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

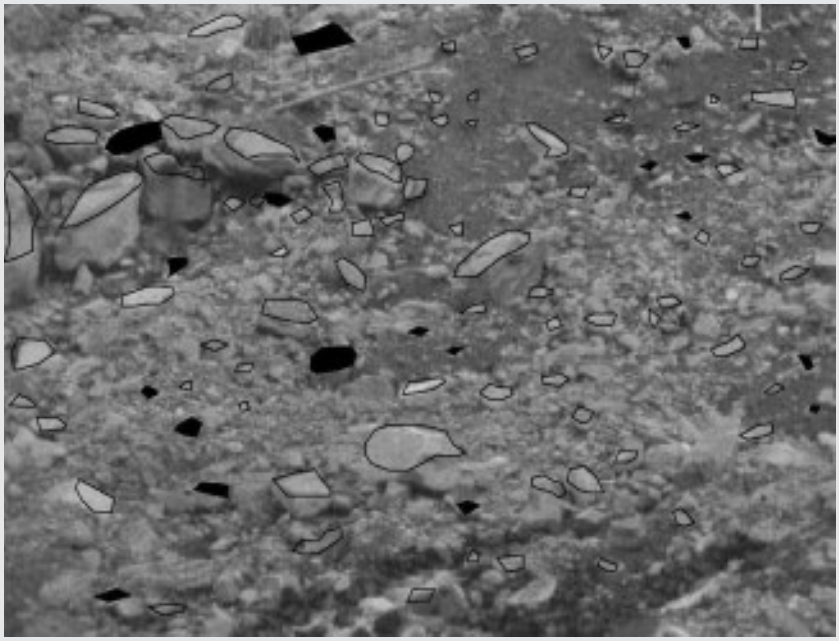
10.1 Issues with correlation and regression of spatial data

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

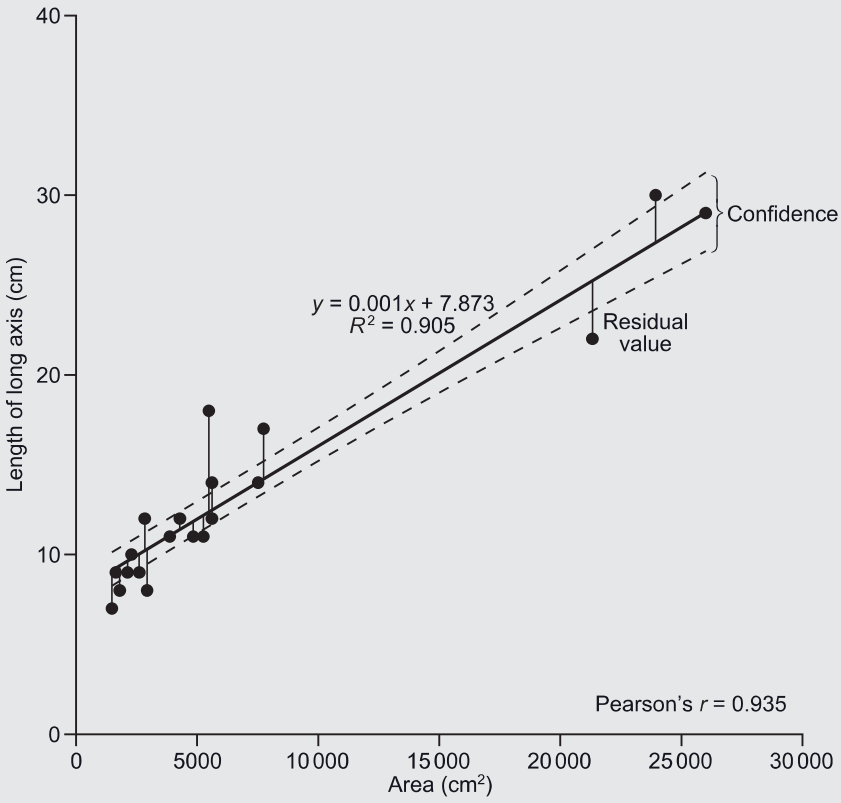
Box 10.1: Spatial autocorrelation.



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

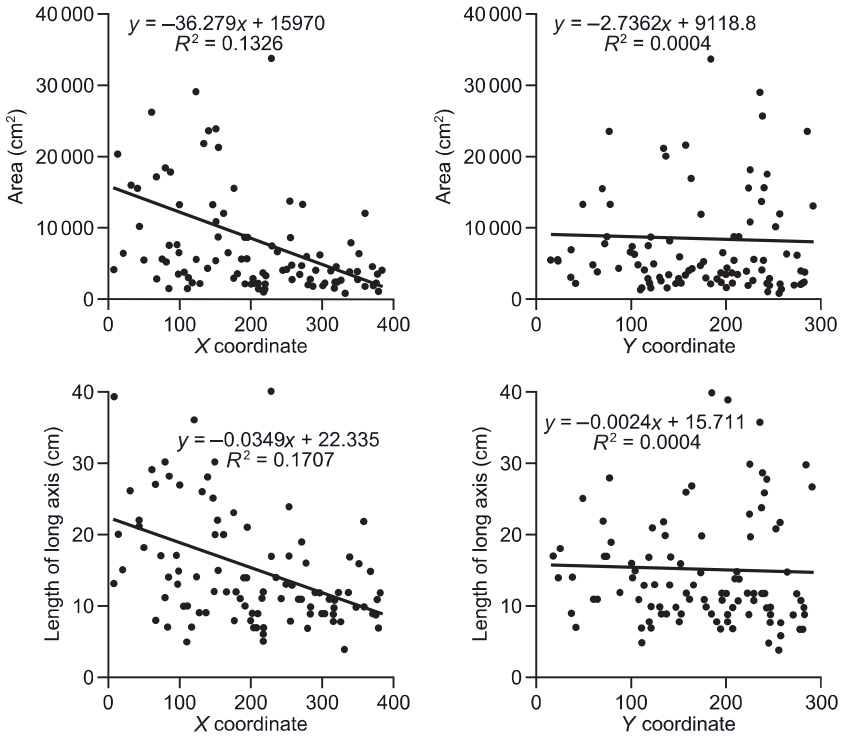


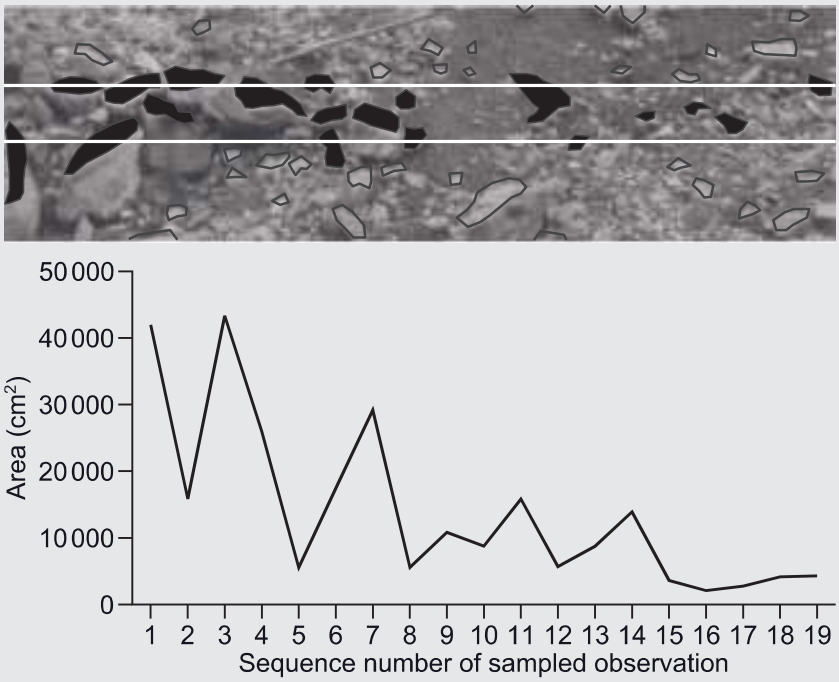
Figure 10.1 Scatter plots of full sample of moraine material by area and length of long axis against *X* and *Y* spatial coordinates.

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.

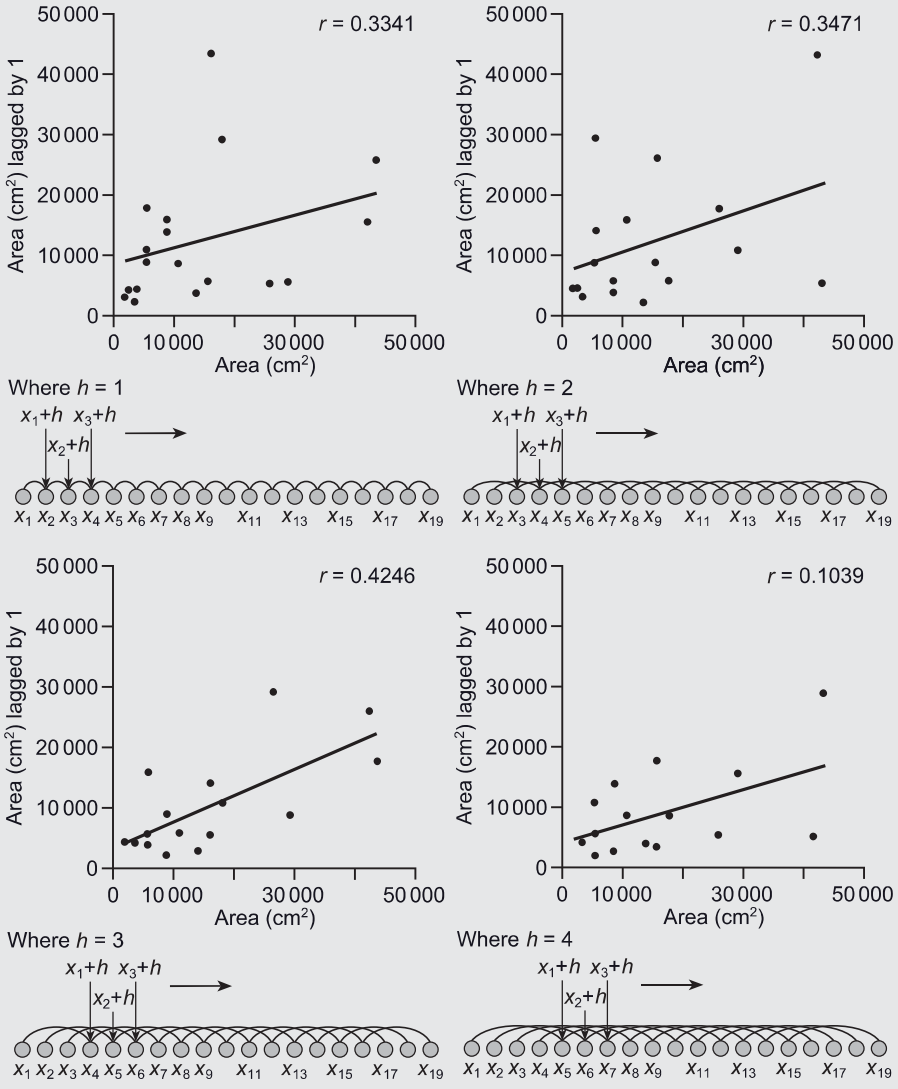
Box 10.2b: Linking data values by spatial lags.

		$x_{1\ldots n} + h$	$x_{1\ldots n} + h$	$x_{1\ldots n} + h$	$x_{1\ldots n} + h$
		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

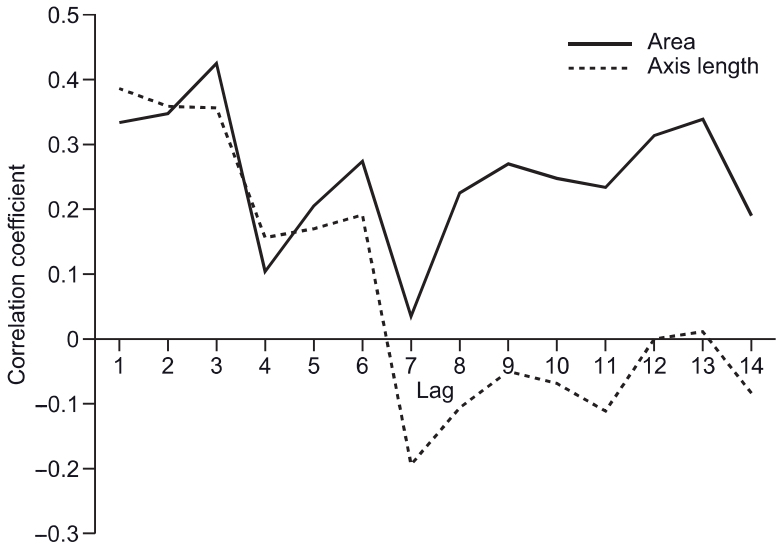


Figure 10.2 Correlograms for area and length of the long axis of moraine material intersecting with transect.

