20. Radial basis functions and archaeological surfaces

J.G.B. Haigh
Department of Mathematics, University of Bradford, Bradford. BD7 1DP, U.K.

20.1 Introduction

The ability to recreate surfaces is an increasingly important requirement of modern archaeology. Examples of such applications occur in the examination of physical topography in the vicinity of a site, in the detailed surface survey of a site to provide evidence of earlier structure, and in the interpretation of geophysical and other quantitative surveys. In all these examples the aim is to construct a surface above a reference plane, which can generally be taken to be a local horizontal plane and be represented by a geographical map or plan. There are, however, significant variations in the requirements and methodology for different applications.

The standard problem is to construct a continuous surface from a finite and discrete set of data. In topographical survey, the discrete data relate to an actual physical surface, and the aim is to reproduce that surface as accurately as possible, since errors could have unpredictable effects at a later stage of the application. In other forms of quantitative survey, the discrete data may represent the totality of available information, and the surface is constructed primarily as an aid to visualising and interpreting the data; in such circumstances, the main consideration is not of accuracy but of utility in interpretation.

Another important variation is in the distribution of the data stations over the reference plane. If the stations are situated at the vertices of a regular grid, as is usually the case in geophysical survey, then a number of standard, reliable methods are available to interpolate between them and define a function of two variables from which the surface may be constructed. If the stations are irregularly distributed over the reference plane, however, the methodology is much less well defined; a large number of techniques are available, most of which can be expected to work well in certain circumstances, but few of which are reliable in all situations.

The present paper is mainly concerned with the reconstruction of a topographical surface (a digital terrain model or DTM) from a contour map of the relevant area. This problem is of particular interest to the author (Haigh 1989), since a DTM is required for the accurate rectification of aerial photographs of uneven terrain. It is inconvenient to use a regular grid of stations, when heights are readily available only at points on the recorded contours on the map. Consequently, the DTM must be constructed from an irregular distribution of points selected on the available contours.

As a contrasting problem, consideration will be given to the construction of a continuous distribution from an irregular set of samples over the surface of a bounded site. Such a distribution might represent the density of some class of artefact, and it may be assumed that the density becomes zero beyond the site boundaries. In this problem, the emphasis is on the satisfactory representation of the data, and accuracy is not a prime requirement. The two problems represent opposite extremes of requirement, and a method which has been tested on both should be applicable to many intermediate problems.

It must be emphasised that this paper is concerned with the mathematical construction of surfaces, and not with their subsequent graphical representation. The choice of representation — contours, grey scales, coloured scales, dot densities, wire-frame diagrams, or synthetic illumination — is largely a matter of personal taste, moderated by available computing power. Many graphical techniques cannot work directly from an irregular data set, but require a regular grid of values to be created as an intermediate step. In the problem of constructing DTMs, graphical representation is largely irrelevant, the prime consideration being the accuracy of results from the rectification of aerial photographs.

20.2 The mathematical problem

The main problem to be considered in this paper is: Given a set of n data points (x_i, y_i), $i = 1,\ldots, n$, irregularly distributed over some region of the Euclidean plane, and a dependent data value u_i associated with each point, how can the data be interpolated to associate a dependent value with each point of the region? In other words: How can the function f be constructed to define the dependent variable

$$u = f(x,y)$$

over the whole region, so that

$$u_i = f(x_i, y_i) , \qquad i = 1,\ldots, n,$$

at each of the data points (x_i, y_i)?

In seeking a solution to this problem in an archaeological context, consideration should be given to three different criteria: *reliability*, *realism*, and *robustness*. *Reliability* indicates that the method works in every appropriate case; a method which works well in some cases, but fails in other, apparently similar, cases cannot be satisfactory for an archaeologist whose knowledge of mathematics is limited. *Realism* means that the results should give a satisfactory reproduction of the intended surface; when constructing a DTM, the dependent variable u should reproduce the actual ground height to within required accuracy. *Robustness* indicates that the results should not be sensitive to the precise details of the data set; data sets which are intended to provide similar information about the same surface should give precisely similar results.

The literature contains a large number of attempted solutions to the problem. Broadly speaking, the methods fall into three different categories:

triangulation techniques, *locally fitted functions*, and *global functions*.

