10

Correlation and regression of spatial data

The issues that can arise when applying correlation and regression analysis techniques to data relating to observations that are located at specific places in space or occur on fixed occasions in time are examined in this chapter. Indices for quantifying global spatial autocorrelation and local spatial association are explored together with an introduction to the relatively advanced techniques of trend surface analysis and geographically weighted regression. Students often develop a level of confidence with the correlation and regression techniques covered in previous chapters, but the issues associated with applying these to spatially autocorrelated data are sometimes neglected. This chapter shows how relatively simple measures can be calculated and in some cases tested statistically to avoid the pitfalls of unwittingly ignoring the lack of independence in spatial data by students and researchers in Geography, Earth and Environmental Science and related disciplines.

Learning outcomes

This chapter will enable readers to:

- describe the characteristics and implications of spatial autocorrelation;
- calculate and apply suitable indices to measure the global and local effects of spatial autocorrelation;
- consider how to incorporate these measures when analysing geographical datasets in an independent research investigation in Geography, Earth Science and related disciplines.

Practical Statistics for Geographers and Earth Scientists Nigel Walford

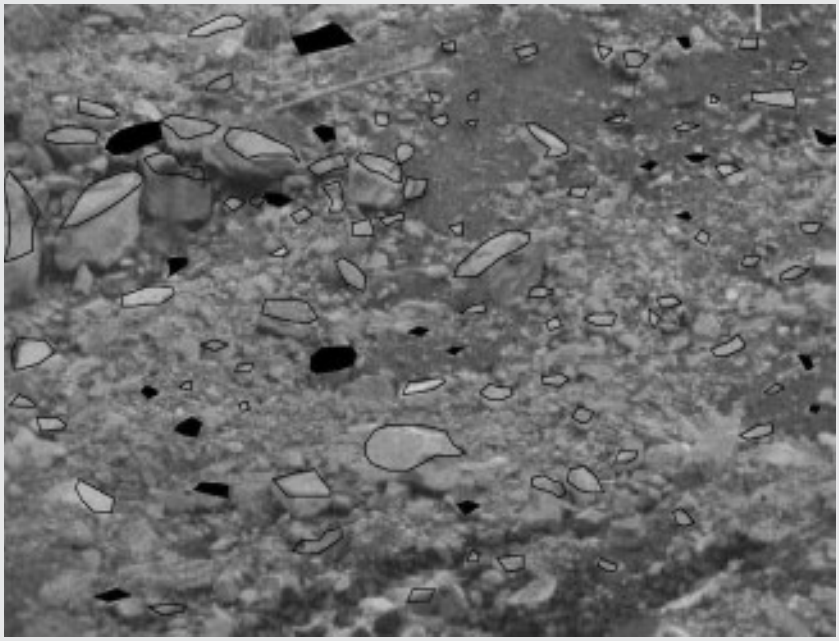
10.1 Issues with correlation and regression of spatial data

Correlation and regression analysis are often used to investigate research questions in the geographical sciences, although there are some important issues that need to be considered when the variables and attributes relate to spatial entities. Some applications of correlation and regression may be carried out in a particular geographical context, such as with respect to businesses operating in a certain city region in Human Geography or to the concentration of pollutants in a particular river system in Environmental Science. Provided that the spatial distribution of the population and sample of observations are only of incidental interest and they are independent of each other, both types of statistical analysis can be carried out with relative ease. However, once the spatial location of the entities starts to be regarded as relevant to the investigation, for example the distribution of businesses in relation to each other or to some other place, such as the centre of the city or the sites along the river channels where water samples are selected in relation to land use, then some issues overshadow the application of correlation and regression as described in the previous chapters.

The origin of these problems arises from the fact the individual entities that make up a given collection of spatial units (points, lines and areas) are rarely, if ever, entirely independent of each other. Yet a fundamental assumption of correlation and regression is that the values possessed by each observation in respect of the variables and attributes being analysed should be independent. If any dependence between the entities is ignored then its effect on the results of the correlation and regression will be undetected. For example, it might have artificially increased or decreased the value of the correlation coefficient, thus indicating a stronger or weaker relationship than is really present. Similarly, it might have affected the form of the regression equation and could lead to unreliable predicted values for the dependent variable. The possible problems that might arise from a lack of independence between spatial features is illustrated in Box 10.1 with respect to a section of the moraine where the material has emerged from Les Bossons Glacier near Chamonix in France and been transported and been deposited on the sandur plain. There is a mixture of sizes of material in the area of moraine shown and it clear from a superficial examination that the different-sized material is not randomly distributed. There are clumps of individual boulders, stones and pebbles together with finer material not visible in the image. The upward facing surface of a random sample of these boulders, stones and pebbles has been digitized and shown on a 'map' superimposed on the image. The sampled items have been measured in respect of their surface area and the length of their long axis.

Regression and correlation analyses have been carried out on a subsample of these items (in the interests of limiting the calculations shown) with surface area as the independent and axis length as the dependent variables. The results (Pearson's correlation coefficient and linear regression equation) are shown on the scatter plot in Box 10.1. These suggest a very strong positive relationship between the variables

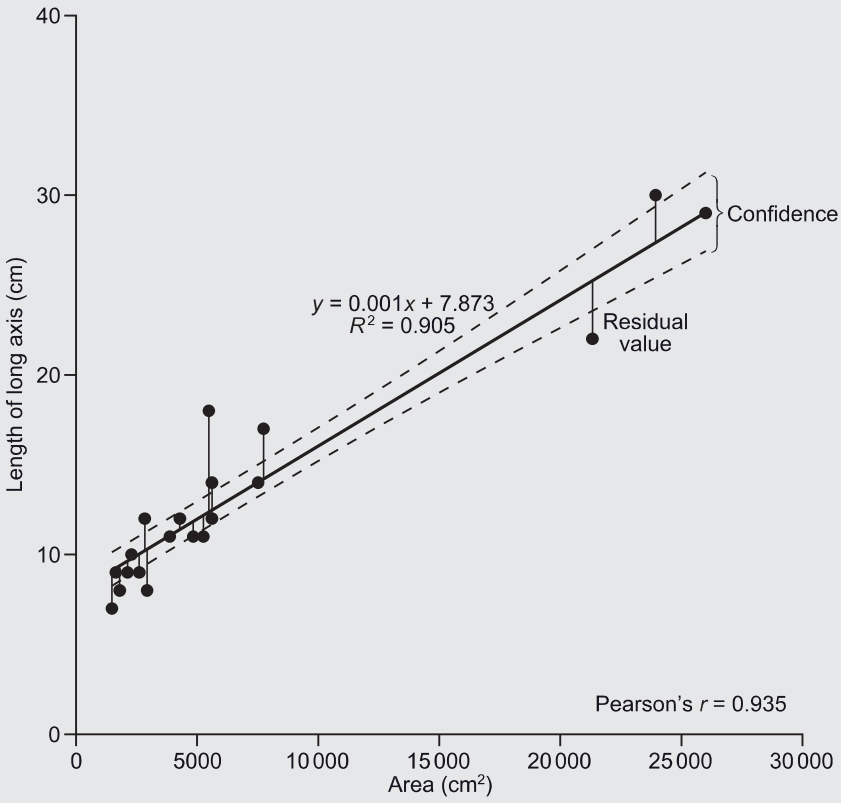
Box 10.1: Spatial autocorrelation.