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

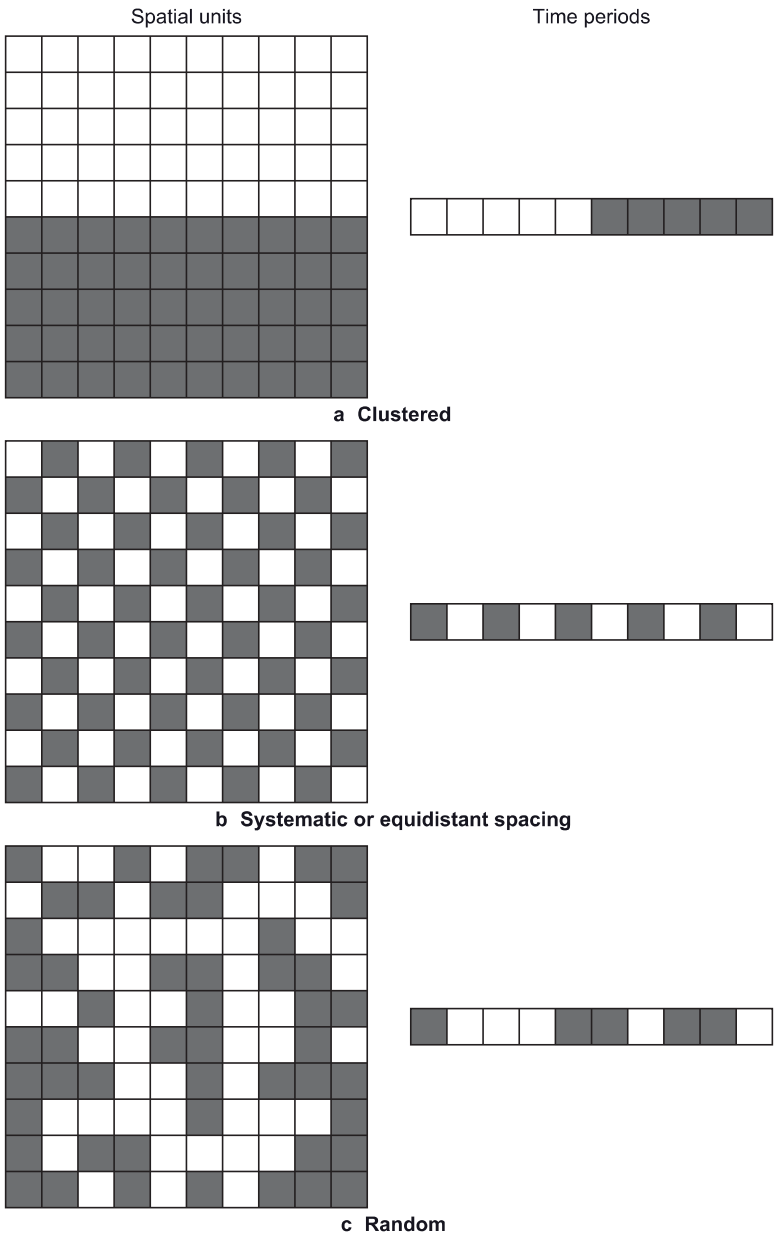


Figure 10.3 Binary join count patterns for regular grid squares and linear time series.

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:

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

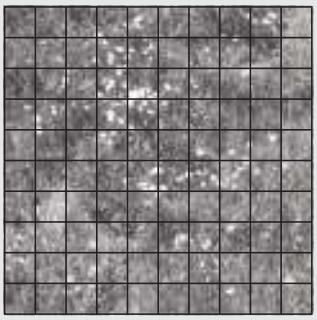
Standard deviation of expected B–W joins:

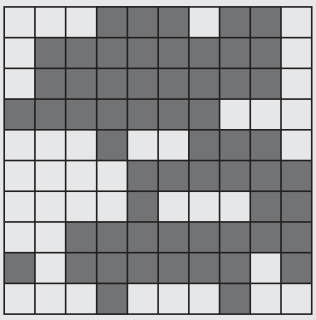
$$\sigma_{BW} = \sqrt{\frac{2(J+K)n_B n_W}{n(n-1)} + 4\lfloor J(J-1)-2K \rfloor\left(\frac{n_B(n_B-1)n_W(n_W-1)}{n(n-1)(n-2)(n-3)}\right) - 4\left(\frac{Jn_B n_W}{n(n-1)}\right)^2}$$

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.

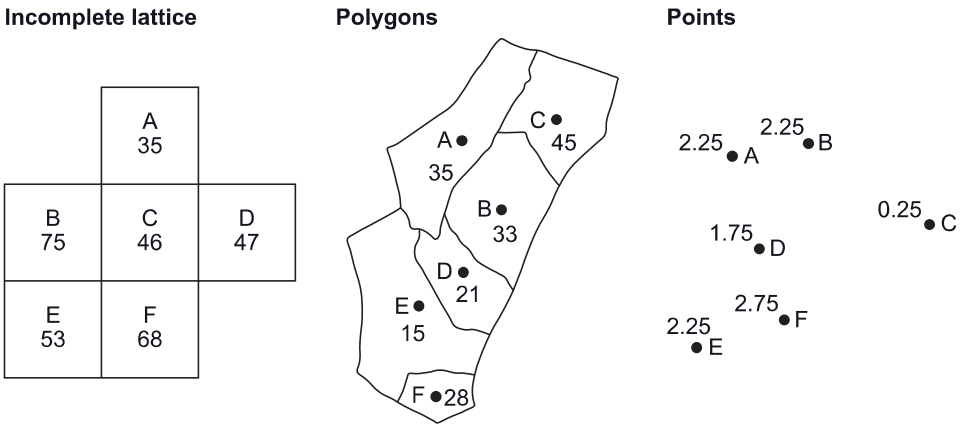
The application in Box 10.3 examines the application of JCS in respect of the presence or absence of *dianthus gratianopoltanus* on the side of the Mont Cantal in the southern Auvergne region of France. The expected counts are calculated under the assumption of nonfree sampling, since there is no *a 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

dichotomous categories. The difference between the presentation of the raw data for Moran's *I* compared with JCS is that rather than tabulating count statistics the data are organized in a three-column format where the first two columns contain *X* and *Y* coordinates relating to row and column numbers, the grid references of points features or the centroids of polygons. Figure 10.4 illustrates the procedure with respect to an incomplete lattice, irregular polygons and points. The data tables shows the *X* and *Y* coordinates or row and column numbers and the third column contains the *Z* data values corresponding to these locations, which may be decimal values or integer counts, as in Figure 10.4. Although this process retains references to the spatial



Tabulated data

	X	Y	z
A	1	2	35
B	2	1	75
C	2	2	46
D	2	3	47
E	3	1	53
F	3	2	68

Tabulated data

	X	Y	z
A	3616	1720	35
B	3620	1721	33
C	3618	1720	45
D	3616	1716	15
E	3617	1717	21
F	3616	1714	28

Tabulated data

	X	Y	z
A	5181	1673	2.25
B	5182	1673	0.75
C	5183	1672	0.25
D	5182	1672	1.75
E	5181	1671	2.25
F	5182	1671	2.75

Adjacency matrix

	A	B	C	D	E	F
A	0	0	1	0	0	0
B	0	0	1	0	1	0
C	1	1	0	1	0	1
D	0	0	1	0	0	0
E	0	1	0	0	0	1
F	0	0	1	0	1	0

Adjacency matrix

	A	B	C	D	E	F
A	0	1	1	1	1	0
B	1	0	1	1	0	0
C	1	1	0	0	0	0
D	0	0	0	0	1	1
E	1	1	0	1	0	0
F	0	0	0	0	1	0

Nearest neighbour

	A	B	C	D	E	F
A	0	1	0	0	0	0
B	1	0	0	0	0	0
C	0	1	0	0	0	0
D	0	0	0	0	0	1
E	0	0	0	0	0	0
F	0	0	0	1	0	0

Figure 10.4 Tabular representation of irregular lattice, polygon and point feature data and weights matrices.

location of the features it discards information about their topological connections, in other words whether one unit is adjacent to another. Since such information is a vital component in the analysis of spatial patterns, it is necessary to create a **weights matrix** (W) that records which units share a common boundary in the case of area data or are nearest neighbours to each other in the case of point feature data. Figure 10.4 includes the first-order weights matrices for the three types of spatial data. Normally these weights matrices are obtained for 1st-order neighbours, but 2nd-, 3rd- or higher-order neighbours can also be used. The weights matrix can incorporate rook's and queen's adjacencies, which can be weighted to denote their relative importance if required.

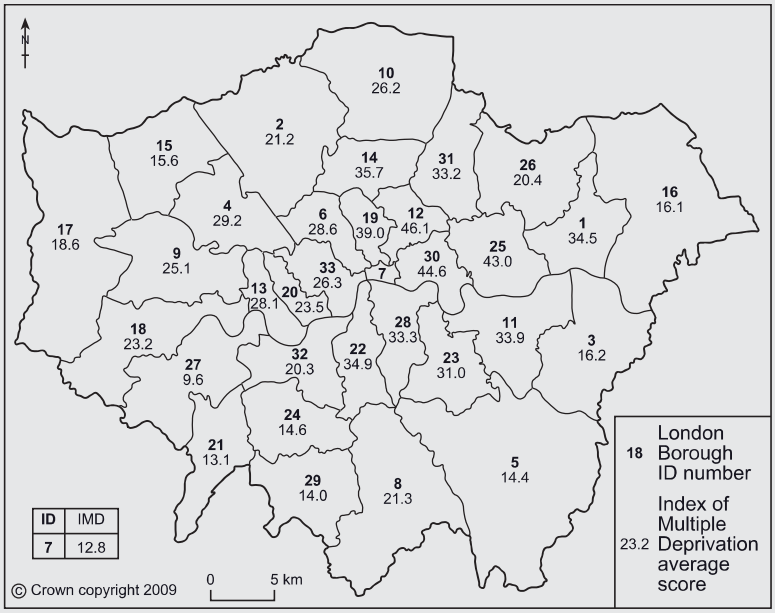
Pairs of adjacent features in a spatial pattern displaying or possessing spatial autocorrelation will both have positive **or** negative values for the variable under investigation and if these values are relatively large, it will indicate stronger rather than weaker spatial autocorrelation. In contrast, pairs of neighbouring features where one has a positive and the other a negative value suggest the absence of spatial autocorrelation. These statements are effectively another way of describing Tobler's first law of Geography. Moran's *I* encapsulates the essence of these statements in a single index value. The sequence of calculations to compute Moran's *I* is somewhat protracted and involves the manipulation of data in matrix format. The first stage is subtraction of the overall mean of the variable (Z) from the data values of each spatial feature and then multiplying the result for each pair of features. The spatial weights are used to select which pairs of features are included and excluded from the final calculation of the index: those with a 0 in the weights matrix (denoting nonadjacency) are omitted, whereas those with a 1 are included. Summing these values and dividing by the sum of the weights produces a covariance. This forms the numerator in the equation for Moran's *I,* which is divided by the variance of the data to produce the index value that lies within the range −1.0 to +1.0. Values towards the extremes of this range, respectively, indicate negative and positive spatial autocorrelation, whereas a value around zero generally its absence.

Box 10.4 illustrates the application of Moran's *I* in relation to the mean scores on the UK government's 2007 Index of Multiple Deprivation for the 33 London Boroughs. The IMD is computed from a series of different variables within domains covering such areas as income, employment and social conditions. The map of the IMD for the London local authorities suggests that higher mean scores were computed for many of the inner London Boroughs, whereas several of those more suburban ones had lower deprivation overall. Some degree of positive spatial autocorrelation seems to exist with high values clustered together near the centre and lower ones in outer areas. The results from the analysis provide moderate support for this claim, since Moran's *I* computes as 0.305. One possible explanation for this outcome is the presence of one relatively small authority in the centre, the City of London, with a very low index value (12.54) in comparison with its seven adjacent neighbours, which apart from one have IMD values over 25.00. The Moran's *I* index has been recomputed leaving out the City of London Borough and the effect is to

Box 10.4a: Moran's I Index.

Moran's I: $$I = \frac{1}{p} \frac{\sum_{I=1}^{n} \sum_{j=1 j \neq i}^{n} w_{i,j}(z_i - \overline{z})(z_j - \overline{z})}{\sum_{i=1}^{n} (z_i - \overline{z})^2}$$

where $$p = \sum_i \sum_j w_{ij} / n$$



raise the value to 0.479, which seems to confirm the initial visual impression of strong positive spatial autocorrelation.

There are two approaches to testing the significance of Moran's I and similar global spatial statistics such as **Geary's C** and the Getis and Ord's **G statistic**. One approach, following the standard assumption of nonspatial statistics, is to assume that the calculated statistic or index value is from a Normal Distribution of such quantities computed from a series of independent and identical samples or selections. The notion of sampling with respect to spatial features was examined in Section 6.2.2. The second approach is rather more empirical in nature and views the particular set of data values that has arisen as just one of all the possible random distributions of observed data values across the set of zones. The number of random distributions equals the factorial of the number of spatial features, for example if there are four

Box 10.4b: Application of Moran's *I* index.