For methods in the first category, the plane region of interest is divided into disjoint subregions, and a separate function is defined over each subregion. With irregular data, the most convenient subregion is a triangle, with vertices at three of the data points. For the technique to work satisfactorily, the division into triangular subregions must follow unambiguous criteria. It has been shown (Green & Sibson 1978) that Delaunay triangulation satisfies an optimal set of such criteria; Delaunay triangulation is the dual operation of Dirichlet tessellation, whose results are often known to geographers and archaeologists as Thiessen polygons. It is customary to represent the surface as a plane facet over each triangle, thereby guaranteeing the continuity of the interpolant f, but not of its derivatives. This method is reliable, as it always works, but neither realistic, since the facets can deviate considerably from the true surface, nor robust, since small changes in the data set can cause significant changes in the triangulation. It is possible to fit more complicated splines over each triangular subregion, thereby achieving higher-order continuity, but such methods are quite difficult, and not often used.

The *locally fitted functions* are defined for each data point, fitting their own point and approximating its neighbours. The value of the dependent variable u at an intermediate point is determined as a weighted average of neighbouring local functions. Shepard (1968) has published a method of this type; the author has devised independently a similar method for the construction of DTMs (Haigh 1989). Local function methods are quite reliable and realistic, provided that some care is taken over the choice of the data set, principally by ensuring a fairly uniform spread of data points. However, they are not robust, since a poor spread of data can lead to disastrous results.

The majority of *global function* methods are approximation, rather than interpolation, techniques, since a function with a limited number of parameters is fitted to the data, possibly on the basis of a least-squares criterion. Such methods may not produce a realistic DTM, since it is difficult to devise a mathematical function which can reproduce all the intricacies of a typical landscape. In recent years, however, a global function method for interpolation has emerged which seems worthy of consideration for archaeological surfaces, and the remainder of this paper is devoted to a discussion of the technique.

20.3 Radial basis functions

The original use of radial basis functions for data interpolation is attributed to Hardy (1971), and is described in the following paragraphs.

Select some suitable function $\phi(r)$ of radial distance r from the origin:

$$r = \sqrt{x^2 + y^2}.$$

In place of the argument r, substitute the radial distance from any one (x_j, y_j) of the n data points:

$$r = \| x - x_j \|_2 = \sqrt{(x - x_j)^2 + (y - y_j)^2} ,$$

to give

$$\phi(\| x - x_j \|_2) .$$

In this notation, the vector symbol x represents the coordinate pair (x, y), and the notation $\| \ \|_2$ denotes the *Euclidean square norm*, which is effectively equivalent to the Euclidean distance between two points. Both the notation and the interpolation technique are applicable to spaces of higher dimension, but only the two-dimensional case is discussed here.

The function $f(x)$ is now defined as a linear combination of the n functions $\phi(\|x - x_j\|_2)$ introduced above:

$$f(x) = \sum_{j=1}^{n} \lambda_j \phi(\| x - x_j \|_2) , \qquad (1)$$

where each coefficient λ_j is an undetermined parameter. Since the function f is to be an interpolant, it must take the correct value at every data point. Thus

$$u_i = f(x_i) ,$$

and

$$\sum_{j=1}^{n} \lambda_j \phi(\| x_i - x_j \|_2) = u_i , \qquad i = 1,..., n . \qquad (2)$$

Equation (2) is a system of n linear equations in n unknowns λ_i. In principle, it should possess a unique solution, provided that the $n \times n$ matrix H of values $\phi(\|x_i - x_j\|_2)$ is non-singular; i.e. provided that

$$\det[\phi(\| x_i - x_j \|_2)] \neq 0 .$$

The resultant values λ_i are substituted into equation (1) to give the required interpolant, which has now been expressed on the basis of the radial functions $\phi(r)$, where $r = \phi(\|x - x_j\|_2)$, explaining the use of the term *radial basis functions*.

Having experimented with several different forms for the basis function ϕ, Hardy (1971) declared that the most successful one was the *multiquadric*

$$\phi(r) \equiv \sqrt{r^2 + c^2} ,$$

where c is an adjustable constant. Mathematicians appear not to have taken much interest in radial basis functions until the 1980s, when Franke (1982) re-examined Hardy's results and confirmed empirically the usefulness of the multiquadrics.