Subsample of debris shown with dark shading.

A lack of independence in the data values for a collection of n observations is likely to mean that there is some systematic pattern in the size of the residuals along the regression line. It could be that the lower and upper ends of the range of values for the variable x produce larger residuals and so a poorer prediction of the dependent variable in regression, or perhaps there is a repeating pattern of large and small residuals along the range x values: either way, these patterns indicate the presence of autocorrelation in the data.

The image of part of the moraine of Les Bossons Glacier suggests that the size and long axis length of debris material is not distributed randomly. For example, there seems to be a group of large boulders towards the upper left and a relatively larger number of small items in the upper and central right areas. The 20 debris items shaded black have been randomly selected as a subsample of the full 100 boulders, stones and pebbles in the full sample. Their surface area has been measured and simple linear regression analysis has been applied to examine the supposed relationship that hypothesizes area as an explanatory variable in respect of axis length. The calculations for the regression analysis have not been included since the standard procedures discussed in Chapter 9 have been followed. The regression equation is $\hat{y} = 7.873 + 0.0001x$ and with $r^2 = 0.905$ there is a strong indication that surface area has significant explanatory power in respect of the long axis length. The residuals from this regression analysis seem to display some systematic pattern along the regression line with smaller residuals at the lower end of the range of x values. The residuals seem to become progressively larger towards the upper end.



Linear regression line, 95% confidence intervals and residuals.

(+0.951) and with r^2 equal to 0.905, there is some indication that surface area explains 90.5 per cent of the variability in axis length. The scatter plot also shows the residuals of the sampled data points as vertical lines connected to the regression line, which represents the predicted value of the dependent variable for the known values of surface area (dependent). These reveal an interesting feature: generally speaking the residuals are smaller for sampled stones that were towards the lower end of the surface area scale. The confidence limits support this notion since they are curving away from the regression lines towards the upper end of the independent variable axis. In other words, there appears to be a relationship between successive values of the residuals along the regression line and they vary in a systematic way. This might not be a problem if the different sizes of moraine material were randomly distributed across the area, but the image clearly shows this is not the case. Separate subsamples of material from the different parts of the moraine could potentially produce contrasting and even contradictory results from their respective correlation and regression analyses.

10.2 Spatial and temporal autocorrelation

The example in Box 10.1 illustrates a problem known as **spatial autocorrelation**. Correlation analysis as outlined previously concentrates on the strength and direction of the relationship between two variables for either a population or a sample of observations, but it does not take into account the relationship between the individual entities. So far, we have ignored the possibility that one observation possessing a certain value for the X variable might have some bearing on the value of X (or Y) of other observations. Rather than the observations being independent they might be **interdependent**. Autocorrelation occurs when some or all of the observations are related to each other. Spatial autocorrelation arises when it is locational proximity that results in observations being related and temporal autocorrelation when closeness together in time is the cause. Spatial and temporal autocorrelation are most commonly positive in nature in the sense that the observations possess similar values for attributes and variables, where as negative autocorrelation, when spatially or temporally close observations have dissimilar values, is rarer but by no means unknown. Many geographical phenomena display positive spatial autocorrelation, for example people living in housing on the same street and soil samples taken from the same field, are likely to be more similar to each other than they are to the same types of observation from locations that are further apart.

The underlying reason why this might be a problem is illustrated in Figure 10.1, which shows scatter plots for the complete sample of boulders, stones and pebbles on the Bossons Glacier moraine. Rather than plotting the dependent variable (length of long axis) against the independent one (area), the upper and lower pairs of plots, respectively show these plotted against the X and Y coordinates of the locations of the sampled debris. There are a number of important features to note from these scatter plots. First, the r^2 values are relatively low, which indicates that the X and Y coordinates do not provide a strong explanation for variability in area or length. Secondly, the relationships are all negative, although the slope of the regression line is much higher in the case of the X coordinates. However, perhaps the most striking feature is that there are some clumps of data points where there are groups of observations that have very similar coordinates and area or length values. One clear example of this is to be found just above the centre of the horizontal axis of the upper-left plot where there is a group of 11 observations with low area values and X coordinates around 200. Spatial autocorrelation extends the general concept of autocorrelation in two ways: first that adjacent values are strongly related and second that randomly arranged values indicate the absence of autocorrelation.

Where else are there clumps of data points in Figure 10.1? What are the combinations of variable and coordinate values at these locations?

Understanding of spatial autocorrelation owes much to earlier work concerned with **time-series analysis** and the fact that geographical investigations are often focused not

