8

Spatial analysis

8.1 Introduction

Spatial analysis lies at the core of GIS and builds on a long history of quantitative methods in archaeology. Many of the foundations of spatial analysis were established by quantitative geographers in the 1950s and 1960s, and adopted and modified by archaeologists in the 1970s and 1980s. For a variety of reasons, spatial analysis fell out of fashion both in archaeology and in the other social sciences. In part this was because of the perceived overgeneralisation of certain types of mathematical models, but also because of a shift towards more contextually orientated and relativist studies of human behaviour. Recently, however, there has been a renewed interest in the techniques of spatial analysis for understanding the spatial organisation of human behaviour that takes on board these criticisms. In the last decade there have been several advances within the social sciences, particularly geography and economics, in their ability to reveal and interpret complex patterns of human behaviour at a variety of scales, from the local to the general, using spatial statistics. Archaeology has participated somewhat less in these recent developments, although there is a growing literature that demonstrates a renewed interest in the application of these techniques to the study of past human behaviour. In this chapter we review some historically important methods (e.g. linear regression, spatial autocorrelation, cluster analysis) and also highlight more recent advances in the application of spatial analysis to archaeology (e.g. Ripley's K, kernel density estimates, linear logistic regression). Readers requiring more in-depth discussion of methods of spatial analysis are advised to consult the sources that we have made use of for this review, particularly Bailey and Gatrell (1995); Fotheringham *et al.* (2000b); Rogerson (2001) and Haining (2003).

8.2 Linear regression

Linear regression has long been a staple of quantitative analysis. It is used to model the relationship between two continuous variables, and is one of the more important methods in spatial statistics. Consequently we have described the technique, and some associated potential pitfalls, in some detail.

Linear relationships between two quantitative variables may be expressed in terms of the degree of *correlation*, of which there are three basic possibilities: positive correlation, zero correlation or negative correlation. Two variables are said to be positively correlated when there is a simultaneous increase in value between two numerical variables (Fig. 8.1a) and negatively correlated when one variable

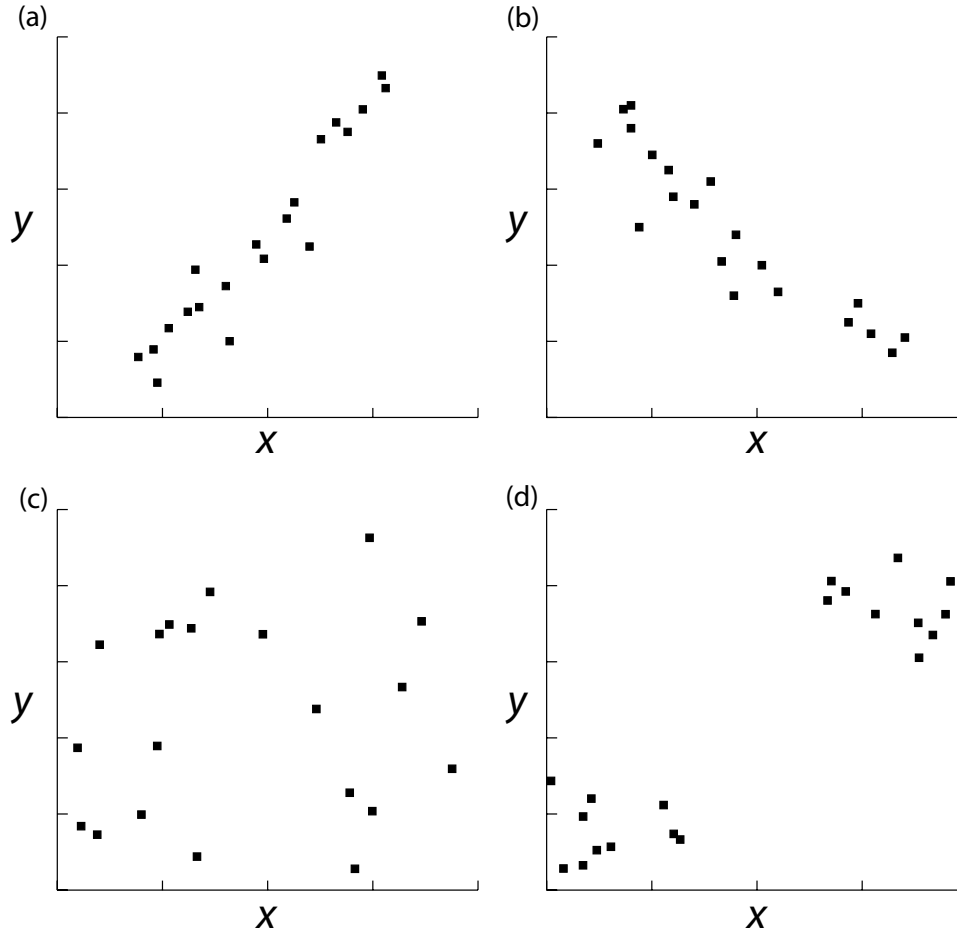

Fig. 8.1 Idealised correlation patterns: (a) positive, (b) negative, (c) zero, (d) spurious positive correlation in an uneven, clustered, dataset.

increases while the other decreases (Fig. 8.1b). Zero correlation occurs when there is no relationship between the two variables (Fig. 8.1c). Figure 8.1(d) shows a more complex relationship between two variables where there is a degree of positive correlation, although there is also a strong clustering pattern and *heteroscedasticity* (unevenness) in the data distribution that reduces the predictive value of the model. We will examine the first two examples initially, then return to cases of zero or spurious correlation later.

When examining the type and strength of correlation between two variables, one variable is considered to be *dependent* and the other *independent*. When plotted on an x, y-graph, the independent variable is plotted on the x-axis and the dependent variable is plotted on the y-axis. The difference between the dependent and independent variables is important and can be thought of as close to that of cause and effect.

To use a well-known archaeological example, the proportion of a particular type of raw material can often be shown to decline with the distance from the source of the raw material – i.e. distance and proportion are negatively correlated, as predicted by Renfrew's 'law' of monotonic decrement (Renfrew and Dixon 1976). In this case, it is the proportion of material that acts as the dependent variable, as its value is determined by its distance from the source. In situations where the suspected causal relationship is more ambiguous, for example between the number of artefacts and the size of an archaeological site, it is possible to speak of *interdependence*. In these cases, which variable is x and which is y is only significant in terms of what the regression analysis is specifically attempting to model.

While it is acceptable simply to describe the relationship between two quantitative variables as either positively or negatively correlated, it is often more useful to express this in terms of the strength of the relationship. The standard measure of the linear correlation between two variables is the *Pearson correlation coefficient*, symbolised by r and given by

$$r = \frac{\sum(x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \times \sum(y_i - \bar{y})^2}} \tag{8.1}$$

where $\bar{x}$ and $\bar{y}$ are the mean values of the independent and dependent variables. Values of r range from $+1.0$ for a perfect positive correlation to -1.0 for a perfect negative correlation. The midpoint, $r = 0.0$, denotes a complete absence of correlation between two variables. For example, the two variables listed in columns one and two in Table 8.1 possess a correlation coefficient of $+0.96$, meaning that they are highly positively correlated – as the x-value increases, so too does y in a highly predictable manner.

The correlation coefficient is usefully visualised as a 'line of regression' placed to minimise the sum of the vertical distances (actually the sum of the squared distances) from each point (the 'residuals') as illustrated in Fig. 8.2. In contrast to r, which simply describes the strength of the correlation and whether it is positive or negative, the r^2-value gives a better indication of the predictive power of the independent variable and can be interpreted as a proportion of the variation in the values of y that are determined by x. To convert this to a more tangible example, imagine that x and y are taken to refer, respectively, to site size and artefact count (so that artefact count is acting as the dependent variable). A correlation coefficient of 0.96 converts to a coefficient of determination, r^2, of 0.88. This indicates that 88 per cent of the variation in artefact count can be explained simply by site size. Note that it would be acceptable to turn this around and recompute the correlation coefficient using site size as the dependent variable (i.e. as y), if the purpose of the analysis was to predict site size on the basis of artefact count.

Two quantities of the line of regression, its slope (a), which defines the rate of change and the point at which the line crosses the y-axis (b, called the *intercept*),

Table 8.1 *Sample x- and y-values and the calculations for deriving r in Eq. 8.1 and Fig. 8.2*

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x}) \times (y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
	0.17	365	−0.32	−563.15	180.21	0.102	317 137.92
	0.21	401	−0.28	−527.15	147.60	0.078	277 887.12
	0.22	243	−0.27	−685.15	184.99	0.073	469 430.52
	0.25	502	−0.24	−426.15	102.28	0.058	181 603.82
	0.30	580	−0.19	−348.15	66.15	0.036	121 208.42
	0.32	780	−0.17	−148.15	25.19	0.029	21 948.42
	0.33	602	−0.16	−326.15	52.18	0.026	106 373.82
	0.40	702	−0.09	−226.15	20.35	0.008	51 143.82
	0.41	440	−0.08	−488.15	39.05	0.006	238 290.42
	0.48	900	−0.01	−28.15	0.28	0.000	792.42
	0.50	832	0.01	−96.15	−0.96	0.000	9 244.82
	0.56	1 023	0.07	94.85	6.64	0.005	8 996.52
	0.58	1 100	0.09	171.85	15.47	0.008	29 532.42
	0.62	890	0.13	−38.15	−4.96	0.017	1 455.42
	0.65	1 400	0.16	471.85	75.50	0.026	222 642.42
	0.69	1 480	0.20	551.85	110.37	0.040	304 538.42
	0.72	1 435	0.23	506.85	116.58	0.053	256 896.92
	0.76	1 542	0.27	613.85	165.74	0.073	376 811.82
	0.81	1 703	0.32	774.85	247.95	0.103	600 392.52
	0.82	1 643	0.33	714.85	235.90	0.109	511 010.52
Mean	0.49	928.15					
Sum	9.8	18 563			1 786.51	0.849	4 107 338.50

are collectively referred to as the *regression constants* and are given by:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{8.2}$$

and:

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} \tag{8.3}$$

The slope and intercept values allow for predictions to be made for y for any given value of x, as given by:

$$y = a + b \times x \tag{8.4}$$

https://doi.org/10.1017/CBO9780511807459.008 Published online by Cambridge University Press

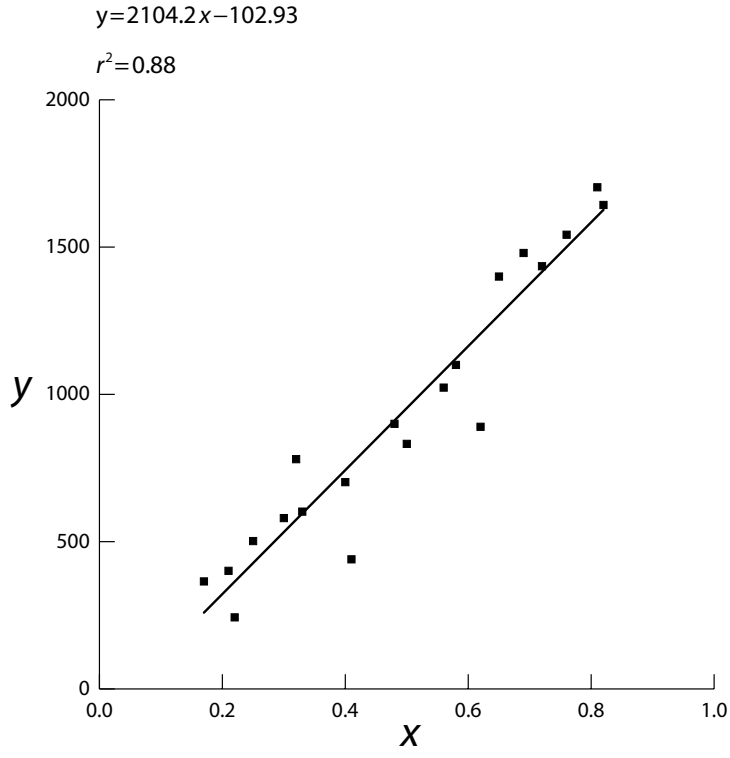

Fig. 8.2 A line of regression fitted to the x- and y-values in Table 8.1, shown with the coefficient of determination (r^2) and the regression equation.

For example, if Fig. 8.2 showed the correlation between the size of a site in hectares (x) and the number of artefacts recovered from surface collection (y) and $a = -102.93$ and $b = 2104.2$, then it would be possible to predict the number of artefacts from a site of 0.9 ha by substituting these values into Eq. 8.4: $y = (-102.93) + 2104.2 \times 0.9 = 1790.9$.

