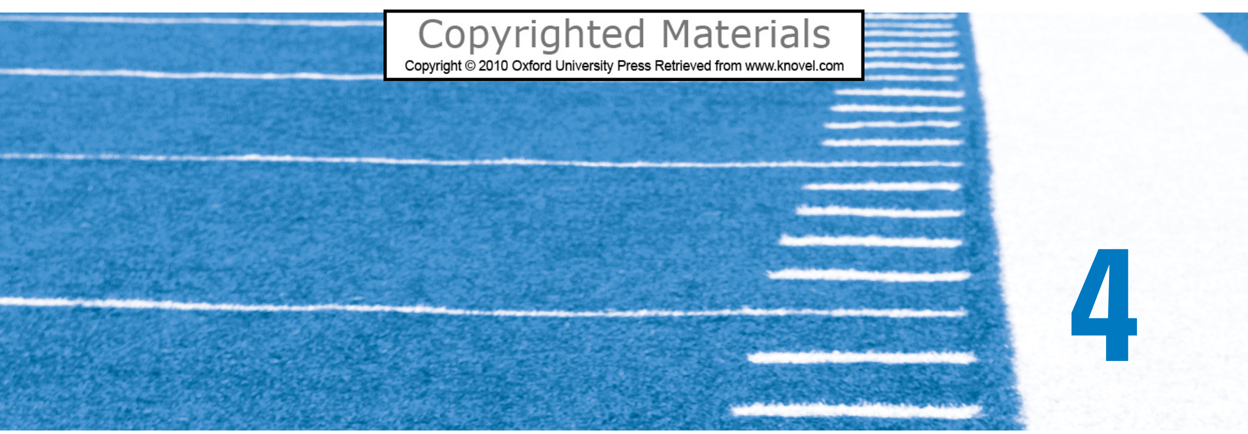


4

Key concepts 3
Spatial data analysis

4.1 Introduction

In the previous chapter, a key focus was on introducing some methods for aspatial data analysis. This chapter builds on that previous discussion and discusses some ways of extracting information from spatially referenced (mappable) data. The chapter details how basic measurements, including lengths, perimeters, and areas, are made in a GIS. Following this, the generation of buffers (distance bands around specified objects) is detailed. Approaches for use with both vector and raster representations of features are considered. Next, the idea of moving windows—whereby some operation (such as computing the mean average) is conducted using local subsets of the data—is introduced. The subject of geographical weights, another core component of many methods detailed in the book, is outlined next. With such approaches, the tendency for nearby values to be more similar than those that are more distant is taken into account. A section on spatial dependence and spatial autocorrelation discusses spatial patterns and some ways of analysing such patterns. The basis of spatial dependence is that values close together in space tend to be more similar than those that are farther apart. This principle has been referred to as the 'first law of geography' (the concept is outlined by Tobler, 1970) and this concept is central to many methods for the analysis of spatial data. The need to consider spatial scale and the form of zones (where used) in any analysis is discussed in the following section. A small section on merging polygons then follows. Finally, the key themes are revisited and summarized.

In short, the objective of the chapter is to introduce the basic components of some key tools for the analysis of spatial data. Once these ideas (and those presented in the previous two chapters) are grasped, all of the background necessary to understand (at a simple level) the rest of the material presented in the book will have been developed. The key components of the chapter can be summarized as follows:

- Measuring distances.
- Measuring lengths and perimeters.
- Buffers—measuring zones of fixed distances around objects.
- Moving windows—mapping how values change from place to place. Moving windows are used in many contexts, including ascertaining the gradient or aspect of the terrain locally.
- Geographical weighting—for a given location, giving more influence to close-by values (e.g. estimating the mean at a given location, but giving greater weight to close-by values than to values further away).
- Spatial dependence and spatial autocorrelation—measuring the degree of similarity in neighbouring values (or values separated by a particular distance).
- The ecological fallacy and the modifiable areal unit problem—making inferences from aggregated data (e.g. numbers of people aged over 65 in an area) and considering changes in results due to changes in the size and shape of zones (e.g. using large administrative zones or smaller zones that fit within them).
- Merging polygons—joining subregions to form new larger regions.

4.2 Distances

Much of spatial data analysis relies on measuring straight line (Euclidean) distances between different locations. The distance, d, between point i and point j is calculated using Pythagoras' theorem:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{4.1}$$

where location i has the coordinates x_i, y_i and location j has the coordinates x_j, y_j. In words, the squared difference between the two x coordinates and the squared difference between the two y coordinates are calculated and added together, and the square root of the product is taken. As an example, take a location with an x coordinate of 10 and a y coordinate of 15, and a second location with an x coordinate of 22 and a y coordinate of 19. The distance between these two locations is obtained from:

$$\sqrt{(10-22)^2 + (15-19)^2} = \sqrt{144+16} = \sqrt{160} = 12.649$$

In some cases, straight line distances may not be meaningful. As well as Euclidean distances, Manhattan distances (also referred to as taxicab distances) are also widely used. These refer to distances along grids and the name derives from the grid-like configuration of streets in Manhattan, New York. Manhattan distances are the sum of distances along each grid segment connecting the start and end points. Network distances are distances along networks. An example is distances along a road network, using the kind of vector structures detailed in Section 2.2. This book details a variety of ways of representing distances. These include friction surfaces, whereby the 'cost' of moving over a particular area of land is taken into account (see Section 10.6).

4.3 Measuring lengths and perimeters

With raster grids, lengths can be measured along cells (as outlined above) or in terms of Euclidean distances between, for example, cell centroids. In the former case, measurement requires information on the spatial resolution of cells and their number. In a simple case, if we are measuring the length along the side of five cells and their spatial resolution is 10 m, then clearly the distance is $5\times10=50$ m. Another way of dealing with distance travelled over raster grids, the use of friction (cost) surfaces, is discussed in Section 10.6. Measurement of lengths of vector features is discussed next.

4.3.1 Length of vector features

Lines can be measured simply by calculating the length of each line segment using Pythagoras' theorem (see Section 4.2) and summing the lengths of each segment that makes up a line. Perimeters of polygons can be measured in the same way by working from one polygon node, around the polygon, and back to the same node. Measurement of line lengths is a common task in GIS contexts. As an example, applications concerning road networks (see Chapter 6) often make use of information on the length of road networks.