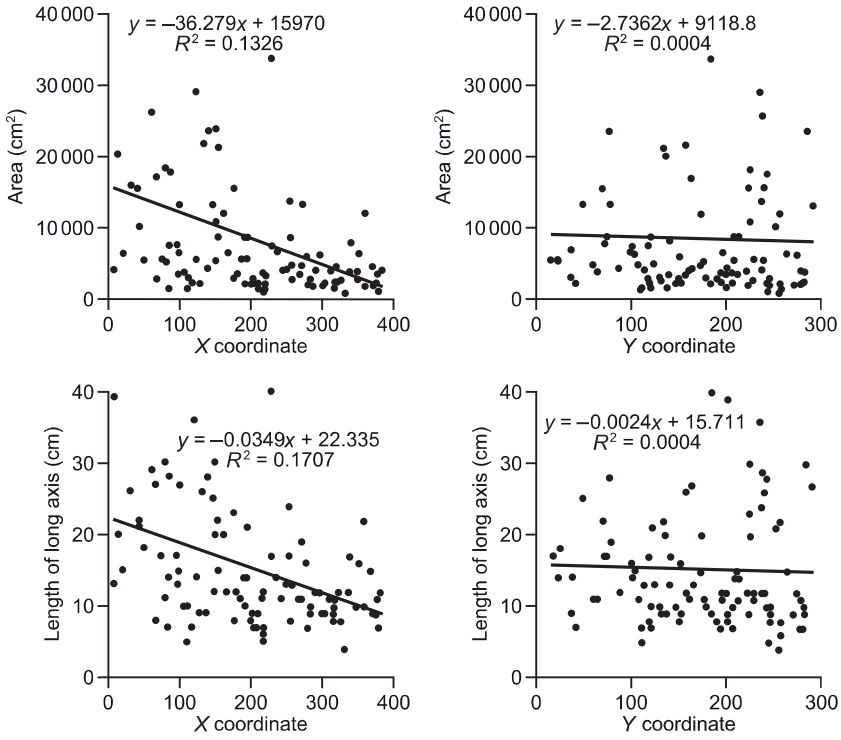


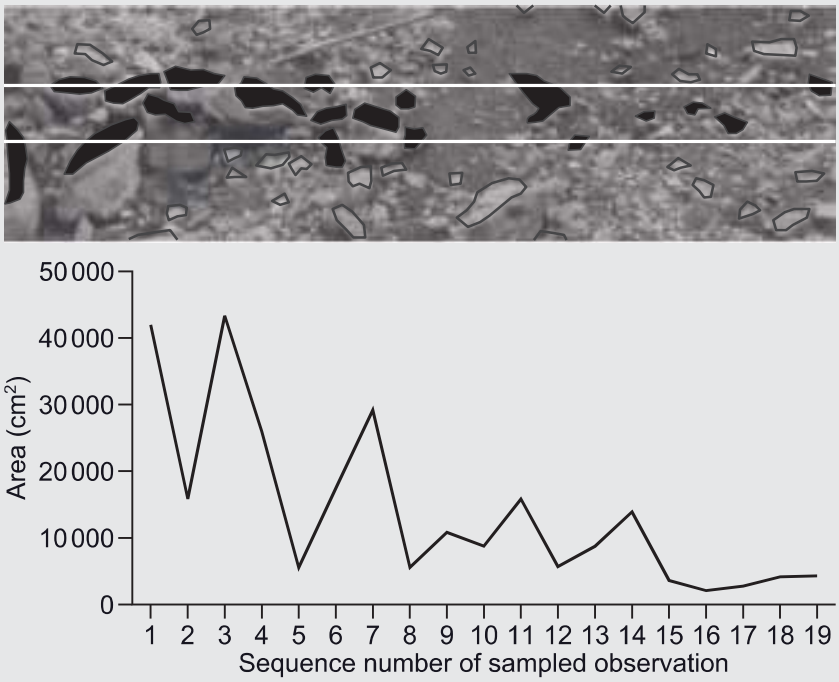
Figure 10.1 Scatter plots of full sample of moraine material by area and length of long axis against *X* and *Y* spatial coordinates.

only on spatial occurrences of phenomena but also the measurement of variables as they change over time. For example, human geographers might be interested in how deprivation is distributed spatially **and** temporally between different census areas. The concept of covariance, the way in which two independent pairs of data values for variables *X* and *Y* vary jointly, was introduced in Chapter 8 as the starting point for understanding correlation. Dividing the covariance by the product of the squares roots of the variances of *X* and *Y* produces the Pearson's correlation coefficient (*r*). This effectively standardizes the value of the coefficient to lie within the range –1.0 to +1.0. In Box 10.1 we focused on the relationship between the area and long axis length of boulders, stones and pebbles on part of the Les Bossons Glacier moraine, but suppose we were interested in a set of *n* values for one of the variables, say surface area, measured in respect of the spatially contiguous debris over the surface of the moraine. Box 10.2 illustrates the effects of spatial autocorrelation by examining the **spatial contiguity** with respect to the subset of all 19 items (boulders, pebbles and stones) lying partly or wholly within a transect across the surface. The series of four scatter plots in Box 10.2b are known as h-scatter plots, where *h* refers to the spatial lag between data values. When such lags are used in time-series analysis the length of time periods or intervals is often constant throughout the sequence, for example daily amounts of precipitation,

Box 10.2a: Spatial lags

Serial correlation coefficient for lag 1: $$r_{.1} = \frac{\sum_{h=1}^{n-1}(x_h - \bar{x}_{.1})(x_{h+1} - \bar{x}_{.2})}{\sqrt{\sum_{h=1}^{n-1}(x_h - \bar{x}_{.1})^2}\sqrt{\sum_{h=1}^{n-1}(x_{h+1} - \bar{x}_{.2})^2}}$$

Serial correlation coefficient for k lags: $$r_k = \frac{\sum_{h=1}^{n-k}(x_h - \bar{x})(x_{h+k} - \bar{x})}{\sum_{h=1}^{n}(x_h - \bar{x})^2}$$



Transect through sample of debris on Les Bossons Glacier moraine.

The data values for variables measured in respect of observations that are located in space may be related to each other and display positive or negative autocorrelation. A transect has been superimposed on the top of the image representing the part of Les Bossons Glacier's moraine and the boulders, pebbles and stones intersecting with this area have been identified and numbered 1 to 19 in sequence from left to right. This example deals with objects located irregularly in space, but the procedure could as easily be applied to regularly spaced features, for example items that are a fixed distance apart.

Autocorrelation can be examined by means of the serial correlation coefficient where there are k lags and each lag is identified as h units (e.g. h = 1,2,3 up to k) or t time periods in the case of time-series analysis. The 19 values for the variable measuring the surface area of these

objects, denoted as x, have been tabulated in Box 10.2b and labelled as x_1 to x_{19}. In the second column of data values they have been shifted up by one row, thus pairing the data value for one object with the next in the sequence. Lag 2 works in a similar way, but pairs one data value with the next but one in the sequence, and so on for however many lags are required. Once the data values have been paired in this way the Pearson's Correlation coefficients are calculated and these have been shown in h-scatter plots for spatial lags 1 to 4.

The r coefficients show an increase through lags 1, 2 and 3 (0.3341, 0.3471 and 0.4246) and them decline to 0.1039 for lag 4. This indicates that spatial autocorrelation in respect of area for these observations starts to reduce after spatial lag 3.