Predictions also have an associated standard error, calculated as

$$s_{y-\hat{y}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} \tag{8.5}$$

where $\hat{y}$ is the predicted value of the dependent variable. Using the data from Table 8.2, the standard error of the prediction is the square root of 348 999.1 divided by 18, which equals 139.2. One standard error is roughly equivalent to 68 per cent of observations if the residuals are normally distributed. In this example, assuming that this is the case, then the prediction for the number of artefacts on a site 0.9 ha in size is 1790.9 $\pm$ 139.2, which has a 68 per cent probability of being correct. If

https://doi.org/10.1017/CBO9780511807459.008 Published online by Cambridge University Press

Table 8.2 *Data for the calculation of standard error for the regression analysis of the variables given in Table 8.1*

y_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
365	254.8	110.2	12 147.6
401	339.0	62.0	3 850.0
243	360.0	−117.0	13 687.6
502	423.1	78.9	6 222.1
580	528.3	51.7	2 669.8
780	570.4	209.6	43 926.3
602	591.5	10.5	111.2
702	738.8	−36.8	1 350.6
440	759.8	−319.8	102 266.9
900	907.1	−7.1	50.2
832	949.2	−117.2	13 728.8
1 023	1 075.4	−52.4	2 748.1
1 100	1 117.5	−17.5	306.5
890	1 201.7	−311.7	9 7140.7
1 400	1 264.8	135.2	18 279.0
1 480	1 349.0	131.0	17 169.4
1 435	1 412.1	22.9	524.7
1 542	1 496.3	45.7	2 092.0
1 703	1 601.5	101.5	10 307.9
1 643	1 622.5	20.5	419.7
Sum			348 999.1

greater accuracy is needed, then doubling the standard error to ±278.4 provides a 95 per cent probability of being correct.

The Pearson correlation coefficient (r) and the coefficient of determination (r^2) are included in nearly all computer statistical packages, including Microsoft Excel, SPSS, S-Plus and R. However, the use of Pearson's r for describing relationships can be problematic and it is worth reviewing the major pitfalls into which unexperienced users often stumble. Firstly, *it is unwise to assume a causal relationship solely on the basis of an observed correlation*. For example, while it may be generally observed that there is a strong positive correlation between the size of an archaeological site and the age of the director of the excavation, there is no causal relationship in either direction between these two variables. In cases where a causal relationship is suspected, it is always worth investigating the possibility that there are intermediary variables (e.g. in the sequence of: age → seniority → size of funding grant → size of archaeological site).

Regression analysis also depends on a number of assumptions that must be shown to be true before any measured correlation can be shown to be meaningful (cf. Shennan 1988, pp. 139–142). For example, the predictive value of the statistic is dependent on the variation around the line of regression being *homoscedastic* or evenly distributed. If this is not the case the variation is described as *heteroscedastic*.

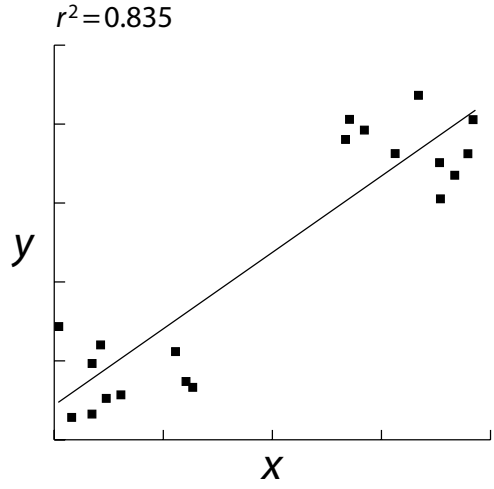

Fig. 8.3 A line of regression fitted to a heteroscedastic point distribution. There is a strong positive correlation ($r = 0.91$), but the coefficient of determination ($r^2 = 0.84$) is meaningless.

Although a heteroscedastic distribution can be subject to regression and a formula obtained, the results are largely meaningless. Figure 8.3, for example, shows two random clusters of points that individually have an r^2 of 0 but collectively have an r^2 of 0.835. The heteroscedastic nature of the distribution, however, means that x has very little predictive value for y. Similarly meaningless results can occur when one or two outliers from a random distribution result in a line of regression with a strong positive or negative value. In these cases, visual inspection of the scatterplot is essential to ensure that the distribution of points is evenly spread along the x- and y-values. If this is not the case and clustering or outliers are evident, then the latter should be removed (and separately accounted for) and clusters investigated separately. Even if obvious outliers are not present, analysis of residuals in the ways described by Shennan (1988, pp. 139–144) can provide considerable insight into the structure of a linear relationship.

Thirdly, linear regression attempts to model linear relationships between variables, but a non-linear trend might be apparent in the data, as in Fig. 8.4. When visual inspection of an x, y-plot suggests a non-linear pattern, it is appropriate to transform one or both variables prior to performing a linear regression, especially as this will not reduce the predictive nature of the model. Common transformations include the square, the square root, the natural log or $\log_{10}$ of one or both variables. Experimentation with different transformations is often needed to obtain the optimum correlation coefficient with data that exhibit curvilinear tendencies. Shennan (1988, pp. 135–165) provides a comprehensive discussion of transformations of variables to improve linear regression.

Fourthly, one assumption of regression analysis as applied to spatial data is that each of the two observations on the x- and y-variables should be independent and

https://doi.org/10.1017/CBO9780511807459.008 Published online by Cambridge University Press

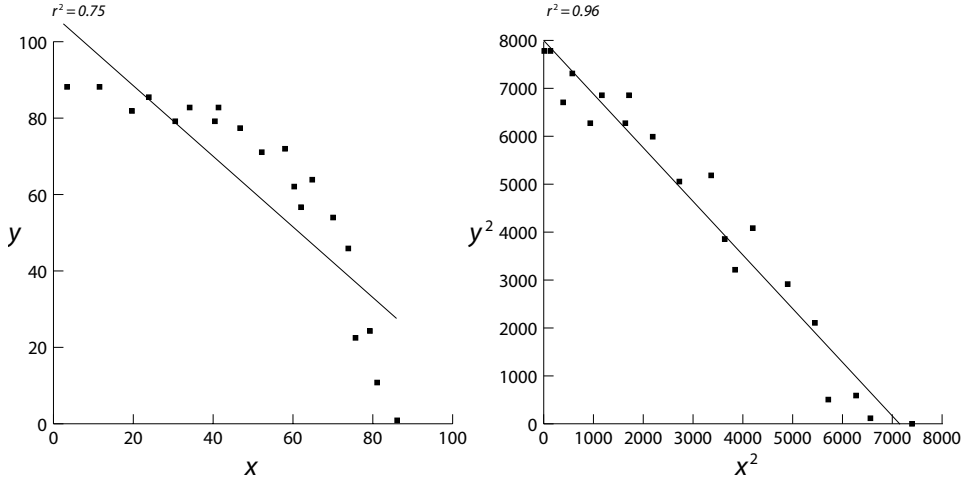

Fig. 8.4 Transforming a variable to improve the correlation coefficient. The scatterplot on the left has a slight curvilinear trend, which can be transformed to a linear trend by squaring the x- and y-variables, as shown on the right. The regression equation for the transformed data therefore becomes $y^2 = a + b \times x^2$, or $y = \sqrt{a + b \times x^2}$.

not spatially autocorrelated (see Section 8.3), nor should the residuals be spatially autocorrelated (Rogerson 2001, p. 154). If both x and y have high spatial autocorrelation then the variance and standard error of r, which is a function of the number of observations as defined in Eq. (8.5), will be underestimated because Pearson's r assumes that each pair of observations are independent of each other (Haining 2003, p. 279). The correlation coefficient will as a consequence be overestimated. Although an assessment of autocorrelation using methods such as Moran's I can be made on each of the two variables prior to the regression, spatial dependence can also be assessed by mapping the residuals from a regression analysis and visually searching for evidence of spatial autocorrelation (Fotheringham *et al*. 2000b, pp. 162–165).

For example, imagine that the relationship between the amount of prehistoric and medieval pottery from a sample of surface collection areas was being investigated (Table 8.3). There is a null hypothesis of no correlation (i.e. higher or lower amounts of prehistoric pottery have no bearing on how much medieval pottery is recovered, and vice versa). A Pearson's r analysis (8.1) returns a value of 0.4 with medieval pottery as the dependent variable, suggesting a slight but definite positive correlation between the two pottery types, perhaps indicating that the prehistoric and medieval sites overlap to a certain degree.

A plot of the residuals of y, however, shows considerable positive spatial autocorrelation (Fig. 8.5). High positive deviations between predicted and observed medieval pottery cluster in the centre and upper right, and high negative deviations only appear in the upper left and the bottom of the survey area. This indicates that the observations are not independent. This in turn warns that the results of the

Table 8.3 *Counts of prehistoric and medieval pottery recovered from ten surface collection areas*

Area	Prehistoric	Medieval	Predicted	Residual
1	30	2	21.1	−19.1
2	6	3	8.4	−5.4
3	56	56	34.9	21.1
4	42	23	27.5	−4.5
5	21	45	16.4	28.6
6	59	12	36.6	−24.6
7	56	65	34.9	30.1
8	21	30	16.4	13.6
9	43	9	28.1	−19.1
10	33	2	22.7	−20.7
Sum	367	247	247.0	0.0

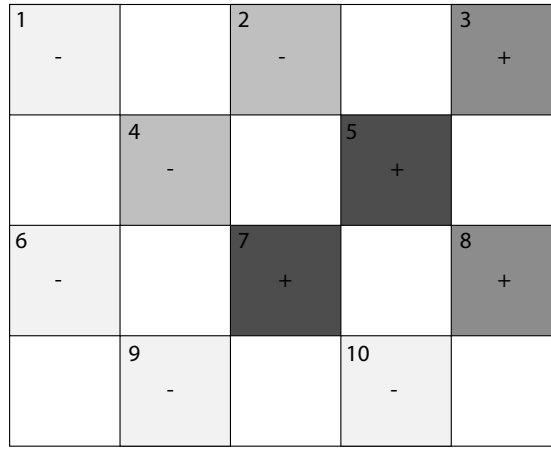

Fig. 8.5 A plot of the residuals of predicted versus actual medieval pottery. Numbers in upper left of squares refer to individual surface collection areas as defined in Table 8.3.

correlation analysis are not necessarily valid and that r will be overestimated; thus any judgement of the significance of the correlation becomes problematic.

When a plot of the residuals exhibits positive spatial autocorrelation then the regression analysis may be improved either by investigating additional explanatory variables by producing *added-variable plots* (Haining 1990), by using the technique of *spatial regression* (Rogerson 2001, pp. 187–188), modelling the residuals as a function of the surrounding residuals (Bailey and Gatrell 1995), or by employing a geographically weighted regression (GWR) technique as developed by Fotheringham and colleagues (Fotheringham *et al*. 1998). Extensive discussion and many worked examples of GWR are provided in Fotheringham *et al*. (2002a).

https://doi.org/10.1017/CBO9780511807459.008 Published online by Cambridge University Press

Alternatively, it is possible to control for the influence of autocorrelated variables on the sample by first establishing the number of spatially independent pairs of observations (n') from all observations n and using only the former to establish the significance of the correlation (Clifford and Richardson 1985; Haining 2003, pp. 278–279). These and other common problems with regression analysis are summarised in Table 8.4.

8.3 Spatial autocorrelation

The term 'spatial autocorrelation' refers to the degree of correlation between pairs of observed values and the distance between those observations in spatial distributions (Cliff and Ord 1981). Positive spatial autocorrelation describes a state where attribute values exhibit a tendency to be more similar the closer they are together (e.g. such as elevation, where the closer two sample points are together, the more likely they are to share a similar elevation). If there is no apparent relationship between spatial proximity and attribute value, then the distribution exhibits zero spatial autocorrelation. Negative spatial autocorrelation occurs when similar attribute values are located away from each other (Worboys 1995, pp. 157–158).