4.4 Measuring areas

Measurement of areas with raster grids is straightforward. If n cells belong to a given class then the area of a cell (given by the spatial resolution squared) is simply multiplied by n to get the total area covered by pixels in that class. Measurement of the areas of vector polygons is outlined in the following section. Many applications require information on the areas of zones. As an example, to compute population density in an area both the total population and the area of the zone are required.

4.4.1 Areas of polygons

The area of a polygon can be calculated by:

$$A = 0.5 \times \sum_{i=1}^{n} y_i \times (x_{i+1} - x_{i-1}) \tag{4.2}$$

where x_i and y_i are the x and y locations for node i. In Figure 4.1 a simple polygon feature is shown. The x and y coordinates of its nodes and the calculations following the equation are given in Table 4.1. Note that x_{i+1} refers to the next node in the list and x_{i-1} is the previous one. For the first node (node 1) the previous node is the last node in the list (in this example, node 5).

As an example, we take the y coordinate of node 1 and multiply it by the product of the x coordinate of the next node (obviously, node 2) minus the x coordinate of the previous node (node 5 in this case). Next, we take the y coordinate of node 2 and multiply it by the product of the x coordinate of the next node (node 3) minus the x coordinate of the previous node (node 1 in this case). This is done for each node and the results summed and multiplied by 0.5. Note that the procedure should be followed in a clockwise direction, if it isn't then the area returned will be negative.

In this case the area, A, is given by $0.5 \times 48 = 24$.

The procedure works for any polygon, whatever its degree of complexity. Calculation of areas of polygons is also demonstrated by Kitchin and Tate (2000) and Wise (2002).

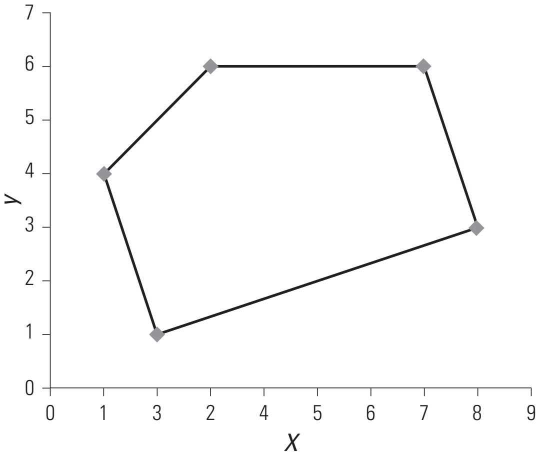


Figure 4.1 Simple polygon feature.

Table 4.1 Simple polygon nodes and area calculations

Node	x_i	y_i	$y_i \times (x_{i+1} - x_{i-1})$
1	2	1	$1 \times (1-8) = -7$
2	1	4	$4 \times (3-2) = 4$
3	3	6	$6 \times (7-1) = 36$
4	7	6	$6 \times (8-3) = 30$
5	8	3	$3 \times (2-7) = -15$
		Sum	48

4.5 Distances from objects: buffers

In applications where selection of objects within a set distance of other objects is the concern, generation of a buffer polygon is a likely step. Buffers are widely used in site-selection projects and in many other contexts where straight line (Euclidean) distances are meaningful. Cases where distance is more logically measured along networks (as would be the case when the distance between places by road is measured), rather than as a straight line between start and end points, are dealt with in Chapter 6.

4.5.1 Vector buffers

A buffer polygon represents the area within a specified distance of an object—that is, if the buffer is computed for a distance of 5 km, the buffer polygon represents a distance of 5 km from the object of interest. Figure 4.2 gives an example of a buffer polygon around a linear feature. The object could also be a point (in which case, the buffer is a circle) or a polygon. Overlay operators, which could be used to identify areas or objects falling within buffer polygons, are discussed in Chapter 5.

One method for generating buffers entails moving a circle with the required radius along the feature to be buffered. The internal boundaries of the overlapping circles can then be dissolved (see Section 4.10 for a discussion about dissolving internal boundaries). Often, buffers are computed for several distance bands and items that fall within each band can be identified using an overlay operator, as detailed in the following chapter. Such an operation is routine in most GIS software. It is also possible to vary the width of the buffer to take into account specific local characteristics. For example, buffers may be wider in areas with steep slopes or in areas that are environmentally sensitive.

4.5.2 Raster proximity

Buffers can also be represented as raster grids. For an input of cells that are coded as locations of interest, a buffer grid can be generated. In such an output, cells within the specified distance of the objects will be given one value (e.g. 1) while cells outside of that area will be given another value (e.g. 0). With the raster model, proximity is often measured directly and each cell in the output records the distance from the cell or cells of interest in the input grid. Figure 4.3A shows such a grid with cell values representing the distance of cells from a linear feature running from the top left to the bottom right of the grid—the feature corresponds to the cells with zero distance

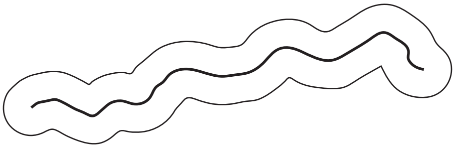


Figure 4.2 Buffer (polygon with light line) around a linear feature (heavy line).