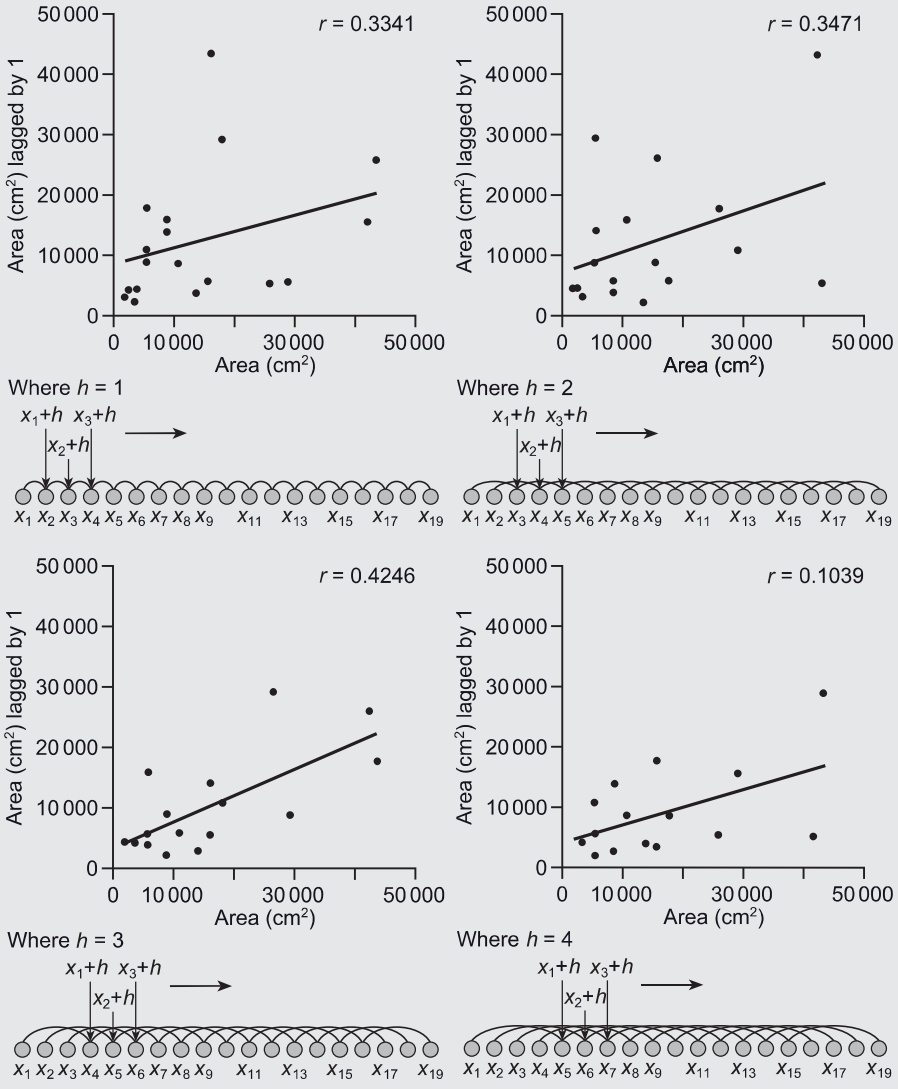
Box 10.2b: Linking data values by spatial lags.

		$x_{1\ldots n} + h$	$x_{1\ldots n} + h$	$x_{1\ldots n} + h$	$x_{1\ldots n} + h$
		Lag 1 ($h = 1$)	Lag 2 ($h = 1$)	Lag 3 ($h = 1$)	Lag 4 ($h = 1$)
x_1	41 928.88	15 889.19	43 323.43	26 015.02	5536.74
x_2	15 889.19	43 323.43	26 015.02	5536.74	17 865.73
x_3	43 323.43	26 015.02	5536.74	17 865.73	29 164.62
x_4	26 015.02	5536.74	17 865.73	29 164.62	5620.59
x_5	5536.74	17 865.73	29 164.62	5620.59	10 889.42
x_6	17 865.73	29 164.62	5620.59	10 889.42	8805.97
x_7	29 164.62	5620.59	10 889.42	8805.97	15 814.79
x_8	5620.59	10 889.42	8805.97	15 814.79	5703.54
x_9	10 889.42	8805.97	15 814.79	5703.54	8 95.05
x_{10}	8805.97	15 814.79	5703.54	8795.05	13 891.54
x_{11}	15 814.79	5703.54	8795.05	13 891.54	3669.44
x_{12}	5703.54	8 95.05	13 891.54	3669.44	2116.38
x_{13}	8795.05	13 891.54	3669.44	2116.38	2824.25
x_{14}	13 891.54	3669.44	2116.38	2824.25	4148.04
x_{15}	3669.44	2116.38	2824.25	4148.04	4276.09
x_{16}	2116.38	2824.25	4148.04	4276.09	
x_{17}	2824.25	4148.04	4276.09		
x_{18}	4148.04	4276.09			
x_{19}	4276.09				

whereas spatial lags can be regular or irregular. In Box 10.2 the separate items of moraine debris in the transect are not located at a regular distance apart, but are lagged according to their sequential spatial contiguity or neighbourliness.

The series of correlation coefficients for the lagged variable should be approximately zero if they were calculated for a random set of data values and plotting a **correlogram** is a useful way of examining whether this is the case when the spacing of the observations is equal. Although the observations in our example are not spaced

Box 10.2c: Spatially lagged scatter plots of area for moraine objects lying within transect.



at a regular distance apart across the transect, they are in a unitary sequence relating to the first, second, third and so on up to $n - 1$ nearest neighbours. It is therefore not entirely inappropriate to plot the series of correlation coefficients for the lags as a correlogram. Figure 10.2 shows the correlograms for the area and axis length variables