Having an understanding of the spatial autocorrelation of a data distribution provides important supporting information for certain types of modelling procedures. In particular, the linear regression of autocorrelated data is problematic for reasons described in the previous section. On the other hand, interpolation is only a valid exercise for data with some degree of positive autocorrelation. While this can be assumed for many environmental phenomena, such as elevation, rainfall, temperature, etc., it cannot be assumed for anthropogenic data. Creating a continuous surface of artefact densities from, for example, a sample of testpits is only useful if the sample data show some degree of positive autocorrelation (which is assessed within the technique of kriging as described in Chapter 6). Thus interpolation methods that incorporate measures of autocorrelation into their procedures – such as the technique of kriging described in Chapter 6 – typically produce continuous surfaces that are more accurate than other methods.

More generally, there has been some optimism that measures of spatial autocorrelation may have wider application in archaeology (Williams 1993), but thus far the most successful applications have been constrained to the analysis of Mayan terminal monument dates (Premo 2004).

The most common method of measuring autocorrelation is using *Moran's I* statistic (Moran 1950):

$$I = \left(\frac{n}{\sum_{i}^{n} \sum_{j}^{n} w_{ij}} \right) \left[\frac{\sum_{i}^{n} \sum_{j}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i}^{n} (x_i - \bar{x})^2} \right] \tag{8.6}$$

https://doi.org/10.1017/CBO9780511807459.008 Published online by Cambridge University Press

Table 8.4 *Some common problems, consequences and solutions with regression analysis*

Problem	Consequences	Diagnostic	Corrective action
Residuals non-normal	Inferential test is likely to be invalid	Shapiro–Wilks test (Chapter 7)	Transform y-values
Heteroscedastic	Biased estimation of error variance and invalid inference	Plot of residuals against y	Transform y-values
Non-independent variables	Underestimation of variance and invalid inference	Moran's I	GWR, added-variable plots, spatial regression
Non-linear relationship	Poor fit and non-independent residuals	Scatter plot	Transform y- and/or x-variable
Outliers	Can severely affect model estimates and fit	Scatter plot	Delete outliers
Non-interval or ratio data	Linear regression not valid		Logistic regression

Source: Adapted from Haining (1990, pp. 332–333); Rogerson (2001, p. 146).

where subscripts i and j refer to the spatial objects of which there are n, $\bar{x}$ is the mean of all attributes and w_{ij} is a weighting function to reduce the impact of distant points. If the variable of interest x is first transformed to a z-score $\{z = (x - \bar{x})/s\}$ then formula can be simplified to (Rogerson 2001, p. 167):

$$I = \frac{\sum_i \sum_j w_{ij} z_i z_j}{(n-1) \sum_i \sum_j w_{ij}} \tag{8.7}$$

The weighting function w_{ij} is most often an inverse distance measure ($\frac{1}{d_{ij}}$). For area data a measure of *binary connectivity*, where $w_{ij} = 1$ if i and j are adjacent and $w_{ij} = 0$ if not, is frequently used instead (Rogerson 2001, p. 167).

The expected value of Moran's I, if there is no spatial autocorrelation, is defined by $E(I)$:

$$E(I) = -\frac{1}{n-1} \tag{8.8}$$

Values of I larger than $E(I)$ indicate positive autocorrelation and values lower indicate negative autocorrelation (Fotheringham *et al.* 2000a). The statistical significance of any departure from this expected value can be tested using an assumption of normality (i.e. that the values of x_i are drawn from a normal population). In cases where n is 'large' the standardised statistic

$$Z = \frac{I - E(I)}{\sqrt{\text{var}(I)}} \tag{8.9}$$

https://doi.org/10.1017/CBO9780511807459.008 Published online by Cambridge University Press

can be used where the variance of I under an assumption of normality is

$$\text{var}(I) = \frac{n^2(n-1)S_1 - n(n-1)S_2 + 2(n-2)S_0^2}{(n+1)(n-1)^2 S_0} \tag{8.10}$$

where

$$S_0 = \sum_i^n \sum_{j \neq i}^n w_{ij} \tag{8.11}$$

$$S_1 = \frac{\sum_i^n \sum_{j \neq i}^n (w_{ij} + w_{ji})^2}{2} \tag{8.12}$$

$$S_2 = \sum_k^n \left(\sum_j^n w_{ij} + \sum_j^n w_{ji} \right)^2 \tag{8.13}$$

Less restrictively, an assumption of randomisation can be used (i.e. where observed I is compared to an expected I if x_i was randomly distributed) (Hodder and Orton 1976, p. 178). In this case, the variance of I is given by

$$\text{var}(I) = \frac{nS_4 - S_3 S_5}{(n-1)(n-2)(n-3)\left(\sum_i^n \sum_j^n w_{ij} \right)^2} \tag{8.14}$$

where

$$S_3 = \frac{n^{-1} \sum_i^n (x_i - \bar{x})^4}{\left(n^{-1} \sum_i^n (x_i - \bar{x})^2 \right)^2} \tag{8.15}$$

$$S_4 = (n^2 - 3n + 3)S_1 - nS_2 + 3\left(\sum_i^n \sum_j^n w_{ij} \right)^2 \tag{8.16}$$

$$S_5 = S_1 - 2ns_1 + 6\left(\sum_i^n \sum_j^n w_{ij} \right)^2 \tag{8.17}$$

The variance can then be used in (8.9) to calculate a Z-value, which can then be compared to a normal distribution for significance (Fotheringham *et al*. 2000b, p. 204). In cases where n is 'small' then it may be necessary to simulate the parameters of I using Monte-Carlo methods (see Box 8.1) for which Fotheringham *et al*. (2000b, pp. 204–209) provide a worked example.

https://doi.org/10.1017/CBO9780511807459.008 Published online by Cambridge University Press

Box 8.1 Monte-Carlo simulation

Monte-Carlo simulation predates the rise of computing, but it only really became an important form of statistical sampling in the second half of the the last century, particularly in physics (Robert and Casella 2004). More recently it has emerged as an increasingly important method of statistical sampling in the social sciences, as it provides a way of estimating the parameters of complex populations. Monte-Carlo simulation thus has an important role to play in GIS.

The basis of the technique is common to that of statistical sampling: that a random sample of individuals from a population will show some correspondence to the population parameters, and thus the latter can be estimated from the sample. In many cases where populations are large and potentially diverse it is unclear how a random sample should be generated, how many samples should be taken or whether any given random sample is at all representative of the population. Monte-Carlo simulation reduces this uncertainty by taking repeated random samples (often 1000 or more). It is then possible to examine the distribution of values of some statistic (usually the mean) across the samples.

For example, a common archaeological GIS problem is whether the average viewshed size of a sample of archaeological sites is different from the average viewshed size of the background landscape. Viewshed calculations are computationally intensive, and thus even for a modestly sized study area it is unrealistic to attempt to establish the viewshed size for every cell. Taking a random sample of points in the landscape and comparing this with the archaeological sample is the only realistic option, but then it may not be clear whether the random sample is representative of the background population. A Monte-Carlo simulation approach to this problem will take several random samples and so provide a better estimate of the population parameters.

A common starting point for Monte-Carlo simulation is to take 1000 simple random samples each consisting of 1000 individuals and then to average the results. Note that this is computationally equivalent to taking 1 000 000 samples, which may exceed the population size! In cases like this it is acceptable to reduce the size and number of the samples; for example, Lake and Woodman (2000) used 100 random samples of 30 locations to estimate the parameters of the viewshed characteristics of their study area (see Chapter 3). A result is significant if the statistic for the sites falls on the edge of or outside the range of values of the statistics for the non-site samples.

Obviously this is not a task that can be performed manually – ideally it requires a simple program that selects the random samples, calculates their viewshed size and stores the results in a log file. Repetitive tasks like these are relatively easy to program in, for example, Visual Basic for ArcGIS. Alternatively, sets of random x, y-locations can be generated in a spreadsheet or statistical package such as R, saved as text files, and then imported into the GIS to be used as the basis for the calculations.

In other applications, such as spatial analysis, Monte-Carlo techniques are used to establish what a random distribution actually looks like. This is important in Ripley's K (this chapter) where a known distribution of points must be compared to a random distribution to establish whether or not it is distinctive (i.e. is clustered, or regular). Although still computationally intensive, 1000 samples of 1000 points is a good starting point as it increases the confidence that the characteristics of a random sample are accurately estimated.

Robert and Casella (2004) provide a good introduction to Monte-Carlo methods.

Although Moran's I can be calculated by hand, it soon becomes unwieldy when there are more than ten or so objects. It is much easier to use a statistical computer package, particularly when computing the significance of the statistic (as the previous formulae might suggest!). Modules for calculating Moran's I are readily available for many popular GIS packages (e.g. Spatial Statistics for ArcView (Monk 2001), r.moran for GRASS and AUTOCORR for Idrisi) and are included in dedicated geostatistical packages such as G+.[1] It is also included in the freely available CrimeStat[2] spatial statistics program (Levine 2002).

The degree of positive spatial autocorrelation in a spatial dataset may assist with the interpretation of certain anthropogenic phenomena. For example, the spatial structure of 'event horizons' such as the spread of agriculture (Sokal *et al*. 1989; Gkiasta *et al*. 2003) or the collapse of the Classical Mayan state (Kvamme 1990d; Williams 1993; Neiman 1997), depend to a large extent on establishing positive autocorrelation between the dates and locations at which the event is first observed. A recent use of measures of spatial autocorrelation in archaeology can be found in (Premo 2004), who applies autocorrelation statistics to contextual terminal dates of Mayan sites within local neighbourhoods to provide further insight into Mayan 'collapse'. While examples such as these do depend on adequately sampled data, there are certainly many more potential archaeological applications of the analysis of autocorrelation than have yet been realised.

8.4 Cluster analysis

Archaeologists frequently use points to represent the location of artefacts, features and sites. The analysis of point distribution patterns is therefore an important tool for describing, interpreting and explaining the spatial characteristics of these phenomena. Point distribution patterns are often described in terms of their configuration vis-a-vis three idealised states – namely random, clustered or regular (Fig. 8.6a–c). In reality, spatial arrangements, whether artefactual or settlement, can rarely be so simply described. Analysis of distribution patterns needs to be sensitive to the fact

[1] www.gammadesign.com/.
[2] www.icpsr.umich.edu/NACJD/crimestat.html.

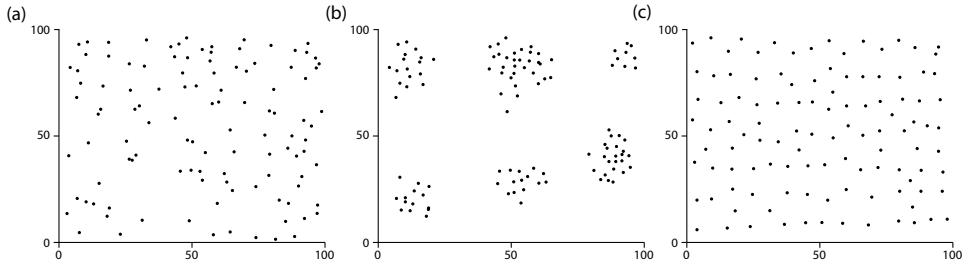

Fig. 8.6 Idealised point distributions: (a) nearly random, (b) nearly clustered, (c) nearly regular.

that several different smaller-scale patterns may exist within a study area and that different types of patterning often exist at different spatiotemporal scales.

In the case of settlement pattern analysis, regular spacing of sites has been taken to reflect either a form of competition between settlements, the existence of site catchments, or a combination of both as a result of demographic growth from an initial random distribution (Hodder and Orton 1976, pp. 54–85; Perlès 2001, pp. 132–147). Clustering of sites may result from a number of factors, but localised distribution of resources and the emergence of polities or regional centres have often been highlighted (Roberts 1996, pp. 15–37; Ladefoged and Pearson 2000). In contrast, random distributions have usually been treated as the statistical null hypothesis, though several commentators provide good examples of how apparently random distributions can be conditioned by less-obvious environmental, biological and social variables (Maschner and Stein 1995; Woodman 2000b; Daniel 2001). In general terms, the interpretation of archaeological settlement distributions is in need of new theory building coupled with renewed empirical and experimental investigation. Recent work by Premo (2004) on Mayan site distribution provides a good example of such an approach.

One major issue in the analysis of point distributions is the effect that the size of the study area has on the detection and characterisation of patterning. Figure 8.7 shows how adjusting the scale of analysis has a major influence on both the homogeneity, intensity and clustering tendencies of point distributions. In the entire study area, $A1$, the pattern is homogenous with a clustered structure (i.e. clustering occurs relatively evenly) such that a frequency distribution of the distances to each point's nearest neighbour would be normally distributed. At smaller scales, for example in area $A2$, the pattern is heterogeneous with a strong left to right gradient. A neighbourhood density function would be positively skewed with a bimodal tendency. Area $A3$ is similarly heterogeneous, although its density value is significantly lower than $A2$. Area $A4$ has a high intensity and homogenous distribution, although here it is far more regular than seen elsewhere.