Micchelli (1986) extended an earlier result of Schoenberg to prove the *Schoenberg-Micchelli theorem*, which provides a set of sufficient conditions on the function ϕ to ensure that the matrix H is non-singular. Among the functional forms that satisfy the Schoenberg-Micchelli theorem are the following:

$\phi(r) \equiv r$	: linear;
$\phi(r) \equiv r^3$	: cubic;
$\phi(r) \equiv (r^2 + c^2)^{1/2}$	: multiquadric;
$\phi(r) \equiv (r^2 + c^2)^{-1/2}$	: inverse multiquadric;
$\phi(r) \equiv r^2 \ln r$	: Duchon's thin-plate splines;
$\phi(r) \equiv \exp(-r^2/2\sigma^2)$	: gaussian.

The quadratic r^2 is a notable absentee from this list; in fact, together with other even powers, it does not satisfy the conditions of the Schoenberg-Micchelli theorem. A small experiment soon convinced the author that quadratics do not make satisfactory radial basis functions!

The linear function in two dimensions is a cone with its apex at the origin $r = 0$. Consequently its first derivatives have a discontinuity there. The significance of the multiquadric is that the constant c, which can be quite small, removes the discontinuity at the origin, and turns the cone into a hyperboloid. At first sight the multiquadric may seem to be a surprising choice of radial basis function, since it increases indefinitely for large values of r. In fact, the awkward behaviour at large r seems to be the very property which makes it ideally suitable for the creation of surfaces, since the interpolant is dominated by the smoothing influence of more distant data points, with local points providing only a slight correction. Consequently the summed interpolant is a smoothly varying function, tightly fitted to the data points, with little tendency to overshoot or to produce other artefacts of calculation.

20.4 Application to DTMs

The author has used multiquadric basis functions to create DTMs for the rectification of aerial photographs of uneven terrain, replacing the locally fitted functions used in earlier versions of his program (Haigh 1989). He has applied the method to several different sites, and has created several different DTMs for each site, varying the number and distribution of the data points. The effects of removing a few points at random from a data set were studied, as were the effects of adding miscellaneous points. The results were judged mainly on the basis of the accuracy of the rectification, comparing features from different photographs with each other and with the map. Tests were also made by producing raster graphics, showing contours which could be compared with those of the original map. Examination of the results, in the light of the three criteria of section 20.2, showed that the method is very reliable. It produced good results for each of the sites, and no unreasonable precautions were required in preparing the data set.

The results were also realistic. If care was taken to ensure that a particular detail of the landscape was incorporated into the data set, then the detail was apparent in the interpolated DTM. This effect was demonstrated most clearly by showing the contours as raster graphics, but it was also revealed in the successful rectification of features from areas with complex topography. It is only necessary to provide as many data points as are essential to indicate the shapes of the contours. In level areas, with widely spaced and gentle contours, only a few data points need be provided; in hilly areas, where the contours are dense and sharply curved, a large number of data points should be provided. The method has no difficulty in coping with the consequent variation in the density of data points.

The results are also very robust. Applying DTMs based on different data sets has no discernible effect on the rectification, provided that sufficient detail is incorporated into each DTM. If a topographical feature is not apparent in the data set, then photographic details situated there cannot be correctly rectified. When reasonable care is taken in preparing the DTM, the results can be expected to be satisfyingly accurate.

On all three criteria, the results were better than those from any other method of interpolation known to the author, who recommends the method to users of rectification software. As with any numerical method, certain pitfalls must be avoided, some of which are discussed in section 20.6 below, but clear guidelines can be laid down, allowing the general user to achieve satisfactory results in every case.

An incidental advantage of multiquadric basis functions is that the interpolated surface remains valid right up to the limits of the region. There is no tendency suddenly to fall off the edge of the model, as often occurs with other methods. In fact, the model may be extrapolated as a pleasing surface well beyond the bounds of the data set, but the extrapolated surface cannot be relied upon in applications where accuracy is a prime requirement.

The author has experimented with other forms of radial basis function, most notably the inverse multiquadric. The inverse multiquadric, which decreases with increasing r, tends to stretch up to each relatively high data value, and down to each low one. The result is an effect of pimples and dimples, which does not satisfy the realism criterion, and is disastrous for rectification. The pimples and dimples can be counteracted by increasing the value of the parameter c but, unless the data are distributed evenly, it is difficult to find a value of c which is appropriate over the whole region.

