6 Practical experience in creating digital terrain models

John G.B. Haigh

6.1 INTRODUCTION

In ideal circumstances, aerial archaeologists would obtain their photographs in stereo pairs, and would reconstruct the subjects in three dimensions with the aid of stereoscopic devices such as analytical plotters. In practice, a large proportion of aerial photographs are obtained as single images, grabbed as opportunity allows, and the use of analytical plotters is well beyond the budget of the majority of archaeologists.

When the subject is a set of markings on the ground surface, such as cropmarks, it may be analysed and reconstructed from a single photograph, provided that a sufficiently accurate model of the ground surface is available. Furthermore, such an analysis need not involve the use of expensive equipment; personal computers, within the range now generally available at reasonable prices, are perfectly adequate for the task. It is only necessary to attach a small digitising tablet, by means of which information may be transferred from the photograph to the computer (Haigh 1989).

Aerial archaeologists are then faced with the problem of where to obtain an accurate model of the ground surface. Assuming that they are not willing to carry out their own detailed survey of the site, and to do so would defeat the purpose of aerial photography as a rapid technique of large–scale preliminary survey, the most likely source of information is a contour map. In Great Britain, the best contour maps are generally the 1:10000 Series of the Ordnance Survey; doubtless maps of similar standard are available in most industrialised countries.

When the map is placed on the digitising tablet, the locations of points situated on contours of known height may be transferred to the computer. In this manner, an assemblage of co–ordinates of discrete points within the ground surface may be created. Such an assemblage is known as a *digital terrain model*, or DTM, of the ground surface. The number and spacing of the points necessary to create a satisfactory DTM will be discussed in Section 6.4.

In order to reconstruct the subject of the photograph, it is necessary to estimate the ground height not only at the discrete data points of the DTM, but also at all intermediate points and, in particular, at points between the contours of the original map. The process of estimating the height at intermediate points is known as *Interpolation.* Interpolation between irregularly scattered data points in two or more dimensions is usually more difficult than in one dimension, since the concept of ordering, which is the basis of many one–dimension methods, can no longer be applied.

A large number of mathematical techniques have been advocated for interpolation in two or more dimensions, of which one, the use of *radial basis functions* (RBFs), has recently been discussed in the context of archaeological applications by this author (Haigh 1992); a number of background references are cited in the earlier paper. The purpose of this paper is specifically to discuss how RBFs may best be adapted to the problem of creating DTMs, and to give some indication of the reliability of the results, in the light of the author's recent experience.

6.2 INTERPOLATION USING RADIAL BASIS FUNCTIONS

A radial basis function is essentially a function $\phi(r)$ of the radial distance r from the origin of co–ordinates. The distance r may be taken in any number of dimensions, but in this paper it will be confined to two, corresponding to the co–ordinates of a map. Using the vector displacement x to denote the location of a point with co–ordinates (x, y), it is possible to write

[1] $$r = \|\mathrm{x}\|_2 = \sqrt{x^2 + y^2}$$

Here the double vertical lines effectively denote Euclidean distance in the plane of the map, but the notation also applies in spaces of higher dimension.

The DTM, obtained from the digitising tablet, constitutes a set of n data points x_i, at each of which the ground height u_i is known, so that the data set consists of n triads of three–dimensional co–ordinates (x_i, y_i, u_i). With each data point may be associated a RBF, whose argument is the radial distance from the data point concerned:

[2] $$\phi\left(\|\mathrm{x} - \mathrm{x}_j\|_2\right) = \phi\left(\sqrt{(x - x_j)^2 + (y - y_j)^2}\right)$$

The same functional form ϕ is retained for all the data points.

The interpolant f is now taken to be a linear combination of the RBFs for all the different data points:

[3] $$f(\mathrm{x}) = \sum_{j=1}^{n} \lambda_j \phi\left(\|\mathrm{x} - \mathrm{x}_j\|_2\right)$$

The parameters λ_i are constants whose values have to be determined. Since f is intended to be an interpolant, rather than an approximant, it must match the height u_i at each of the data points:

[4] $$\sum_{j=1}^{n} \lambda_j \, \phi\left(\|\mathrm{x}_i - \mathrm{x}_j\|_2\right) = u_i, \quad i = 1, \ldots, n$$

Eq. [4] represents a set of n linear equations in the n unknown parameters n_i. In general, such a set of equations may be solved exactly, and the solution may be substituted into eq. [3] to give the precise functional form for the interpolant f.