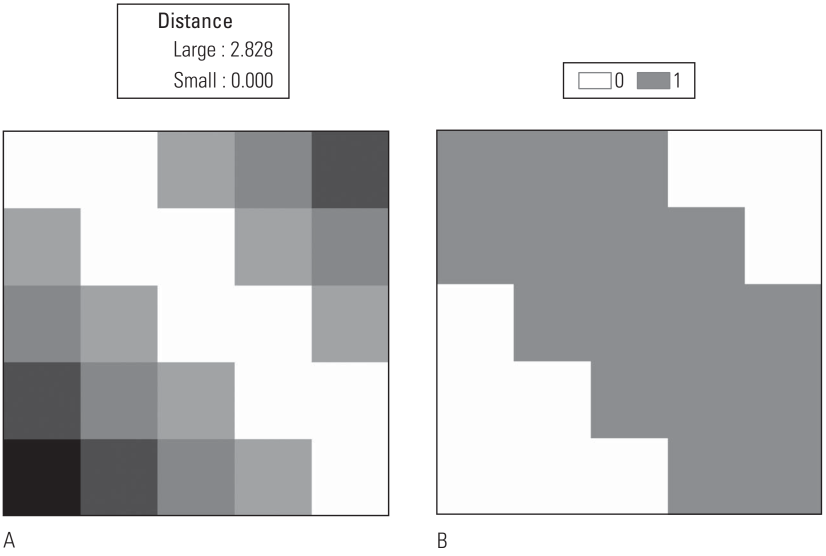


Figure 4.3 Distances from linear feature: (A) distance, (B) binary map for distances less than or equal to 1.4 units (value of 1) or greater (0).

values. A buffer can easily be generated from this grid using a simple classification procedure and Figure 4.3B shows a buffer for distances of less than or equal to 1.4 units. With this approach, all cells with distances of less than, or equal to, the specified amount are coded '1' and all cells with distances greater than this value are coded '0'. It is then straightforward to select all cells in a second image which fall within the buffer defined in the first image. Section 10.2 shows how this can be done. Note that a raster proximity map can be generated directly from vector data, as well as from particular cells in another raster grid.

The following section deals with a key concept in spatial data analysis, the moving window.

4.6 Moving windows: basic statistics in subregions

In many cases, spatial variables have different properties at different locations (for example, values tend to be large in some areas and small in others). In such cases, it is useful to be able to account for these differences and moving windows offer one solution. The idea of the moving window is core to spatial data analysis. In simple terms a moving window represents a region covering part of the entire study area and this region or 'window' is moved from one location to another. Usually, the window is circular or square in shape. In many applications, the window moves in regular steps across the study region, with some operation (e.g. the calculation of the mean average of values in the window) conducted at each location. Such an approach to computing

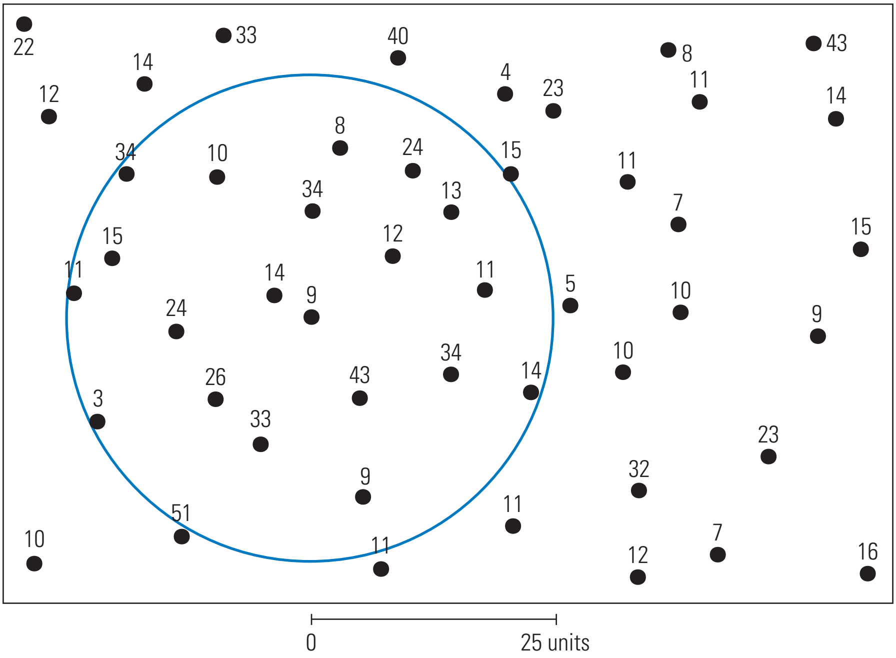


Figure 4.4 Moving window centred on one point: radius of 25 units.

the mean average can be given as follows, for a moving window which is a circle with a radius of 25 units (e.g. metres):

1. Go to location *i*.
2. Calculate the mean average of all values around location *i* that fall within 25 m of that location.
3. Make $i = i + 1$ (the next location is often at some fixed distance away and in a predetermined direction; the locations may be nodes of a regular grid) and go to step 1.

In Figure 4.4, a window with a radius of 25 units is centred on one observation. The window could be centred anywhere in the study region—the location of an observation or elsewhere. In this case, the window contains 20 values and the sum of these values is 381. The mean of these values is $^{381}/_{20} = 19.05$. The approach is straightforward to apply whether the data are regularly spaced (like a grid) or irregularly spaced (like, for example, rain gauges tend to be).

Moving windows are used widely in image processing. Usually each cell in an image (raster grid) is visited in turn and some statistic is computed using that cell and its immediate neighbours (note that this is termed a 'focal operator', as described in Sections 10.3 and 10.4). The mean in a moving window might be used to smooth an image (reduce the effect of local outliers) or the standard deviation in a moving window could be used to highlight the edges of features in an image. A focal operator is illustrated in Figure 4.5, where the mean average of neighbouring pixels is computed. Note that when such procedures are employed, the output image often has fewer rows and columns than the input image. In the example below, the window is 3×3 pixels in size and the mean is only computed where there are neighbours on all sides of a pixel. When the window is centred on pixels at the edge of the image, there are fewer than 3×3 pixels and so no value is computed. The moving window statistic could still be calculated from the smaller number of pixels, but in many cases the procedure employed in this example is followed.