The Moran's *I* index is an adaptable statistic for measuring spatial autocorrelation that is widely used in a range of disciplines. The variable being analysed (Z) has subscripts i and j to distinguish between the different observations in a pair. Binary weights are used in this example denoting whether any given pair of areas share a common boundary and the subscripts in the weight term $w_{i,j}$ also refer to a pair of observations i and j in a set with n features overall. The individual values of X are usually adjusted by subtracting the overall mean of the variable before multiplying pairs of values together. Variance/covariance-like quantities are computed, the former from the sum of the squared products in the C matrix that fall along the diagonal and the latter from the nondiagonal elements where there is a join between the spatial features represented by the row and column.

The Moran's *I* index has been applied to the average score variable in the 2007 Index of Multiple Deprivation with respect to the 33 local authorities (boroughs) in London. Visual inspection of the pattern of data values for these areas on the map suggests that lower scores (less deprivation) are found in the more peripheral zones and higher ones in the centre. The tabulated data in Box 10.4c records the mean index value is 25.68 and the following adjacency matrix of 0s and 1s shows that the minimum and maximum numbers of joins between areas are 3 and 7 with a total of 164. The values along the diagonal in the matrix used to compute the variance-like quantity are unshaded and the values for those pairs of local authorities that are adjacent (i.e. have a 1 in the weights matrix) are shaded in a darker grey. The Moran's *I* index in this application is 0.305, which indicates a moderate positive spatial autocorrelation. The randomization significance testing procedure has been applied and indicates that this index value is significant at the 0.05 level with 9999 permutations used to produce the reference distribution.

areas the number of permutations is 24 (4*x*3*x*2*x*1), whereas there are 8,683,317,618,811,890,000,000,000,000,000,000,000,000 (33!) ways in which the 33 values of the 2007 Index of Multiple Deprivation could be arranged across the London Boroughs.

Both methods are implemented in various software packages that carry out spatial statistics and in the case of the randomization approach, users are normally asked to specify the number of random permutations to be generated and the spatial index is then calculated for each of these to produce a pseudo-probability distribution. The index computed from the observed data values is compared with this distribution. Both approaches to testing the significance of spatial indices involve converting the observed index value into a Z score in the standard way using the hypothesized or empirically derived population mean and standard deviation. For example, in the randomization approach the mean and standard deviation of the index in the pseudo-probability distribution are used. Box 10.4 includes the results of testing the significance of the observed Moran's *I* value, 0.305, using one of these statistical packages. The probability of the Moran's *I* obtained from the data values can be compared with a significance level that relates to the number of permutations used to generate the

Box 10.4c: Calculation of Moran's *I*.

Id No.	*z*	Id No.	*z*	Id No.	*z*	Id No.	*z*	Id No.	*z*
1	34.49	8	21.31	15	15.59	22	34.94	29	13.98
2	21.16	9	25.10	16	16.07	23	31.04	30	44.64
3	16.21	10	26.19	17	18.56	24	14.62	31	33.19
4	29.22	11	33.94	18	23.20	25	42.95	32	20.34
5	14.36	12	46.10	19	38.96	26	20.36	33	26.30
6	28.62	13	28.07	20	23.51	27	9.55		
7	12.84	14	35.73	21	13.10	28	33.33		
						Mean	$\sum x/n = 847.44/33 = 25.68$		

Adjacency matrix.

$\frac{j}{i}$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	1	0
2	0	0	0	1	0	1	0	0	0	1	0	0	0	1	1	0	0
3	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0
4	0	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0	0
5	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0
6	0	1	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0
7	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
8	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	1
10	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
11	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0
13	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
14	0	1	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0
15	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1
16	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0
18	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1
19	0	0	0	0	0	1	1	0	0	0	0	1	0	1	0	0	0

Adjacency matrix.

$\frac{j}{i}$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
20	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0
23	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
24	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
25	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
26	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
27	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
28	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0
31	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
33	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0
$\sum w_i$	5	5	4	7	6	6	7	4	5	3	6	6	6	6	4	3	3

$\frac{j}{i}$	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33
1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
5	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0
6	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	0	1	0	0	1	0	0	0	0	0	1	0	1	0	0	1
8	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0
9	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
11	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0
12	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0

13	1	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	
14	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
18	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
21	0	0	0	0	0	0	1	0	0	1	0	1	0	0	1	0	
22	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	1	
23	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	
24	0	0	0	1	1	0	0	0	0	0	0	1	0	0	1	0	
25	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	
26	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	
27	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	
28	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	
29	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	
30	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	
31	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	
32	0	0	1	1	1	0	1	0	0	1	0	0	0	0	0	1	
33	0	0	1	0	1	0	0	0	0	0	0	0	0	0	1	0	
$\sum w_i$	4	4	4	4	7	4	5	6	4	4	5	3	6	5	7	6	164

Calculation of variance/covariance-like quantities.

		j	1	2	3	4	5	6	7	8
i	z	$z-\overline{z}$	8.8	−4.5	−9.5	3.5	−11.3	2.9	−12.8	−4.4
1	34.5	8.8	77.5	−39.8	−83.4	31.1	−99.7	25.9	−113.1	−38.5
2	21.2	−4.5	−39.8	20.5	42.9	−16.0	51.2	−13.3	58.1	19.8
3	16.2	−9.5	−83.4	42.9	89.8	−33.5	107.3	−27.8	121.7	41.4
4	29.2	3.5	31.1	−16.0	−33.5	12.5	−40.0	10.4	−45.4	−15.5
5	14.4	−11.3	−99.7	51.2	107.3	−40.0	128.2	−33.2	145.4	49.5
6	28.6	2.9	25.9	−13.3	−27.8	10.4	−33.2	8.6	−37.7	−12.8
7	12.8	−12.8	−113.1	58.1	121.7	−45.4	145.4	−37.7	165.0	56.2
8	21.3	−4.4	−38.5	19.8	41.4	−15.5	49.5	−12.8	56.2	19.1

Calculation of variance/covariance-like quantities.

		j	1	2	3	4	5	6	7	8
i	z	$z-\overline{z}$	8.8	−4.5	−9.5	3.5	−11.3	2.9	−12.8	−4.4
9	25.1	−0.6	−5.1	2.6	5.5	−2.1	6.6	−1.7	7.5	2.6
10	26.2	0.5	4.5	−2.3	−4.8	1.8	−5.7	1.5	−6.5	−2.2
11	33.9	8.3	72.7	−37.3	−78.2	29.2	−93.5	24.2	−106.0	−36.1
12	46.1	20.4	179.8	−92.4	−193.4	72.2	−231.2	59.9	−262.2	−89.3
13	28.1	2.4	21.0	−10.8	−22.6	8.4	−27.0	7.0	−30.6	−10.4
14	35.7	10.0	88.5	−45.4	−95.2	35.5	−113.8	29.5	−129.0	−43.9
15	15.6	−10.1	−88.9	45.7	95.6	−35.7	114.3	−29.6	129.6	44.1
16	16.1	−9.6	−84.7	43.5	91.1	−34.0	108.9	−28.2	123.5	42.1
17	18.6	−7.1	−62.7	32.2	67.5	−25.2	80.7	−20.9	91.5	31.2
18	23.2	−2.5	−21.9	11.2	23.5	−8.8	28.1	−7.3	31.9	10.9
19	39.0	13.3	116.9	−60.1	−125.8	46.9	−150.3	39.0	−170.5	−58.1
20	23.5	−2.2	−19.1	9.8	20.6	−7.7	24.6	−6.4	27.9	9.5
21	13.1	−12.6	−110.8	56.9	119.2	−44.5	142.5	−36.9	161.6	55.0
22	34.9	9.3	81.5	−41.9	−87.7	32.7	−104.8	27.2	−118.9	−40.5
23	31.0	5.4	47.2	−24.2	−50.7	18.9	−60.7	15.7	−68.8	−23.4
24	14.6	−11.1	−97.4	50.1	104.8	−39.1	125.3	−32.5	142.1	48.4
25	43.0	17.3	152.0	−78.1	−163.6	61.1	−195.5	50.7	−221.8	−75.5
26	20.4	−5.3	−46.9	24.1	50.4	−18.8	60.3	−15.6	68.4	23.3
27	9.6	−16.1	−142.1	73.0	152.9	−57.1	182.7	−47.4	207.2	70.6
28	33.3	7.6	67.3	−34.6	−72.4	27.0	−86.6	22.4	−98.2	−33.4
29	14.0	−11.7	−103.1	52.9	110.9	−41.4	132.5	−34.4	150.3	51.2
30	44.6	19.0	166.9	−85.8	−179.6	67.0	−214.7	55.7	−243.5	−82.9
31	33.2	7.5	66.1	−34.0	−71.1	26.5	−85.0	22.0	−96.4	−32.8
32	20.3	−5.3	−47.1	24.2	50.6	−18.9	60.5	−15.7	68.6	23.4
33	26.3	0.6	5.4	−2.8	−5.8	2.2	−7.0	1.8	−7.9	−2.7

		j	9	10	11	12	13	14	15	16
i	z	$z-\bar{z}$	−0.6	0.5	8.3	20.4	2.4	10.0	−10.1	−9.6
1	34.5	8.8	−5.1	4.5	72.7	179.8	21.0	88.5	−88.9	−84.7
2	21.2	−4.5	2.6	−2.3	−37.3	−92.4	−10.8	−45.4	45.7	43.5
3	16.2	−9.5	5.5	−4.8	−78.2	−193.4	−22.6	−95.2	95.6	91.1
4	29.2	3.5	−2.1	1.8	29.2	72.2	8.4	35.5	−35.7	−34.0
5	14.4	−11.3	6.6	−5.7	−93.5	−231.2	−27.0	−113.8	114.3	108.9
6	28.6	2.9	−1.7	1.5	24.2	59.9	7.0	29.5	−29.6	−28.2
7	12.8	−12.8	7.5	−6.5	−106.0	−262.2	−30.6	−129.0	129.6	123.5
8	21.3	−4.4	2.6	−2.2	−36.1	−89.3	−10.4	−43.9	44.1	42.1
9	25.1	−0.6	0.3	−0.3	−4.8	−11.9	−1.4	−5.9	5.9	5.6
10	26.2	0.5	−0.3	0.3	4.2	10.3	1.2	5.1	−5.1	−4.9
11	33.9	8.3	−4.8	4.2	68.2	168.6	19.7	82.9	−83.3	−79.4
12	46.1	20.4	−11.9	10.3	168.6	416.8	48.7	205.1	−206.1	−196.3
13	28.1	2.4	−1.4	1.2	19.7	48.7	5.7	24.0	−24.1	−22.9
14	35.7	10.0	−5.9	5.1	82.9	205.1	24.0	100.9	−101.4	−96.6
15	15.6	−10.1	5.9	−5.1	−83.3	−206.1	−24.1	−101.4	101.9	97.0
16	16.1	−9.6	5.6	−4.9	−79.4	−196.3	−22.9	−96.6	97.0	92.4
17	18.6	−7.1	4.2	−3.6	−58.8	−145.4	−17.0	−71.6	71.9	68.5
18	23.2	−2.5	1.5	−1.3	−20.5	−50.7	−5.9	−25.0	25.1	23.9
19	39.0	13.3	−7.8	6.7	109.6	271.0	31.7	133.4	−134.0	−127.6
20	23.5	−2.2	1.3	−1.1	−17.9	−44.4	−5.2	−21.8	21.9	20.9
21	13.1	−12.6	7.3	−6.4	−103.9	−256.9	−30.0	−126.4	127.0	121.0
22	34.9	9.3	−5.4	4.7	76.4	189.0	22.1	93.0	−93.4	−89.0
23	31.0	5.4	−3.1	2.7	44.2	109.4	12.8	53.8	−54.1	−51.5
24	14.6	−11.1	6.5	−5.6	−91.3	−225.9	−26.4	−111.1	111.7	106.4
25	43.0	17.3	−10.1	8.7	142.6	352.5	41.2	173.5	−174.3	−166.0
26	20.4	−5.3	3.1	−2.7	−44.0	−108.7	−12.7	−53.5	53.7	51.2
27	9.6	−16.1	9.4	−8.2	−133.2	−329.4	−38.5	−162.1	162.9	155.1
28	33.3	7.6	−4.5	3.9	63.1	156.1	18.2	76.8	−77.2	−73.5
29	14.0	−11.7	6.8	−5.9	−96.6	−238.9	−27.9	−117.6	118.1	112.5
30	44.6	19.0	−11.1	9.6	156.5	387.0	45.2	190.4	−191.3	−182.2
31	33.2	7.5	−4.4	3.8	62.0	153.2	17.9	75.4	−75.8	−72.2
32	20.3	−5.3	3.1	−2.7	−44.1	−109.1	−12.8	−53.7	53.9	51.4
33	26.3	0.6	−0.4	0.3	5.1	12.6	1.5	6.2	−6.2	−5.9