20.5 Application to site distributions

When constructing model surfaces for site distributions, the need for the interpolant ultimately to fall to zero must be taken into account. It is possible to ensure that a function constructed from multiquadrics falls to zero, simply by incorporating into the data set a number of zeros at points beyond the bounds of the site. This technique has been used in creating surfaces to represent mathematical functions, but these are often quite simple, and can be constructed from comparatively few data points. Archaeological distributions may be quite complicated and involve large numbers of data points, requiring many zeros to ensure the correct behaviour at large r. Clearly it should be much simpler to use radial basis functions with the correct asymptotic properties, avoiding the need for additional zeros.

From the list in section 20.3, two forms of radial basis function commend themselves: the inverse multiquadric and the gaussian. Although the inverse multiquadric function approaches zero for large r, the trend (inverse distance) is quite slow for practical purposes, and without extra zeros it may be difficult to confine the distribution within the site boundary. After extensive experiments with inverse multiquadrics, the author has failed to produce a site distribution surface with all the

desired properties, and has concluded they are not suitable for this purpose.

Gaussian functions decay to zero much more rapidly — as rapidly as any function in common use. The main problem is to ensure that the parameter σ, which defines the width of the gaussian function, is sufficiently large to sustain the function from a data point to its neighbours, requiring σ to be of the order of the mean nearest neighbour distance. With an appropriate choice of σ, a very satisfactory surface can be achieved, with a nice decay to zero around the site boundary. If too small a value is chosen for σ, then a pimple effect may occur even more sharply than for the inverse multiquadric. If too large a value for σ is chosen, then some numerical instability may occur, so that the interpolant fluctuates between large positive and large negative values and gives a totally unacceptable surface.

Between the two extremes, there is a range of values over which the results are acceptable and largely insensitive to the choice of σ. Nevertheless, this is not a satisfactory state of affairs, since a general user could have difficulty in selecting a reliable value for σ. The next step may be to investigate whether a suitable value of σ can be set automatically, perhaps from the distribution of nearest neighbour distances. On a site where the data points are unevenly distributed, it may be appropriate to select different values of σ for each data point, smaller values where the density is high, and larger where the density is low. This suggestion moves away from the essential philosophy of radial basis functions, where the same function is used at each data point, but it might prove to be a useful extension.

20.6 Problems

The discussion of section 20.4 indicates that the multiquadric basis model provides an entirely satisfactory technique for the creation of DTMs, while section 20.5 shows that the gaussian basis model could be adapted to create site distribution surfaces. What are the likely difficulties in using radial basis functions?

The most obvious problem is the amount of computing time required. Equation (2) shows that the calculation of the parameters λ_i, which define the interpolant, involves the solution of n equations in n unknowns. Since the matrix H has no special properties which allow time-saving techniques to be applied, equations (2) must be solved by Gaussian elimination or some related method, for which the calculation time is proportional to n^3, and becomes very expensive for large values of n.

Using a PC-386 computer with a 387 coprocessor, the time taken to solve the equations with $n \approx 150$ is around two minutes. For $n \approx 300$, this time can be expected to increase by a factor of eight, to about fifteen minutes. On a basic PC-86 machine without coprocessor, both times will be multiplied by around ten. Clearly these times become prohibitive when n is substantially greater than 150, particularly in the case of photograph rectification, where new DTMs need to be calculated quite frequently. On smaller PC systems, subject to the 640 Kbyte limit, memory requirements place a similar restriction on the size of problem. From practical considerations, therefore, the small-computer user may be restricted to data sets of not much more than 150 points.

A more serious problem has been reported by Light (see the acknowledgement below), in attempting to calculate radial basis functions for much larger data sets. When n is greater than 300, the routine for solving equations (2) may suffer from a form of numerical instability known as *ill-conditioning*, a phenomenon which may cause the routine to produce completely spurious solutions. Although this instability is currently the subject of intensive mathematical investigation, no specific cause has yet been established. It places a very clear restriction on large data sets, but in practice few workers have the computing resources to tackle such large problems.