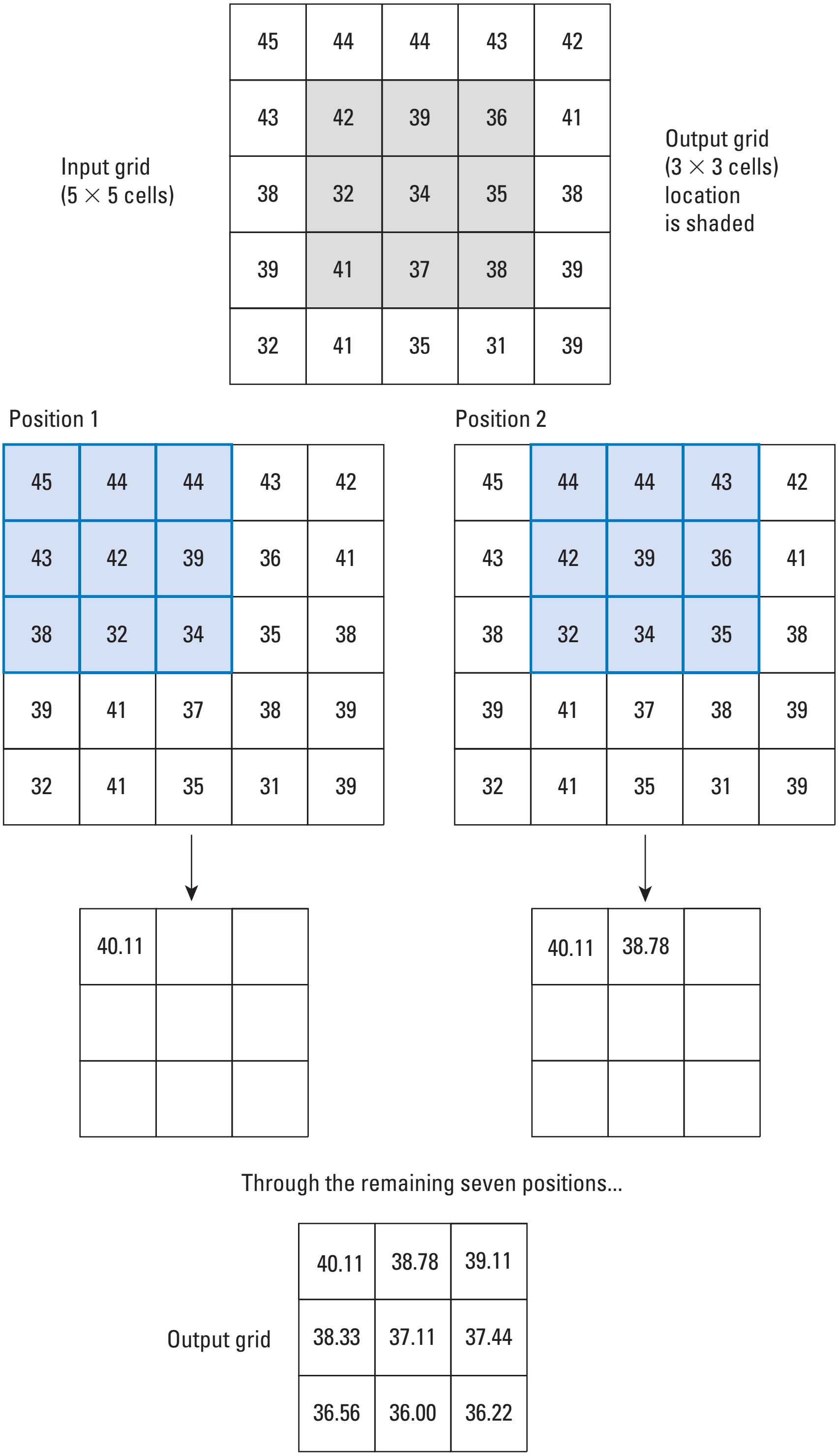
In the case of position 1, the value in the centre of the window is 42 and its neighbouring values are 45, 44, 44, 43, 39, 38, 32, and 34. Adding these values together and dividing the sum by nine gives a value (the mean average) of 40.11, as shown in the top-left cell of the output grid.

The next section extends the moving window idea by treating each of the observations in the window differently according to where they are located.

4.7 Geographical weights

The tendency for observations close together in space to be more similar than observations that are separated by larger distances (see Section 4.1) is often accounted for in spatial analyses. For example, a summary statistic computed in a moving window of a particular size may be based equally on all of the data in the window at a certain position. Alternatively, observations close to the centre of the window may be given more weight (or influence). Logically this is sensible: if the summary statistic is allocated to a point in the centre of the window, it is sensible to allow close-by observations to have most influence on the estimated statistic at that location since these close-by values are most likely to be similar. The objective is to obtain a more reliable statistic as distance to neighbours is taken into account.

Weights can be based on adjacency or they can, for example, be a function of distance. The example application of the Moran's *I* statistic in the following section uses adjacency: neighbouring cells are given a weight of one while all other cells are given a weight of zero, i.e. they are not included in the calculations. Alternatively, all cells (or points/areas) or some subset could be used in the calculations but with larger weights given to cells closer to the cell of interest. There is a large variety of weighting functions that determine how much weight should be given to observations as a function of distance. A simple linear weighting function could be used whereby an observation twice as far away receives half as much weight, e.g. an observation at 10 km receives twice as much weight as an observation 20 km away. In practice, more sophisticated schemes are used for most applications. One well-known weighting function is based on taking the inverse of the squared distance from the location of interest (following the inverse square law). In other words, the weight is a function of (is dependent on) the inverse squared distance, d^{-2} (this can be obtained with $^1/d^2$, as detailed below, and see Section 9.5 for an application of this weighting function). Whatever distance decay weighting



45	44	44	43	42
43	42	39	36	41
38	32	34	35	38
39	41	37	38	39
32	41	35	31	39

Position 1

45	44	44	43	42
43	42	39	36	41
38	32	34	35	38
39	41	37	38	39
32	41	35	31	39

40.11		

Position 2

45	44	44	43	42
43	42	39	36	41
38	32	34	35	38
39	41	37	38	39
32	41	35	31	39

40.11	38.78	

Through the remaining seven positions...

Output grid

40.11	38.78	39.11
38.33	37.11	37.44
36.56	36.00	36.22

Figure 4.5 Mean average computed for a 3×3 pixel moving window.

scheme is used, observations at smaller distances have larger weights than observations at larger distances from the location of interest. Some other weighting functions are described later in this book, but a general summary of distance weighting is provided below.