The dichotomy of dispersion vs. nucleation created by many point-based analyses provides only a coarse characterisation of human settlement patterns. In particular,

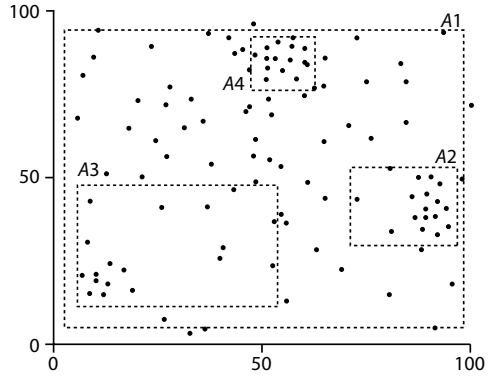

Fig. 8.7 Several smaller-scale point patterns are apparent in the near-random distribution of points in area *A*1. For example, areas *A*2 and *A*3 can be described as containing clustered distributions, while *A*4 is better described as regular.

analyses that are sensitive at only one scale (like nearest neighbour analysis) may overlook more complex, multiscalar, spatial patterns.

8.4.1 *Nearest neighbour analysis*

A favourite, if now old-fashioned, technique used by archaeologists to analyse point distributions is nearest neighbour analysis. Clark and Evans (1954) first explored the utility of nearest neighbour analysis in ecology (Box 8.2), and the New Geographers were soon applying it to human settlement patterns (Dacey 1960; Haggett 1965). Its use for archaeological settlement pattern analysis followed some time later in the early 1970s (Hodder and Hassell 1971a; Clarke 1972; Hodder 1972; Whallon 1974; Washburn 1974; Hodder and Orton 1976) and continued through the 1980s and 1990s. The technique retains its prominence in archaeology both in general textbooks (e.g. Wheatley and Gillings 2002) and in culturally specific studies (e.g. Perlès 1999; Ladefoged and Pearson 2000). Its popularity is a product of two factors: it is straightforward to calculate (see Box 8.2) and it provides an easily interpreted coefficient.

There are, however, several significant limitations to nearest neighbour analysis. It was initially designed to detect spatial patterning between 1st nearest neighbours and thus is not suited to identifying multiscalar effects.

For example, Fig. 8.8 shows a hypothetical distribution of a number of sites represented by points. A single-order nearest neighbour analysis applied to the point distribution in the left panel would detect the presence of clusters, and a *K-means* statistic (described below) could be employed to show that the optimum number of clusters was probably eight. However, neither of these analysis would be able to identify the fact that there is also a higher-order scale producing three clusters. Furthermore, if we include the finer artefact-scale resolution represented on the right panel (rather than just an approximation of the centre of the artefact distribution), then clustering can be shown to exist at three different spatial scales:

Box 8.2 Clark and Evans' nearest neighbour statistic

This useful but problematic statistic is calculated by dividing the mean of the observed distance between each point and its nearest neighbour (denoted by $\bar{R}_o$) by an expected value of R if the distribution was random ($\bar{R}_e$). This latter is estimated using the equation:

$$\bar{R}_e = \frac{1}{2\sqrt{\lambda}} \tag{8.18}$$

where λ is the density of points in the study area (i.e. the mean intensity of points), as given by Eq. (8.22).

The ratio of $\bar{R}_o$ to $\bar{R}_e$ (denoted by R) provides the statistic:

$$R = \frac{\bar{R}_o}{\bar{R}_e} \tag{8.19}$$

If $\bar{R}_o$ and $\bar{R}_e$ are equal – in other words, the observed mean nearest neighbour distance is equivalent to that predicted if the distribution were random – then their ratio (R) will be equal to 1. In a clustered distribution, the mean distance between points will be less than when they are randomly distributed. Thus an R-value less than 1 indicates a clustered distribution. If R is greater than 1 (up to its theoretical maximum of 2.15), this indicates that the points are more regularly spaced.

The significance of R is dependent on the sample size and density of the point distribution. It is known that the variance of mean distances between neighbours in a random distribution is

$$V[R_e] = \frac{4 - \pi}{4 \times \pi \times \lambda \times n} \tag{8.20}$$

where n is the number of points and λ is the mean intensity of points (Rogerson 2001, p. 162). As we can estimate the variance, a z-test can therefore be used to test the null hypothesis of random distribution:

$$z = \frac{(R_o - R_e)}{\sqrt{V[R_e]}} \tag{8.21}$$

Tables of the standard normal distribution can be used to assess significance: z-values of 1.96 or greater indicating significant uniformity and values of -1.96 or lower indicating a significant tendency towards clustering.

(i) artefacts forming sites (clusters i–x); (ii) sites forming primary clusters (clusters 1–8), and (iii) primary clusters forming secondary clusters (clusters A–C).

Increasing the nearest neighbour measurement to the second, third, . . . , nth neighbour may detect clustering at different scales, but the statistical validation of patterning then becomes difficult (Hodder and Orton 1976, p. 41). Nearest neighbour analysis is also significantly influenced by the size of the area to be analysed,

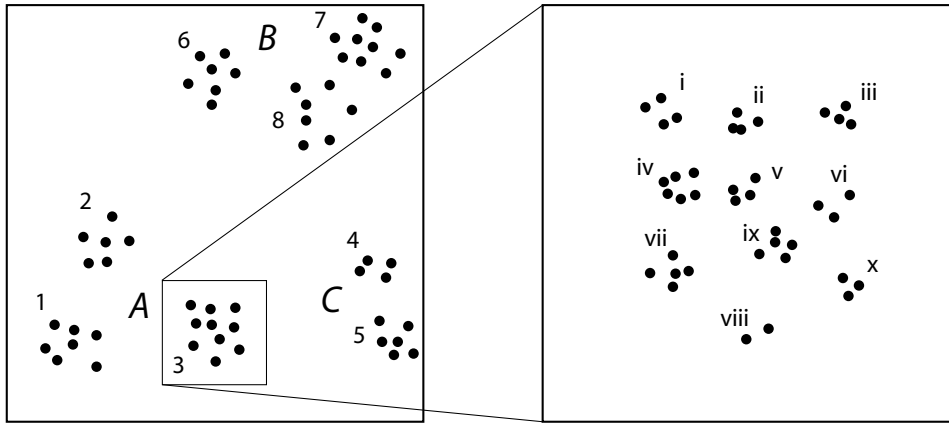

Fig. 8.8 Multiscalar patterning: three large clusters (*A*, *B*, *C*) are each composed of smaller clusters (1–8), which themselves consist of smaller clusters (e.g. i–x for cluster 3).

with regular, random or clustered distributions partially dependent on the shape of the study area. The size of surrounding area included in the analysis also can significantly influence the identification of clustering; the greater the amount of empty space surrounding a central distribution of random points, the more likely it is that the pattern will be identified as clustered. There are 'workarounds' for these problems but the technique nevertheless remains a somewhat blunt instrument with which to describe point distribution patterns.

8.4.2 *Ripley's K*

As GIS-led approaches to the collection and management of archaeological survey data are able to store data at several different scales within the same environment (e.g. artefacts, sites and regions), more sophisticated and spatially sensitive techniques are required to identify and characterise distribution patterns. One technique that addresses some of the inherent problems of nearest neighbour analysis is *Ripley's K-function* (Ripley 1976, 1981). The technique was designed to identify the relative aggregation and segregation of point data at different spatial scales and the shape of the study area has little effect on the assessment of patterning. The statistic is defined for a process of point intensity λ, where $\lambda K(r)$ defines the expected number of neighbours in a circle of radius r at an arbitrary point in the distribution (Pélissier and Goreaud 2001, p. 101). The K-distribution is a cumulative frequency distribution of average point intensity at set intervals of r. Significance intervals are generated by Monte-Carlo simulation of random distributions of the points and a 95 per cent confidence interval can usually be obtained within 1000–5000 iterations (Manly 1991). These estimates can be compared with the observed values of K to provide a statistically robust measure of cluster size and cluster distance in the dataset. For clarity of presentation the cumulative

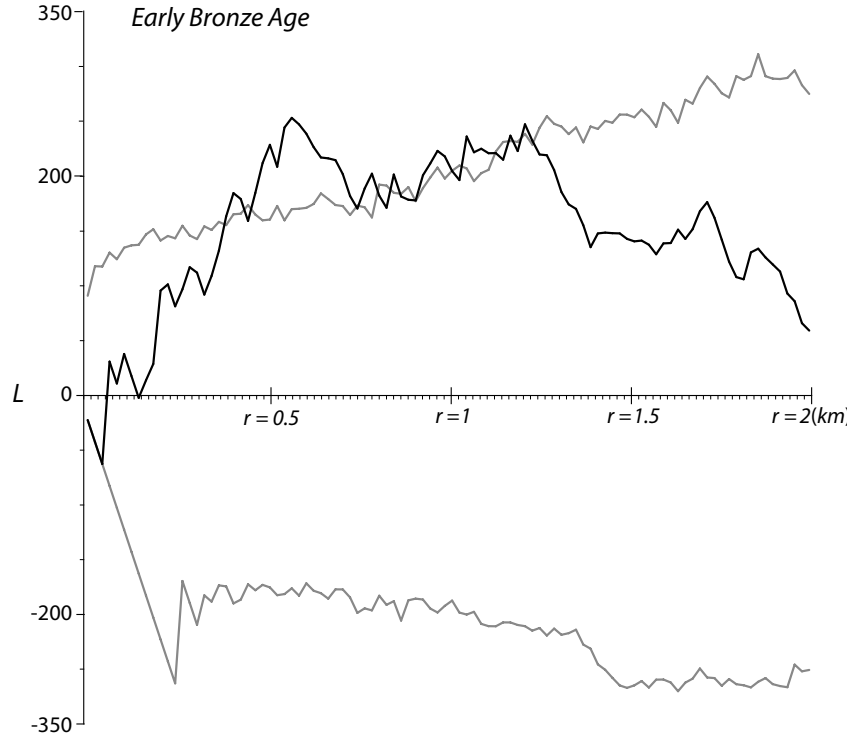

Fig. 8.9 Identification of multiscalar clustering in the Kytheran Early Bronze Age using Ripley's K. The presence of clusters of settlements between 500 and ~700 m is attested by the peak at that position on the x-axis, with significant but less obvious clustering occurring between ~700 and 1250 m (for details see Bevan and Conolly in press).

K-distribution is usually transformed to $L(r) = \sqrt{K(r)/\pi} - r$, where the expectation under randomness ($L(r) = 0$) is a horizontal line (Fig. 8.9). $L(r) < 0$ means that there are fewer than expected neighbours at distance r, suggesting a regular pattern, and $L(r) > 0$, means that there are more neighbours than expected at distance r, indicating a clustered pattern (Pélissier and Goreaud 2001, p. 102).

Ripley's K is available within comprehensive statistical packages such as R, as a module in the freely available spatial statistics package ADE-4[3] (Thioulouse *et al.* 1997) and within the third-party extension for ArcView Spatial Statistics[4] (Monk 2001). Although it is more complex than the Clark and Evans nearest neighbour statistic, and it may take longer to calculate because of the necessity of simulating the parameters of a random distribution, Ripley's K offers a much better route to investigating the spatial structure of point patterns. Bevan and Conolly (in press), for example, have used the technique to investigate the changing nature of settlement

[3] http://pbil.univ-lyon1.fr/ADE-4/.
[4] http://arcscripts.esri.com/.

patterns on the island of Kythera. They were able to show not only the presence of clustering during different phases of settlement on the island, but also how both the structure of the distribution and the size of clusters of settlements changed over time, reflecting different settlement strategies.

8.5 Identifying cluster membership

If clustering is identified in a distribution of point data, the second phase of analysis usually consists of defining the number and location of those clusters. There are a number of techniques that may be used to achieve this objective, which can be divided into three groups depending on whether they use *hierarchical* (Sokal and Sneath 1963; Sneath and Sokal 1973), *partitioning* (Ball and Hall 1970) or *density* methods for cluster definition (Silverman 1986). Hierarchical methods start with individual objects and progressively group them into fewer higher-order clusters so that eventually all objects assume membership of one group. Partitioning methods begin with the complete distribution, and break it into a number of smaller units, while density approaches identify dense concentrations of objects. In this section we describe three approaches: hierarchical cluster analysis, *k*-means partitioning and density analysis.