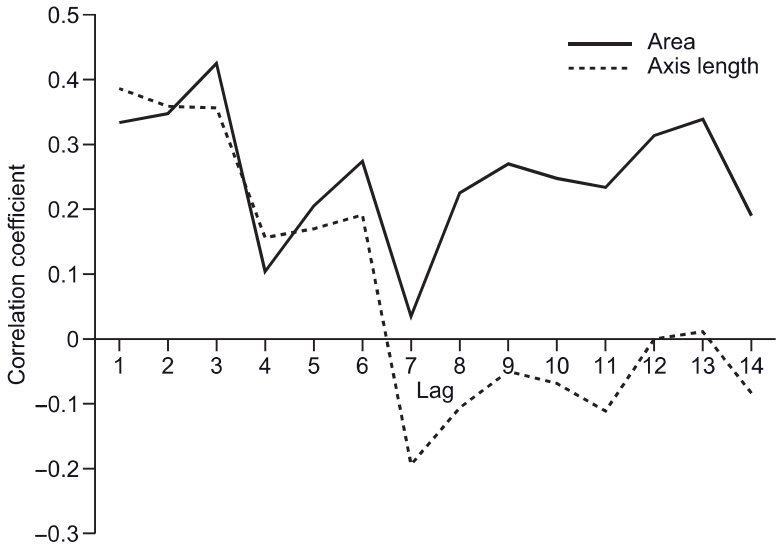


Figure 10.2 Correlograms for area and length of the long axis of moraine material intersecting with transect.

with moderately strong positive correlation coefficients for both variables for spatial lags 1 to 3 followed mainly by a decline until lag 7. Thereafter, the area line continues to record low positive correlation coefficient values, whereas for axis length they are very low negative ones.

This section has introduced some of the ways of examining spatial autocorrelation as though they could simply be migrated across from time-series analysis in an unproblematic fashion. The addition of a transect to the image of the moraine is to some extent an artefact simply being used to illustrate the principles of spatial autocorrelation. There may be some underlying trend in the data values not only in respect of the portion of the moraine shown in the image but also across the area as a whole, which may be connected with distance from the glacier snout, slope angle and other factors. We will return to this issue later when examining the application of trend surface analysis. A further important issue is that a given sequence of measurements may include some rogue values or **outliers**, which distort the overall pattern. The following sections will examine a range of procedures available for examining patterns in spatial data starting with those dealing with global spatial autocorrelation and then moving onto those capable of indicating local spatial association.

10.2.1 Global spatial autocorrelation

Indices of global spatial autocorrelation summarize the extent of this characteristic across the whole of the area under study. There are a number of measures available

that are suited for use with different types of data. The starting point for many of the techniques is that the area or region of interest can be covered by a regular grid of squares or by a set of irregular-shaped polygons. A further factor influencing the choice of technique concerns whether the values are numerical measurements or counts of nominal attributes. The data type presented by the Bossons Glacier moraine example where there are data values for 100 randomly distributed points is also covered. The essential purpose of all the techniques outlined in the following sections is to explore the correlation between the units (areas or points) at different degrees of spatial separation and to produce a measure that is comparable to the serial correlation coefficient used in time-series analysis.

10.2.1.1 Join counts statistics

Join count statistics (JCS) focus on the patterns produced by sets of spatial units that have nominal data values by counting the number of joins or shared boundaries between areal units in different nominal categories. Most applications relate to binary data values, for example the absence or presence of a particular characteristic, although data with more than two classes can be regrouped into a binary form. Perhaps the simplest place to start with exploring JCS is the case where a regular grid of squares has been superimposed over the study area and these squares have been coded with a value of 0 and 1 to denote the binary categories. Chapter 6 discussed three 'standard' ways in which spatial features could be arranged, clustered, equidistant and random. Figure 10.3 illustrates the three situations with respect to a regular grid of 100 squares that belong to binary classes, here shown as either black or white. In Figure 10.3a the 100 squares are split equally in half with all the white ones at the top and the black ones in the bottom. The middle grid has a systematic pattern of white and black squares, rather similar to a chess board, with none of the same coloured squares sharing a boundary and only meeting at the corners. Figure 10.3c, again with half of the squares shaded black and the other half white, shows a random distribution with some same coloured squares sharing edges and others meeting at corners.

These comments have already given a clue as to how we might analyse the different patterns and to decide whether a given pattern is likely to have occurred by chance or randomly. First, consider the situation in time-series analysis, where time periods are usually assumed to form a linear sequence so that one period of 24 hours (a day) is followed by another and so on and each period has one join with its predecessor and one with its successor, apart from those at the arbitrary start and end of the series. If these time periods were classified in a binary fashion (e.g. absence or presence of President Obama's name on the front page of the New York Times over a period of 10 days) they could be represented as a series of black and white squares in one dimension, such as those down the right-hand side of Figure 10.3. The three linear sequences correspond to their grid square counterparts on the left-hand side. The joins between the spatial units (squares in this case) work in two dimensions rather than the one dimension of the time series. Joins between squares in the grid occur in two ways edge to edge and corner to corner, and in an analogy with chess the former are referred to