A variety of different functional forms for the basis function ϕ have been proposed. The discussion in the earlier paper (Haigh 1992) demonstrated that only one functional form appeared to be appropriate for the interpolation of DTMs, namely the *multiquadric function*

[5] $$\phi(r) = \sqrt{r^2 + c^2}$$

Here the non–zero parameter c is introduced in quite an arbitrary manner to avoid an awkward singularity at the origin of the argument r. Although the multiquadric function does not have the some of the mathematical properties one might expect, practical experience demonstrates that it is the right choice for this and many similar problems.

The received wisdom is that the form of the interpolant f is largely insensitive to value of c, but that generally speaking it should be made as large as seems practically feasible. Such a formulation does nothing to define what is the practical upper limit for c, and gives no indication of how c should be preselected within a general computer application. Experience of using multiquadric RBFs in the rectification of aerial photographs has shown that the detailed results are sensitive to the value of c, and that a good choice may be crucial to the success of the technique. Much of this paper will be devoted to discussing how to assign a value to c, and the overall reliability of the RBF technique will be assessed in the context of such an assignment.

6.3 REFINING THE MULTIQUADRIC INTERPOLANT

6.3.1 Fitting deviations from mean height

Before tackling the problem of choosing the parameter c, one small refinement of the RBF technique should be considered. One would expect the fitted interpolant to depend only on the shape of the ground surface, and not on the absolute height. That is to say, the interpolant should be the same, irrespective of whether the section of terrain is near the top of a mountain or deep in a valley, provided that the shape remains constant. This cannot be the case when the interpolant f in eq. [4] is fitted directly to the absolute heights u_i at the data points.

The problem can be overcome by calculating the mean height over all the data points:

[6] $$\bar{u} = \frac{1}{n} \sum_{j=1}^{n} u_j$$

and then fitting the interpolant to the deviations from the mean height, rather to the absolute heights.

Corrected for the subtraction of the mean height, and with the multiquadric basis function of eq. [5] substituted for ϕ in eq. [3], the interpolant f may then be expressed as

[7] $$f(x,y)=\bar{u}+\sum_{j=1}^{n}\lambda_j\sqrt{(x-x_j)^2+(y-y_j)^2+c^2}$$

The parameters λ_i are solutions of the simultaneous linear equations

[8] $$\sum_{j=1}^{n}a_{ij}\lambda_j=u_i-\bar{u},\, i=1,\ldots,n$$

where the coefficients a_{ij} are given by

[9] $$a_{ij}=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2+c^2}$$

Although the introduction of deviations from the mean height is a small change to the overall calculation, it can lead to a perceptible improvement in the results, particularly around the outer fringe of the distribution of data points. Indeed, it sometimes allows the model to be extrapolated a little way beyond the distribution, without encountering excessive errors. Extrapolation should not be relied on as a regular practice, since ground surfaces are always liable to behave in a manner which cannot be predicted mathematically.

It should be noted that taking the mean height into account does not improve the solubility of eq. [8], which depends only on the coefficients a_{ij}, which in turn depend only on the locations of the data points in the plane of the map. This is in contrast to conventional least–squares fitting, where mean values should always be taken into account if the stability of the solution is to be guaranteed. In the present case, it is the quality of the resultant solution which is likely to be improved.

6.3.2 Choosing a value for *c*

The parameter c in eqs [5], [7], and [9] has the dimension of distance. Furthermore, its value should be related to the area density of the data points, a high density of points indicating a small value for c, since the distance between the data points will then be small, and a low density of points indicating a larger value for c. On the basis of this argument, c should behave in much the same manner as the *mean nearest–neighbour distance*, a parameter often used to indicate the density of a distribution of points, with which it has an inverse relationship. It seems reasonable to suggest that c should be taken to be a proportion κ of the mean nearest–neighbour distance:

[10] $$c=\kappa\,\bar{r}_{NN}$$

This conclusion creates some problems. A typical section of terrain is likely to contain some level areas and also some steep hillsides. The density of data points over the hillside areas, where there are many contour lines, is probably higher than over the level areas, where there may be no contour lines at all. Consequently the nearest–neighbour distances may show distinct differences between the two types of area.