8.5.1 *Hierarchical cluster analysis*

This approach to clustering has a very long history of application in archaeology. It works by creating a 'distance matrix' between objects based on their attribute states, and as applied to spatial data the linear distances between all points in the dataset form the basis of the matrix. For example, the relationships between the points in Fig. 8.10 can be converted to a distance matrix (Table 8.5), which can be used to construct a *dendrogram* defining the hierarchical grouping of individual points (Fig. 8.11).

The construction of the dendrogram begins in an agglutinative manner by defining each separate point as a group unto itself, then locating the pair that possess the smallest distance. In this example, this occurs between points 1 and 5, which are 0.7 units apart. The pair of points with the second smallest distance (7 and 9) are then grouped together. Individual points and grouped pairs may link to existing groups on the basis of a specified rule, which defines the cluster method and ultimately the shape of the resulting dendrogram. For example, one of the simplest methods is called *Single-Link Cluster Analysis* (SLCA) which joins points to groups or groups to groups on the basis of a shared level of similarity between any member of the two groups. In this scenario, point 10 therefore would link to the pair (7, 9), and point 4 would link to the pair (1, 5). Additional points are then progressively joined leading to the dendrogram depicted in Fig. 8.11.

Examination of Fig. 8.11 provides some indication of the spatial structure of the point distribution in Fig. 8.10. Two major clusters can be distinguished, one consisting of points 8, 10, 7 and 9, and a second consisting of 6, 3, 2, 4, 1 and 5, with 6 sitting as an outlier in that group. Within the first group, point 8 stands out from the main distribution. With such a small distribution the additional insight

Table 8.5 *A distance matrix for hierarchical cluster analysis*

	1	2	3	4	5	6	7	8	9
2	7.8								
3	7.3	4.3							
4	4.1	3.7	4.5						
5	0.7	7.0	6.7	3.4					
6	7.3	12.1	13.7	9.3	7.5				
7	13.6	19.7	20.7	16.6	14.0	7.7			
8	12.7	19.8	20.0	16.3	13.3	8.8	3.8		
9	14.4	20.7	21.6	17.5	14.9	8.7	1.0	3.7	
10	16.6	22.8	23.8	19.7	17.1	10.8	3.1	5.2	2.2

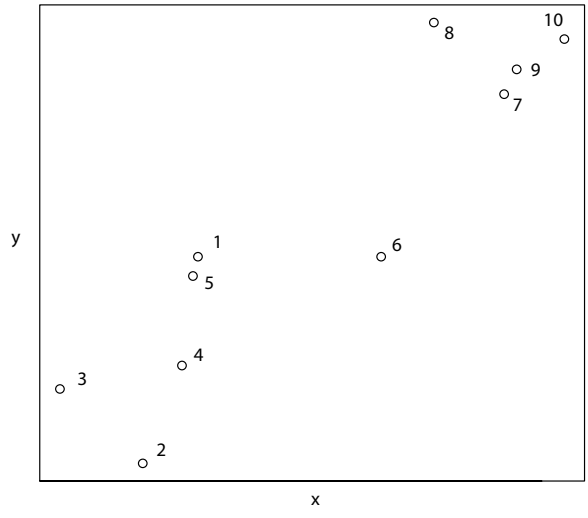

Fig. 8.10 A simple point distribution.

that this brings over visual assessment of the distribution is marginal, although hierarchical analysis applied to larger datasets may help make sense of more complex arrangements. However, a major problem with SLCA is that links between groups are created very easily, so large chains of small clusters often result and outliers are connected to other clusters based on a connection with only one member. Better methods, such as *Average-Link Cluster Analysis*, use average similarity scores of groups to define the level at which additional members cluster. Even more sophisticated approaches to group definition offer further advantages, such as *Ward's Method* which seeks to maximise the homogeneity of clusters by defining clusters so that the *error sum of squares* (ESS) – the sum of the squared distances of all points from the means of the clusters to which they belong – is minimised.

https://doi.org/10.1017/CBO9780511807459.008 Published online by Cambridge University Press

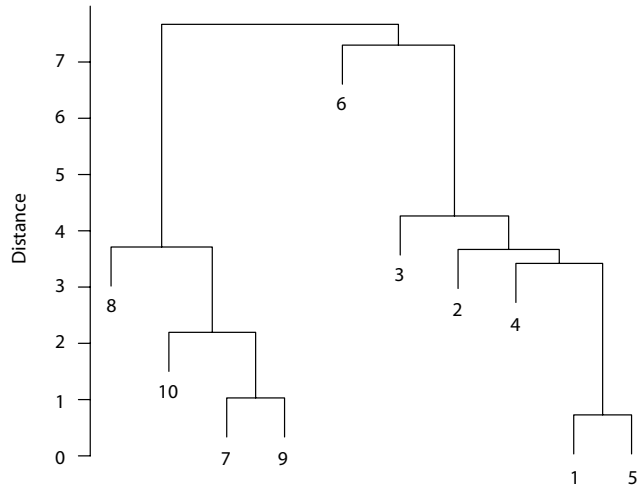

Fig. 8.11 A single-link cluster analysis of the point distribution in Fig. 8.10.

Clusters created using Ward's Method are typically easier to interpret than single- or average-link methods because long chains of small groups occur less frequently, resulting in more homogeneous clusters and a more easily interpreted dendrogram. For example, Ward's (1990) method was used to construct the dendrogram in Fig. 8.12. The results are arguably better than the results from SLCA, in that 3, 2 and 4 form a cluster distinct from 1 and 5, with 6 as an outlier from the latter pair.

Shennan (1988, pp. 212–232) provides a comprehensive review of several hierarchical clustering methods. In general terms, the advantage of hierarchical methods is that it is possible to view clustering at a range of different scales, beginning with small groups of two or more objects and building eventually into larger clusters that include many smaller groups. This advantage, however, can then present difficulties when deciding exactly how many clusters may exist in a dataset, as the number of clusters is defined by an arbitrary level of similarity (or error sum of squares in the case of Ward's method) chosen by the analyst. For this reason it is often wise to avoid dependency on a hierarchical method and instead to use complementary statistics, such as *k*-means analysis, to help ascertain the optimum number of clusters in a dataset. Most popular statistical packages, such as SPSS, S+ and R, offer a range of hierarchical clustering methods.

8.5.2 *k-Means analysis*

When the number of objects is large (e.g. >100), the dendrogram produced from a hierarchical cluster analysis is not easily interpreted and it becomes difficult to ascertain the level of similarity at which cluster groups should be defined. Better are methods that allow the desired number of clusters to be specified beforehand so that a range of solutions can be compared and an optimal solution chosen. *k*-Means analysis is one such method. The major difference between *k*-means and hierarchical

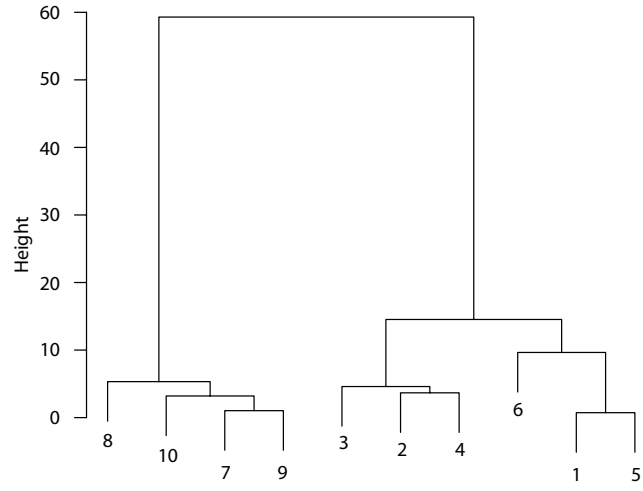

Fig. 8.12 A cluster analysis of the distribution shown in Fig. 8.10 using Ward's method.

clustering is that *k*-means is a *partitioning* clustering technique, because instead of grouping similar objects together, it divides up a group of objects into a specified number of clusters. Cluster centres are initially defined by the selection of random points from the distribution which act as 'seeds', and the remaining objects are then added to the cluster which they are nearest. As new objects are added to a cluster, the centre of the cluster is recalculated, and if a previously assigned object now lies closer to another cluster centre it is reassigned. This *iterative reallocation* is a major strength of the *k*-means technique, although because the seeds that define the clusters are random, different optimum solutions may result and solutions may not be replicable. Once all objects have been allocated, each cluster's sum of squared Euclidean distances (i.e. the squared distance between each object and the centre of its cluster) is calculated to provide an assessment of the clustering solution.

One way to determine the optimum number of clusters is to examine the rate of decrease in the total sum of squared distances over the increasing number of cluster solutions. As the number of cluster solutions increases towards the number of points in the distribution, the sum of squares reduces towards 0. The rate of decline, although generally exponential, reduces at points where increasing the number of clusters does not drastically alter the total sum of squares. The way that this is usually measured is to plot the natural log of the percentage of the total sum of squares for an increasing number of clusters (k). In situations where the distributions are reasonably highly clustered, this can be a useful technique for identifying the optimum number and membership of groups. One good example of this is the *k*-means analysis of distribution of medieval castles on Okinawa Island by Ladefoged and Pearson (2000), which suggested that the optimum clustering solution was three (Fig. 8.13).

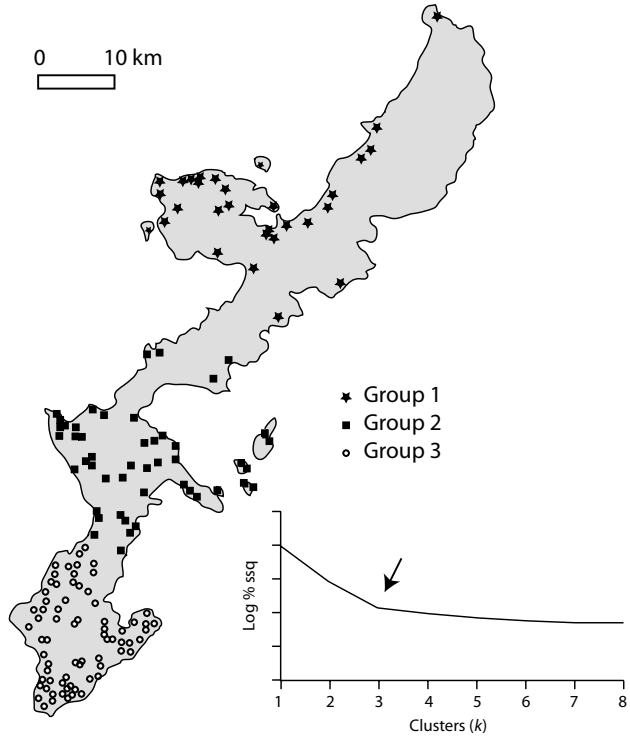

Fig. 8.13 A *k*-means cluster analysis of medieval castles on Okinawa Island, Japan, with the 'elbow' at solution 3 in the associated graph of the decreasing per cent sum of squares (redrawn from Ladefoged and Pearson 2000, Figs. 2–4).

More complex distributions prove difficult to cluster in such a straightforward manner. For example, the distribution of 1795 points in Fig. 8.14 represent the spatial arrangement of stone artefacts from Trench 4b from the Lower Palaeolithic site of Boxgrove, England (Roberts and Parfitt 1999, Fig. 279). The Clark and Evans *R*-statistic is 0.69, with a *z*-value of 24.8, which allows a null hypothesis of randomness to be safely rejected. The question remains as to the number and location of cluster groups in this distribution, which presents significant challenges because of the diffuse nature of the distribution pattern and the lack of clear lines of division between higher and lower density areas.