The choice of the parameter c in the multiquadric presents a minor problem. The received opinion is that the calculated interpolant f is largely insensitive to c, and a few simple tests readily demonstrate this to be the case. However, c must be large enough to smooth away the singularity at the apex of the cone, and at the same time small enough not to have a substantial effect at neighbouring data points. Generally speaking, a satisfactory value of c may be estimated from the density of the data points. However, when rectifying an aerial photograph, the data points have to be projected onto the plane of the photograph, before the DTM is recalculated. The density of the projected points is not entirely obvious, and it becomes more difficult to predict a satisfactory value for c.

Their apparent success in a wide range of interpolation problems has made radial basis functions a subject of great mathematical interest. Mathematicians are seeking to understand the reasons for their success, and to establish the limits of their utility. For instance, substantial effort is devoted to determining whether the interpolant converges to the underlying function as the number of data points increases. Although this is clearly an important problem, it may not be directly relevant to many archaeological situations, where the available data is strictly limited; when a DTM is constructed from a contour map, there is little point in considering the effect of adding data points between the available contours.

A data set of up to 150 points is large enough to provide a satisfactory DTM for the majority of sites in aerial archaeology, or to use as a sampling scheme for a small site in surface survey. However, if a user wishes to extend the DTM to cover a system of sites, or to set up a sampling scheme for an entire landscape, the limit in the size of the data set becomes critical, and ways should be sought to circumvent it. It may be possible to break a large data set into patches, providing a separate interpolant for each patch, and matching interpolants smoothly across the boundaries. With an irregular pattern of data points, it may be difficult to decide on a suitable system of patches, and to locate the boundaries correctly. The extension of the

method to large data sets must remain a matter for future investigation.

20.7 Conclusions

Radial basis functions provide an excellent method of calculating an interpolant for a small set of data points in two-dimensional space. The results in setting up DTMs for the rectification of aerial photographs have been outstanding, and the use of multiquadric basis functions for this purpose can be recommended wholeheartedly. The method is entirely reliable and has given realistic and robust results in each case on which it was tested. The only proviso must be in ensuring that the data set is confined to not much more than 150 points, since otherwise time and memory requirements will become excessive, and eventually problems of numerical stability may be encountered. In practice, the topography for the majority of sites can be described quite adequately with such a limited number of data points, and it is unlikely that an aerial archaeologist would wish to exceed it for a single site.

The use of a radial basis model in cases where the interpolant should decay to zero beyond the boundaries of the region is a little more problematic. Gaussian basis functions seem to provide a possible solution, but care must be taken to ensure that their width parameter is sufficiently large to provide a good overlap between adjoining functions, but not so large that numerical problems are encountered at the edge of the region. It may be appropriate to estimate a width parameter for each individual basis function, dependent upon the distance of the neighbouring data points.

Although the use of small data sets is adequate for many applications, there will be cases where users wish to extend a model for a single site into a group of sites, into a landscape, and possibly into a GIS model. It is important that means should be investigated of extending the radial basis model to large data sets, without encountering the problems associated with large numbers of data points.

For small data sets, however, especially in cases where it is appropriate to use multiquadric functions, the author supports the general enthusiasm for radial basis functions, and commends their use to anyone who is concerned with the construction of surfaces from irregular data points.

Acknowledgements

The author is grateful to Professor Peter Graves-Morris, who first drew his attention to the use of radial basis functions, and to Dr Will Light of Lancaster University, whose seminar at Bradford elucidated so many salient features of the topic.

References

FRANKE, R. 1982. "Scattered data interpolation: tests of some methods", *Math. Comp.*, **38**: 181–200.

GREEN, P.J. & R. SIBSON 1978. "Computing Dirichlet tessellations in the plane", *Comput. J.*, **21**(2): 168–173.

HAIGH, J.G.B. 1989. "Rectification of aerial photographs by means of desk-top systems", in S.P.Q. Rahtz & J.D. Richards (eds.), *Computer Applications and Quantitative Methods in Archaeology 1989*, British Archaeological Reports (International Series) **548**, Oxford, British Archaeological Reports: 111–119.

HARDY, R.L. 1971. "Multiquadric equations of topography and other irregular surfaces", *J. Geophys. Res.*, **76**: 1905–1915.

MICCHELLI, C.A. 1986, "Interpolation of scattered data: distance matrices and conditionally positive definite functions", *Constructive Approximation*, **2**: 11–22.

SHEPARD, D. 1968, "A two-dimensional interpolation function for irregularly spaced data", *Proc. 23rd Nat. Conf. ACM*: 517–523.