CHAPTER NINE

Trend surface and interpolation

"Everything is related to everything else, but near things are more related than distant things." (Tobler 1970)

Interpolation refers to the process of making mathematical guesses about the values of a variable from an incomplete set of those values. Spatial interpolators use the spatial characteristics of a variable, and the observed values of that variable to make a guess as to what the value of the variable is in other (unobserved) locations. For example, if a regularly spaced auger survey revealed an easily identifiable horizon, then it may be desirable to try to produce a contour map of that deposit from the points of the survey. In this case a spatial interpolator would have to be used to predict the values of the variable (in this case the topographic height of deposit X) at all the unsampled locations.

Chapter 5 has already introduced one specific type of interpolation problem: that of estimating topographic height from contours. However, we noted then that interpolation has both a far wider range of applications than merely to elevation data, and also encompasses a very wide range of methods. In this chapter we will review some of the main methods for interpolating continuous values from point data, and comment on their applications to archaeological situations.

There are a wide variety of procedures for spatial interpolation, each with different characteristics. It is important to note that the choice of interpolation procedure should be dependent on the nature of the problem. In addition, a sound understanding of how each interpolator works is essential if they are to be used effectively.

It is possible to use the spatial allocation methods discussed in Chapter 7 as crude spatial interpolators. Point data can be used to generate a Voronoi tessellation, and the values of the points allocated to the polygon defined by it (the effect of this can be seen in Figure 9.7, top left). The major disadvantage of this is that it does not provide a continuous result. Many archaeological variables, such as artefact density or dates, do not conform to finite boundaries and although there are methods for interpolating from area data to continuous variables—see *e.g.* Burrough (1986)—these are rarely, if ever, used in archaeology.

The remainder of this chapter provides a brief overview of some of the methods available for interpolation of points into density estimates and of ratio or higher scale observations into continuous surfaces. We have already discussed one continuous surface, topographic height, but other examples could include any archaeological or environmental value that can be measured on an ordinal scale or higher and for which we would like to estimate values at unobserved locations. Examples might include artefact density, rainfall, ratio of deciduous to coniferous species and many more.

These are illustrated with images showing the results of applying different procedures to a dataset of 176 points for which there is a measurement of some continuous variable (originally elevation in metres, but we will imagine that it could be anything because some procedures would not be appropriate for elevation data), and for which there is also a record of presence/absence of some diagnostic

characteristic at that location. For clarity, the same dataset has been used throughout the chapter, although in places it is treated simply as a set of points.

Some of this chapter resorts to a rather formal, mathematical notation to describe the methods, although we have omitted considerable detail for the sake of brevity. We realise that we have not been particularly consistent in the level of mathematical detail included, but chose to give quite a lot of the formal notation for some procedures in preference to providing a little maths for all of them. A reader who does not have enough familiarity to follow the mathematical notation should still be able to follow the main differences between these approaches described in the text, and if more detail is required then a more specific text should be consulted. Bailey and Gatrell (1995) covers the vast majority of the methods described in this chapter with particular emphasis on geostatistical methods. This text also has the considerable advantage that it includes exercises, datasets and software allowing the reader to gain hands-on experience of the procedures without needing access to expensive hardware and software. Alternatively (or additionally) Isaaks and Srivastava (1989) provide a comprehensive and easily followed introduction to applied geostatistics that would form the ideal starting point for further study.

9.1 CHARACTERISTICS OF INTERPOLATORS

One characteristic of an interpolator is whether or not it requires that the result passes through the observed data points. Procedures that do are *exact interpolators* as opposed to *approximators,* in which the result may be different at the sampled locations to the observed values of the samples.

Another important characteristic is that some procedures produce results whose derivatives are continuous, while others produce surfaces whose derivatives are discontinuous, in other words the surface is allowed to change slope abruptly.

Most importantly, some interpolators are *constrained*, so that the range of values that may be interpolated is restricted, while others are *unconstrained* where the interpolated values might theoretically take on any value. In some instances, particularly in situations where you are predicting things that cannot be negative, the use of unconstrained procedures should obviously be avoided.

We might also classify interpolation approaches by the extent to which they can be thought of as *global* or *local* procedures. Global approaches derive a single function that is mapped across the whole region, while local interpolators break up the surface into smaller areas and derive functions to describe those areas. Global approaches usually produce a smoother surface and, because all the values contribute to the entire surface, a change in one input value affects the whole surface. Local interpolators may generate a surface that is less smooth, but alterations to input values only affect the result within a particular smaller region.

Although we pay little specific attention to it in this chapter, readers should also be aware that some of the procedures could be influenced by *edge effects*. These occur in situations where estimations rely on a neighbourhood or region surrounding a point, or on the number of points close to an unknown location. In both of these cases the estimation will not be possible close to the edge of the study area and, although various compensations are used to allow for this, it may still produce unpredictable results near the edge of the study area.

9.2 POINT DATA

There are many occasions when archaeologists have spatial observations that are not measured on a numerical scale at all. Instead, we may have points that represent observations that something is merely present at a particular location. Examples of this kind of data include sites at a regional scale or artefacts within smaller analytical units. Storing this kind of data within a spatial database is unproblematic, as it consists of simple point entities to which we may also attach other attributes such as size or period (for sites) or object type and classification (for artefacts).

Archaeologists have often used methods of 'contouring' this type of data into density or other continuous surfaces. Although several procedures exist to do this, careful thought needs to be given to whether or not it is an appropriate or meaningful thing to do: it *may* be useful to 'interpolate' from the observations in order to estimate the density of similar artefacts or sites that might be expected to occur at unsurveyed locations. On the other hand, the interpolation of artefact densities within a wholly surveyed area into a continuous product is of entirely dubious utility. We have seen this done—for example using the densities of artefacts within features to create a surface that includes excavated but empty parts of a site—but would caution strongly against it unless there are strong analytical reasons to do so. The 'interpolated' values do not represent an estimate of anything real, because we already *know* where all the artefacts actually are and we are, essentially, generating a wholly artificial surface. Very often, archaeologists would be better advised to concentrate on the presentation of the original data values using choropleth maps, symbols or charts (to show proportions) rather than creating dubious 'interpolated' continuous products. With that proviso, we can note that there are several methods of converting presence/absence data stored as points into an ordinal or continuous product.