For the rectification of aerial photographs, the variation in density is likely to be further emphasised, since the DTM is constructed not over the plane of the map itself, but over the projection of the ground surface through the lens on to the focal plane of the camera. Consequently, the effect of perspective increases the density of points in the more distant parts of the subject, and decreases it in the nearer parts.

Consideration of the last two paragraphs suggests that there is a need for a formula for c which is not constant throughout the model, but takes into account the variation in density. The obvious solution is to take c not as a proportion of the mean nearest–neighbour distance, fixed throughout the model, but as a proportion of the actual nearest–neighbour distance for each data point. Then c is no longer a fixed parameter for the entire model, but takes a specific value c_i associated with each of the data points.

Eq. [7] for the interpolant is then replaced by

[11] $$f(x,y)=\bar{u}+\sum_{j=1}^{n}\lambda_j\sqrt{(x-x_j)^2+(y-y_j)^2+c_j^2}$$

The parameters λ_i are solutions of eq. [8], but eq. [9] for the coefficients is replaced by

[12] $$a_{ij}=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2+c_j^2}$$

where the parameters c_j are given by

[13] $$c_j=\kappa\,d_{(NN)j}$$

Although these changes appear to provide a straightforward answer to the problems, some attention should be given to their consequences. The first point is that the RBFs are no longer the same for every data point, as eq. [2] specifies they should be, but vary in detail from one data point to another; this could be seen as destroying the pure concept of RBFs. The second point is that the matrix of the coefficients a_{ij} is no longer exactly symmetric, which could significantly affect some of its properties. After extensive testing, the author has been unable to detect any undesirable results from either of these two points, but there remains a remote possibility of problems emerging in the future.

6.3.3 Choosing a value for κ

In order to implement eqs [11]–[13], it is necessary to choose a value for the constant of proportionality κ. A useful analogy is to think of the fitting of multiquadric basis functions as stretching a rubber sheet over the data points. In the neighbourhood of each data point the function takes the form of a circular tent, possibly inverted, with a rounded peak.

A small value for κ, and consequently for the parameters c_i, indicates that the rubber sheet is very thin, so that it stretches very tightly between the data points; the peaks will then be quite prominent, and the overall appearance will not be very smooth.

A larger value for κ indicates that the sheet is quite thick, so that it bends gently round each data point, to take on an overall smooth appearance. The problem here is that the thick sheet cannot bend very readily, and may have difficulty in moulding itself to neighbouring data points. Consequently it may show a tendency to overshoot the range of heights indicated in the data, to form itself into creases, and possibly even to tear.

The author decided to examine a number of DTMs for values of κ in the range from 0.1 to 1.0. On intuitive grounds, it seemed that anything less than 0.1 would fail to perform its aim of smoothing out the peak of the multiquadric basis function; anything greater than 1.0 would cause too great an interaction between neighbouring points or, in other words, make the rubber sheet too thick.

The results were examined by displaying the interpolated ground heights as colours on a VGA screen. With the standard sixteen colours available, the effect was that of a coloured contour diagram, which could be transferred to a laser printer in the form of a simulated grey scale. The output was arranged so that certain of the displayed contours corresponded exactly to those of the original map. When the data points covered fewer than eight distinct contours, some intermediate contours were generated on the computer display, taking advantage of the full range of sixteen available shades.

With values of κ less than 0.3, the contours reproduced in steeply sloping areas often show a scalloped effect, as they wander off the direct line between data points. This is the effect of pulling a thin sheet too tightly over the points, and justifies the received wisdom of making the value of c as large as reasonably possible. The effect is illustrated in Figure 6.1, which shows a DTM for a section of terrain, reconstructed from a set of 92 data points, using $\kappa = 0.2$. The scalloping can be seen quite distinctly, particularly along the sides of the valley at the bottom right–hand corner of the figure. One intermediate contour has been inserted between each pair of contours on the original map; the inserted contours tend to be smoother than the reconstructions of the originals.

As the value κ of is increased, the scalloping becomes gradually less prominent. For values of κ greater than about 0.7, new problems arise, but they may not be quite so obvious in a contour diagram. In fact, ridges and valleys become more emphasised, and may eventually become grossly exaggerated. This can be very noticeable during the rectification of aerial photographs, when the model may overcorrect for the effects of high ground. Something of these effects can be seen in Figure 6.2, which shows the DTM reconstructed from the same data set as Figure 6.1, using $\kappa = 0.8$. The contours show signs of unnatural distortion in various parts of the figure, particularly along the spur running up from the bottom left–hand corner, and on the hillside in the upper left. There is little sign of scalloping, however, and most of the contours are remarkably smooth.