The inverse distance weighting scheme is illustrated now. The weight for location *i* can be given by w_{ij}, indicating the weight of sample *i* with respect to location *j*. The inverse distance weight is given by:

$$w_{ij} = d_{ij}^{-k} \tag{4.3}$$

which indicates that the weight for location *i* with respect to location *j* is obtained by raising the distance *d* between locations *i* and *j* (i.e. d_{ij}) to the power $-k$. As noted above in the case of $k=2$, this is obtained with $^1/_{d^k}$. The inverse distance weighting scheme is illustrated in Figure 4.6. The value of the exponent determines the degree of weighting by distance. With larger exponent values, the weights decline more sharply with distance, whereas with smaller exponent values distant observations receive, relatively, larger weights. Note that, with an exponent of zero, all of the weights are equal to one. An application of the inverse distance weighting scheme is outlined below.

Different forms of weighting scheme have found favour in particular contexts. For example, the Gaussian weighting scheme described in Section 8.4 has been used for weighting observations as a part of a method called geographically weighted regression, which is detailed in Section 8.5.3, while the quartic weighting scheme (see Section 7.3.2) has been used for point pattern analysis. Inverse distance weighting is the basis of a spatial interpolation method (a method for predicting values at unsampled locations), which is discussed in Section 9.5 and is illustrated briefly here.

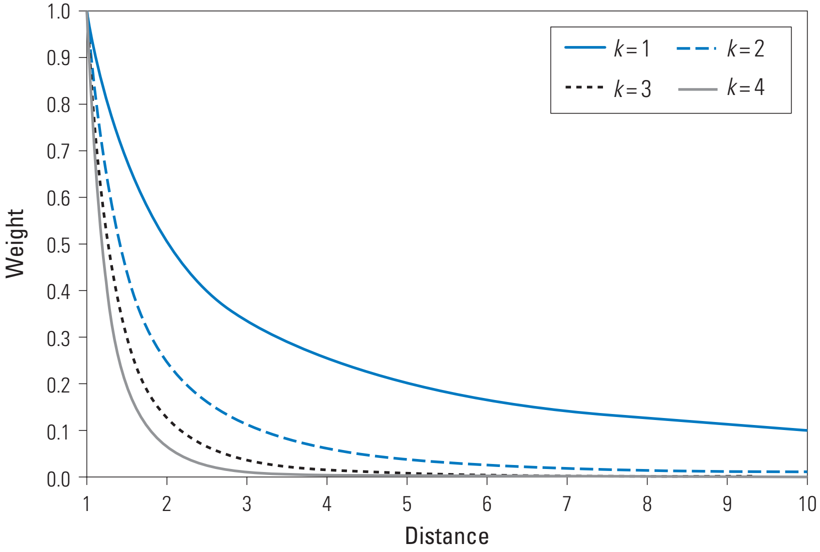


Figure 4.6 Inverse distance weighting scheme for exponents (*k*) of 1, 2, 3, and 4.

Any standard statistic can be geographically weighted (see Fotheringham *et al.* (2002) for more information). As an example of a geographical weighting scheme in practice, obtaining the locally weighted mean using inverse distance is illustrated below. The locally weighted mean is given by:

$$\bar{z}_i = \frac{\sum_{j=1}^{n} z_j w_{ij}}{\sum_{j=1}^{n} w_{ij}} \tag{4.4}$$

Recall from Equation 3.3 that $\bar{z}$ indicates the mean of z, so $\bar{z}_i$ indicates the mean at location i; w_{ij} indicates, as before, the weight for the distance between observations i and j. In the case where all the weights are one, Equation 4.4 corresponds to the standard mean average (it is then the sum of the values divided by the number of observations).

In Table 4.2, a set of observations (which can be treated as measurements of precipitation in millimetres for illustrative purposes) and their distance from a fixed location are given. In this case, the value at the fixed location is unknown and it will be

Table 4.2 Observations (j), distance from observation 1 (d_{ij}), weights (w_{ij}), and weights multiplied by values ($z_j w_{ij}$)

			k=1		k=2		k=3	
j	d_{ij}	z_j	w_{ij}	$z_j w_{ij}$	w_{ij}	$z_j w_{ij}$	w_{ij}	$z_j w_{ij}$
1	4.404	14	0.227	3.179	0.052	0.722	0.012	0.164
2	9.699	43	0.103	4.434	0.011	0.457	0.001	0.047
3	10.408	12	0.096	1.153	0.009	0.111	0.001	0.011
4	10.871	34	0.092	3.127	0.008	0.288	0.001	0.026
5	12.958	26	0.077	2.007	0.006	0.155	0.000	0.012
6	13.959	24	0.072	1.719	0.005	0.123	0.000	0.009
7	14.066	33	0.071	2.346	0.005	0.167	0.000	0.012
8	15.506	34	0.064	2.193	0.004	0.141	0.000	0.009
9	17.256	10	0.058	0.579	0.003	0.034	0.000	0.002
10	17.606	8	0.057	0.454	0.003	0.026	0.000	0.001
11	18.018	13	0.055	0.721	0.003	0.040	0.000	0.002
12	18.025	11	0.055	0.610	0.003	0.034	0.000	0.002
13	18.285	24	0.055	1.313	0.003	0.072	0.000	0.004
14	19.253	9	0.052	0.467	0.003	0.024	0.000	0.001
15	21.335	15	0.047	0.703	0.002	0.033	0.000	0.002
16	23.845	14	0.042	0.587	0.002	0.025	0.000	0.001
17	23.988	34	0.042	1.417	0.002	0.059	0.000	0.002
18	24.464	3	0.041	0.123	0.002	0.005	0.000	0.000
19	24.522	11	0.041	0.449	0.002	0.018	0.000	0.001
	Sum	372	1.347	27.582	0.128	2.533	0.017	0.309
	Mean	19.579		20.474		19.844		17.804

Weights are obtained using the inverse distance weighting scheme with exponents of 1, 2, and 3.