The only clear 'elbow' is at the two-cluster solution, as shown in Fig. 8.15. The result is a distribution divided in two with the dividing line roughly corresponding to an area of reduced density running through the middle (Fig. 8.16). There are, however, difficulties with this solution that highlight some of the problems with *k*-means analysis. In particular, the dividing line between the left and right clusters appears to be less than optimally positioned. This has arisen because of the relatively simple manner in which *k*-means calculates centroids as the mean of the *x*- and *y*-coordinates.

https://doi.org/10.1017/CBO9780511807459.008 Published online by Cambridge University Press

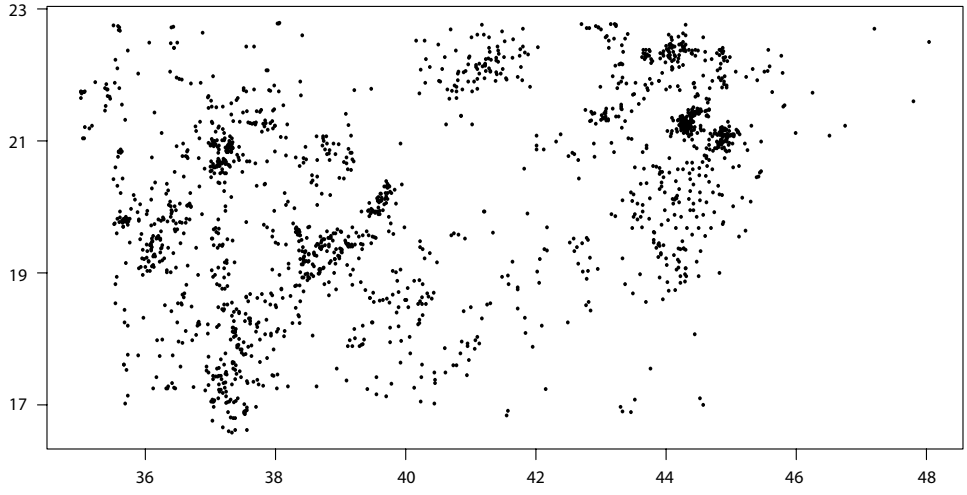

Fig. 8.14 Distribution of stone artefacts from horse-butchery trench 4b at the Lower Palaeolithic site of Boxgrove, England (Roberts and Parfitt 1999, Fig. 279). Source: The Boxgrove Project. Used with permission.

In some solutions this can result in points that lie towards the peripheries of cluster groups being placed into different clusters even though they form a coherent group. These difficulties are to a large extent created because of the nature of this point distribution, which is characterised by a number of small high-density clusters interspersed with a lower density 'carpet' that forms diffuse clusters that bleed into each other. Identifying the optimum number of clusters, and their membership, can be very difficult in these situations, although subdividing the distribution and applying *k*-means to each subunit may help. In general, however, complex spatial distributions such as this example are difficult to cluster using *k*-means, and may not usefully contribute to the interpretation and understanding of the behaviour that created the dataset.

More sophisticated partitioning algorithms like *Partitioning Around Medoids* (PAM; Kaufman and Rousseeuw 1990; van de Laan *et al*. 2002), may provide additional insight. Alternatively, examining cluster patterning through the generation of density measurements can be informative.

8.6 Density analysis

There are many distributions where clustering is evident, but the definition of membership is very difficult to define because of the quantity of points, or because of the 'fuzzy' boundaries of concentrations. In these situations, the problem of how best to define cluster location and size may benefit from approaches that describe the changing density (or, more properly, the *intensity*) of material. These approaches fall under a category of spatial modelling called *intensity analysis* and allow archaeologists to describe and visualise the changing frequency of observations that occur

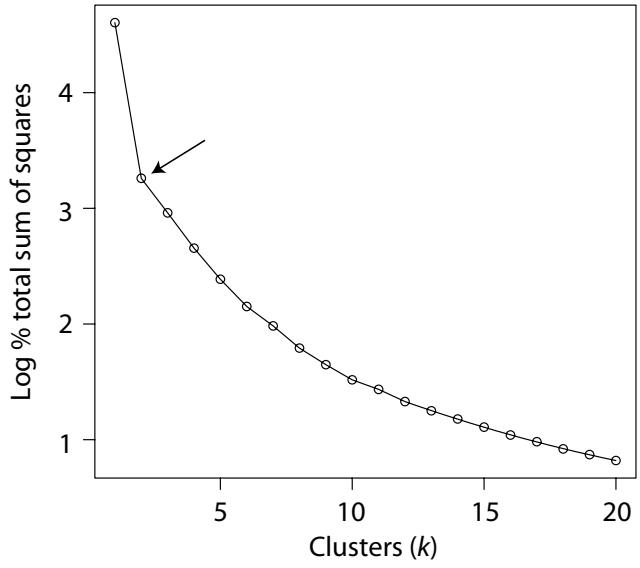

Fig. 8.15 Rate of change of sum of squares (expressed as log percentage of total sum of squares) for 1–20 cluster solutions. The deviation ('elbow') in the rate of change at $k = 2$ (arrow) suggests two clusters is an optimum solution.

within a given area, often to compare different phenomena within the same area or against the same phenomenon in different areas.

If $N(a)$ denotes the number of observations of a phenomenon occurring in A, then the *mean intensity* of the phenomenon, M_1 is given by:

$$M_1 = \frac{N(a)}{A} \tag{8.22}$$

For example, the mean surface intensity for a distribution of 1795 artefacts in a study area of 62 m^2 is calculated as $1795/62 = 28.95$ artefacts per square meter. The same principle could be applied to subregions of the study area, for example, individual 1-m squares, and the first-order intensity function calculated for each square. In this case, the intensity is given by $\lambda(x)$ as we are calculating it for a subset of a.

More usefully, however, the size of the observation area can be reduced to a 'moving window' to derive measures of local density, $\lambda_n(x_i)$, on a regular lattice across the study area. This can be formally expressed by the equation:

$$\lambda_n(x_i) = \frac{N[C(x_i, r)]}{\pi \times r_i^2} \tag{8.23}$$

where x_i is the location at which the intensity is being calculated, and N is the number of artefacts in $C(x_i, r)$, which is a circle of radius r around point x_i. This calculation is referred to as the *naive estimator* (Fotheringham *et al*. 2000a, p. 147).

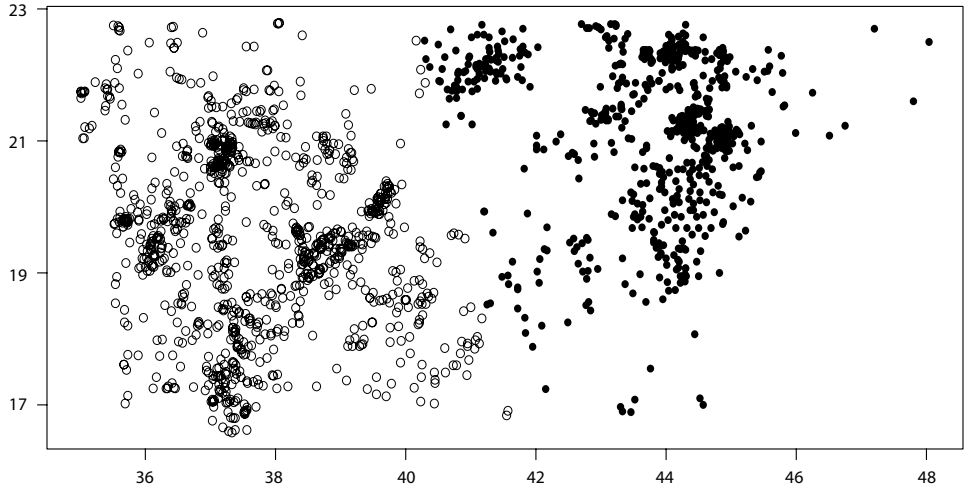

Fig. 8.16 The two-cluster solution of a *k*-means cluster analysis.

Obviously the area of observation (r) and the grid cell size will have a pronounced influence on the resulting density surface. Some insight might be gained by comparing localised surfaces (i.e. smaller search radii) against generalised surfaces (i.e. larger search radii) to investigate small-scale versus large-scale influences on distribution patterns (Fig. 8.17).

Note that as the search area decreases, local densities will increase for smaller clusters – e.g. 5 artefacts falling within a search radius of 0.25 m will result in a local density of $5/(\pi \times 0.25^2) = 25.5$, whereas 5 artefacts falling in a search radius of 1 m would produce a local density of $5/(\pi \times 1^2) = 1.2$.

8.6.1 *Kernel density estimates*

A more sophisticated density measure called *kernel density estimation* (KDE) produces smoother and more readily interpreted results than simple density techniques (Silverman 1986).

Kernel density estimation is a non-parametric technique in which a two-dimensional probability density function (the 'kernel') is placed across the observed data points to create a smooth approximation of its distribution from the centre of the point outwards. The two parameters that can be manipulated are the shape of the kernel placed over each data point (although in many GIS packages this is set to a quadratic function and cannot be changed) and the variance (or radius) of the kernel, referred to as the bandwidth and denoted by h. The density value for each cell is then established by adding together values of the density distributions (each of which will be a fraction of 1, unless the data points represent populations) that overlie that grid cell. Experimentation with different values of h is advised, and more detailed guidelines can be found in Wand and Jones (1995); archaeological

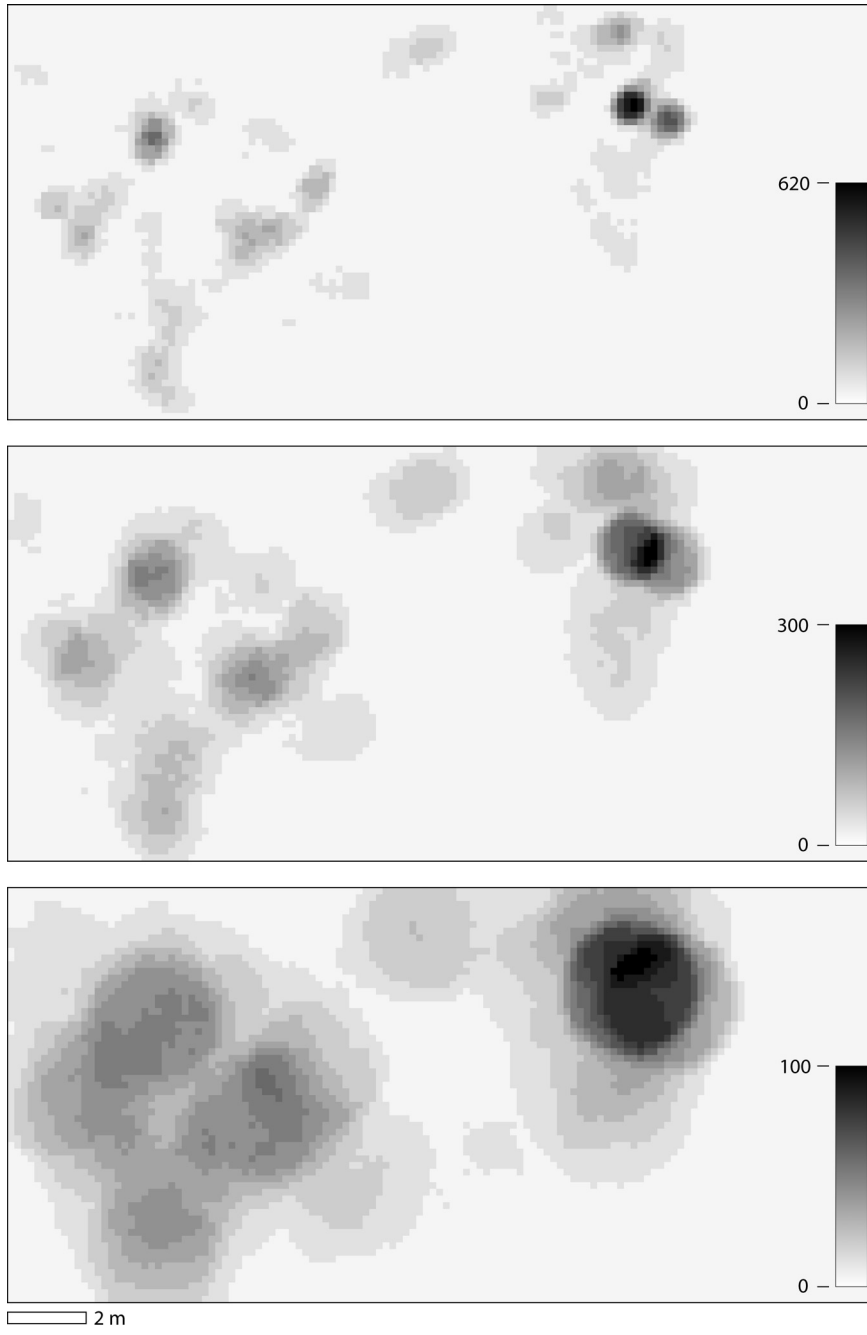

Fig. 8.17 Three intensity surfaces of the artefact distribution in Fig. 8.14: 0.25-m radius (top), 0.5-m radius (middle), 1-m radius (bottom). All calculated on a 10-cm grid. Intensity values are expressed in artefacts per square meter.

https://doi.org/10.1017/CBO9780511807459.008 Published online by Cambridge University Press

applications are described in Beardah and Baxter (1996) and Beardah (1999). In general, using too wide a radius will result in an overly smoothed distribution, whereas too narrow a radius will produce peaks around data clusters that may not reflect the actual distribution. Figure 8.18 shows the result of KDE using the same point data as for the simple density calculations shown in Fig. 8.17. The result is a smoother and more easily interpreted continuous surface for cluster identification.