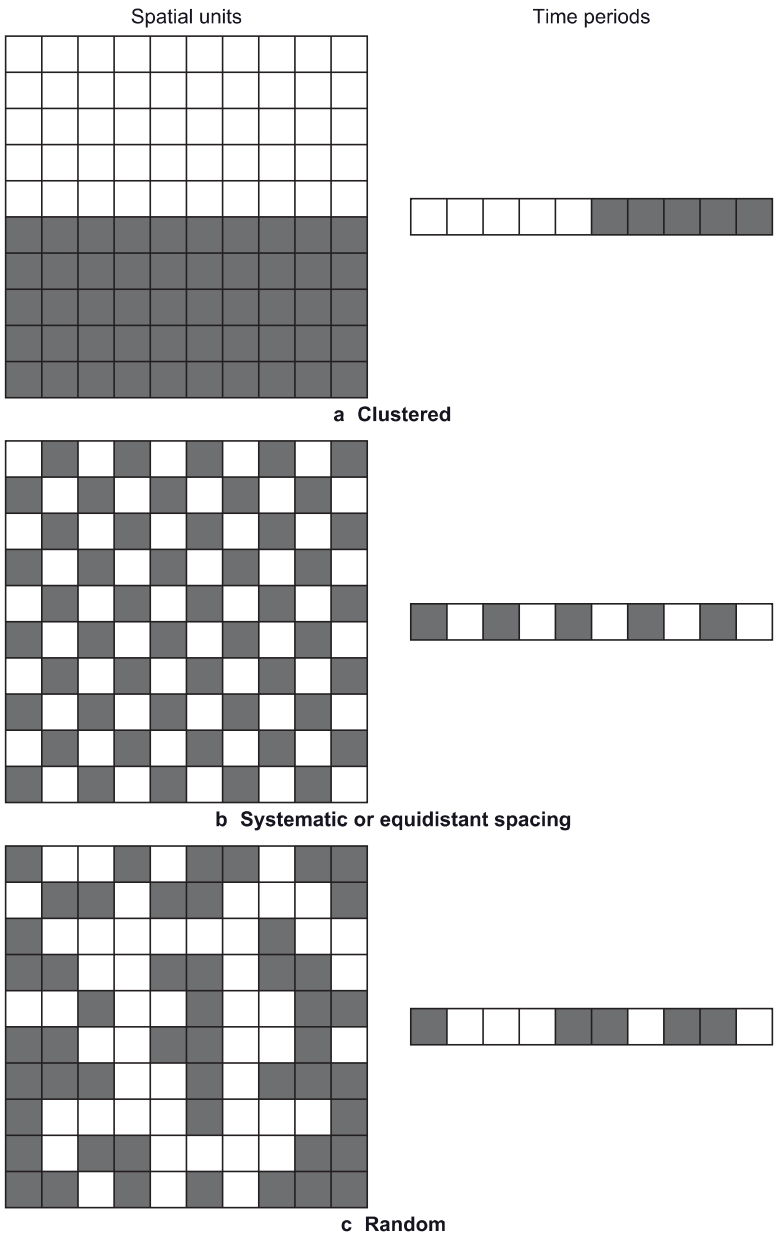


Figure 10.3 Binary join count patterns for regular grid squares and linear time series.

as rook's and the latter as queen's moves. This produces up to eight possible joins for each square with those around the edge and at the corners of the overall grid covering the study area having less. The edge effects can be disproportionately important if the size of the study area or the number of grid squares is relatively small, although this obviously raises the question of how small is small. Given that each square has the binary codes 0 or 1, where they join there are four possible combinations: 1–1, 0–0, 1–0 and 0–1. Counting the number of joins of these different types indicates the origin of join count statistics.

How many rook's and queen's joins does each corner square have in the grids down the left-hand side of Figure 10.3? How many of both types of join does each of the four squares at the centre of these grids have?

Each corner square has two adjacent squares, all of the other squares along the boundaries or sides of the grid have three rook adjacencies and all of the remaining squares have four. A 10×10 grid of 100 squares will therefore have 4 corner squares (8 joins), 32 side squares (96 joins) and 64 inner squares (256 joins) summing to 360, but because this has double counted joins between adjacent squares the sum is halved to give a total of 180. The 100 squares in the grids in Figure 10.3 are equally divided between black and white, therefore each join combination (1–1, 0–0, 1–0 and 0–1) has an equal probability of occurrence and we would expect there to be 45 joins of each type (180/4). However, we are interested in the deviation from the two extreme situations of perfect separation (Figure 10.3a) and regularity (Figure 10.3b), which respectively have 10 and 180 0–1 and 1–0 joins. The random pattern shown in Figure 10.3c has 92 0–1 and 1–0 joins, which is slightly more than the expected total of 90, but is the difference more or less than might have occurred through chance or sampling error. These comments indicate that the empirical count of each adjacency combination, with 0–1 and 1–0 being taken together since they are equally indicative of a mixed pattern, should be tested for their significance. This can be achieved by converting the difference between the observed and expected frequency into a Z score having calculated the standard deviation of the expected number of counts corresponding to each combination.

One complicating factor should be noted before examining the application of JCS, which relates to whether the data has been obtained by means of free sampling with replacement or nonfree sampling without replacement. Mention of sampling might seem a little odd, since the regular grids shown in Figure 10.3 have squares that cover all of the study area. So in what way have these data been sampled? In Chapter 2 we saw that the main difference between sampling with and without replacement when using nonspatial statistics is that the probability of an item being selected changes as each additional entity enters the sample from the population. Here, the issue concerns whether the probability that any particular square in the grid will be black or white. If this probability can be determined *a priori*, for example from published figures for another location, in other words without reference to the empirical data for the study

Box 10.3a: Join count statistics

Free sampling with replacement

Expected number of B–B joins: $E_{BB} = Jp^2$

Expected number of B–W joins: $E_{BW} = 2Jpq$

Expected number of W–W joins: $E_{WW} = Jq^2$

Standard deviation of expected B–B joins: $\sigma_{BB} = \sqrt{Jp^2 + 2Kp^3 - (J+2K)p^4}$

Standard deviation of expected B–W joins: $\sigma_{BW} = \sqrt{(2J + \sum L(L+1)pq - 4(J + \sum L(L-1)p^2q^2}$

Standard deviation of expected W–W joins: $\sigma_{WW} = \sqrt{Jq^2 + 2Kq^3 - (J+2K)q^4}$

Free sample without replacement

Expected number of B–B joins: $E_{BB} = J\dfrac{n_W(n_W-1)}{n(n-1)}$

Expected number of B–W joins: $E_{BW} = 2J\dfrac{n_B n_W}{n(n-1)}$

Expected number of W–W joins: $E_{WW} = J\dfrac{n_B(n_B-1)}{n(n-1)}$

Standard deviation of expected B–B joins:

$$\sigma_{BB} = \sqrt{E_{BB} + 2K\frac{n_B(n_B-1)(n_B-2)}{n(n-1)(n-2)} + [J(J-1)-2K]\frac{n_B(n_B-1)(n_B-2)(n_B-3)}{n(n-1)(n-2)(n-3)} - (E_{BB}^2)}$$

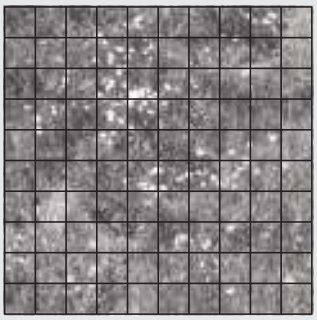
Standard deviation of expected B–W joins:

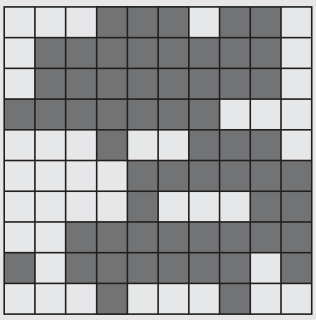
$$\sigma_{BW} = \sqrt{\frac{2(J+K)n_B n_W}{n(n-1)} + 4\lfloor J(J-1)-2K \rfloor\left(\frac{n_B(n_B-1)n_W(n_W-1)}{n(n-1)(n-2)(n-3)}\right) - 4\left(\frac{Jn_B n_W}{n(n-1)}\right)^2}$$

Standard deviation of expected W–W joins:

$$\sigma_{WW} = \sqrt{E_{WW} + 2K\frac{n_W(n_W-1)(n_W-2)}{n(n-1)(n-2)} + [J(J-1)-2K]\frac{n_W(n_W-1)(n_W-2)(n_W-3)}{n(n-1)(n-2)(n-3)} - (E_{WW}^2)}$$

Z test statistic: $Z = (O_{BW} - E_{BW})/\sigma_{BW}$





W–W	B–B	W–B/B–W
2	3	4
0	7	2
0	7	2
2	6	1
3	2	4
3	5	1
5	1	3
1	7	1
0	5	4
5	0	4

W–W	5	5	2	1	0	0	0	0	1	4
B–B	0	2	3	6	6	4	5	5	5	3
W–B/B–W	4	2	4	2	3	5	4	4	3	2

39		
	82	
		59

Box 10.3b: Application of join counts statistics.