The most straightforward method is to generate a grid in which the attribute of the grid cells is the density of the points that occur within them. This is essentially the same as the quadrat approach to point pattern analysis that we discussed in Chapter 6 and presents the same problems, notably of choosing an appropriate grid/quadrat size. It requires a grid resolution sufficient to ensure that many grid cells contain several points, but not so large that the subtle variations in artefact densities are obscured by aggregation. Different grid cell sizes may appear to produce rather different patterns, so it is important to pay careful attention to this, ideally generating densities with a number of different cell sizes. This method is also not optimal because the shape of the grids/quadrats, almost always squares, can affect the density surface.

This approach has been followed in many archaeological examples—see *e.g.* Ebert, Camilli and Berman (1996:32) for an example using densities of lithic flakes and angular debris) but its popularity may derive more from the intuitive familiarity that fieldworkers have with the way that it works than any inherent methodological advantages. It is, after all, a computational expression of what archaeologists have done for many years with fieldwalking data and paper—see *e.g.* papers in Haselgrove, Millett and Smith (1985).

Another approach is to generate a circular area of radius r centred on each point and give it, as an attribute, the density derived from its source point (in other words $1/\pi r^2$ where r is the radius of the circle). These circles can be added

together to give what is sometimes referred to as a *simple density* operator. As the radius for the values increases, the density surface becomes more and more generalised. Figure 9.1 shows clearly how this affects a given set of points. It will be apparent that the density surface produced can be as local or as general as we wish and care should therefore be taken to select an appropriate radius for the task in hand. Although this takes account of the arbitrary nature of square sample units, it unfortunately should also be clear that the greater the search radius becomes the more influenced the resulting surface will be by *edge effects* although it is possible to adjust the initial density values of those circles which are only partially within the area. The bottom two images of Figure 9.1 were calculated without making any adjustments in this way and are therefore significantly influenced by the edges of the study area.

A related approach is the use of Kernel Density Estimates (KDE). These have recently been introduced to archaeology—see *e.g.* Baxter, Beardah and Wright (1995), Beardah (1999)—although not specifically for spatial interpolation. KDE operate in a similar manner to simple density estimates except that the circle centred at each point is replaced by a density function called the *kernel*. In simple terms, the kernel describes a 'bump' at each data point and these can be added together to arrive at a density estimate. The kernel itself can be any bivariate probability density function which is symmetric about the origin and which can be controlled with some bandwidth parameter. Just as with the radius of the circles, the bandwidth parameter (sometimes parameters) will determine the smoothness and other characteristics of the surface. Further details have been omitted here for the sake of brevity, but Bailey and Gatrell (1995:84-88) would make a good starting point for further reading on the use of KDE as a spatial interpolator.

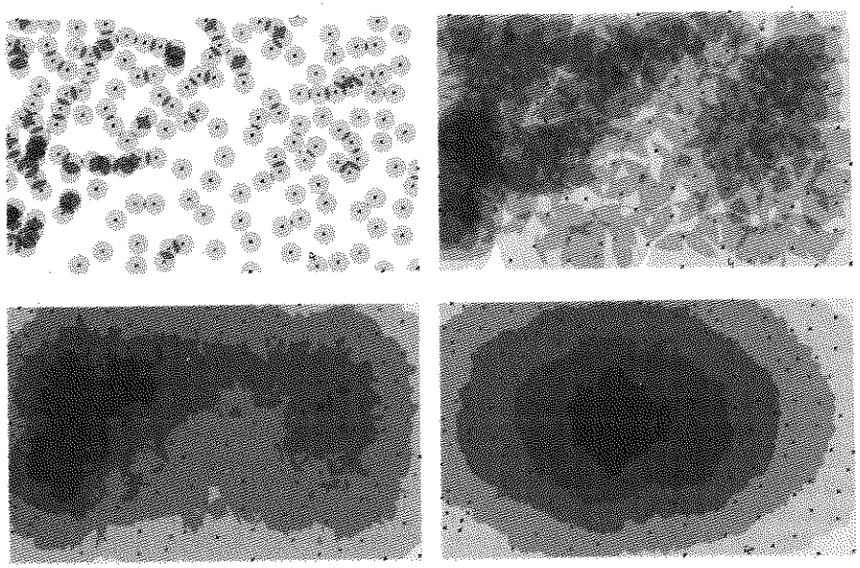

Figure 9.1 Simple density estimates for a point data set using values of 150m (top left), 500m (top right), 1000m (bottom left), and 2500m (bottom right) for the radius of the circles. Grey scale varies. Generated with ArcView.

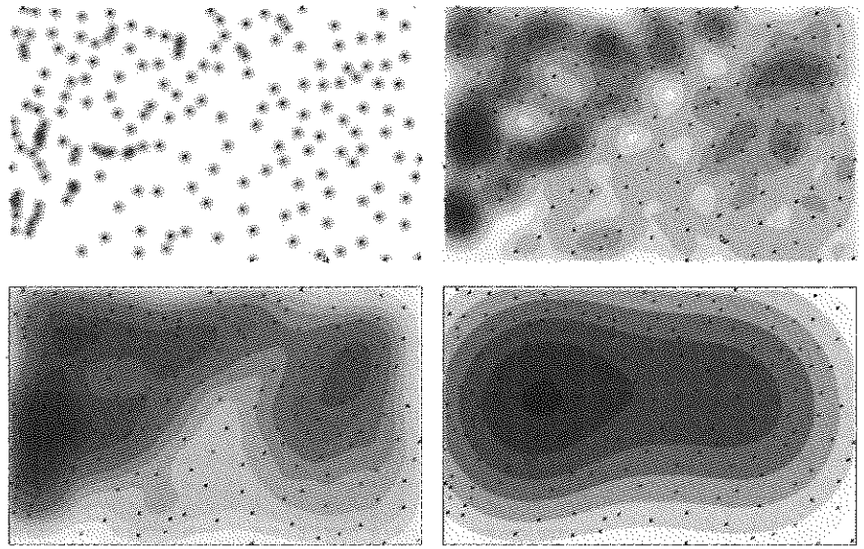

Figure 9.2 Kernel density estimates for a point data set using values of 150 (top left) 500 (top right) 1000 (bottom left) and 2500 (bottom right) for the 'radius' of the density function. Grey scale varies. Generated with ArcView spatial analyst.