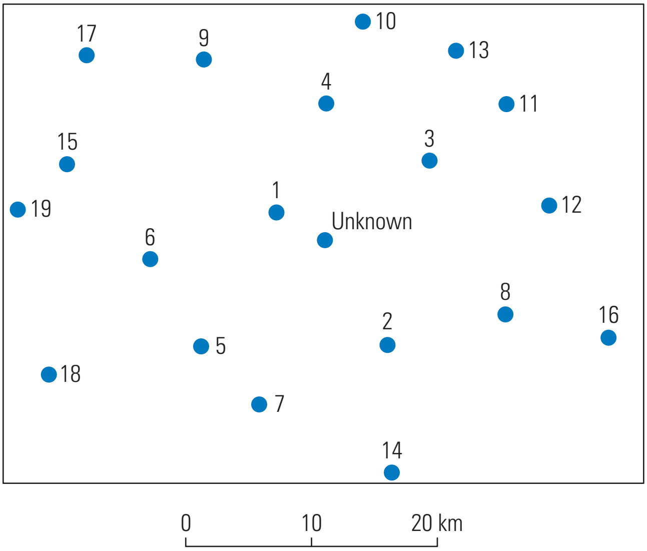


Figure 4.7 Locations of observations listed in Table 4.2.

predicted using inverse distance weighting. Figure 4.7 shows the locations of the observations, along with the location for which a prediction will be made. The weights, obtained using the inverse distance weighting scheme detailed above (with exponents of 1, 2, and 3), are given for each distance. The weights for each location are then multiplied by the value at that location. As an example, following Equation 4.4 (and with weights as defined in Equation 4.3) for an exponent of 1, the products of the multiplications are summed, giving a value of 27.582. The weight values are also summed, giving a value of 1.347. The weighted mean is then obtained as $^{27.582}/_{1.347}=20.474$. The mean obtained without geographical weighting (i.e. with all weights equal to 1) is 19.579.

Weighted means (or other statistics) can be calculated anywhere: at the location of an observation or anywhere else. Section 8.4 demonstrates the geographically weighted mean using another weighting scheme. Fotheringham *et al.* (2002) discuss a range of geographically weighted statistics. Geographical weights will be encountered throughout this book.

Selection of a weighting scheme is usually arbitrary, but there may be characteristics of a data set that guide selection. In essence, the choice of weighting function (and, where relevant, parameters like the bandwidth, see Sections 7.3 and 8.4) should be determined either through experimentation (e.g. for interpolation, which weighting function leads to the most accurate predictions) or knowledge of the process of interest—if we know something about the scale of variation (see Section 2.7) this may inform our choice of weighting scheme.

The focus now moves from measurement of distances to characterizing the spatial structure of values (i.e. how similar neighbouring values are to one another).

4.8 Spatial dependence and spatial autocorrelation

A key concern in spatial data analysis is to examine spatial patterning in the variable or variables of interest. For example, are values of a particular variable large in some areas and small in others? Also, do similar values tend to cluster or are values visually erratic? The term 'spatial dependence' refers to the dependence of neighbouring values on one another (Haining, 2003). As outlined at the start of this chapter, the basis of spatial dependence is that values close together in space tend to be more similar than those that are farther apart. The 'first law of geography' (Tobler, 1970) is a key concept in geography in general and spatial data analysis in particular. In the context of statistical measurement, this idea is related to spatial autocorrelation—the degree to which a variable is spatially correlated with itself. A measure of spatial autocorrelation may suggest spatial dependence (i.e. neighbouring values are similar—positive spatial autocorrelation) or spatial independence (neighbouring values are dissimilar—negative spatial autocorrelation).

There is a range of measures of spatial autocorrelation. The joins count approach is one means of summarizing the tendency of neighbouring observations to be the same (O'Sullivan and Unwin, 2002). The measure of spatial autocorrelation encountered most frequently in the spatial analysis literature is the *I* coefficient proposed by Moran (Moran, 1950; Cliff and Ord, 1973). It is given by

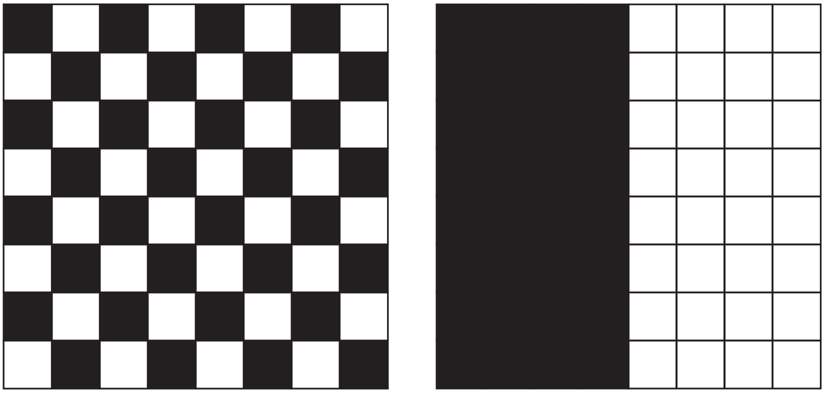
$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\right)} \tag{4.5}$$

where the values y_i (of which there are n) have the mean $\bar{y}$ and the proximity between locations i and j is given by w_{ij}. As before, this is a geographical weight and is often set to 1 when locations i and j are neighbours and 0 when they are not. Note that here y is a data value and not a coordinate. Elsewhere in this book, z is used to represent data values but y is used here to distinguish the use of z as a deviation of y from its mean in the local spatial autocorrelation measures detailed in Section 8.4.1. Equation 4.5 includes double summations (note that single summation was introduced with respect to Equation 3.1), that is:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}$$

This means start with $i=1$ and $j=1$, next, $i=1$ and $j=2$ then $i=1$ and $j=3$, and so on until $j=n$. After that point, $i=2$ and we work through all values of j until all combinations of i and j have been accounted for. At each stage, the computed values are added to the values obtained previously. In this way all combinations of i and j are included. With the numerator of Equation 4.5

$$n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})$$



Negative spatial autocorrelation
$I = -1.000$

Positive spatial autocorrelation
$I = 0.865$

Figure 4.8 Spatial autocorrelation: rook's case contiguity. Black cells have a value of 1, white cells have a value of 0.

the calculations are computed for every combination of *i* and *j* and the results of each calculation are added together, with the product being multiplied by the weight w_{ij}. The procedure to calculate the Moran's *I* value is demonstrated below. Negative values of *I* indicate negative spatial autocorrelation—neighbouring values tend to be different. Positive values of *I* indicate positive spatial autocorrelation—neighbouring values tend to be similar. Values of *I* close to zero indicate that there is no structure. This section first outlines the basic concepts, then gives a small, fully worked numerical example, and finally gives an example using a larger grid.

