. Hypothetical source and target zones.

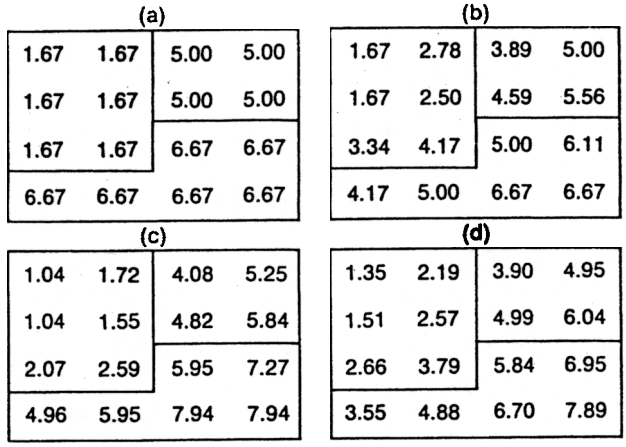


Figure 6. Pycnophylactic interpolation.

each zone and convert to population values. 5) Compare the actual source zone values (P_k) with the predicted ones (P_k^*) and adjust the grid cell values by multiplying by the ratio between them, (P_k/P_k^*). For example, source zone A has a predicted population value of 16.13 after step (4). The new value for cell (1, 1) becomes $d_{ij} \times P_k/P_k^* = 1.67 \times (10/16.13) = 1.04$ (Figure 6c). This step enforces the pycnophylactic condition, whereas step (3) enforces the smoothing condition. 6) The process repeats until either there is no significant difference between the actual and the predicted population values or until there are no significant changes of grid values compared with the last iteration. Figure 6d gives the final grid values after 10 interactions. 7) Finally, simply aggregate the grid cells into target zone boundaries and sum the grid values. Target zone D in this case has a value of 17.74.

Tobler (1979) used a different algorithm for pycnophylactic interpolation. Compared with the above algorithm, Tobler's algorithm is more complicated but is believed to provide a faster convergence. Notice that the example given here is solely for demonstrating the general procedures involved in pycnophylactic interpolation. In real applications, a much finer lattice should be used to assure the maintenance of both the pycnophylactic and the smoothing conditions.

• **AM/FM International.** A new, not-for-profit, educational institution has been formed for persons interested in utility mapping, distribution engineering, city and county mapping, geographic facilities management, and other applications of computer graphics and database systems to manage spatial data. AM/FM International, short for Automated Mapping and Facilities Management International, is concerned principally with information exchange. It plans to publish a newsletter and to offer conferences and workshops. For further information, write:

AM/FM International
5680 South Big Canon Drive
Englewood, CO 80111

[Source: Brimmer Sherman]

• **Delaware Valley Map Society.** Formed in May 1983, the Delaware Valley Map Society is an organization for all persons in the Greater Philadelphia area interested in maps, ancient or modern. Neophytes or experts are welcome. Meetings will include informal discussions, lectures, and trips to sites of interest. For further information, write:

Delaware Valley Map Society
33 Benezet Street
Philadelphia, PA 19118

[Source: David J. Cuff]

• **News Deadline.** News items to be included in the April 1984 issue must be received by the Editor no later than December 15, 1983.

[Source: *Ed.*]