Thus values of κ are required somewhere intermediate between 0.3 and 0.7, sufficiently large to overcome the scalloping effect, but not so large as to cause exaggeration of peaks and troughs. Somewhat arbitrarily, the author has settled on a value of 0.4, but would be prepared to accept 0.5. The important thing is that detailed examination limits the desirable values of κ and, by implication, of c to quite a narrow range, in contradiction to the received wisdom that the results are insensitive to the choice of c. The DTM reconstructed from the same data set as Figure 6.1, using $\kappa = 0.4$, is shown in Figure 6.3; although there are some signs of scalloping, the extreme effects of Figures 6.1 and 6.2 have been avoided successfully.

6.4. HOW MANY DATA POINTS?

6.4.1 The computational expense of RBF techniques

As the author pointed out in his earlier paper (Haigh 1992), a major disadvantage of RBF techniques is that they are expensive both in computation time and in memory requirements, principally because of the need to invert a large matrix.

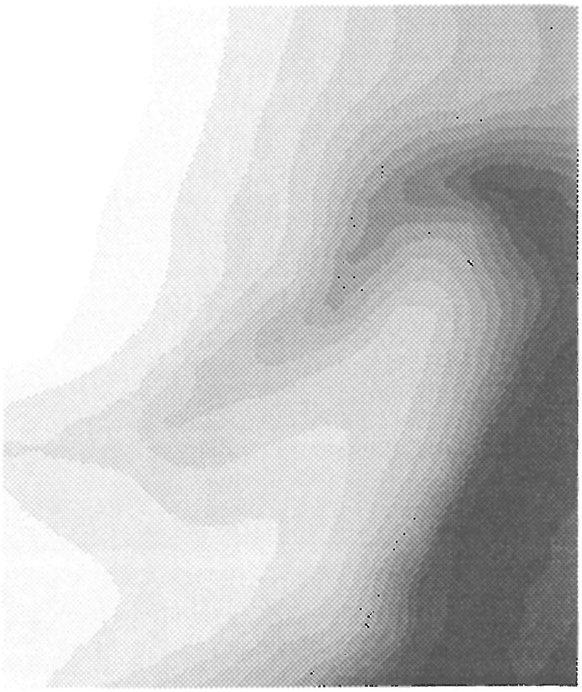

Figure 6.1: A shaded contour diagram, representing a DTM constructed from 92 data points on the map contours, with $\kappa = 0.2$*. Light shades indicate high ground, and dark shades low ground,*

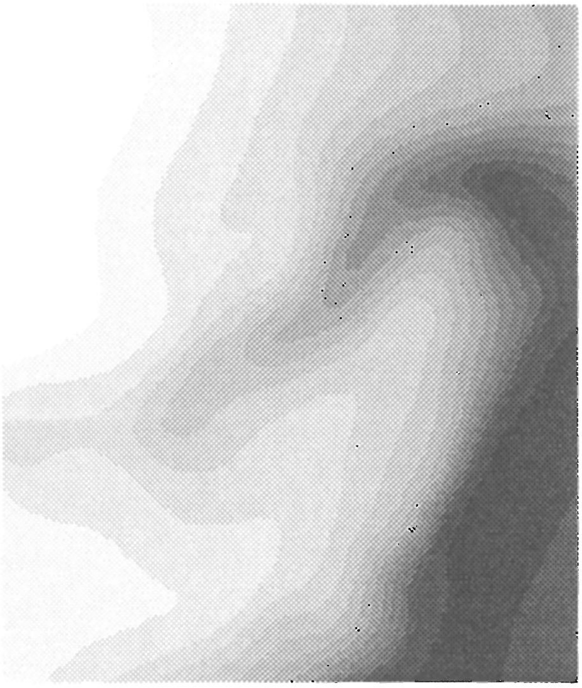

Figure 6.2: A DTM constructed from the same data set as Figure 6.1, with $\kappa = 0.8$*.*

As a general principle, the computation time may be taken to be proportional to n^3 and the memory requirement to be proportional to n^2, where n is the number of data points. A time of 40 seconds was quoted to construct a DTM from 100 data points, using a PC–compatible computer with 386 processor and 387 coprocessor running at 20 MHz. Such timings would inhibit calculations with large numbers of data points, particularly on slow machines, and on those which do not possess a floating–point coprocessor.