		j	17	18	19	20	21	22	23	24
i	z	$z-\bar{z}$	−7.1	−2.5	13.3	−2.2	−12.6	9.3	5.4	−11.1
1	34.5	8.8	−62.7	−21.9	116.9	−19.1	−110.8	81.5	47.2	−97.4
2	21.2	−4.5	32.2	11.2	−60.1	9.8	56.9	−41.9	−24.2	50.1
3	16.2	−9.5	67.5	23.5	−125.8	20.6	119.2	−87.7	−50.7	104.8
4	29.2	3.5	−25.2	−8.8	46.9	−7.7	−44.5	32.7	18.9	−39.1
5	14.4	−11.3	80.7	28.1	−150.3	24.6	142.5	−104.8	−60.7	125.3
6	28.6	2.9	−20.9	−7.3	39.0	−6.4	−36.9	27.2	15.7	−32.5
7	12.8	−12.8	91.5	31.9	−170.5	27.9	161.6	−118.9	−68.8	142.1
8	21.3	−4.4	31.2	10.9	−58.1	9.5	55.0	−40.5	−23.4	48.4
9	25.1	−0.6	4.2	1.5	−7.8	1.3	7.3	−5.4	−3.1	6.5
10	26.2	0.5	−3.6	−1.3	6.7	−1.1	−6.4	4.7	2.7	−5.6
11	33.9	8.3	−58.8	−20.5	109.6	−17.9	−103.9	76.4	44.2	−91.3
12	46.1	20.4	−145.4	−50.7	271.0	−44.4	−256.9	189.0	109.4	−225.9
13	28.1	2.4	−17.0	−5.9	31.7	−5.2	−30.0	22.1	12.8	−26.4
14	35.7	10.0	−71.6	−25.0	133.4	−21.8	−126.4	93.0	53.8	−111.1
15	15.6	−10.1	71.9	25.1	−134.0	21.9	127.0	−93.4	−54.1	111.7
16	16.1	−9.6	68.5	23.9	−127.6	20.9	121.0	−89.0	−51.5	106.4
17	18.6	−7.1	50.7	17.7	−94.6	15.5	89.6	−65.9	−38.2	78.8
18	23.2	−2.5	17.7	6.2	−33.0	5.4	31.3	−23.0	−13.3	27.5
19	39.0	13.3	−94.6	−33.0	176.3	−28.9	−167.1	122.9	71.1	−146.9
20	23.5	−2.2	15.5	5.4	−28.9	4.7	27.4	−20.1	−11.6	24.1
21	13.1	−12.6	89.6	31.3	−167.1	27.4	158.4	−116.5	−67.4	139.2
22	34.9	9.3	−65.9	−23.0	122.9	−20.1	−116.5	85.7	49.6	−102.4
23	31.0	5.4	−38.2	−13.3	71.1	−11.6	−67.4	49.6	28.7	−59.3
24	14.6	−11.1	78.8	27.5	−146.9	24.1	139.2	−102.4	−59.3	122.4
25	43.0	17.3	−123.0	−42.9	229.2	−37.5	−217.3	159.8	92.5	−191.0
26	20.4	−5.3	37.9	13.2	−70.7	11.6	67.0	−49.3	−28.5	58.9
27	9.6	−16.1	114.9	40.1	−214.2	35.1	203.0	−149.3	−86.4	178.5
28	33.3	7.6	−54.5	−19.0	101.5	−16.6	−96.2	70.8	41.0	−84.6
29	14.0	−11.7	83.4	29.1	−155.4	25.4	147.3	−108.3	−62.7	129.5
30	44.6	19.0	−135.0	−47.1	251.7	−41.2	−238.5	175.5	101.5	−209.7
31	33.2	7.5	−53.5	−18.6	99.7	−16.3	−94.5	69.5	40.2	−83.0
32	20.3	−5.3	38.1	13.3	−70.9	11.6	67.2	−49.5	−28.6	59.1
33	26.3	0.6	−4.4	−1.5	8.2	−1.3	−7.8	5.7	3.3	−6.8

		j	25	26	27	28	29	30	31	32
i	z	$z-\bar{z}$	17.3	−5.3	−16.1	7.6	−11.7	19.0	7.5	−5.3
1	34.5	8.8	152.0	−46.9	−142.1	67.3	−103.1	166.9	66.1	−47.1
2	21.2	−4.5	−78.1	24.1	73.0	−34.6	52.9	−85.8	−34.0	24.2
3	16.2	−9.5	−163.6	50.4	152.9	−72.4	110.9	−179.6	−71.1	50.6
4	29.2	3.5	61.1	−18.8	−57.1	27.0	−41.4	67.0	26.5	−18.9
5	14.4	−11.3	−195.5	60.3	182.7	−86.6	132.5	−214.7	−85.0	60.5
6	28.6	2.9	50.7	−15.6	−47.4	22.4	−34.4	55.7	22.0	−15.7
7	12.8	−12.8	−221.8	68.4	207.2	−98.2	150.3	−243.5	−96.4	68.6
8	21.3	−4.4	−75.5	23.3	70.6	−33.4	51.2	−82.9	−32.8	23.4
9	25.1	−0.6	−10.1	3.1	9.4	−4.5	6.8	−11.1	−4.4	3.1
10	26.2	0.5	8.7	−2.7	−8.2	3.9	−5.9	9.6	3.8	−2.7
11	33.9	8.3	142.6	−44.0	−133.2	63.1	−96.6	156.5	62.0	−44.1
12	46.1	20.4	352.5	−108.7	−329.4	156.1	−238.9	387.0	153.2	−109.1
13	28.1	2.4	41.2	−12.7	−38.5	18.2	−27.9	45.2	17.9	−12.8
14	35.7	10.0	173.5	−53.5	−162.1	76.8	−117.6	190.4	75.4	−53.7
15	15.6	−10.1	−174.3	53.7	162.9	−77.2	118.1	−191.3	−75.8	53.9
16	16.1	−9.6	−166.0	51.2	155.1	−73.5	112.5	−182.2	−72.2	51.4
17	18.6	−7.1	−123.0	37.9	114.9	−54.5	83.4	−135.0	−53.5	38.1
18	23.2	−2.5	−42.9	13.2	40.1	−19.0	29.1	−47.1	−18.6	13.3
19	39.0	13.3	229.2	−70.7	−214.2	101.5	−155.4	251.7	99.7	−70.9
20	23.5	−2.2	−37.5	11.6	35.1	−16.6	25.4	−41.2	−16.3	11.6
21	13.1	−12.6	−217.3	67.0	203.0	−96.2	147.3	−238.5	−94.5	67.2
22	34.9	9.3	159.8	−49.3	−149.3	70.8	−108.3	175.5	69.5	−49.5
23	31.0	5.4	92.5	−28.5	−86.4	41.0	−62.7	101.5	40.2	−28.6
24	14.6	−11.1	−191.0	58.9	178.5	−84.6	129.5	−209.7	−83.0	59.1
25	43.0	17.3	298.1	−91.9	−278.6	132.0	−202.1	327.3	129.6	−92.3
26	20.4	−5.3	−91.9	28.3	85.9	−40.7	62.3	−100.9	−40.0	28.5
27	9.6	−16.1	−278.6	85.9	260.3	−123.4	188.8	−305.8	−121.1	86.2
28	33.3	7.6	132.0	−40.7	−123.4	58.5	−89.5	144.9	57.4	−40.9

		j	25	26	27	28	29	30	31	32
i	z	$z-\bar{z}$	17.3	−5.3	−16.1	7.6	−11.7	19.0	7.5	−5.3
29	14.0	−11.7	−202.1	62.3	188.8	−89.5	137.0	−221.9	−87.9	62.5
30	44.6	19.0	327.3	−100.9	−305.8	144.9	−221.9	359.3	142.3	−101.3
31	33.2	7.5	129.6	−40.0	−121.1	57.4	−87.9	142.3	56.3	−40.1
32	20.3	−5.3	−92.3	28.5	86.2	−40.9	62.5	−101.3	−40.1	28.6
33	26.3	0.6	10.6	−3.3	−9.9	4.7	−7.2	11.7	4.6	−3.3

		j	33		
i	z	$z-\bar{z}$	0.6	$(z_i-\bar{z})^2$	Row sum $w_{i,j}(z_i-\bar{z})(z_j-\bar{z})$
1	34.5	8.8	5.4	77.5	9.78
2	21.2	−4.5	−2.8	20.5	−31.35
3	16.2	−9.5	−5.8	89.8	36.72
4	29.2	3.5	2.2	12.5	−40.44
5	14.4	−11.3	−7.0	128.2	−188.73
6	28.6	2.9	1.8	8.6	29.67
7	12.8	−12.8	−7.9	165.0	−938.92
8	21.3	−4.4	−2.7	19.1	108.63
9	25.1	−0.6	−0.4	0.3	8.05
10	26.2	0.5	0.3	0.3	6.59
11	33.9	8.3	5.1	68.2	244.27
12	46.1	20.4	12.6	416.8	1106.69
13	28.1	2.4	1.5	5.7	−55.32
14	35.7	10.0	6.2	100.9	403.02
15	15.6	−10.1	−6.2	101.9	87.77
16	16.1	−9.6	−5.9	92.4	57.60
17	18.6	−7.1	−4.4	50.7	93.76
18	23.2	−2.5	−1.5	6.2	53.29

19	39.0	13.3	8.2	176.3	272.88
20	23.5	−2.2	−1.3	4.7	−2.60
21	13.1	−12.6	−7.8	158.4	556.78
22	34.9	9.3	5.7	85.7	−339.58
23	31.0	5.4	3.3	28.7	126.05
24	14.6	−11.1	−6.8	122.4	273.83
25	43.0	17.3	10.6	298.1	1012.08
26	20.4	−5.3	−3.3	28.3	−127.58
27	9.6	−16.1	−9.9	260.3	290.82
28	33.3	7.6	4.7	58.5	71.88
29	14.0	−11.7	−7.2	137.0	327.96
30	44.6	19.0	11.7	359.3	873.81
31	33.2	7.5	4.6	56.3	322.09
32	20.3	−5.3	−3.3	28.6	158.70
33	26.3	0.6	0.4	0.4	−2.85
			$\sum_{1}^{33}(z_i-\overline{z})^2=3167.58$		$\sum_{i=1}^{33}\sum_{j=1}^{33}w_{i,j}(z_i-\overline{z})(z-\overline{z})=4805.35$
Moran's I			$I=\frac{1}{p}\frac{\sum_{I=1}^{n}\sum_{j=1 j\neq i}^{n}w_{i,j}(z_i-\overline{z})(z_j-\overline{z})}{\sum_{i=1}^{n}(z_i-\overline{z})^2}$		$\frac{33(4805.35)}{164(3167.58)}=0.305$
Z test statistic of Moran's I			$Z=(i_O-\mu_e)/\sigma_e$		$(0.305-(-0.0297))/0.1045=2.63$
Probability			Randomization applied with 9999 permutations $p=0.004$		

reference distribution, for example 99 and 999 are associated with the 0.01 and 0.001 significance levels, respectively.

Adjacency and contiguity are clearly important concepts in understanding the principles underlying Join Counts Statistics, the Global Moran's *I* index and other similar measures of spatial autocorrelation. Another important concept is distance between features, both points and areas, with the centroid usually marking the location in the latter case. Features that are located further apart may be expected to exert less influence on each other in respect of contributing towards spatial autocorrelation compared with those that are closer together. This presumption leads to the use of inverse distance weighting as a way of taking into account the diminishing or decaying effect of distance on the data values of features that are further apart. Spatial weighting methods are based on contiguity (e.g. rook's and queen's adjacency for polygons) or distance using polygons' centroids or user-defined *X*, *Y* coordinate pairs. Other methods of weighting data values focus on each point (centroid) location and then average over a prespecified number of nearest neighbours. The outcome of spatial weighting is a spatially lagged variable that is an essential requirement for testing autocorrelation and carrying out spatial regression. The application of Moran's *I* in Box 10.4 used a contiguity weight (i.e. adjacent boundaries) in respect of irregular polygons.

A useful way of visualizing the extent of spatial autocorrelation is by means of a Moran's *I* scatter plot (MSP). Figure 10.5 plots the standardized and spatially lagged values of the 2007 IMD for the 33 London Boroughs. The MSP is divided into four segments centred on the mean of the two variables that can be summarized as follows:

- Low–Low: spatial units where standardized and lagged values are low;
- Low–High: spatial units where standardized value is low and lagged value is high;
- High–Low: spatial units where standardized value is high and lagged value is low;
- High–High: spatial units where lagged and standardized values are high.