KDE produces significantly smoother surfaces than the other methods described. Moreover, the kernel can be made asymmetrical so that point distributions that seem to cluster in a directional way can have a more appropriate form of density function. Figure 9.2 shows how changing the form of the kernel—in this case to make the bumps wider—can allow tuning of the resulting density estimate from a highly local surface that reflects smaller clusters of points to a far more general approximator which reflects larger scale structure in the data. Unless great care is taken, however, KDE—like simple density estimates—can be significantly affected by edge effects.

9.3 TREND SURFACE ANALYSIS

One of the simplest and most widely used methods for interpolating data measured at ordinal scale or higher is *trend surface analysis*, which is essentially a polynomial regression technique extended from two to three dimensions. In trend surface analysis, an assumption is made that there is some underlying trend that can be modelled by a polynomial mathematical function, and that the observations are therefore the sum of this and a random error. Trend surface analysis is similar to regression analysis, but extended to 3D (see Figure 9.3). In a simple linear regression, the relationship between a dependent variable Z and independent X would be estimated as:

$$z = a + bx \tag{9.1}$$

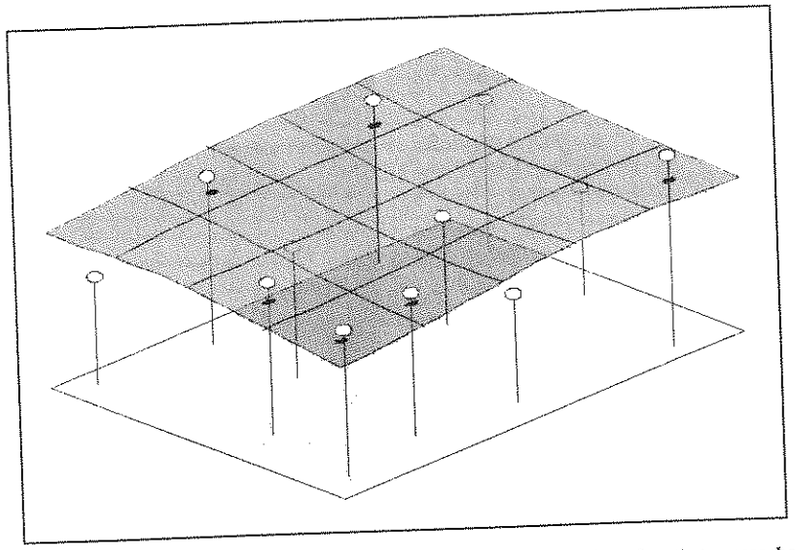

Figure 9.3 Trend surface analysis: the aim is to generate a 'best fit' smooth surface that approaches the data points, and the surface is unlikely to actually pass through any of the points.

where a is the intercept, and b the slope. In trend surface analysis, it is assumed that the spatial variable (z) is the dependent variable, while the co-ordinates x and y are independent variables. In a linear trend surface, the equation relating z to x and y is a surface of the form:

$$z = f(x, y) = a + bx + cy \tag{9.2}$$

and the coefficients b and c are chosen by the least squares method, which defines the best 'fit' of a surface to the points as that which minimises the sum of the squared deviations from the surface.

As with regression analysis, different orders of polynomial equations can also be used to describe surfaces with increasing complexity. For example, a quadratic equation (an equation that includes terms to the power of 2) describes a simple hill or valley:

$$z = a + bx + cy + dx^2 + exy + fy^2 \tag{9.3}$$

The highest power of the equation is referred to as the *degree* of the equation, and in general, surfaces that are described by equations of degree n can have at most n-1 inflexions. A cubic surface (degree 3) can therefore have one maximum and one minimum:

$$z = a + bx + cy + dx^2 + exy + fy^2 + gx^3 + hx^2y + ixy^2 + jy^3 \tag{9.4}$$

Because continuous surfaces cannot be adequately represented in GIS systems, most implementations of trend surface analysis are in raster-based

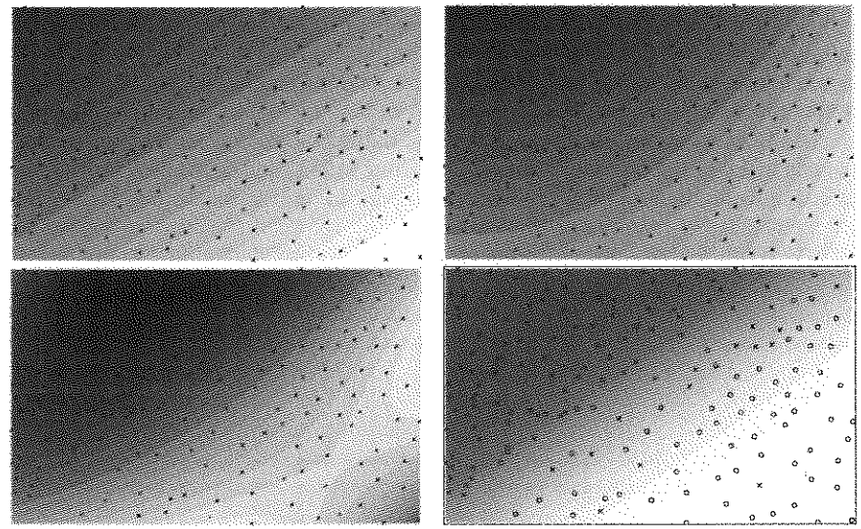

Figure 9.4 Example trend surfaces. Linear (top right), quadratic (top left) and cubic (bottom right) surfaces generated from the same values and a logistic trend surface (bottom left) from the same points, but different attributes, shown as crosses for presence and zeros for absence. Generated with ArcView spatial analyst.

systems, with the output taking the form of a discrete approximation of the continuous surface obtained by calculating the value on the surface of the centre of each grid cell, or the average of the four corners of each grid cell.

One disadvantage of trend surfaces is that the surface described is unconstrained. In order to satisfy the least squares criterion, it can be necessary for intervening values to be interpolated as very high or very low, sometimes orders of magnitude outside the range of the data. Higher order polynomials, in particular, should therefore be used with extreme caution. It should also be noted that values interpolated outside the area defined by the convex hull of the data points—the area bounded by a line connecting the outermost points—should be treated with extreme scepticism, if not discarded as a matter of course. Higher order polynomial trend surfaces can also be very computationally expensive, because their estimation requires the solution of a large number of simultaneous equations.