Two recent developments have improved the situation considerably. The first is that the author has adopted *LU*–decomposition as the standard method of matrix inversion. This has proved to be much faster than Gaussian elimination; although the number of floating–point operations is precisely the same in both methods, *LU*–decomposition reduces the number of accesses to individual elements of the matrix, and hence the number of fixed–point operations. As a result, the above time of 40 seconds can be reduced to around 24 seconds, with even greater savings for larger values of n, since the computation time is now rather better than n^3. Unfortunately, users who do not possess a coprocessor cannot expect such a dramatic improvement.

The second development is the availability of the 486DX processor at reasonable prices. A PC–computer running such a processor at 50 MHz is approximately six times faster than the 386 machine quoted above. Hence the DTM construction for 100 data points may be reduced to about 4 seconds, and one for 300 data points should take considerably less than 2 minutes. In fact, the last calculation would exceed the MS–DOS restriction to 640 KB of memory, but hardware and software developments will soon make that restriction obsolete.

6.4.2 Recommendations for reducing the data set

Any users, equipped with modern personal computers, should have the opportunity to construct DTMs from large data sets, if they so wish. The question then remains as to whether it is desirable to use very large data sets. The author believes that it is neither advantageous nor desirable, other than in exceptional circumstances. In order to minimise the number of data points, the following recommendations should be followed:

- Do not construct a DTM larger than is necessary for the purposes of rectification of photographs of a particular site;
- When two sites are recognised as separate entities, even if closely contiguous, construct a separate DTM for each;

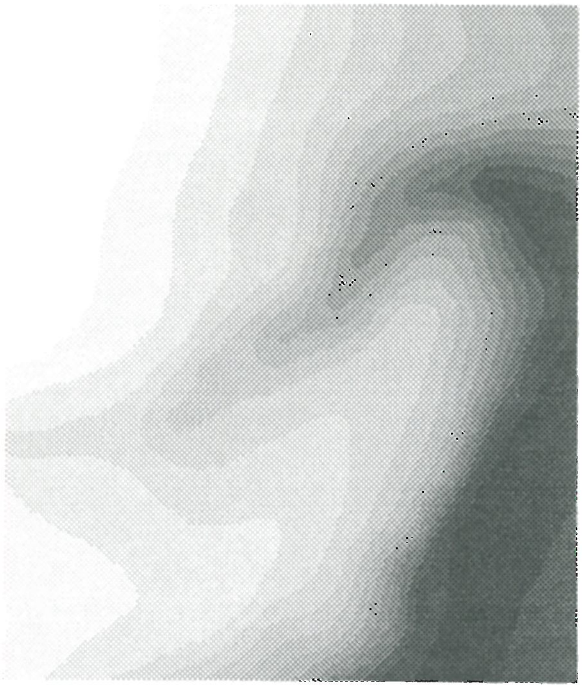

Figure 6.3: A DTM constructed from the same data set as Figure 6.1, with $\kappa = 0.4$.

Figure 6.4: A DTM for the same site as Figure 6.1, but constructed from 155 data points on the map contours, with $\kappa = 0.4$.

- When the above recommendations are followed, a DTM covering an area of 600 × 600 metres should be adequate for the majority of archaeological aerial photographs;
- The spacing of data points along contours should be roughly equal to the distance between contours (the user will have to exercise some judgement in areas where the density of contours changes rapidly);
- In areas where the ground slopes steeply, and the contours are densely packed, it is usually possible to omit some contours without losing essential features of the landscape;
- Do not try to include unnecessary detail in the DTM, but at the same time watch out for any minor features which are likely to have a significant effect on the rectification.

When these recommendations are followed, it has been found that around 150 data points are quite adequate for the majority of sites. Decreasing the spacing between data points along contours, and bringing in extra contours in regions of steep slope, are likely to produce at best only a marginal improvement in the final results. Users with access to the faster, modern, personal computers should be able to construct reliable DTMs in reasonable computation time, and without too much personal effort. Those who still have to use older and slower machine must give careful attention to minimising the number of data points.