In the Low–Low quadrant of Figure 10.5 the point (borough) with lowest combination of values (−1.24 for standardized IMD and −1.08 for lagged IMD) is coincidentally Kingston upon Thames. The map in Box 10.4a shows this borough to have an IMD value of 13.1 and its four contiguous neighbours (Richmond upon Thames (9.6), Merton (14.6), Sutton (14.0) and Wandsworth (20.3)) also have comparatively low values. Although Richmond upon Thames has the lowest IMD value of all the boroughs, two of its neighbours have comparatively high values (Hounslow at 23.2 and

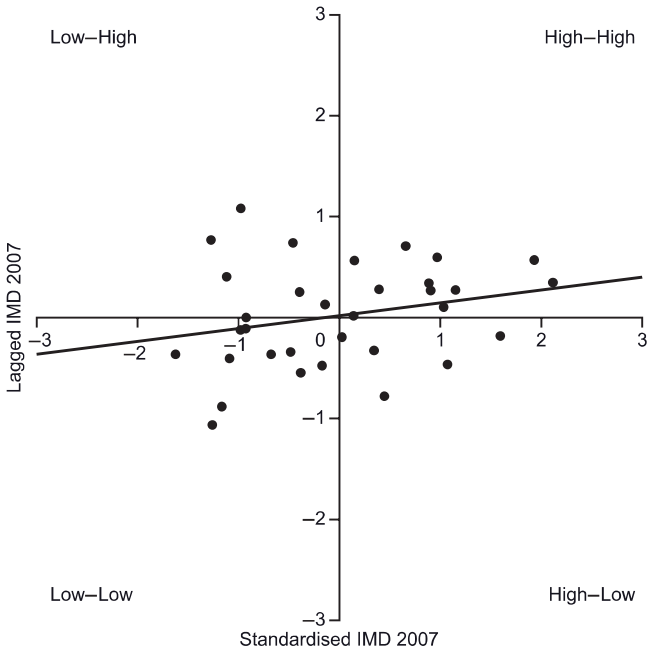


Figure 10.5 Moran's *I* scatter plot.

Hammersmith and Fulham at 28.1 and the influence of these outweighs its contiguity with Kingston upon Thames and Wandsworth. At the other end of the scale in the High-High quadrant are Newham and Tower Hamlets, respectively, with standardized and spatially lagged IMD values of 1.92 and 0.59, and 2.10 and 0.37.

What are the 'raw' IMD values of Newham's 6 adjacent boroughs and what are the values for the 6 contiguous boroughs of Tower Hamlets? Note: the ID numbers of Newham and Tower Hamlets are 25 and 30, respectively.

10.2.2 Local Indicators of Spatial Association (LISA)

Moran's *I* index is a global measure of the extent of spatial autocorrelation across a study area and quantifies the degree to which features that are spatially proximate have similar values. However, it is entirely possible for local 'pockets' of positive and negative spatial autocorrelation to exist that at least partially cancel each other out and lead to a deflated index in comparison with what might have been obtained had the study area been divided into subareas and separate indices computed for these. Such variability in the distribution of spatial autocorrelation can be quantified

by **Local Indicators of Spatial Association** (**LISA**), these focus on the extent to which features that are close to a specific point have similar values. The last part of the covariance/variance computation tabulation in Box 10.4c showed the row sums (S) obtained by summing the nondiagonal elements that were shaded dark grey (i.e. the spatial features are joined). If each of these S values is standardized by dividing by the sum of the squared deviations along the diagonal elements (the variance-like quantity) and the results are multiplied by the total number of spatial units (n), the figures obtained represent the local contribution of each row (feature) to the global spatial autocorrelation. These provide another way of computing the overall Moran's I index, since the sum of these local components divided by the total number of joins equals the global I computed from the definitional equation given in Box 10.4a:

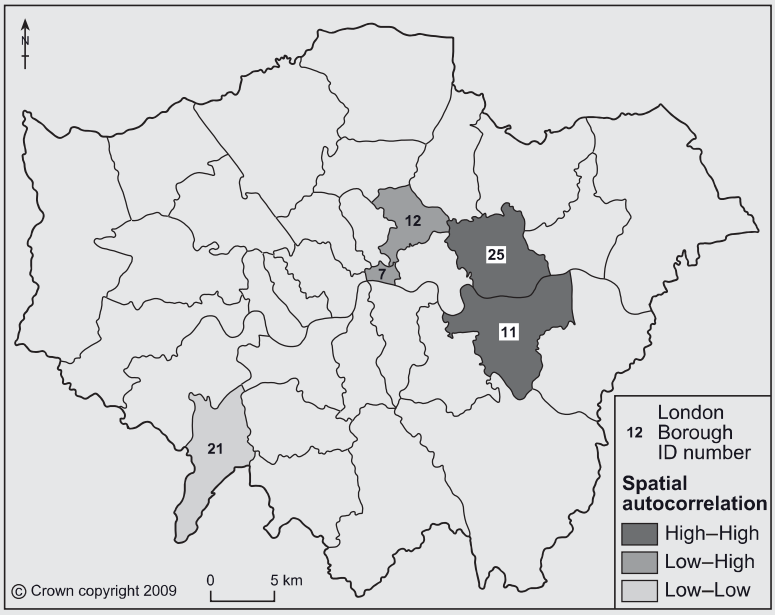
$$I = 50.06/164 = 0.305$$

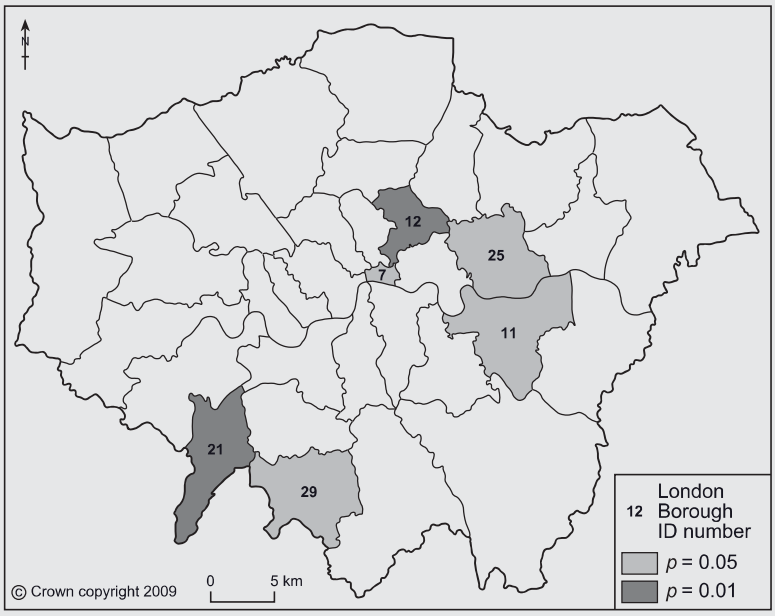
These local components are the LISA values, which can be mapped and tested for significant differences between areas. Rather than using these 'raw' LISA values they are often standardized by dividing by the number of joins possessed by each feature (row): thus the raw LISA values of a feature with four joins is divided by 4. When this has been done for each row in the matrix the results are in the form of a local average. Again, the row sums (S) can be divided by the sum of the squared deviations along the diagonal and then multiplied by $n - 1$ rather than n to produced standardized LISA values.

The calculation of these LISA values is shown in Box 10.5 with respect to the 2007 Index of Multiple Deprivation mean score for the 33 London Boroughs. The effect of these adjustments is to increase global Moran's I to 0.331 (0.305 previously) and when the City of London area is excluded (see above for justification) Moran's I index becomes 0.663 compared with 0.479. The results of LISA calculation can be visualized in a number of ways. Box 10.5a shows the significance and cluster maps relating to the application of Moran's I LISA for the 33 London Boroughs. Each LISA is tested for significance using a randomization process to generate a reference distribution and areas are shaded according to whether their LISA is significantly different from what would be expected at the 0.05, 0.01 and 0.001 levels. The cluster maps include those areas that are significant in the four quadrants of the MSP. Taken together these maps allow the significant combinations of positive and negative local spatial autocorrelation to be discovered. The London Boroughs used in the analyses included in Boxes 10.4 and 10.5 are in some respects rather large and have been used in order to keep the complexity of the calculations to a manageable scale. The analysis would more appropriately be carried out for smaller spatial units such as local authority wards across the whole of London. Arguably, the scale of these units is better suited to reflecting local variations that can easily become lost for relatively large areas. Nevertheless, even at the borough scale there is some evidence of local spatial autocorrelation in certain parts of the Greater London Authority's area.

Box 10.5a: Moran's *I* local indicator of spatial association.

Local Moran's I Indicator of Spatial Association: $I_i = \dfrac{\sum_{j=1}^{n} w_j(z_i - \bar{z})(z_j - \bar{z})}{s_z^2 \sum_{j=1}^{n} w_{ij}}$





Box 10.5b: Application of the Moran's *I* local indicator of spatial association.

The Moran's *I* Local Indicator of Spatial Association focuses attention on the extent of spatial autocorrelation around individual points, including centroids as point locations for polygons, rather than providing a global summary measure. The calculations produce Moran's *I* LISA values for each spatial feature that can be tested for significance using a randomization process to generate a reference distribution. The results of the analysis are shown as two maps: the cluster map that shows combinations of high and low indices; and the significance map that shows whether a local index value is significant at certain levels. The results support the earlier global Moran's I with a pair of contiguous boroughs in South-West London (Kingston upon Thames and Sutton) having Low–Low spatial autocorrelation and Greenwich and Newham east of the centre with the High–High combination. There are two boroughs, the City of London and Hackney, with the Low–High combination. The probability of the index values associated with all of these areas is ≤0.03, which is smaller than the 'standard' 0.05 significance and therefore it is reasonable to conclude that there is significant local spatial autocorrelation in these parts of London.

10.3 Trend surface analysis

The most straightforward approach to introducing **trend surface analysis** is to recognize that it comprises a special form of regression. What makes it special is the inclusion of spatial location in terms of *X*, *Y* coordinates as independent variables recording the perpendicular spatial dimensions on a regular square grid. The purpose of the analysis is to define a surface that best fits these locations in space where the third dimension or the dependent variable, usually represented by the letter *Z*, denotes the height of the surface. The origins of the analysis lie in representing the physical surface of the Earth in which case the third dimension is elevation above a fixed datum level. However, surfaces can be in principle produced for any dependent variable whose values can be theoretically conceived as varying and thus producing a surface in space. For example, land values might be expected to vary across space with peaks and troughs occurring at different places. Similarly, physical variables such as measurements of pressure and water vapour in the troposphere, the lowest layer in the Earth's atmosphere, differ in a similar way. It is the relative concentration of different values, such as high or low amounts of water vapour that creates 'spatial features' in the atmosphere (e.g. clouds).

The starting point is a horizontal regular square grid in two dimensions (*X* and *Y*) and the end result is estimated values for the dependent or *Z* variable for the set of evenly spaced points on the grid. Connecting these estimated values together produces a visualization of the surface. There are two main approaches to deriving a trend surface known as global and local fit and these are connected with the difference between quantifying global and local spatial autocorrelation examined previously.

Box 10.5c: Calculation of local indicator of spatial association and significance testing.