Given that it is not generally possible or advisable to use higher order polynomials for trend surface analysis, the surfaces described are almost always relatively simple (see Figure 9.4 for examples). As such, trend surface analysis is most meaningfully applied to modelling the spatial distribution of variables that are likely to have relatively simple forms, and it is not appropriate for approximating complex surfaces such as topography.

An advantage of the method is that, like two-dimensional regression analysis, it is possible to analyse the residuals (the distance from observations to surface) to obtain an estimate (r^2) of how well the surface fits the data. It is also useful to map the residuals because spatial patterning within them may reveal systematic problems with the process.

Applications of trend surface analysis in archaeology have a fairly long pedigree. Hodder and Orton (1976) describe the technique and present studies including a trend surface analysis of the distribution of length/width indices of Bagterp spearheads in northern Europe, and of percentages of Oxford pottery in southern Britain.

Another example of the 'global' trend surface approach can be found in the analysis of Lowland Classic Maya sites by Bove (1981) and subsequently Kvamme (1990d). Both identify spatial autocorrelation in the terminal dates of Maya settlement sites (see boxed section in chapter 6) and then use a polynomial trend surface to aid further investigation and explanation of the trend. More recently Neiman (1997) has also investigated the terminal dates of Classic Maya sites using a *loess* model of the same data. Loess is a variation on trend surface analysis that provides a more local and robust estimation of trend—see *e.g.* Cleveland (1993), (cited in Neiman 1997), for discussion. Neiman's interpretation is complex but uses the resulting loess trend surface in comparison to mean annual rainfall for the Maya Lowlands to argue that the collapse phenomenon was caused, ultimately, by ecological disaster rather than drought or invasion. Interestingly, Neiman also turns to variogram methods to understand the structure of the dataset (see below).

Finally, it is possible to fit a probability model (usually a logistic equation) to a set of data presence/absence observations. This is conceptually very similar to 'classic' trend surface analysis but provides output scaled between 0 and 1 and which can be appropriately interpreted as an estimate of the probability that any location will contain one or other of the input categories.

9.4 APPROACHES THAT USE TRIANGULATION

Often simply referred to as 'contouring', because contours can easily be generated by joining together points of equal value on the sides of the triangles, this approach involves generating a triangulation from all the known data values. This is usually done by deriving a Delaunay triangulation, as described in Chapter 5 in the discussion of Triangulated Irregular Network models of elevation. Where locations lie away from data points but within the triangulation—in other words within the *convex hull* of the data points—then their values may be obtained from the geometry of the relevant triangular face (see Figure 9.5).

When unknown values are assumed to lie on flat triangular facets, then the operation is linear, constrained, and behaves in a very predictable manner because unknown values must lie within the range of the three nearest known values (assuming that the interpolation is restricted to the convex hull of the data points). It produces a 'faceted' model that is continuous, although its derivatives are likely to be non-continuous.

If, instead of triangular facets, small polynomial 'patches' are used in order to produce a result with continuous derivatives, then the approach is sometimes referred to as 'non-linear contouring'. This provides a smooth, non-faceted result, and the higher the order of the patch, the higher the order of derivatives that will be continuous. However, it removes the constraint that the unknown values must lie between the three defining data points. Inappropriate choice of parameters or data points can therefore lead to strange and extreme 'folds' in the result. As with

higher order polynomial trend surfaces, therefore, the use of non-linear contouring should be treated with caution.

All such approaches are *exact* interpolators because the resulting surface must pass through the data points, and are generally extremely *local*. Unknown values are wholly defined by the closest three known values, and are unaffected by any other data points. In some applications this might be desirable because deviant

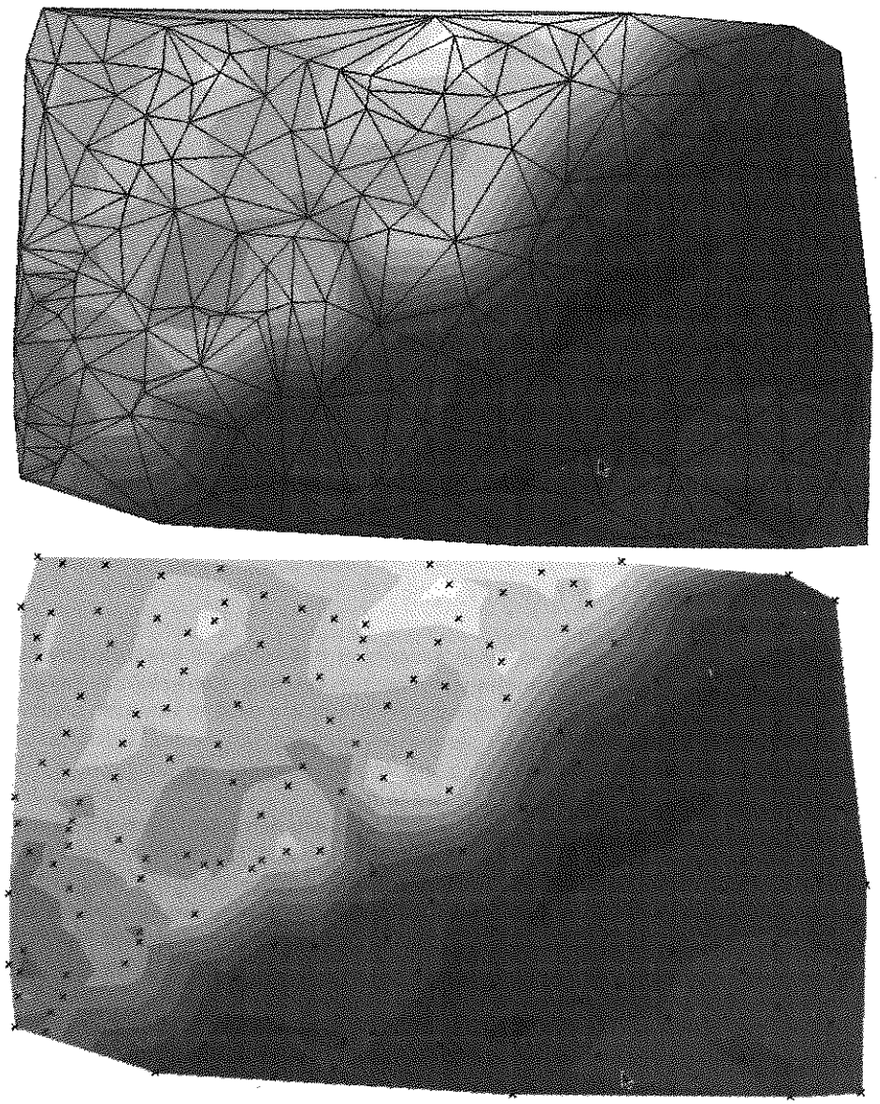

Figure 9.5 Interpolation using Delaunay triangulation. Higher values are shown as lighter colours. The rather angular result is clearer in the bottom image, which also shows the data points from which the surface derives. Note the interpolated values are limited to the convex hull of the data points.