Join Count Statistics work by examining the amount of separation between individual spatial units in respect of the nominal categories to which they have been assigned as the result of the distribution of some phenomenon. The procedure involves counting the number of joins between spatial units (grid squares in this example) that fall into the different possible combinations (0–0, 1–1 and 1–0/0–1). 0 denotes the absence of the phenomenon and 1 its presence, which are here represented as White and Black squares. The total number of cells in the grid (n) divides between n_W and n_B, where the subscripts denote the type of square. The total number of joins is identified as J and K is defined as $K = \sum J_i(J_i - 1)/2$ where the subscript i refers to individual squares from 1 to n. The probabilities of presence and absence of the phenomenon in any individual cells are referred to as p and q, respectively. These probabilities, which are used to calculate the expected numbers of joins in the different combinations, are determined in one of two ways depending upon whether free (with replacement) or nonfree (without replacement) sampling is used. The difference between these relates to whether the probabilities are defined by *a priori* reasoning or by *a posteriori* empirical evidence. The present application is typical in so far as nonfree sampling is assumed.

This application of JCS concerns the distribution of *dianthus gratianopoltanus* on part of the slope of Mont Cantal in the Auvergne. A 10×10 square grid has been superimposed over the area and the presence or absence of the species is shown by black and white shading. There were 59 black and 41 white squares. The calculations in Box 10.3d show that the expected number of BB joins was 29.82, of BW or WB was 87.96 and 62.22 for WW. The observed numbers were, respectively, 82, 39 and 59 (see above). The Null Hypothesis when testing JCS is that the spatial pattern of the phenomena in the grid squares is random, while the Alternative Hypothesis states it is either clustered or dispersed. Given the differences between these figures it is not surprising that the Z tests indicate that they are significant at the 0.05 level and the spatial pattern is not likely to have occurred by chance. It is reasonable to conclude that there is significant spatial autocorrelation in the distribution of *dianthus gratianopoltanus* in this area.

The key stages in applying Join Count Statistics are:

Tabulate the individual squares in the grid and count the different types of join for each: this can be a laborious process, since it involves inspecting each square and determining the code (0/1 or B/W) of all of its neighbours;

Calculate the values J and K, and count the numbers of observed BB, BW/WB and WW joins: these calculations are illustrated below;

Calculate the counts and standard deviations of the expected number of BB, BW/WB and WW joins: these are obtained by applying the equations appropriate to free or nonfree sampling

Calculate the Z test statistics for each type of join: the Z test statistics are calculated in a similar way to other tests and in this application are 2.54 (BB), 4.49 (BW/WB) and 6.22 (WW);

State Null Hypotheses and significance level: each Null Hypothesis for the three types of join states that the difference between the observed and expected counts is not significantly greater than would occur by chance at the 0.05 level of significance.

Determine the probabilities of the Z test statistics: the probabilities are equal to or less than 0.01;

Accept or reject the Null Hypothesis: each of the Null Hypotheses should be rejected at the 0.05 level of significance and the Alternative Hypotheses are therefore accepted leading to the conclusion that the spatial pattern tends toward being clustered.

Box 10.3c: Calculation of Join Count Statistics and significance testing.

$i = 1 \ldots n$	WW	BB	WB/BW	J_i	$J_i(J_i - 1)$
1	2			2	2
2	2		1	3	6
3	1		2	3	6
4		2	1	3	6
5		3		3	6
6		2	1	3	6
7			3	3	6
8		2	1	3	6
9		2	1	3	6
10	1		1	2	2
11	2		1	3	6
12		2	2	4	12
13		3	1	4	12
14		4		4	12
15		4		4	12
16		4		4	12
17		3	1	4	12
18		4		4	12
19		3	1	4	12
20	2		1	3	6
21	1		2	3	6
22		3	1	4	12
23		4		4	12
24		4		4	12
25		4		4	12
26		4		4	12
27		4		4	12
28		3	1	4	12
29		2	2	4	12
30	2		1	3	6
31		1	2	3	6
32		3	1	4	12

33		3	1	4	12
34		4		4	12
35		3	1	4	12
36		3	1	4	12
37		3	1	4	12
38	1		3	4	12
39	2		2	4	12
40	3			3	6
41	2		1	3	6
42	3		1	4	12
43	2		2	4	12
44		1	3	4	12
45	1		3	4	12
46	1		3	4	12
47		3	1	4	12
48		3	1	4	12
49		2	2	4	12
50	1		2	3	6
51	3			3	6
52	4			4	12
53	4			4	12
54	2		2	4	12
55		2	2	4	12
56		2	2	4	12
57		3	1	4	12
58		3	1	4	12
59		4		4	12
60		2	1	3	6
61	3			3	6
62	4			4	12
63	3		1	4	12
64	2		2	4	12
65		2	2	4	12
66	1		3	4	12
67	2		2	4	12
68	1		3	4	12

$i = 1 \ldots n$	WW	BB	WB/BW	J_i	$J_i(J_i - 1)$
69		3	1	4	12
70		3		3	6
71	2		1	3	6
72	3		1	4	12
73		2	2	4	12
74		3	1	4	12
75		4		4	12
76		3	1	4	12
77		3	1	4	12
78		3	1	4	12
79		3	1	4	12
80		3		3	6
81			3	3	6
82	2		2	4	12
83		2	2	4	12
84		4		4	12
85		3	1	4	12
86		3	1	4	12
87		3	1	4	12
88		3	1	4	12
89	1		3	4	12
90		1	2	3	6
91	1		1	2	2
92	3			3	6
93	1		2	3	6
94		1	2	3	6
95	1		2	3	6
96	2		1	3	6
97	1		2	3	6
98		1	2	3	6
99	2		1	3	6
100	1		1	2	2