Calculation of variance/covariance-like quantities.

		j	1	2	3	4	5	6	7	8
i	z	$z-\bar{z}$	8.8	−4.5	−9.5	3.5	−11.3	2.9	−12.8	−4.4
1	34.5	8.8	77.5	0.0	−16.7	0.0	0.0	0.0	0.0	0.0
2	21.2	−4.5	0.0	20.5	0.0	−3.2	0.0	−2.7	0.0	0.0
3	16.2	−9.5	−20.9	0.0	89.8	0.0	26.8	0.0	0.0	0.0
4	29.2	3.5	0.0	−2.3	0.0	12.5	0.0	1.5	0.0	0.0
5	14.4	−11.3	0.0	0.0	17.9	0.0	128.2	0.0	0.0	8.3
6	28.6	2.9	0.0	−2.2	0.0	1.7	0.0	8.6	−6.3	0.0
7	12.8	−12.8	0.0	0.0	0.0	0.0	0.0	−5.4	165.0	0.0
8	21.3	−4.4	0.0	0.0	0.0	0.0	12.4	0.0	0.0	19.1
9	25.1	−0.6	0.0	0.0	0.0	−0.4	0.0	0.0	0.0	0.0
10	26.2	0.5	0.0	−0.8	0.0	0.0	0.0	0.0	0.0	0.0
11	33.9	8.3	12.1	0.0	−13.0	0.0	−15.6	0.0	0.0	0.0
12	46.1	20.4	0.0	0.0	0.0	0.0	0.0	0.0	−43.7	0.0
13	28.1	2.4	0.0	0.0	0.0	1.4	0.0	0.0	0.0	0.0
14	35.7	10.0	0.0	−7.6	0.0	0.0	0.0	4.9	0.0	0.0
15	15.6	−10.1	0.0	11.4	0.0	−8.9	0.0	0.0	0.0	0.0
16	16.1	−9.6	−28.2	0.0	30.4	0.0	0.0	0.0	0.0	0.0
17	18.6	−7.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18	23.2	−2.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19	39.0	13.3	0.0	0.0	0.0	0.0	0.0	9.7	−42.6	0.0
20	23.5	−2.2	0.0	0.0	0.0	−1.9	0.0	0.0	0.0	0.0
21	13.1	−12.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22	34.9	9.3	0.0	0.0	0.0	0.0	−15.0	0.0	−17.0	−5.8
23	31.0	5.4	0.0	0.0	0.0	0.0	−15.2	0.0	0.0	0.0
24	14.6	−11.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	9.7
25	43.0	17.3	25.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
26	20.4	−5.3	−11.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
27	9.6	−16.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
28	33.3	7.6	0.0	0.0	0.0	0.0	−17.3	0.0	−19.6	0.0

		j	1	2	3	4	5	6	7	8
i	z	$z-\bar{z}$	8.8	−4.5	−9.5	3.5	−11.3	2.9	−12.8	−4.4
29	14.0	−11.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	17.1
30	44.6	19.0	0.0	0.0	0.0	0.0	0.0	0.0	−40.6	0.0
31	33.2	7.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
32	20.3	−5.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
33	26.3	0.6	0.0	0.0	0.0	0.4	0.0	0.3	−1.3	0.0
		j	9	10	11	12	13	14	15	16
i	z	$z-\bar{z}$	−0.6	0.5	8.3	20.4	2.4	10.0	−10.1	−9.6
1	34.5	8.8	0.0	0.0	14.5	0.0	0.0	0.0	0.0	−16.9
2	21.2	−4.5	0.0	−0.5	0.0	0.0	0.0	−9.1	9.1	0.0
3	16.2	−9.5	0.0	0.0	−19.6	0.0	0.0	0.0	0.0	22.8
4	29.2	3.5	−0.3	0.0	0.0	0.0	1.2	0.0	−5.1	0.0
5	14.4	−11.3	0.0	0.0	−15.6	0.0	0.0	0.0	0.0	0.0
6	28.6	2.9	0.0	0.0	0.0	0.0	0.0	4.9	0.0	0.0
7	12.8	−12.8	0.0	0.0	0.0	−37.5	0.0	0.0	0.0	0.0
8	21.3	−4.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	25.1	−0.6	0.3	0.0	0.0	0.0	−0.3	0.0	1.2	0.0
10	26.2	0.5	0.0	0.3	0.0	0.0	0.0	1.7	0.0	0.0
11	33.9	8.3	0.0	0.0	68.2	0.0	0.0	0.0	0.0	0.0
12	46.1	20.4	0.0	0.0	0.0	416.8	0.0	34.2	0.0	0.0
13	28.1	2.4	−0.2	0.0	0.0	0.0	5.7	0.0	0.0	0.0
14	35.7	10.0	0.0	0.8	0.0	34.2	0.0	100.9	0.0	0.0
15	15.6	−10.1	1.5	0.0	0.0	0.0	0.0	0.0	101.9	0,0
16	16.1	−9.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	92.4
17	18.6	−7.1	1.4	0.0	0.0	0.0	0.0	0.0	24.0	0.0
18	23.2	−2.5	0.4	0.0	0.0	0.0	−1.5	0.0	0.0	0.0
19	39.0	13.3	0.0	0.0	0.0	67.8	0.0	33.3	0.0	0.0
20	23.5	−2.2	0.0	0.0	0.0	0.0	−1.3	0.0	0.0	0.0
21	13.1	−12.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22	34.9	9.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
23	31.0	5.4	0.0	0.0	11.1	0.0	0.0	0.0	0.0	0.0
24	14.6	−11.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

25	43.0	17.3	0.0	0.0	23.8	58.8	0.0	0.0	0.0	0.0
26	20.4	−5.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	12.8
27	9.6	−16.1	0.0	0.0	0.0	0.0	−9.6	0.0	0.0	0.0
28	33.3	7.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
29	14.0	−11.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
30	44.6	19.0	0.0	0.0	26.1	64.5	0.0	0.0	0.0	0.0
31	33.2	7.5	0.0	0.8	0.0	30.6	0.0	15.1	0.0	0.0
32	20.3	−5.3	0.0	0.0	0.0	0.0	−1.8	0.0	0.0	0.0
33	26.3	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

		j	17	18	19	20	21	22	23	24
i	z	$z-\overline{z}$	−7.1	−2.5	13.3	−2.2	−12.6	9.3	5.4	−11.1
1	34.5	8.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	21.2	−4.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	16.2	−9.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	29.2	3.5	0.0	0.0	0.0	−1.1	0.0	0.0	0.0	0.0
5	14.4	−11.3	0.0	0.0	0.0	0.0	0.0	−17.5	−10.1	0.0
6	28.6	2.9	0.0	0.0	6.5	0.0	0.0	0.0	0.0	0.0
7	12.8	−12.8	0.0	0.0	−24.4	0.0	0.0	−17.0	0.0	0.0
8	21.3	−4.4	0.0	0.0	0.0	0.0	0.0	−10.1	0.0	12.1
9	25.1	−0.6	0.8	0.3	0.0	0.0	0.0	0.0	0.0	0.0
10	26.2	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	33.9	8.3	0.0	0.0	0.0	0.0	0.0	0.0	7.4	0.0
12	46.1	20.4	0.0	0.0	45.2	0.0	0.0	0.0	0.0	0.0
13	28.1	2.4	0.0	−1.0	0.0	−0.9	0.0	0.0	0.0	0.0
14	35.7	10.0	0.0	0.0	22.2	0.0	0.0	0.0	0.0	0.0
15	15.6	−10.1	18.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16	16.1	−9.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
17	18.6	−7.1	50.7	4.4	0.0	0.0	0.0	0.0	0.0	0.0
18	23.2	−2.5	4.4	6.2	0.0	0.0	0.0	0.0	0.0	0.0
19	39.0	13.3	0.0	0.0	176.3	0.0	0.0	0.0	0.0	0.0
20	23.5	−2.2	0.0	0.0	0.0	4.7	0.0	0.0	0.0	0.0
21	13.1	−12.6	0.0	0.0	0.0	0.0	158.4	0.0	0.0	34.8

		j	1	2	3	4	5	6	7	8
i	z	$z-\bar{z}$	8.8	−4.5	−9.5	3.5	−11.3	2.9	−12.8	−4.4
22	34.9	9.3	0.0	0.0	0.0	0.0	0.0	85.7	0.0	−14.6
23	31.0	5.4	0.0	0.0	0.0	0.0	0.0	0.0	28.7	0.0
24	14.6	−11.1	0.0	0.0	0.0	0.0	27.8	−20.5	0.0	122.4
25	43.0	17.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
26	20.4	−5.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
27	9.6	−16.1	0.0	10.0	0.0	0.0	50.8	0.0	0.0	0.0
28	33.3	7.6	0.0	0.0	0.0	0.0	0.0	14.2	8.2	0.0
29	14.0	−11.7	0.0	0.0	0.0	0.0	49.1	0.0	0.0	43.2
30	44.6	19.0	0.0	0.0	0.0	0.0	0.0	0.0	16.9	0.0
31	33.2	7.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
32	20.3	−5.3	0.0	0.0	0.0	1.7	9.6	−7.1	0.0	8.4
33	26.3	0.6	0.0	0.0	0.0	−0.2	0.0	1.0	0.0	0.0
		j	25	26	27	28	29	30	31	32
i	z	$z-\bar{z}$	17.3	−5.3	−16.1	7.6	−11.7	19.0	7.5	−5.3
1	34.5	8.8	30.4	−9.4	0.0	0.0	0.0	0.0	0.0	0.0
2	21.2	−4.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	16.2	−9.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	29.2	3.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	14.4	−11.3	0.0	0.0	0.0	−14.4	0.0	0.0	0.0	0.0
6	28.6	2.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	12.8	−12.8	0.0	0.0	0.0	−14.0	0.0	−34.8	0.0	0.0
8	21.3	−4.4	0.0	0.0	0.0	0.0	12.8	0.0	0.0	0.0
9	25.1	−0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	26.2	0.5	0.0	0.0	0.0	0.0	0.0	0.0	1.3	0.0
11	33.9	8.3	23.8	0.0	0.0	0.0	0.0	26.1	0.0	0.0
12	46.1	20.4	58.8	0.0	0.0	0.0	0.0	64.5	25.5	0.0
13	28.1	2.4	0.0	0.0	−6.4	0.0	0.0	0.0	0.0	−2.1

14	35.7	10.0	0.0	0.0	0.0	0.0	0.0	0.0	12.6	0.0
15	15.6	−10.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16	16.1	−9.6	0.0	17.1	0.0	0.0	0.0	0.0	0.0	0.0
17	18.6	−7.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
18	23.2	−2.5	0.0	0.0	10.0	0.0	0.0	0.0	0.0	0.0
19	39.0	13.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20	23.5	−2.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.9
21	13.1	−12.6	0.0	0.0	50.8	0.0	36.8	0.0	0.0	16.8
22	34.9	9.3	0.0	0.0	0.0	10.1	0.0	0.0	0.0	−7.1
23	31.0	5.4	0.0	0.0	0.0	10.2	0.0	25.4	0.0	0.0
24	14.6	−11.1	0.0	0.0	0.0	0.0	25.9	0.0	0.0	11.8
25	43.0	17.3	298.1	−15.3	0.0	0.0	0.0	54.5	21.6	0.0
26	20.4	−5.3	−23.0	28.3	0.0	0.0	0.0	0.0	−10.0	0.0
27	9.6	−16.1	0.0	0.0	260.3	0.0	0.0	0.0	0.0	21.6
28	33.3	7.6	0.0	0.0	0.0	58.5	0.0	29.0	0.0	0.0
29	14.0	−11.7	0.0	0.0	0.0	0.0	137.0	0.0	0.0	0.0
30	44.6	19.0	54.5	0.0	0.0	24.2	0.0	359.3	0.0	0.0
31	33.2	7.5	25.9	−8.0	0.0	0.0	0.0	0.0	56.3	0.0
32	20.3	−5.3	0.0	0.0	12.3	0.0	0.0	0.0	0.0	28.6
33	26.3	0.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	−0.5

		j	33	Row sum	$\frac{S}{\sum(z_1-\bar{z})^2}$	$\frac{S}{\sum(z_1-\bar{z})^2}n-1$
i	z	$z-\bar{z}$	0.6	(S)		
1	34.5	8.8	0.0	1.96	0.001	0.02
2	21.2	−4.5	0.0	−6.27	−0.002	−0.06
3	16.2	−9.5	0.0	9.18	0.003	0.09
4	29.2	3.5	0.3	−5.78	−0.002	−0.06
5	14.4	−11.3	0.0	−31.46	−0.010	−0.32
6	28.6	2.9	0.3	4.95	0.002	0.05
7	12.8	−12.8	−1.1	−134.13	−0.042	−1.36
8	21.3	−4.4	0.0	27.16	0.009	0.27

		j	1	2	3	4	5	6	7	8
i	z	$z-\overline{z}$	8.8	−4.5	−9.5	3.5	−11.3	2.9	−12.8	−4.4
9	25.1	−0.6	0.0	1.61	0.001			0.02		
10	26.2	0.5	0.0	2.20	0.001			0.02		
11	33.9	8.3	0.0	40.71	0.013			0.41		
12	46.1	20.4	0.0	184.45	0.058			1.86		
13	28.1	2.4	0.0	−9.22	−0.003			−0.09		
14	35.7	10.0	0.0	67.17	0.021			0.68		
15	15.6	−10.1	0.0	21.94	0.007			0.22		
16	16.1	−9.6	0.0	19.20	0.006			0.19		
17	18.6	−7.1	0.0	31.25	0.010			0.32		
18	23.2	−2.5	0.0	13.32	0.004			0.13		
19	39.0	13.3	0.0	68.22	0.022			0.69		
20	23.5	−2.2	−0.3	−0.65	0.000			−0.01		
21	13.1	−12.6	0.0	139.20	0.044			1.41		
22	34.9	9.3	0.8	−48.51	−0.015			−0.49		
23	31.0	5.4	0.0	31.51	0.010			0.32		
24	14.6	−11.1	0.0	54.77	0.017			0.55		
25	43.0	17.3	0.0	168.68	0.053			1.70		
26	20.4	−5.3	0.0	−31.90	−0.010			−0.32		
27	9.6	−16.1	0.0	72.71	0.023			0.73		
28	33.3	7.6	0.0	14.38	0.005			0.15		
29	14.0	−11.7	0.0	109.32	0.035			1.10		
30	44.6	19.0	0.0	145.64	0.046			1.47		
31	33.2	7.5	0.0	64.42	0.020			0.65		
32	20.3	−5.3	−0.5	22.67	0.007			0.23		
33	26.3	0.6	0.4	−0.48	0.000			0.00		
				1048.21	0.331			10.589		

10.3.1 Fitting a global surface

Global fitting produces one mathematical function that describes the entire surface of the study area on the basis of estimating Z values for the nodes on the grid in a single operation, whereas local fitting derives multiple equations based on using a subset of points around successive individual nodes in the grid. The regression analysis techniques examined in Chapter 9 produce equations that define the line providing the best fit to the known data values, although it is acknowledged that these are unlikely to provide a 100 per cent accurate prediction of the dependent variable. The same caveat applies with trend surface analysis and Figure 10.6 illustrates the link between prediction in linear and polynomial regression and trend surface analysis. The regression lines may be thought of as transecting the surface along a particular trajectory. There are known data values above and below the first-, second- and third-degree polynomial regression lines on the left of Figure 10.6, but the lines show the overall slope. The three dimensional representations on the right reveal that the known data values fall above or below (respectively solid and open circles) the surface.