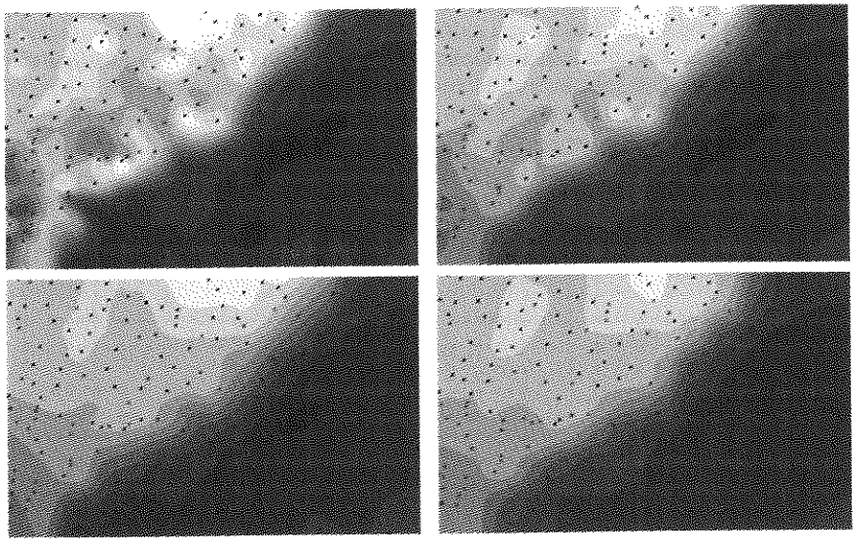

Figure 9.6 Interpolation with splines showing the effect of increasing tension. Values for 'tension' are 0.00025 (top left) 500 (top right) 10,000 (bottom left) and 10,000,000 (bottom right). Generated with ArcView Spatial Analyst.

values will be easily identified but it also means that these approaches can be particularly susceptible to 'noise' in the sample data.

9.5 APPROACHES THAT USE SPLINES

The term *spline* was originally used to describe a flexible ruler used for drawing curves. In contemporary terminology, *spline* curves are interpolators that are related to trend surfaces in that polynomial equations are used to describe a curve. The key difference is that splines are piecewise functions: polynomial functions are fit to small numbers of points which are joined together by constraining the derivatives of the curve where the functions join.

Because the function is, essentially, a local interpolator the polynomials used can be low-order equations, commonly quadratic or cubic. The resulting surface is also controlled by the severity of the constraint placed on the join. A quadratic spline must be constrained so that the first derivative is continuous, a cubic so that at least the second derivative is continuous and so on. More complex forms of splines that 'blend together' individual polynomial equations can also be used, the widely used Bezier- or B-splines being a good example.

Various methods can be used to control the nature of the surface that is fit to the data points. As with numerical averaging (see below) it is common for software to allow the user to choose how many points are used in the spline and the nearest n points to the unknown value are then used to generate the spline patches. Lower

values of n result in a more complex surface while larger values provide a greater degree of smoothing. Additionally, a *tension* value may be applied to the interpolation. Increasing this has an effect similar to increasing the tautness of a rubber sheet, in that the complexity of the surface is reduced and the surface becomes more of an approximator. Lower tension values permit the spline to follow the data points more closely, but very low values can lead to the same kinds of anomalous 'folds' that can result from higher order polynomial equations. Figure 9.6 illustrates the effect of increasing the tension parameter while holding the number of points used constant.

The main advantage of splines is that they can model complex surfaces with relatively low computational expense. This means that they can be quick and visually effective interpolators and, because they are inherently continuous, derivatives such as slope and aspect can be easily calculated. Their main disadvantages are that no estimates of error are given, and that the continuity can mask abrupt changes in the data.

9.6 NUMERICAL APPROXIMATION

An alternative strategy for interpolating values between known locations is to use numerical approximation procedures. These do not constrain the result to pass through the data points, but instead use the points to approximate the values within both sampled and unsampled locations. Using only the x dimension (for clarity) a weighted average interpolator can be written as:

$$\hat{Z}(x) = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{9.5}$$

where n is the number of points we wish to use to estimate from, and λ_i refers to the weighting given to each point. A requirement of the method is that these weights add up to one, so that:

$$\sum \lambda_i = 1 \tag{9.6}$$

The points used for the interpolation can be obtained in several ways. Commonly they can be selected either by specifying that the nearest n points are used for each unknown value or, alternatively, by specifying a distance d within which all points are used. The weightings λ_i are calculated from $\Phi(d(x,x_i))$ where $\Phi(d)$ must become very large as d becomes very small.

The most widely applied approach is that adopted by Shapiro (1993) within the GRASS GIS (r.surf.idw, r.surf.idw2 and s.surf.idw), referred to as *inverse distance weighting*, and is similar to procedures available within a number of other systems. In this interpolator $\Phi(d)$ is the inverse squared distance weighting, so that the interpolator becomes:

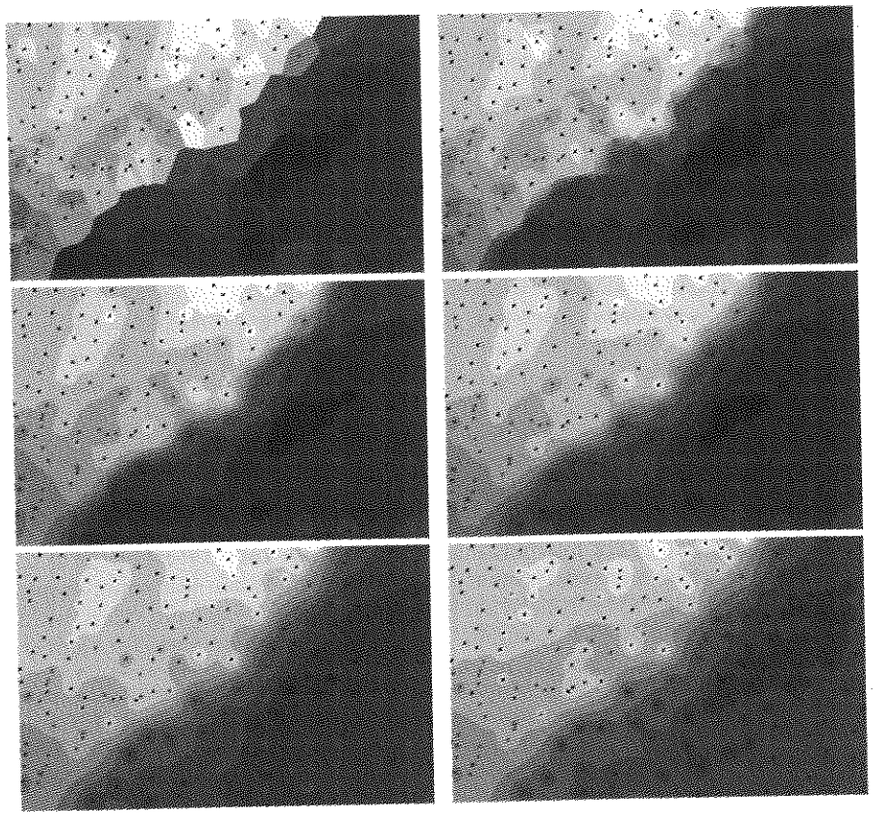

Figure 9.7 Interpolation with Inverse Distance Weighting, showing the effect of increasing n (number of points used). Values are 1, 2 (top) 5, 10 (middle) 50 and 176 (bottom). Grey scale is the same for all images in this case. Generated with ArcView Spatial Analyst.

$$\hat{Z}(x_j) = \frac{\sum_{i=1}^{n} \sqrt{d_{ij}} . Z(x_i)}{\sum_{i=1}^{n} \sqrt{d_{ij}}} \qquad (9.7)$$

This works on the principle that unknown values are more likely to resemble near values than distant ones, so the values of each point used to estimate the unknown value are weighted in inverse proportion to their distance from the unknown value. The actual weighting for each data point is the square root of the distance to the unknown point, and these weighted values are then averaged to produce an estimate of the value at the unknown location. The effect is a simple, generally robust approximation procedure for the interpolation of surfaces from a wide variety of point data.

In contrast to the contouring methods described above, interpolation by this method does not rely only on the nearest three data points. Instead it introduces an element of smoothing, the degree of which is controlled by the number of neighbouring points used in the calculation (or the radius d within which points are included in the estimate). The larger the number of points used, or the larger the distance, the more of the individual variation within the observations will be smoothed out between the data points. This is, of course, a more general formulation of the inverse distance weighting interpolation we described for interpolation from digitised contours. In that case the points used were defined as eight points found by following cardinal directions to the nearest contour lines.

Figure 9.7 shows the effect of choosing different values of n for inverse distance weighting. Note how the surface for a value of 1 is essentially a Voronoi tessellation, with height values attached to the polygons. Surfaces for $n = 2$ upwards become steadily smoother while the very high values of n tend to show a very smooth background between the data points, with the points themselves appearing as 'peaks' above this. It is usually necessary to experiment to obtain the best compromise between the 'blocky' result obtained with low values of n (or low distances for point inclusion) and the 'peaky' result obtained with very high values.

9.7 GEOSTATISTICS AND KRIGING

The main problem with the use of numerical average methods is that they provide no means for deciding what is the right number of points n, or the ideal weights λ_i for any given interpolation. They also provide no indication about how reliable each of the interpolated values is, although it should be clear that the further from the data points the less reliable the interpolated values become.

The term *geostatistics* refers to an approach to spatial variables that addresses these concerns. It was developed by the mathematician George Matheron and D.G. Krige (a South African mining engineer) and it has found wide application in groundwater modelling, soil science, geology, geography and, increasingly, archaeology. The approach is based on the recognition that many spatial variables are best described by a stochastic surface. This is achieved using *regionalized variable theory* (Matheron 1971) that divides the variable into *structure* and *noise*, so that any variable is composed of three components:

1. a structural component described by a function m(x);
2. a random component spatially correlated with (1) called ε'(x);
3. a random noise or residual error not spatially correlated with (1) called ε''.

These are shown diagrammatically in Figure 9.8, which also serves to illustrate the general idea of Kriging, which is to obtain the form of the trend function m(x) from the observations. Using these assumptions, and using only one spatial dimension for clarity, the value of a variable Z at position x can be given by:

$$Z(x) = m(x) + \varepsilon'(x) + \varepsilon'' \qquad (9.8)$$

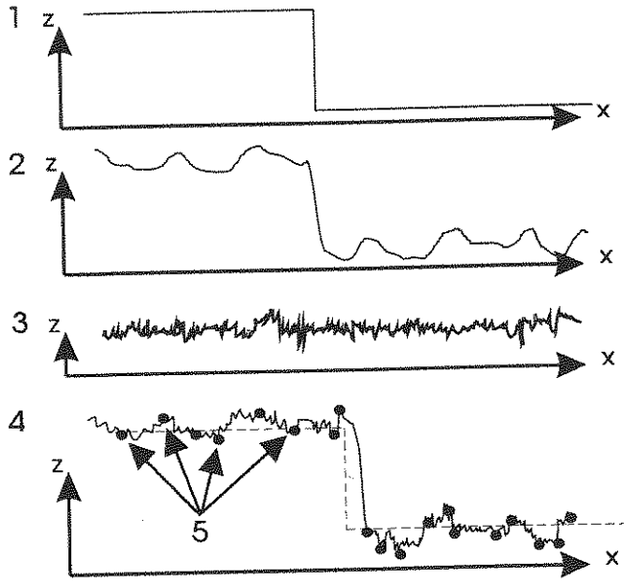

Figure 9.8 Three components of a regionalized variable: (1) structural component, (2) random spatially correlated component and (3) random noise element. (4) Shows how the observations (5) might relate to the other components.

Kriging or *optimal interpolation* is a technique that extends the weighted average method to the optimisation of the values of the weights λ_i to produce the 'Best Linear Unbiased Estimate' (BLUE) of a variable at any given location. It can be used for the same type of data but produces better results in complex spatial variables. The following discussion is heavily based on Burrough (1986:155-165).