8.7 Local functions

Finally, simple density measures can be replaced by other neighbourhood functions to produce continuous surfaces that show the changing nature of an attribute. For example, if the count of artefacts is replaced with an attribute variable, y, then a local estimate of y at point x is given by:

$$\hat{y}(x_i) = \frac{\sum y(r_i)}{n(r_i)} \tag{8.24}$$

This returns a continuous surface showing the mean of the variable. This interpolation is local because values are estimated using the values surrounding each cell and it can be contrasted to the global trend surface of artefact sizes in Fig. 6.1. However, in common with the other density functions, this method is highly influenced by edge effects and should be interpreted with caution.

In certain situations it might be important to identify whether a local region deviates from the global trend. This might occur, for example, with survey data if one was interested in exploring whether some enumeration unit (e.g. fields) had neighbours with higher artefact densities than the global pattern and could be defined as a 'hot spot'. An appropriate method to answer this question is Getis's G_j^* statistic (Ord and Getis 1995):

$$G_i^* = \frac{\sum_j w_{ij}(d)x_j - W_i^*\bar{x}}{s[(nS_{1i}^* - W_i^{*2})/(n-1)]^{1/2}} \tag{8.25}$$

where s is the sample standard deviation of the observation values x, and $w_{ij}(d) = 1$ if region j is within a distance d from region i (Rogerson 2001, p. 174). Finally:

$$W_i^* = \sum_j w_{ij}(d) \tag{8.26}$$

and

$$S_{1i}^* = \sum_j w_{ij}^2 \tag{8.27}$$

For example, consider the distribution of artefacts recovered from the five fields in Fig. 8.19. The question is whether the high values around field 5 represent a

https://doi.org/10.1017/CBO9780511807459.008 Published online by Cambridge University Press

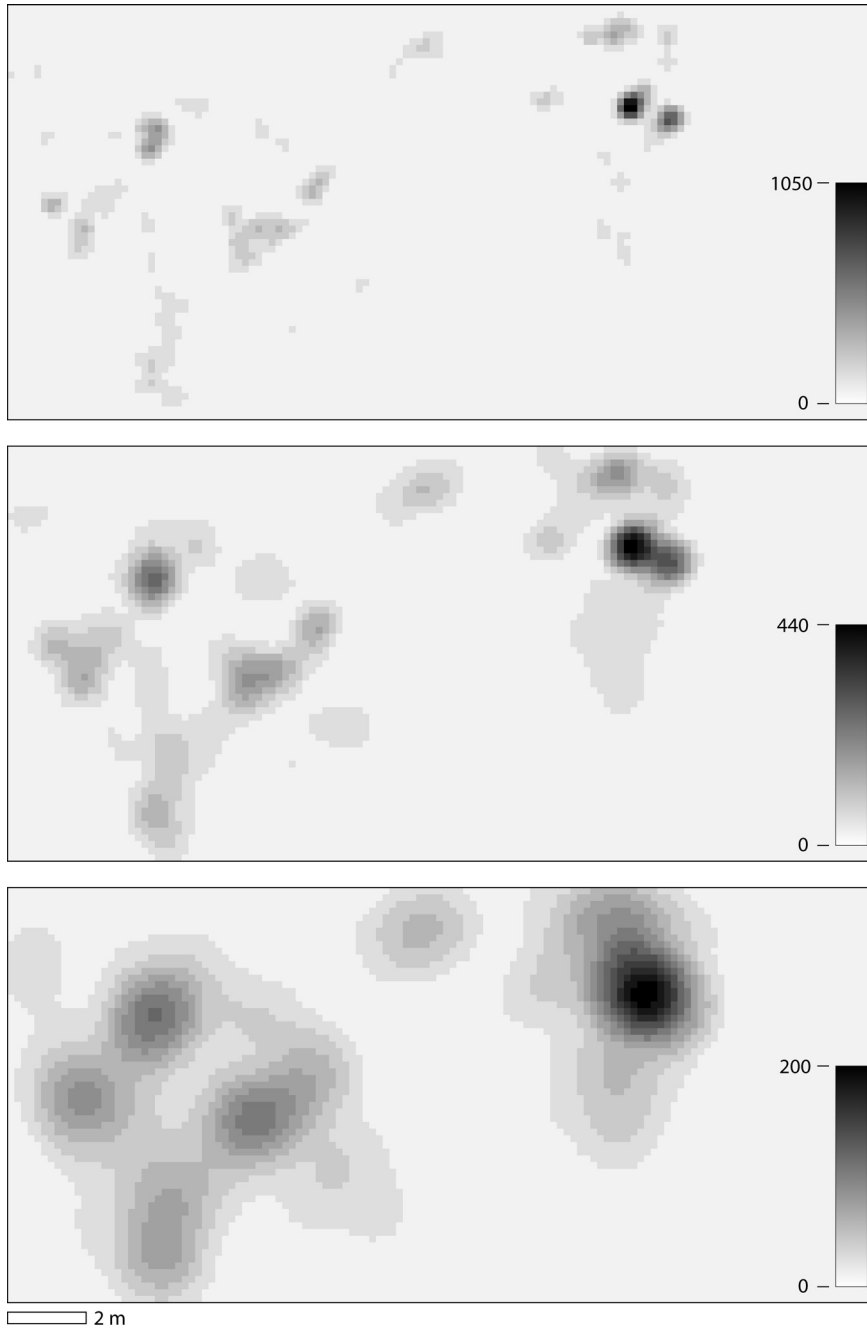

Fig. 8.18 Kernel density estimates of the artefact distribution in Fig. 8.14 using a diameter of 0.25-m radius (top), 0.5-m radius (middle), 1-m radius (bottom). Densities expressed as artefacts per square meter. Compare to Fig. 8.17.

https://doi.org/10.1017/CBO9780511807459.008 Published online by Cambridge University Press

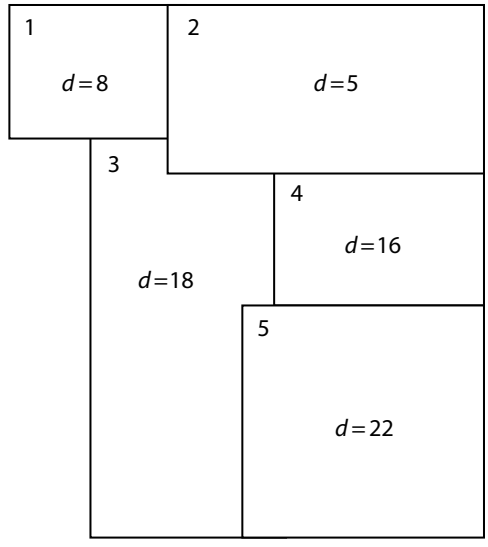

Fig. 8.19 Distribution of artefact densities (d) in five fields. Does field 5 and its two neighbours represent a local 'hot spot'?

statistically significant local cluster distinct from the global pattern. From these observations the values in Table 8.6 can be established to complete Eq. 8.25 so that:

$$G_i^* = \frac{56 - 3 \times 13.8}{7.1\sqrt{(5 \times 3 - 9)/4}} = \frac{14.6}{8.7} = 1.67$$

The statistic G_i^* can be taken as a standard normal random variable with a mean of 0 and a variance of 1 (Rogerson 2001, p. 174). We can therefore use the normal (z) distribution to establish whether the calculated value of G_i^* is sufficiently large that it falls within the critical region for a specified significance level, thus allowing the null hypothesis to be rejected. For a one-sided test, if $p < 0.05$ then $z = 1.645$, which is less than our calculated G_i^*. We can therefore reject the null hypothesis and conclude that field 5 is located in an area of locally high values.

8.8 Predictive modelling

The term 'predictive modelling' refers to the method of predicting the value (or probability of occurrence) of a dependent variable in an unsampled location using one or more independent variables. As applied to archaeology, it is most closely associated with attempts to predict the probability of archaeological settlements occurring in unsampled landscapes on the basis of quantitative assessment of the locational characteristics of settlements in a surveyed area (Kvamme 1983; Judge and Sebastian 1988; Kvamme 1990a; Westcott and Brandon 2000). In this narrow sense, predictive modelling is subject to the charge of environmental determinism

Table 8.6 *A distance matrix for Getis's* G_i^* *statistic*

ij	$w_{ij}(d)$	x_j	$w_{ij}(d)x_j$
5, 5	1	22	22
5, 4	1	16	16
5, 3	1	18	18
5, 2	0	5	0
5, 1	0	8	0
Sum	3	69.0	56
Mean		13.8	
St. dev.		7.1	

for its reliance on a limited range of environmental variables (Gaffney and van Leusen 1995). While it is certainly true that environmental variables influence the choice of settlement location, it is also true that these are not the only factors that people consider when choosing where and how to settle a landscape. Several writers have shown how cultural factors ranging from the influence of 'supernatural' phenomena to the location of pre-existing settlements can play an important role in influencing the human use of space (Ingold 1993; Tilley 1994, 1996; Bradley 1998, 2000; Barrett 1999; Tilley and Bennet 2001). The integration of experiential variables in GIS to improve understanding of the multiple factors that influence human settlement location remains a major challenge. To date most work in this vein has concerned visibility, although as we discuss in Chapter 10, concern with visibility does not automatically overcome the charge of environmental determinism. In the meantime, active research in developing predictive models remains focused on environmental and ecological variables, no doubt in part because of the ease with which these can be measured.

Despite the above critique, we suggest that predictive modelling can be genuinely informative in two situations. The first is cultural resource management (CRM), where it is often necessary to predict the presence of archaeological material in order to prevent or mitigate damage from construction or agricultural practices. From a CRM perspective what normally matters is whether the prediction is correct, not whether it contributes to an explanation of site location. Suffice it to say that numerous examples of predictive modelling have shown that environmental variables such as relief, soil type, drainage and permeability, slope, aspect, distance to water, etc., do – at times only moderately but occasionally significantly – improve archaeologists' ability to predict the occurrence of settlements (Kvamme 1983, 1985, 1990a; Duncan and Beckman 2000; Warren and Asch 2000; Woodman 2000b). The second situation where predictive modelling is useful is in understanding the extent to which site location may have been influenced by a complex interplay of environmental factors (cf. Wheatley 2004). In this case the goal is ultimately

explanation, but providing one remains alert to the fact that correlation does not necessarily imply a causal relationship, predictive modelling provides a valuable tool for identifying multivariate patterning.

The construction of a predictive model is usually undertaken in four stages: data collection, statistical analysis, application of the model and finally model validation (cf. Duncan and Beckman 2000, p. 36; Warren and Asch 2000, p. 13). Note that validation may lead to further refinement. We consider each stage in turn.

8.8.1 *Data collection*

Predictive modelling works on the assumption that it is possible to differentiate between areas of the landscape that have evidence of past occupation (i.e. 'sites') and areas of the landscape that do not ('non-sites') on the basis of one or more landscape attributes. It follows that the construction of a predictive model requires information about the location of sites and about the distribution of the relevant landscape attribute values.

Ideally, site and non-site locations are established by a programme of random or possibly cluster sampling of the landscape, as is common practice in north America (e.g. Kohler and Parker 1986; Kvamme 1992a). This approach has the virtue of allowing estimation of the actual frequency of sites, which in turn allows one to make absolute predictions about the presence or absence of sites. The alternative approach, known as case control, typically makes use of existing data about the presence or absence of sites, and is more common in Europe, where extensive sampling of the landscape is often not practical (Woodman 2000b). Case-control data only support relative rather than absolute predictions of site presence, that is, statements of the form that it is 2.5 times more likely that there is a site at location A than there is a site at location B. Note that it is best to avoid the practice of identifying sites and then randomly picking other locations to serve as non-sites, since if the latter have not been examined they may in fact contain sites, which will then inevitably undermine the ability of the model to identify landscape attribute values that discriminate between site and non-site locations. However the site and non-site locations are identified, it is common practice to split them into a *training sample* used to build the model and a *testing sample* withheld for the purpose of testing its accuracy. This procedure is known as *split sampling* and it is usual to place 50 per cent of locations in the training sample and 50 per cent in the testing sample. When the locations have been obtained by cluster sampling then they should be split by cluster rather than by individual location (Kvamme 1988, p. 395).