The fitting of a global trend surface is achieved through a form of polynomial regression using the least squares method to minimize the sum of the squared deviations from the surface at the known, sampled locations (control points). The equation can be used to estimate the values of the dependent variable at any point and this is commonly carried out for the nodes in the regular grid. The polynomial equation obtained by means of least squares fitting provides the best approximation of the surface from the available data for the control points. The process of creating a surface from the equation is commonly known as **interpolation**, since it involves determining or interpolating previously unknown Z values for the nodes and connecting these together usually by means of a 'wire frame' in order to visualize the surface, although strictly speaking this term should be reserved for dealing with smoothing local variation. Figure 10.7 illustrates the outcome of this process with respect to an irregular sample of data points for elevation on the South Downs in South-East England. The points from which the polynomial equation for the surface was generated are identified by white markings at the peaks.

Despite the general use of polynomial regression-type equations to fit a global surface, there are some limitations to this approach. These are in some respects extensions of the same problems that were identified with respect to using regression to predict nonspatial distributed dependent variables. One of the limitations noted with respect to simple linear regression (first-order polynomial) is that it may be inappropriate to define the relationship as a straight line, especially if visualization of the empirical data suggested some curvilinear connection. It would be just as nonsensical to argue that all surfaces are flat sloping planes. However, moving to a surface based on the second-order polynomial only introduces one maximum or minimum location on the surface and the third order only provides for one peak and one trough. Thus, global fitting does not necessarily produce a very realistic surface. Another problem identified with regression in statistical analysis was that prediction of the

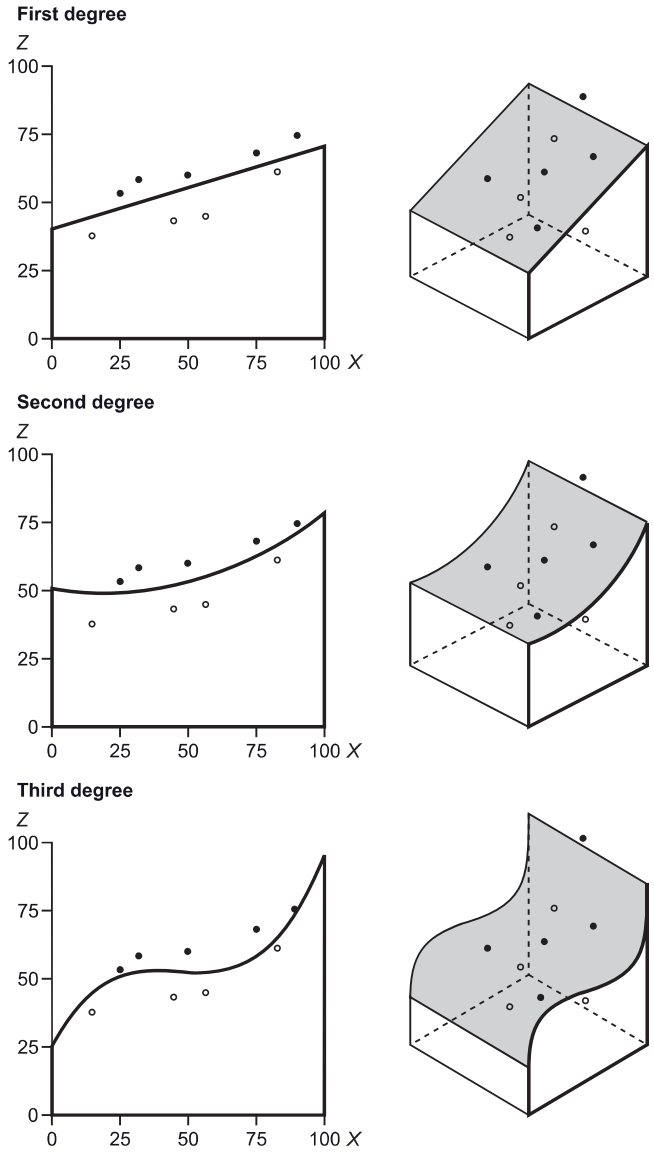


Figure 10.6 Comparison of regression lines and surfaces.

dependent variable much beyond the range of the known values of the independent variable(s) is potentially inaccurate. Similarly, the extension of a trend surface to beyond the area for which there are known data points could also be misleading. Such gaps in the observed or measured data can occur around the edges or in parts of a study area where there is a dearth of control points. More realistic surfaces may be obtained by increasing the order of the polynomial equation beyond three to four, five, six or more, although the calculations involved are computationally taxing.

Figure 10.7 Trend surface for part of South Downs, East Sussex, England.

Box 10.6 illustrates selected aspects of the calculations involved in producing a first-order polynomial trend surface (i.e. a linear surface) in respect of the 2007 Index of Multiple Deprivation for the 33 London Boroughs. The measurement of spatial autocorrelation with Moran's *I* has already shown this to be high in some boroughs just to the east of the City of London and low in others towards the south west. The equation has been computed using geostatistical software and the predicted *Z* values and the residuals are tabulated in Box 10.6c. The percentile maps (Box 10.6a) divide the predicted figures for the linear surface and the residuals into six groups and emphasize the importance of very low and very high values. The linear surface tracks north east to south west and simplifies the spatial pattern. The residuals reveal the highest positive difference was in Hackney and three of its neighbours (Newham, Lambeth and Tower Hamlets) in the central area. The largest negative residual was in Havering on the eastern edge with the next two being the City of London and Redbridge.

10.3.2 Dealing with local variation in a surface

The global fitted trend surface relates to the concept of regional features. Returning to the example of creating a surface using atmospheric pressure and water vapour to identify features in the troposphere, the cyclones or anticyclones and clouds can be viewed as regional or large-scale features. However, at a smaller scale there will often be local variation in the variables producing highs and lows in the overall feature. Fitting the polynomial equation results in a surface, whereas the residuals, the differences between the fitted surface and known values may be thought of as local

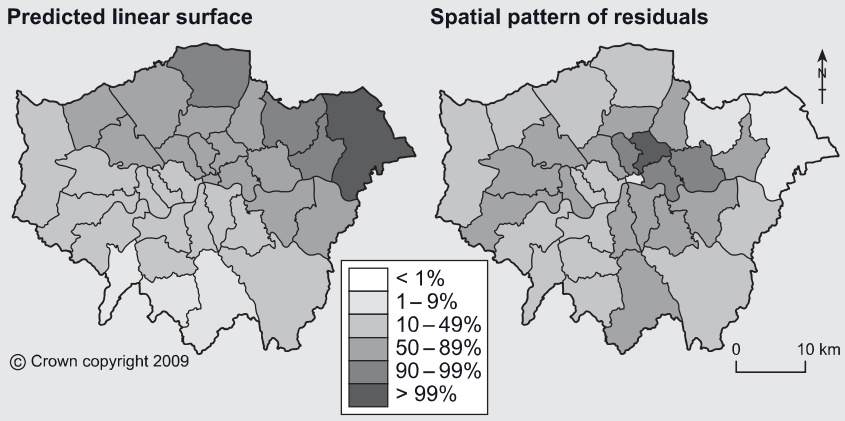
Box 10.6a: Trend surface analysis.

Polynomial equation for fitting a trend surface: $\hat{Z}(x, y) = \sum_{i=1}^{M} a_i x^{\beta 1 i} y^{\beta 2 i}$

Minimization of error of estimate: $E = \sum_{i=1}^{L} (Z(x_i, y_i) - Z_i)^2 \rightarrow \text{minimum}$

Linear trend (flat dipping plane): $\hat{Z}(x, y) = a_0 + a_1 x + a_2 y$

Quadratic trend (one maximum or minimum): $\hat{Z}(x, y) = a_0 + a_1 x + a_2 y + a_3 x^2 + a_4 xy + a_5 y^2$



Box 10.6b: Application of the trend surface analysis.

The terms x and y are coordinates on the plane surface and M is the number of degrees or orders of the polynomial equation. The series of coefficients (a_0, a_1, … a_i) are obtained by minimizing the error of the estimation where L is the number of sample or control points. A linear trend surface has been produced using spatial analysis software (Geoda) for the London Boroughs in respect of their average score on the 2007 IMD using rook spatial contiguity weights. The percentile maps in Box 10.6a show the predicted surface and the residuals emphasizing the extreme cases. The specific 1st-order polynomial equation that best fits the empirical data is shown in Box 10.6c and this has been used to calculate the predicted values of the IMD average score for the 500 m grid squares covering the area (Box 10.6d). This simplifies the spatial pattern and provides a clear visualization of the north east to south west trending surface.

Box 10.6c: Calculation of linear trend surface.

i	X	Y	Z	$\hat{Z}$	e
1	547 980	186 083	34.5	30.50	−3.99
2	524 392	191 070	21.2	27.61	6.45
3	548 572	175 109	16.2	26.58	10.37
4	520 575	185 876	29.2	24.93	−4.29
5	542 070	166 316	14.4	22.04	7.68
6	527 756	184 528	28.6	25.87	−2.75
7	532 573	181 257	12.8	25.63	12.79
8	533 193	164 481	21.3	19.59	−1.72
9	516 188	181 612	25.1	22.48	−2.62
10	532 197	195 153	26.2	30.67	4.48
11	542 484	175 829	33.9	25.62	−8.32
12	533 820	185 439	46.1	27.42	−18.68
13	523 067	179 688	28.1	23.15	−4.92
14	531 141	189 534	35.7	28.39	−7.34
15	515 258	189 084	15.6	25.05	9.46
16	553 902	186 883	16.1	31.98	15.91
17	507 523	183 686	18.6	21.51	2.95
18	514 876	176 035	23.2	20.17	−3.03
19	531 241	185 097	39.0	26.78	−12.18
20	525 389	180 022	23.5	23.74	0.23
21	519 598	167 257	13.1	17.89	4.79
22	530 881	173 991	34.9	22.62	−12.32
23	537 779	174 133	31.0	24.06	−6.98
24	526 431	169 390	14.6	20.04	5.42
25	540 804	184 171	43.0	28.36	−14.59
26	543 517	189 417	20.4	30.83	10.47
27	517 275	173 187	9.6	19.60	10.05
28	533 757	176 088	33.3	23.97	−9.36
29	526 245	164 539	14.0	18.22	4.24
30	536 246	181 938	44.6	26.62	−18.02
31	538 073	189 802	33.2	29.88	−3.31
32	526 925	173 681	20.3	21.72	1.38
33	526 765	181 473	26.3	24.55	−1.75
First-order polynomial equation (linear surface)		$\hat{Z}(x, y) = a_0 + a_1 x + a_2 y$		147.72 + 0.00020x + 0.00037y	

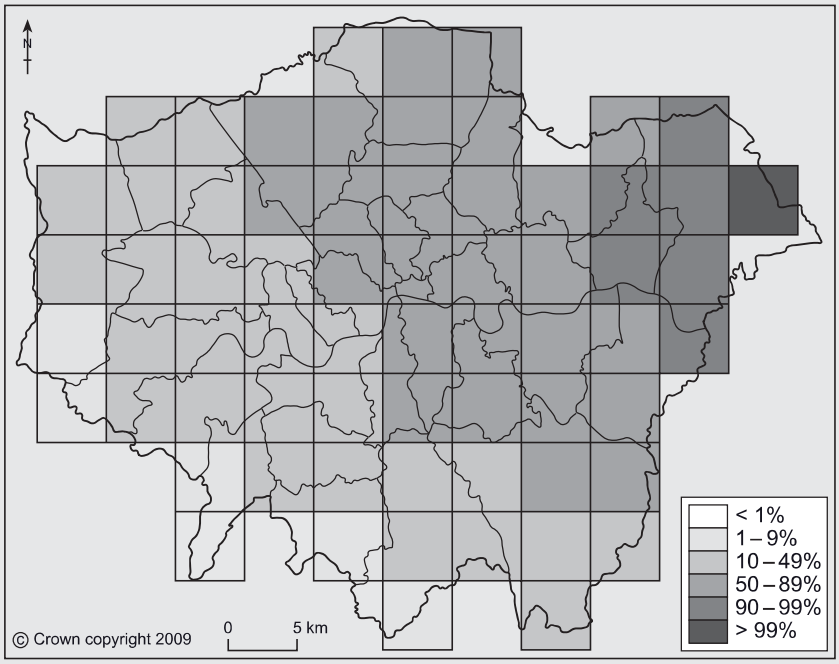
Box 10.6d: Calculation of linear trend surface for grid squares.