Central to the success of geostatistics is the *semivariogram*. This relates the distance between each pair of points (the *lag*) to the influence that these points have upon one another. This is used to determine the underlying structural properties of the variable. In the simplest case of a spatial variable that does not change over the region of interest, then $m(x)$ will be the mean value of the sampling area. In this case the average difference between any two places x and $x+h$ will be zero:

$$E\left[Z(x)-Z(x+h)\right]=0 \tag{9.9}$$

It is assumed that the variance of the differences in Z at the observation points depends only on the spatial distance between the observations so that:

$$E\left[\left\{Z(x)-Z(x+h)\right\}^2\right]=E\left[\left\{\varepsilon'(x)-\varepsilon'(x+h)\right\}^2\right]=2\gamma(h) \tag{9.10}$$

where $\gamma(h)$ is a function known as the *semivariance*—a measure of the extent to which distance h is related to the similarity of any two points. The semivariance

can be obtained by estimation from the sample data. If n is the number of *pairs* of points, each separated by distance h, then the estimated value of $\gamma(h)$— written $\hat{\gamma}(h)$— is as follows:

$$\hat{\gamma}(h)=\frac{1}{2n}\sum_{i=1}^{n}\left\{Z(x_i)-Z(x_i-h)\right\}^2 \tag{9.11}$$

This value can then be calculated for all possible lags within the data set, and plotting this estimated value of semivariance against the sample spacing h (often called the 'lag') gives the *semivariogram.*

The main characteristics of the semivariogram are the curve, which often rises at low values of h until it reaches a plateau area referred to as the *sill*. The lag at which the curve stops rising and levels off into the sill is referred to as the *range*. A curve can be fit to the observations by a variety of mathematical models. Commonly spherical, exponential, linear, gaussian equations or combinations of these are used. Models can also be *bounded*, if they reach a sill, or *unbounded* if they do not. The range is important because it represents the maximum lag at which an observation will be likely to influence an interpolated value. Beyond that range, denoted by a, observations cannot be said to influence the interpolated value. An estimate of the uncorrelated noise in the data is obtained from the intercept (the point at which the curve crosses the y-axis). If this *nugget effect* is too great then no model should be fit to the data.

Having obtained the semivariogram for a set of data observations, it is possible to use a technique similar to numerical averaging (see above) to interpolate the unknown values. In the case of *ordinary Kriging*, the weights are obtained from the semivariogram instead of from an arbitrary function. The function to estimate for the value at the unknown location, written $Z(x_0)$, can be written as follows:

$$\hat{Z}(x_0)=\sum_{i=1}^{n}\lambda_i Z(x_i) \tag{9.12}$$

where (as usual) the known values are numbered from 1 to n. The weight to be ascribed to each known value is given by the λ_i symbol and they must sum to one. The values of λ_i are calculated in such a way as to reflect the distance of the estimator points from the unknown value as observed in the semivariogram. This is done by selecting weights that minimise the estimation variance σ_e^2, which is calculated from:

$$\sigma_e^2=\sum_{j=1}^{n}\lambda_j\gamma(x_j,x_0)+\Psi \tag{9.13}$$

where $g(x_j,x_0)$ is the semivariance between the known point and the unknown point (the quantity Y is required for minimalisation). The minimum value of σ_e^2 occurs when:

$$\sum_{j=1}^{n} \lambda_j \gamma(x_i, x_j) + \Psi = \gamma(x_i, x_0) \qquad (9.14)$$

for all values of i. In this equation $g(x_j,x_i)$ is the semivariance between the sampling points. Both this value and the value $g(x_j,x_0)$ are obtained from the model that was fit to the semivariogram.

There are numerous variants of variograms and Kriging, including methods that investigate or take account of anisotropy in spatial variables or that use other spatial information to assist with interpolation (*stratified Kriging* or *co-Kriging*). These are well beyond the scope of this volume, but further details can be found in Burrough (1986:147-150), Bailey and Gatrell (1995) and Isaaks and Srivastava (1989).

One of the advantages of Kriging methods over simple numerical approximation is that the semivariogram provides an estimate of the *residual noise* in the data. If this residual noise is so large that the structure is not visible then no interpolation technique should be attempted. The semivariogram also provides an optimal estimate of the weighting function if an interpolation is appropriate. A useful by-product of the Kriging equations, the *estimation error* (Kriging variance) can also be mapped and can be a useful indicator of how reliable the interpolated values are at any point. Obviously the estimation error rises with distance from the observations, reflecting the fact that the estimation becomes less reliable.

Mapping the Kriging variance also gives an indication of where the interpolation is most reliable. Figure 9.9 shows both the result of ordinary Kriging, using a linear-plus-sill model and also a map of the Kriging variances. As we might expect, this suggests that the interpolation is most reliable closest to the data points but it also reveals several regions (in darker shades) where we may decide that the interpolated values are likely to be poor estimates.

Against that, Kriging is heavily computational, and in many circumstances the results may not be substantially better than those obtained from using a simpler method. Robinson and Zubrow investigated a range of spatial interpolators for simulated archaeological field data and concluded that:

"Kriging is computationally expensive and, if not fine tuned, it works poorly. Incidentally, it is very expensive in time to calculate and is not recommended unless you have special need and special justification. One justification could be concern about edge effects. Kriging is superior to other techniques in using local topography to infer the form along the edge." (Robinson and Zubrow 1999:79)

It is difficult, however, to take such a generalisation entirely seriously because 'Kriging' covers such a wide range of different specific methods. The example contains no examination of the variogram that should accompany a geostatistical analysis and a more considered conclusion might be that the methods are not amenable to 'button pushing' applications. Herein may actually lie the greatest strength of geostatistics within archaeology—it is not possible to apply the methods without gaining a sound understanding of the structure of the data. Contrary to Zubrow and Robinson's conclusions, other archaeologists have found geostatistical methods, including Kriging, to be useful. Ebert (1998) investigated their use in the analysis of real field data from Als, Denmark, and found two key advantages. Firstly, the methods indeed allow archaeologists to study the *structure* of spatial variation through the variogram, and secondly, they provide the ability to