A wide variety of landscape attributes have been used for predictive modelling. The primary datasets often consist of a combination of elevation, soil, hydrology, geology and vegetation maps. Once input into the GIS these can then be used to derive further secondary datasets such as relief for a set of different catchment sizes (i.e. the range of elevation values in a circumscribed area), slope and aspect, distances to annual and permanent streams, soil productivity, erodibility, permeability

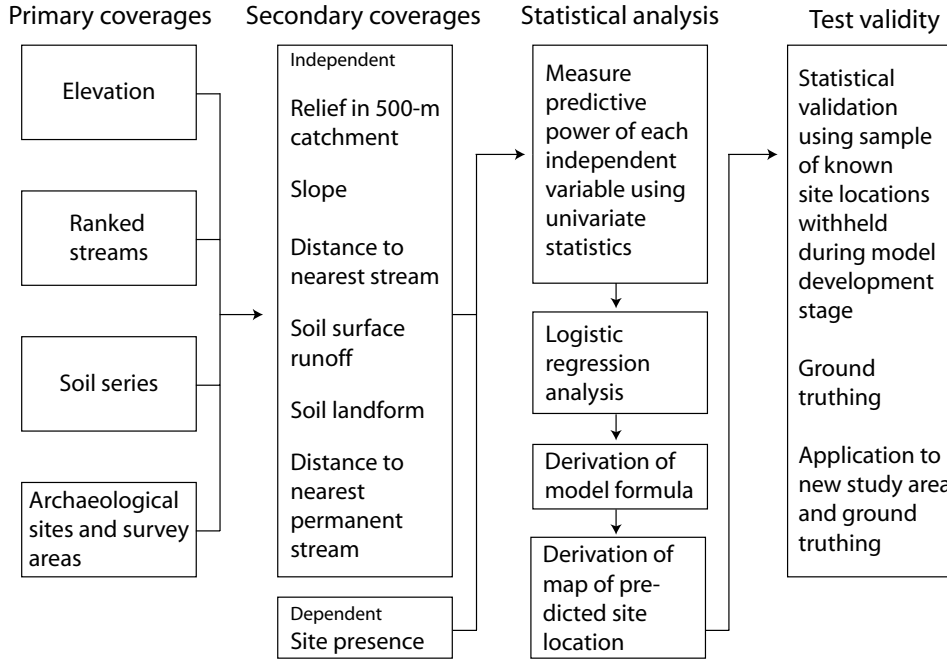

Fig. 8.20 Generalised flowchart of stages in the generation of a predictive model. Adapted from Warren and Asch's flowchart for their predictive model of site location in Montgomery County, Illinois (2000, Fig. 2.4). The six independent variables shown here are only a sample of a much wider range of potentially useful variables. See Warren and Asch (2000) and Woodman (2000b) for examples.

and drainage, vegetation type, exposure, shelter quality, viewshed size, etc. (see Fig. 8.20). Of course, not all the data collected will necessarily turn out to have predictive power.

8.8.2 *Statistical analysis*

Once the data have been assembled, the first task is to identify the landscape attributes that significantly discriminate between the site and non-site locations. This is normally achieved by univariate analysis of each attribute in turn. Depending on the type of data (e.g. nominal in the case of soil types or ratio in the case of distances, elevation and slope) the most appropriate statistical test will probably be one of those described in Chapter 7, such as the chi-squared test, the Kolmogorov–Smirnov test or, if normally distributed, Student's *t*-test. The potentially useful discriminators are those attributes for which it is possible to reject the null hypothesis that their values at site and non-site locations are drawn from the same population. Before these attributes are finally selected for inclusion in the predictive model it is wise to investigate whether any of them either confound, or interact, with others. Confounding occurs when one attribute can substitute for another in predicting the

presence or absence of a site (Woodman 2000b, p. 452), from which it follows that the confounder need not be included in the model. Interaction occurs when one attribute modifies the relationship between another and the presence or absence of a site (Woodman 2000b). For example, in her predictive model of Mesolithic settlement on Islay (see Chapter 3), Woodman found that there is a low chance of finding sites at locations where water sources are small and elevation is high, but a much greater chance of finding them at locations where water sources are small and elevation is low. When attributes interact in this way they should be replaced by hybrid variables that specifically represent the nature of the interaction. Woodman (2000b, pp. 452–453) describes methods that may be used to identify confounding and interaction.

Once the appropriate attributes have been identified, the next step is to build the predictive model itself, for which the preferred technique is *logistic regression analysis* (Stopher and Meyburg 1979; Hosmer and Lemeshow 1989; Menard 2001). Logistic regression differs from linear regression in two ways that make it particularly well suited to predictive modelling. The first is that logistic regression is able to use a combination of variables of different scales (i.e. a mix of nominal, ordinal, interval and/or ratio data). The second is that logistic regression seeks to fit an 'S'-shaped probability curve (hence 'logistic'). The 'S'-shaped curve allows the predicted probabilities of site presence to switch fairly rapidly from low to high, thus avoiding the long sequence of intermediate values (representing uncertainty) that would be produced by a normal linear function. In the case of archaeological predictive modelling, the probability curve is fitted along an axis of discrimination determined by differentially weighting the contribution of the chosen attributes in such a way as to maximise the difference between site and non-site locations (Warren and Asch 2000, pp. 6–9). A number of statistical packages offer the ability to perform this task, including SPSS, R and S+. What all of them output is an intercept a, and a series of regression coefficients $b_1, b_2, \ldots, b_n$ that determine the weighting applied to each of the n attributes $x_1, x_2, \ldots, x_n$. The predictive model is an equation that takes the form

$$V = a + x_1 b_1 + x_2 b_2 + \cdots + x_n b_n \tag{8.28}$$

where V (often referred to as a 'score') is the log odds of site presence.

8.8.3 *Application*

Once logistic regression has been used to build the model it must be applied on a cell-by-cell basis to the study area. The first task is to calculate the score, V, for every map cell by implementing Eq. 8.28 in map algebra (see Chapter 9). In cases where the variables are ratio scale (e.g. elevation) then the coefficients are applied directly to the variable (e.g. if x_1 refers to elevation and the regression coefficient b_1 is 0.345, then the raster map containing elevation data would be multiplied by 0.345). In cases where nominal scale data are used, then the results of the analysis will define a set of numeric *design variables* for each nominal category that can

be inserted into the map algebraic formula (see Warren and Asch 2000, p. 19, for a concrete example). Once the score, V, has been calculated for each map cell, i, it must then be converted to a probability of site presence, p_i. This is achieved by implementing the following equation (Haining 2003, p. 262) in map algebra:

$$p_i = \frac{V_i}{1 + \exp(V_i)} \tag{8.29}$$

Depending how the training sample was collected, the resulting raster map will provide either an absolute or relative probability of a site existing in each map cell.

8.8.4 *Validation*

The fact that it has been possible to construct a predictive model does not in itself guarantee the accuracy of its predictions. This can be assessed using the testing sample that was withheld from the model-building process. The basic idea is to establish how many of the observed sites from the testing sample fall within the area where sites are predicted to be found. For example, if 16 out of 25 observed sites fall in the area where sites are predicted, then the model could be expressed as correctly predicting site location 64 per cent of the time. In reality, however, matters are not quite so simple, for two main reasons:

Prediction is probabilistic Very few, if any, models predict site occurrence with absolute certainty of presence or absence. Consequently it usually only makes sense to talk about the model correctly predicting site presence at some specified probability, p, between 0.0 and 1.0. Models tend to be more accurate at low probabilities and less accurate at high probabilities.

Non-sites matter Often it is possible to specify a probability for the occurrence of sites that is so low that all observed sites do actually fall within the area where sites are predicted, in other words, so that the model is 100 per cent accurate. However, the corollary is usually that a large number of non-sites also fall in the area where sites are predicted, so the model is very inaccurate at predicting the lack of archaeological sites. This would clearly be very undesirable if the purpose of the model was, for example, to identify a route for a new road that minimised the damage to archaeological sites.

Clearly then, it is important to consider the accuracy of a model with reference to the problem at hand. One method that facilitates this is the production of cumulative per-cent-correct prediction curves for both sites and non-sites (Kvamme 1988). Figure 8.21 shows just such a graph, in which the number of sites falling in areas where they are predicted decreases as the probability of site occurrence increases, while the number of correct non-sites increases as the probability of site occurrence increases. In this case, if it was important to avoid damaging sites then one might choose to avoid areas with even a relatively low probability of site occurrence. However, a further complication which arises at this point is that the relevant area is likely to be so large (since these are cumulative probabilities the area in question includes all locations with a low probability or greater) as to render the prediction virtually worthless. There are at least two solutions to this dilemma. One is to pay

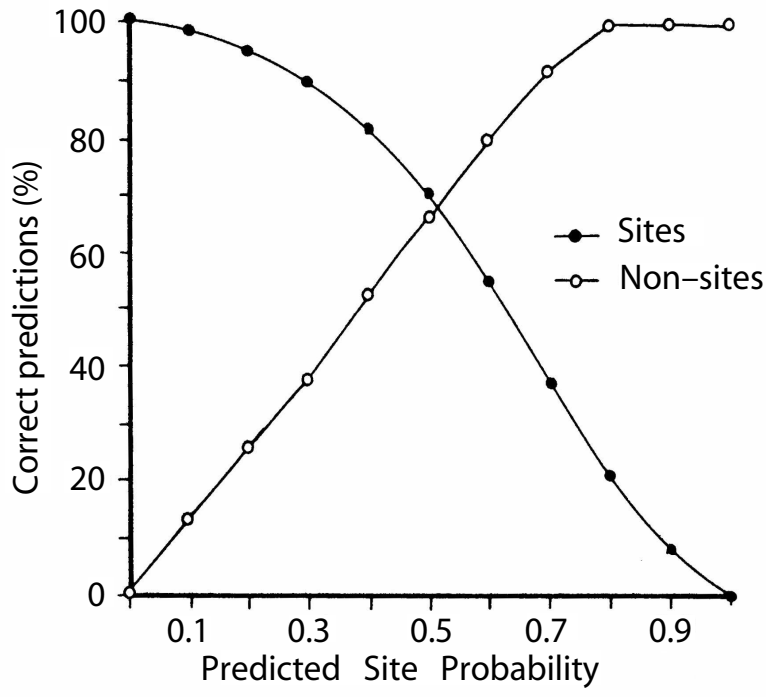

Fig. 8.21 Cumulative per cent correct predictions for model sites and non-sites for all probabilities of occurrence. Reproduced with permission from Kvamme 1988, Fig. 8.11B

attention to the trade-off between correctly predicting site and non-site locations, while another is to examine the predictive *gain* offered by the model. Kvamme (1988, p. 329) defines the gain, G, as:

$$G = 1 - \frac{\%\text{ of total area where sites are predicted}}{\%\text{ of observed sites within area where they are predicted}} \tag{8.30}$$

G, which is calculated for a specified probability of site occurrence, ranges from 1 (high predictive utility) through 0 (no predictive utility) to -1 (the model predicts the reverse of what it is supposed to). The most important property of this measure is that it can distinguish a correct but relatively worthless model from an ostensibly less correct but more useful one. For example, a model that correctly predicts 80 per cent of sites and predicts site occurrence over 70 per cent of the landscape is probably not very useful, which is reflected in the low gain of 0.13. On the other hand, a model that correctly predicts 70 per cent of sites and predicts site occurrence over a mere 5 per cent of the landscape would provide a better basis

https://doi.org/10.1017/CBO9780511807459.008 Published online by Cambridge University Press

for many decisions, which is reflected in the gain of 0.93. Further suggestions for testing predictive models can be found in Kvamme (1988).

8.9 Conclusion

This chapter has introduced some techniques useful for investigating the patterns and relationships in spatial datasets. It is impossible to do full justice to the very large literature on techniques of spatial analysis, so we have chosen to highlight approaches that are within the grasp of the majority of archaeologists, whether numerically inclined or not. Although this chapter has reviewed a number of 'traditional' techniques – nearest neighbour, hierarchical and k-means cluster analysis, for example – we have also highlighted recent approaches to the explication of spatial processes, such as Ripley's K. We encourage further investigation of modern spatial analytical techniques as described by Haining (2003) and Fotheringham *et al*. (2000a).