1	527 500	197 500	32.76	34	512 500	182 500	22.77
2	532 500	202 500	36.09	35	517 500	187 500	26.10
3	537 500	207 500	39.42	36	522 500	192 500	29.43
4	512 500	192 500	27.20	37	527 500	197 500	32.76
5	517 500	197 500	30.53	38	532 500	202 500	36.09
6	522 500	202 500	33.86	39	537 500	207 500	39.42
7	527 500	207 500	37.19	40	542 500	212 500	42.75
8	532 500	212 500	40.52	41	547 500	217 500	46.08
9	537 500	217 500	43.85	42	552 500	222 500	49.41
10	542 500	222 500	47.18	43	507 500	172 500	17.23
11	547 500	227 500	50.51	44	512 500	177 500	20.56
12	507 500	187 500	23.87	45	517 500	182 500	23.89
13	512 500	192 500	27.20	46	522 500	187 500	27.22
14	517 500	197 500	30.53	47	527 500	192 500	30.55
15	522 500	202 500	33.86	48	532 500	197 500	33.88
16	527 500	207 500	37.19	49	537 500	202 500	37.21
17	532 500	212 500	40.52	50	542 500	207 500	40.54
18	537 500	217 500	43.85	51	547 500	212 500	43.87
19	542 500	222 500	47.18	52	517 500	167 500	17.24
20	547 500	227 500	50.51	53	522 500	172 500	20.57
21	552 500	232 500	53.84	54	527 500	177 500	23.90
22	557 500	237 500	57.17	55	532 500	182 500	27.23
23	507 500	182 500	21.66	56	537 500	187 500	30.56
24	512 500	187 500	24.99	57	542 500	192 500	33.89
25	517 500	192 500	28.32	58	547 500	197 500	37.22
26	522 500	197 500	31.65	59	517 500	162 500	15.02
27	527 500	202 500	34.98	60	522 500	167 500	18.35
28	532 500	207 500	38.31	61	527 500	172 500	21.68
29	537 500	212 500	41.64	62	532 500	177 500	25.01
30	542 500	217 500	44.97	63	537 500	182 500	28.34
31	547 500	222 500	48.30	64	542 500	187 500	31.67
32	552 500	227 500	51.63	65	532 500	157 500	16.15
33	507 500	177 500	19.44	66	537 500	162 500	19.48

disturbances. These can be dealt with by a range of techniques that focus on groups of data points in a 'window', frame or **kernel**. These broadly divide into exact methods that produce measured values for a series of points or areas by smoothing the original data and inexact ones that estimate a local trend and include kriging and local trend surface analysis (splines).

The simplest approach is to smooth data value means of a **moving average**, which is similar to the approach often adopted with time-series data. It involves partitioning the data points into groups falling within a frame that moves along a series of data

Box 10.6e: Grid-based linear trend surface for London Boroughs.



points and then interpolating by weighting the values usually according to the midpoint of the window. This is illustrated in Figure 10.8a, where the distances of the Burger King restaurants in Pittsburgh have been interpolated using a 5 km wide frame moving outwards from the city centre and the mean of their daily customer footfall (hypothetical) calculated for each frame (2.5, 3.5, 4.5 … 19.5 km). The averages at these points are shown by the larger solid circles. The points representing the restaurant locations (small solid circles) have been treated as though they all lie along a single, unidirectional *X*-axis, but it would have been possible to use a two-dimensional frame (e.g. a square) and then calculate a weighted average within the areas formed by successive zones moving outwards (large solid circles). The smoothing of the data achieved by a moving average is highly dependent on the size of the frame. Smoothing based on **inverse distance weighting** adjusts the value of each point in an inverse relationship to its distance from the point being estimated. The moving average in Figure 10.8a interpolated values along a one-dimensional axis from the city centre and assumed all points to be located on the eastern side of a north–south line through the centre point. In contrast inverse distance smoothing weights the values according to a predetermined number of nearest neighbours to each point being estimated. The application of inverse distance weighted interpolation to the same data values

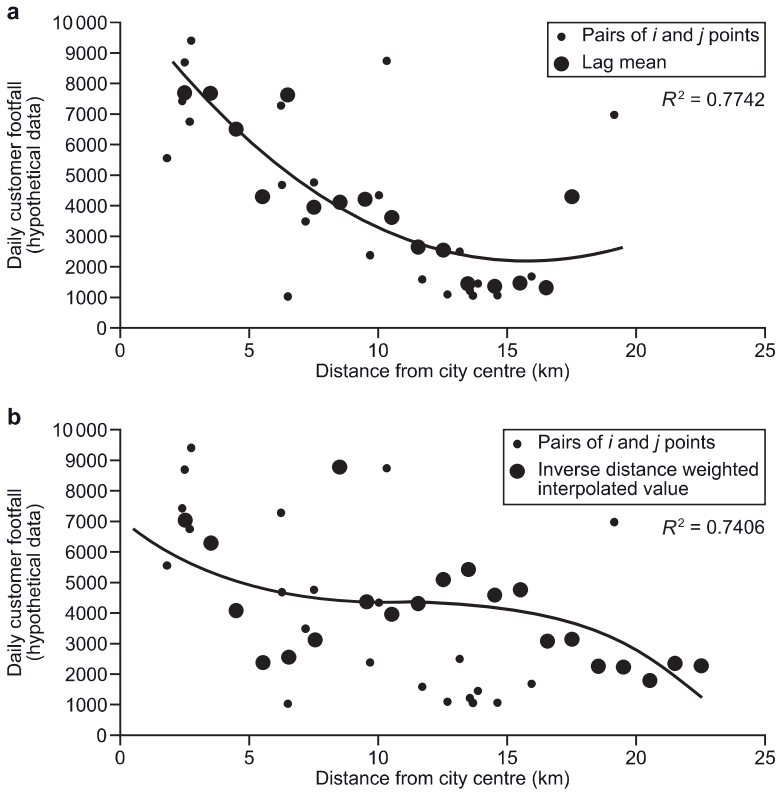


Figure 10.8 Moving average and inverse distance weighted data smoothing.

produces the interpolated values shown in Figure 10.8b (large solid circles). The inverse distance used in this example is $1/d^2$, where d is the distance between the fixed points out from the city centre (2.5, 3.5, 4.5, … 22.5 km) and their four nearest neighbours. The best-fit polynomial regression equation has been included for both sets of smoothed data: these are, respectively, 2nd- and 3rd-degree polynomials.

Another approach to dealing with local variation in a surface is to use **splines**. This involves fitting a polynomial regression equation to discrete groups of the data points along sections of the surface. These splines are then 'tied' together to produce a smooth curve following the overall surface. The points where they connect are known as knots and the polynomial equation for each section (spline) is constrained to predict the same values where they meet. Unlike the moving average, the frames in which the splines are produced do not move across the set of data points but each has a fixed location, although they can have different widths. One similarity with the moving average is that a smaller frame will reflect the local structure of the data values, whereas a wider one will produce a smoother surface.

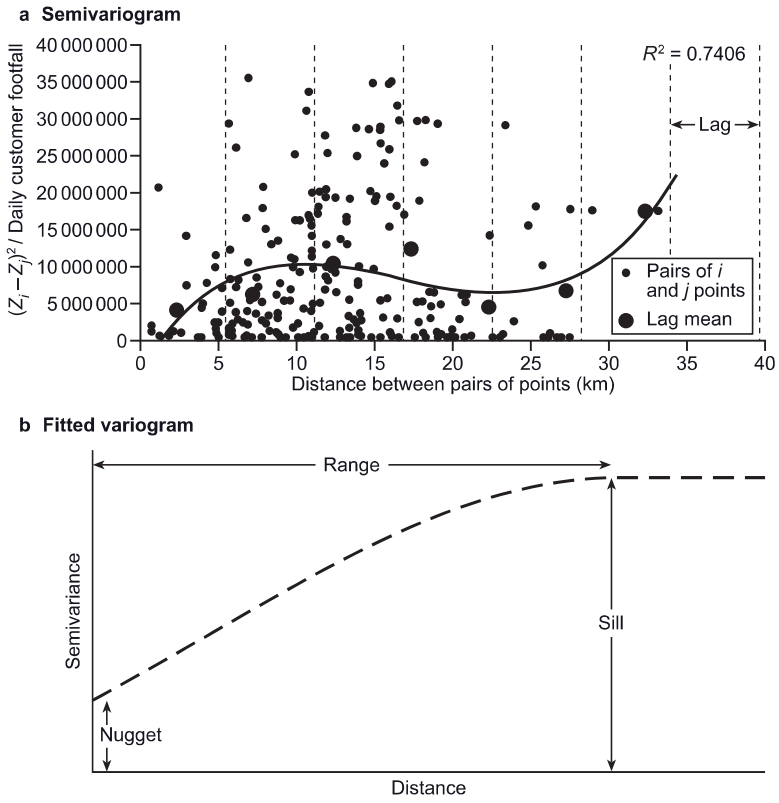


Figure 10.9 Example of semiovariogram and fitted variogram used in Kriging.

One of the most common techniques for dealing with local variation is **kriging**. Trend surface analysis may be used to identify and then remove the overall trend in a surface before using kriging to map short-range variations. Kriging is based on inverse distance weighting and uses the local spatial structure to produce weights from which to predict the values of points. The first stage involves describing the spatial structure by means of a semivariogram, which involves calculating the distance between each pair of points and the square of the difference between their values for the variable *Z*. These are visualized in a scatter plot that includes spatial lags and is known as a **semovariogram**: Figure 10.9 shows the semivariogram for the Burger King restaurants in Pittsburgh with 5 km lags and daily customer footfall as the variable. The mean of each group of data values within a given lag is plotted at the midpoint (large solid circles). The next stage involves summarizing this local spatial variation by a function that best fits the means at the midpoints, which in this example is a 3rd-degree polynomial. The values of neighbouring points are predicted by means of weights derived from the semivariogram. Kriging works by fitting an empirical semi-

variogram to a typical or model variogram. The lower part of Figure 10.9 shows such a model and identifies three main sections of the fitted variogram. The nugget quantifies the uncertainty of the Z values, the range is the section over which there is strong correlation between distance and the Z values and thus represents the distance over which reliable prediction can be made and the semivariance is constant beyond the sill. Kriging is a geostatistical procedure with error estimates being produced that can provide a useful guide as to where more detailed empirical data might be required. This contrasts with the geometrical basis of moving average and inverse distance weighting techniques.

10.4 Concluding remarks

Concern over the presence of significant spatial autocorrelation and contravention of standard assumptions formerly led to a feeling that regression analysis was inappropriate with spatially distributed data. However, over the last few decades three different possibilities have found favour: ignore the spatial autocorrelation especially if it is demonstrably insignificant and carry on with the analysis as normal; acknowledge that the slope coefficient parameter (β) in regression may not apply globally and employ a strategy that allows local best-fit regression lines to be stitched together to produce an overall surface; and incorporate other components in the model that measure the spatial distribution of autocorrelation. This chapter has introduced a selection of spatial analytic and geostatistical techniques that come under the broad heading of **Exploratory Spatial Data Analysis** (ESDA). A focus of attention is the distribution of spatial autocorrelation and recognition that spatial patterns of phenomena and their measured variables often contravene the assumptions of classical statistics. Although trend surface and residuals analysis utilize some of the computation procedures of classical statistics, such as fitting a regression equation, the failure of many standard assumptions to be satisfied, such as independence and conforming to the Normal probability distribution means that it is not necessary to calculate confidence limits for the fitted surface or to apply inferential statistics.

The second of these possibilities, producing and stitching together local best-fit regression equations, has been termed **Geographically Weighted Regression** (GWR) (Fotheringham, Charlton and Brunsdon, 1998; Fotheringham, Brunsdon and Charlton, 2002) and merits further brief discussion, although full details are beyond the scope of this text. GWR allows the estimated slope coefficients to vary spatially across the study area. Unlike the global and local techniques for smoothing data or fitting a surface examined above, GWR seeks to explain a dependent variable in relation to one or more often several independent variables. It therefore corresponds to the classical statistical technique of multiple or multivariate regression. The principal difference is that classical multiple regression produces one equation, with slope coefficients or parameters and an error term that applies across the complete dataset. GWR has the potential to generate separate equations with all these components for

each point or area in the analysis. The analysis proceeds by including the neighbours of each point within a neighbourhood zone usually defined by means of a distance decay function. A weights matrix is defined for each point and the least squares regression is computed. GWR presents a way of quantifying and disentangling complex spatial patterns across a study area and enhances thinking about spatial processes in comparison with applying simple linear regression.