Values of *I* for two different grids are given in Figure 4.8. These examples were computed using the package GeoDa (Anselin *et al.*, 2006) and it should be noted that the results are slightly different to those that would be calculated using Equation 4.5 as GeoDa modifies the form of the weights (in that package the weights are what are termed 'row standardized', i.e. they sum to 1), but that isn't a concern here. In this example, black cells have a value of 1 while white cells have a value of 0. For this example, cells which share an edge with another cell are compared and not cells which share only corners. Through the analogy with movement of pieces in chess, this is called rook's case contiguity. Where cells which share corners (i.e. cells connected diagonally) are also included, this is called queen's case contiguity. Rook's case and queen's case contiguity are illustrated in Figure 4.9. In the case of irregularly shaped zones, rook's case contiguity and queen's case contiguity can also be used, with the latter including zones that are connected only by vertices as well as by edges, while the former includes only zones joined by edges. In packages such as GeoDa, different weighting functions (e.g. rook's case and queen's case) can be used and it is necessary to consider which is used and how this may impact on the results. As well as rook and queen contiguity, other weighting schemes can be used (see Section 4.7).

The case on the left of Figure 4.8 indicates negative spatial autocorrelation—all neighbours of a given cell are, using rook's case contiguity, different to that cell. In the

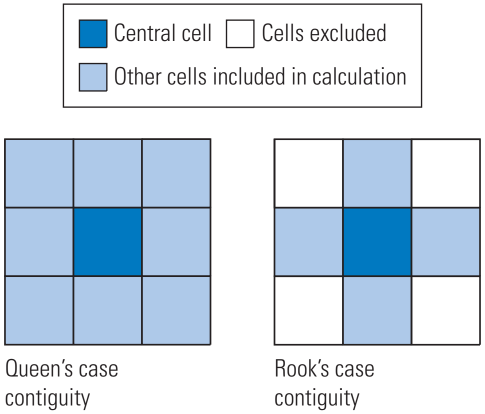


Figure 4.9 Queen's case contiguity and rook's case contiguity.

case on the right, the neighbours of most cells have the same value and, therefore, the values are positively spatially autocorrelated. The only exceptions are the cells in the middle two columns of the grid, which have different values.

Using a small grid of values, Moran's *I* is illustrated below. Note that the method is equally applicable to zones with irregular forms. The sample grid is:

7	8	11
11	9	10
11	12	9

Values that are next to one another along rows or columns (e.g. 7 and 8 or 9 and 10) will be counted as neighbours as will those that are next to one another diagonally (e.g. 8 and 10). As noted above, this is called queen's case contiguity.

First we will calculate $(y_i - \bar{y})(y_j - \bar{y})$—that is, the difference of each value from the mean multiplied by the difference between each neighbouring value and the mean. For example, the value 7 (top left cell) minus the mean (9.778) is –2.778. One of the neighbours of this cell has the value 8 and its difference from the mean is –1.778. We then multiply the two differences together, giving 4.938 (see the last entry of the top row in Table 4.3). This is done for every cell and its neighbours, as shown in Table 4.3. Note that Table 4.3 includes only cells that are neighbours (so the weight in each case is 1). The sum of the products, $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})$, is 3.975.

Next we will calculate $(y_i - \bar{y})^2$, the squared difference between each value and the mean. The results are shown in Table 4.4. The sum of squared differences from the mean $(\sum_{i=1}^{n}(y_i - \bar{y})^2)$ is 21.556.

There are nine observations, $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y}) = 3.975$, the sum of squared differences from the mean is 21.556 and there are 20 adjacencies (the number of rows in Table 4.3 is twice the number of adjacencies). As an example, the cells with values 7 and 8 are neighbours (i.e. they are adjacent to one another). Each adjacency (like all the others) is counted twice as we have 7 paired with 8 and 8 paired with 7.

Table 4.3 Value, value – mean (9.778), neighbour value, neighbour value – mean, product (two differences from mean multiplied together)

Value y_i	Value – mean $y_i - \bar{y}$	Neighbour value y_j	Neighbour – mean $y_j - \bar{y}$	Product $(y_i - \bar{y})\,(y_j - \bar{y})$
7	–2.778	8	–1.778	4.938
7	–2.778	9	–0.778	2.160
7	–2.778	11	1.222	–3.395
11	1.222	7	–2.778	–3.395
11	1.222	8	–1.778	–2.173
11	1.222	9	–0.778	–0.951
11	1.222	12	2.222	2.716
11	1.222	11	1.222	1.494
11	1.222	11	1.222	1.494
11	1.222	9	–0.778	–0.951
11	1.222	12	2.222	2.716
8	–1.778	7	–2.778	4.938
8	–1.778	11	1.222	–2.173
8	–1.778	9	–0.778	1.383
8	–1.778	10	0.222	–0.395
8	–1.778	11	1.222	–2.173
9	–0.778	7	–2.778	2.160
9	–0.778	8	–1.778	1.383
9	–0.778	11	1.222	–0.951
9	–0.778	10	0.222	–0.173
9	–0.778	9	–0.778	0.605
9	–0.778	12	2.222	–1.728
9	–0.778	11	1.222	–0.951
9	–0.778	11	1.222	–0.951
12	2.222	11	1.222	2.716
12	2.222	11	1.222	2.716
12	2.222	9	–0.778	–1.728
12	2.222	10	0.222	0.494
12	2.222	9	–0.778	–1.728
11	1.222	8	–1.778	–2.173
11	1.222	9	–0.778	–0.951
11	1.222	10	0.222	0.272
10	0.222	11	1.222	0.272
10	0.222	8	–1.778	–0.395
10	0.222	9	–0.778	–0.173
10	0.222	12	2.222	0.494
10	0.222	9	–0.778	–0.173
9	–0.778	12	2.222	–1.728
9	–0.778	9	–0.778	0.605
9	–0.778	10	0.222	–0.173