		$\sum J_i/2 = 180$
		180
	$K = \sum J_i(J_i - 1)/2 = 484$	
Expected number of B–B joins	$E_{BB} = J\dfrac{n_W(n_W - 1)}{n(n-1)}$	$180\dfrac{41(40)}{100(99)} = 29.82$
Standard deviation of expected B–B joins	$\sigma_{BB} = \sqrt{\begin{matrix} E_{BB} + 2K\dfrac{n_B(n_B-1)(n_B-2)}{n(n-1)(n-2)} + \\ [J(J-1) - 2K] \\ \dfrac{n_B(n_B-1)(n_B-2)(n_B-3)}{n(n-1)(n-2)(n-3)} - (E_{BB}^2) \end{matrix}}$	$\sqrt{\begin{matrix} 29.83 + 2(484)\dfrac{59(n58)(57)}{100(99)(98)} + \\ [180(179) - 2(484)] \\ \dfrac{59(58)(57)(56)}{100(99)(98)(97)} - 29.82^2 \end{matrix}} = 3.61$
Z test statistic of B–B joins	$Z = (O_{BB} - E_{BB})/\sigma_{BB}$	$(39 - 29.82)/3.61 = 2.54$
Probability	$p = 0.0111$	
Expected number of W–B/B–W joins	$E_{BW} = 2J\dfrac{n_B n_W}{n(n-1)}$	$\dfrac{2(180(59)(41))}{100(100-1)} = 87.96$
Standard deviation of expected B–W joins	$\sigma_{BW} = \sqrt{\begin{matrix} \dfrac{2(J+K)n_B n_W}{n(n-1)} + \\ 4\lfloor J(J-1) - 2K \rfloor \left(\dfrac{n_B(n_B-1)n_W(n_W-1)}{n(n-1)(n-2)(n-3)}\right) - \\ 4\left(\dfrac{Jn_B n_W}{n(n-1)}\right)^2 \end{matrix}}$	$\sqrt{\begin{matrix} \dfrac{2(180+484)59(41)}{100(99)} + \\ 4\lfloor 180(179) - 2(484) \rfloor \left(\dfrac{59(58)41(40)}{100(99)(98)(97)}\right) - 6.45 \\ 4\left(\dfrac{180(59)(58)n_B n_W}{100(99)}\right)^2 \end{matrix}}$
Z test statistic of B–W joins	$Z = (O_{BW} - E_{BW})/\sigma_{BW}$	$(59 - 87.96)/6.45 = 4.49$
Probability	$p < 0.000$	
Expected number of W–W joins	$E_{WW} = J\dfrac{n_B(n_B - 1)}{n(n-1)}$	$180\dfrac{59(58)}{100(99)} = 62.22$
Standard deviation of expected W–W joins	$\sigma_{WW} = \sqrt{\begin{matrix} E_{WW} + 2K\dfrac{n_W(n_W-1)(n_W-2)}{n(n-1)(n-2)} + \\ [J(J-1) - 2K] \\ \dfrac{n_W(n_W-1)(n_W-2)(n_W-3)}{n(n-1)(n-2)(n-3)} - (E_{WW}^2) \end{matrix}}$	$\sqrt{\begin{matrix} 62.22 + 2(484)\dfrac{41(40)(39)}{100(99)(98)} + \\ [180(179) - 2(484)]\dfrac{41(40)(39)(39)}{100(99)(98)(97)} - 62.22^2 \end{matrix}} = 3.41$
Z test statistic of W–W joins	$Z = (O_{WW} - E_{WW})/\sigma_{WW}$	$(41 - 62.22)/3.41 = 6.22$
Probability	$p < 0.000$	

area, then free sampling applies. Sampling without replacement is much more common and its effect is to alter the expected number of joins in each combination (0–0, 1–1 and 1–0/0–1) from an equal distribution or some other hypothesized values.

The application in Box 10.3 examines the application of JCS in respect of the presence or absence of *dianthus gratianopoltanus* on the side of the Mont Cantal in the southern Auvergne region of France. The expected counts are calculated under the assumption of nonfree sampling, since there is no *a priori* reason to assign specific values to p and q, respectively the probabilities of presence and absence of the species in a square. The data values in this example are nominal codes (1 and 0) relating to the presence and absence of *dianthus gratianopoltanus* and there is no indication of the number of individual plants, whereas examination of the image of the slope in Box 10.3a hints at some variation in the density of occurrence. The observed numbers of B–B, B–W/W–B and W–W joins are all significantly different from what would be expected by chance, therefore it is reasonable to conclude that the spatial pattern is not random, but indicates an underlying process in relation to the distribution of *dianthus gratianopoltanus* in this area.

It should be noted that the spatial lag in this example is 1 (i.e. adjacent grid squares), whereas further analyses could be carried out where the comparison was made between 2nd-order neighbours, this would mean that the counts of B–B, B–W/W–B and W–W combinations were made by 'jumping over' adjacent squares to the next but one. Similarly, queen's move adjacencies could also be included. Finally, it should be noted that grids such as the one used in this example are often placed over a study in a relatively arbitrary fashion and the size and number of grid squares may be chosen for convenience rather than in a more rigorous way. There is no reason why a study area should be constrained so that it is covered by a square or rectangular grid. Suppose our study area is bounded on one or more sides by coastline or river, it is highly unlikely that such natural features of the environment will be delimited by straight lines and some of the cells in the grid or lattice are likely to overlap the coast or river. These units would have a reduced chance of including or excluding the phenomenon under investigation. Examination of the image and superimposed grid in Box 10.3a shows that the size of each flower is relatively small in relation to the size of a grid square. Thus, some squares contain just one occurrence, whereas others have many, yet both are counted as presences of the phenomenon. Smaller grid squares closer to the size of each flower head would perhaps give a more realistic impression of the species' distribution, since isolated occurrences may have distorted the situation.

10.2.1.2 Moran's I *Index*

Some of the issues mentioned at the end of the previous section arise from the rather artificial superimposition of a regular grid or lattice over a study area and that JCS applies to nominal data values. **Moran's *I*** is a widely available technique that can be used when the study is covered by an incomplete regular grid or a set of planar polygons and the data values are real numbers rather than counts of units in nominal