Figure 6.4 shows a DTM for the same section of terrain as the earlier figure, but reconstructed from 155 data points, all taken from the map contours, using $\kappa = 0.4$. The results are not conspicuously better than those in Figure 6.3, where only 92 data points were available. In particular, in the valley running across the centre of the figure, the intermediate contours are not centrally spaced between the original contours. In order to obtain good intermediate contours in the valley, it is necessary to take advantage of the expectation that there should be a uniform slope along the bottom of the valley, and to add some extra data points away from the original contour lines. In Figure 6.5, ten additional points have been taken along the valley, to make a data set of 165 points in all, and the results are greatly improved.

This is one of the few instances where the author has found it necessary to take a judicious sample of points in order to improve the results. In every other case, a selection of points taken from the contour line, in accordance with the recommendations above, has proved satisfactory.

6.4.3 Ill–conditioning

A disturbing report has circulated that the matrix of coefficients, defined by eq. [12], becomes *ill–*

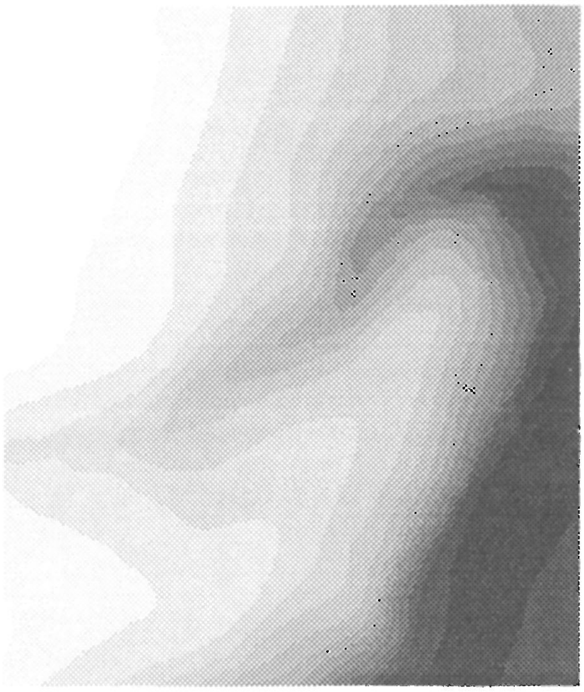

Figure 6.5: A DTM constructed from a similar data set to that of Figure 6.4, but with an additional ten points in the valley, using estimated contour heights; $\kappa = 0.4$.

conditioned when the number of data points is increased to about 300; that is to say, it becomes so close to being singular that its inverse cannot be calculated reliably. If this report is true, then it is impossible to use multiquadric basis functions in constructing DTMs from very large data sets. The author has made many calculations involving 200 data points, and a few informal checks up to about 300 points, but has found no evidence of a problem. Unfortunately, a matrix inversion which simultaneously checks for ill–conditioning involves a much longer calculation than straightforward *LU*–decomposition; consequently the author has not yet had the opportunity to make a large–scale, rigorous investigation of the point.

In spite of his recommendation that there is no need to use such large data sets for the creation of a DTM, the author intends to investigate the problem thoroughly, when the necessary amount of (both actual and processor) time can be devoted to it. It is unsatisfactory to recommend a process which is likely to break down unexpectedly and without explanation. The obvious distortion of the contours in Figure 6.2 may point to an explanation of the ill–conditioning. To follow the analogy introduced at the beginning of Section 6.3.3, the rubber sheet is becoming too thick to bend around the data points satisfactorily. Its effective thickness must be reduced by decreasing the value of κ. The lack of flexibility at large values of κ may have been overlooked by the originators of the report of ill–conditioning.

6.5. TRANSFERRING INFORMATION TO OTHER SYSTEMS

Section 6.4.2 includes the recommendation that a DTM should be constructed only for the region local to the particular site under study. It is now recognised that archaeological data should not be confined simply to local purposes, but should be made available within a global framework. How can the information implicit in the DTM be passed to large–scale systems, such as CAD or GIS?

This question can perhaps be answered at two levels. The calculations discussed here are used as part of the author's software package, which produces vectorised outlines of archaeological features. Provision has been made for the outlines to be passed to AutoCAD, through which they can be reconstructed in two or three dimensions, and treated by all the facilities of the AutoCAD package. Although the transfer routines are specific to AutoCAD, there is no reason why similar routines should not be provided for any other package, in the field of CAD, or database, or GIS.