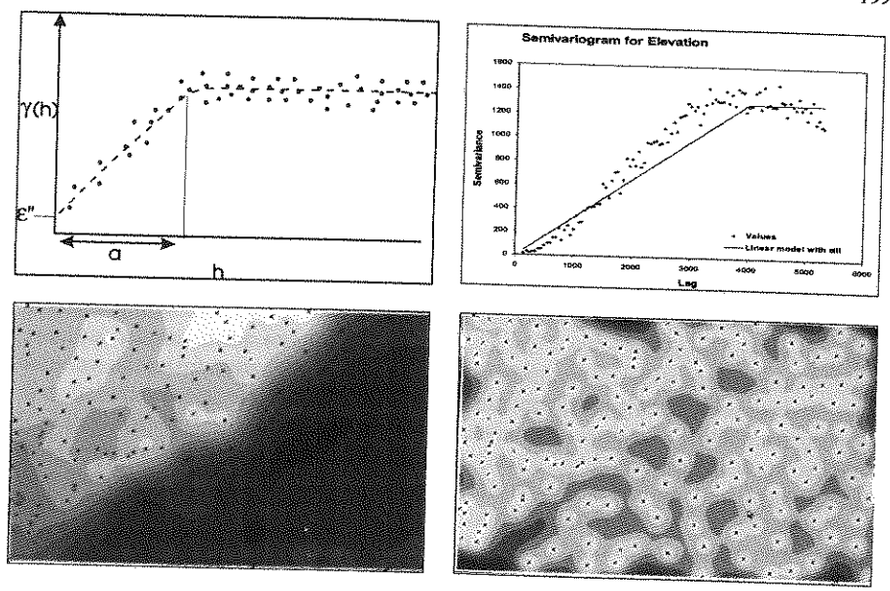

Figure 9.9 The use of the semivariogram in Kriging. Idealised semivariogram (top left) and the actual variogram for the elevation data (top right) with a linear-plus-sill model. Bottom left shows the result of ordinary Kriging using this model and bottom right shows the Kriging variance (higher values are darker). Generated with Marco Boeringa's Kriging Interpolator extension developed for the hydrology department of Amsterdam Water Supply.

produce interpolated maps of spatial distributions and to have confidence in those estimates.

9.8 SUMMARY

This chapter has introduced another large field of spatial technology, commenting on a range of different possible methods for interpolation of continuous variables from point data. It is not possible to 'recommend' one method over another for the analysis of archaeological data because the different characteristics of these methods mean that they are applicable to different archaeological questions. The important thing is to understand the assumptions and nature of the method that is being used, and be certain that it is appropriate for both the data and, more importantly, the archaeological question being posed of it.

We have described several approaches to creating 'density' surfaces from unclassified point observations, although we have also suggested that caution must be exercised whenever continuous or ordinal surfaces are produced from densities. Again, the issue here is that the result must be in some sense meaningful or assist in some way in the interpretation of the distribution. Too often, in published examples, it is not clear that either of these is true.

Trend surface analysis is suitable for the analysis of phenomena that can genuinely be assumed to exhibit simple types of variation over space: 'wave of advance' models, for example, or the quantifiable effects of economic geography.

If it is possible to justify the assumption that observations are the product of a single, relatively simple trend variable plus random noise then the method is suitable.

Where observations are generally free from error, and the closest fit of surface to observations is therefore required, methods based on triangulation will give the most local estimate although linear methods produce rather angular results that generate discontinuous derivatives.

In situations where the modelled surface needs to be complex, and we wish to treat our observations as samples of some underlying variable then more exact interpolators such as inverse distance weighting or spline-based methods are to be preferred. These will give a better estimate of the depth of an underlying layer, or of the changing values of pottery density in a settlement site, for example.

Finally, where a thorough investigation of the spatial structure of the variable is required then archaeologists may choose to invest both time and computation in geostatistics. While Kriging methods undoubtedly provide the best estimates of complex spatial variables, this is certainly at the cost of great complexity. The mathematical and statistical underpinnings of regionalised variable theory and Kriging have been barely touched upon in this chapter, and at the present time, the number of archaeologists who will have the time or inclination to properly understand these methods is very small indeed.

CHAPTER TEN

Visibility analysis and archaeology

"There's more to seeing than meets the eyball." (Chalmers 1999:5)

Ideas of visibility and intervisibility have always been important within archaeological analysis and interpretation. Whether as a factor in the location of 'defended' sites or in the frequently noticed choice of location for prehistoric funerary mounds on 'false crests', the level of visibility of single sites, or intervisibility within a group of sites has long been acknowledged as having played a role in the structuring of archaeological landscapes. What distinguishes recent trends is the gradual move away from little more than a tacit acknowledgement of the importance of visual factors, to instead assigning visibility a central analytical rôle. Visibility, usually equated with perception, has increasingly been viewed as an important correlate to more abstracted approaches to the study of the environment.

Whilst such visual phenomena have long been regarded as important, the acts of 'seeing' and 'looking' have been very hard to operationalise in any traditional, methodological sense. Concepts such as 'panoramic', 'prominent' and 'hidden' may well be useful heuristics, but are very hard to investigate in practice. Visual characteristics have often been mentioned but rarely explored in any formal and meaningful way. As a result, the incorporation of visibility or intervisibility with archaeological interpretations has tended to be anecdotal, at best. Although a number of techniques were developed through the 1970s and 1980s, it can be argued that the first systematic attempts to exploit the visual characteristics, or properties, of locations came in the early 1990s with the widespread adoption by archaeologists of Geographical Information Systems (GIS).

GIS provides archaeologists with a set of standard functions for calculating line-of-sight products from digital models of surface topography. With the discovery of such standard functions for the characterisation of visual parameters nestling within the generic GIS toolbox, researchers found themselves in the unique position of being able to undertake visibility-based studies in a rapid, quantifiable and repeatable way. Through the selection of the appropriate function or command, the precise area of land visible from a given point in a landscape could be identified and potential lines-of-sight between points evaluated. What is more, the results of such analyses could be expressed using the familiar and intuitive medium of the map sheet, so central to archaeological landscape study.

10.1 THE IMPORTANCE OF VISIBILITY IN ARCHAEOLOGICAL ANALYSIS

The visual appearance of a place is, in most cultures and for most people, the most significant impact a location has upon any individual's many senses. A place may feel cold or hot, may smell or have a particular sound quality, but it is the visual characteristics that are most frequently remembered and referred to, and often form the basis upon which a description of a place is instinctively based. It is this role in