If there is a desire to transfer the DTM itself, then it is probably more useful to transfer it in the form of a regular grid, rather than as scattered data points. It is far easier to construct an interpolant from a regular grid of values. Thus the RBF technique should be used to predict the height at each of the vertices of a regular grid, and the predicted heights should be transferred to the CAD or GIS system. Where a number of overlapping DTMs are in use, for instance, when examining a group of neighbouring sites, weighted averages can be used in the regions of overlap.

6.6. CONCLUSIONS

Eq. [11] represents the form for the interpolant f, developed in this paper in terms of multiquadric basis functions. The parameters λ_i in eq. [11] are the solutions of the n simultaneous equation of eq. [8], where the coefficients a_{ij} are given by eq. [12]; the parameters c_i are expressed in terms of nearest–neighbour distance by eq. [13], where the recommended value for the constant κ is 0.4, or possibly 0.5. This formulation has proved to be a very successful means of constructing DTMs for use in rectifying aerial photographs. In his earlier paper (Haigh 1992), the author has assessed the success of RBF techniques under the headings of *reliability*, *robustness*, and *realism*, and it is useful to apply those concepts to the revised formulation.

The reliability is proved by the fact that the technique has been applied to a large number of sites, without any case of failure. The example shown in Figures 6.1–6.5 was quoted because it gives a fair demonstration of the principles, and of some of the problems that may arise. The shaded contour displays provide clear evidence of success in modelling the terrain. Success in terms of the original objectives is witnessed by the fact that the outlines of archaeological features have been located to a relative accuracy of one or two metres. Similarly good results for the location of features have been obtained from the other sites, but it has not been possible to make a detailed examination of the DTM in every case.

A comparison of Figures 6.3 and 6.4 points to the robustness of the technique, since the final form of the DTM is substantially independent of the precise choice of the underlying data points, provided that the recommendations of Section 6.4.2 are followed. Such robustness is particularly important where the software is used by people who are not familiar with the mathematical principles. If archaeologists err on the side of caution in selecting the data points, they may find that they are wasting some computer time in constructing the DTM, but the results should be entirely satisfactory.

The realism of the technique is indicated partly by the satisfactory reproduction of the map contours, and partly by the fact that intermediate contours are constructed compatibly with human preconception of the appearance of the landscape. When its presence is indicated in more than one map contour, a feature will normally be well represented in the intermediate contours. For instance, when the DTM shown in Figure 6.3 was used in rectifying aerial photographs, archaeological cropmarks on the spur in the lower half of the figure were rectified accurately, showing that the spur itself had been realistically constructed. Clearly problems can arise from time to time, as is illustrated by the discussion of the valley, in Section 6.4.2. Although the nature of the ground surface is such that no mathematical formulation can describe it completely, this method can be seen to cope as well as any.

Overall, the author remains convinced that the RBF technique, appropriately modified, is the most reliable method, available for the construction of DTMs while rectifying aerial photographs. The simplicity of its concept, and the accuracy and realism of the results more than compensate for the expense in computer time. The author has adopted it exclusively for his work with aerial photographs, and recommends that it be considered for other archaeological applications.

Acknowledgement

The author is grateful to the staff of the Aerial Photography Unit, and of the Drawing Office, at the Royal Commission on the Ancient and Historical Monuments in Scotland for their consistent support and collaboration; in particular, for the provision of the material which forms the basis for Figures 6.1–6.5.

References

Haigh, J.G.B.

1989 Rectification of aerial photographs by means of desk–top systems. in S.P.Q. Rahtz & J.Richards (eds.) *Computer Applications in Archaeology 1989*. International Series 548. British Archaeological Reports. pp 111–119.

Haigh, J.G.B.

1992 Radial basis functions and archaeological surfaces. in G. Lock & J. Moffett (eds.) *Computer Applications and Quantitative Methods in Archaeology 1991*. BAR International Series S577. Tempus Reparatum. pp 157–161.

Author's address

John G.B. Haigh
Department of Mathematics
University of Bradford
GB–BD7 1DP Bradford
Telephone +44 (274) 733466 Extension 4278
Fax +44 (274) 305340