Table 4.4 Values, difference from the mean (9.778) and the squared differences

Value y_i	Difference $y_i - \bar{y}$	Squared difference $(y_i - \bar{y})^2$
7	–2.778	7.716
11	1.222	1.494
11	1.222	1.494
8	–1.778	3.160
9	–0.778	0.605
12	2.222	4.938
11	1.222	1.494
10	0.222	0.049
9	–0.778	0.605

The sum of the weights therefore, $\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$ (the right-hand side of the denominator of Equation 4.5), is 40. Moran's I is computed by:

$$I = \frac{9 \times 3.975}{21.556 \times 40} = \frac{35.778}{862.222} = 0.041$$

Since this value is close to 0, this indicates that neighbouring values in the example do not tend to be similar. Another example is given which indicates negative spatial autocorrelation. For the grid:

7	8	6
10	14	7
6	11	9

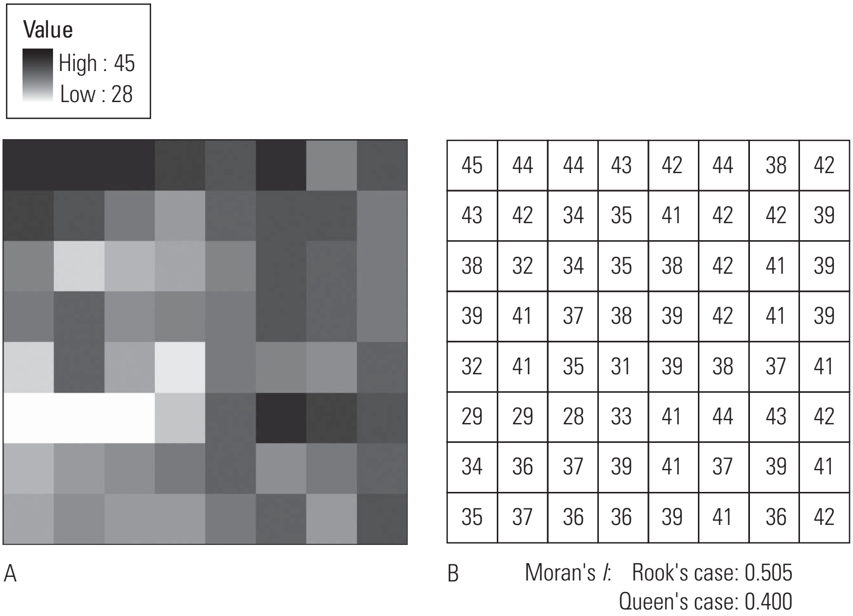
This leads to:

$$I = \frac{9 \times -66.889}{56.000 \times 40} = \frac{-602.000}{2240.000} = -0.269$$

In this case, neighbouring values tend to be dissimilar, thus no clustering of like values is suggested.

It will be obvious that increasing the size of the grid will necessitate the use of a computer to obtain a value of Moran's I. A further example of I is given in Figure 4.10. Like the previous example (Figure 4.9), the calculations were conducted using GeoDa. Figure 4.10 shows the potential difference in results obtained using rook's case (using only cells joined by edges) or queen's case (using cells joined by edges or by corners/vertices).

As noted in Section 3.5, spatial autocorrelation has an impact on standard statistical procedures. In essence, a large sample size gives greater confidence in the inferences we make than a small sample size; this is intuitively obvious. If neighbouring values are similar then the observations are considered dependent on one another. A practical implication is that the degree of confidence in our results will be smaller than the



45	44	44	43	42	44	38	42
43	42	34	35	41	42	42	39
38	32	34	35	38	42	41	39
39	41	37	38	39	42	41	39
32	41	35	31	39	38	37	41
29	29	28	33	41	44	43	42
34	36	37	39	41	37	39	41
35	37	36	36	39	41	36	42

Figure 4.10 Example raster using (A) grey scales and (B) numerical values: spatial autocorrelation for queen's case and rook's case contiguity.

sample size suggests, since the observations are not independent of one another. As noted previously, it is important to remember that we must be wary of such problems when applying standard statistical procedures (e.g. significance tests) in the analysis of spatial data (see Rogerson (2006) for a discussion about this topic).

There is a variety of other methods for measuring the degree of spatial autocorrelation. The purpose of this chapter is only to introduce topics and the analysis of spatial autocorrelation is explored further in Chapter 8.

4.9 The ecological fallacy and the modifiable areal unit problem

It is often necessary to work with spatially aggregated data, for example census zones or cells in remotely sensed images. Such zones are unlikely to be internally homogeneous. For example, a cell in a remotely sensed image has only one value, but in the real world there may be several features in the area covered by the cell. In words, the variation within the cell (or other area) is lost if the area is larger than the individual features it contains. This section explores two sets of concepts that relate to such issues. These are the ecological fallacy and the modifiable areal unit problem.

The ecological fallacy refers to the problem of making inferences about individuals from aggregate data. For example, not all people in one census zone are likely to share the same characteristics. The majority of people in a census zone may be wealthy,

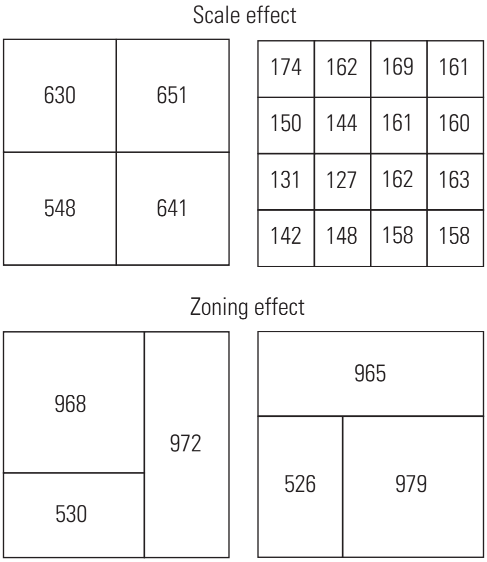


Figure 4.11 The scale and zoning effects.

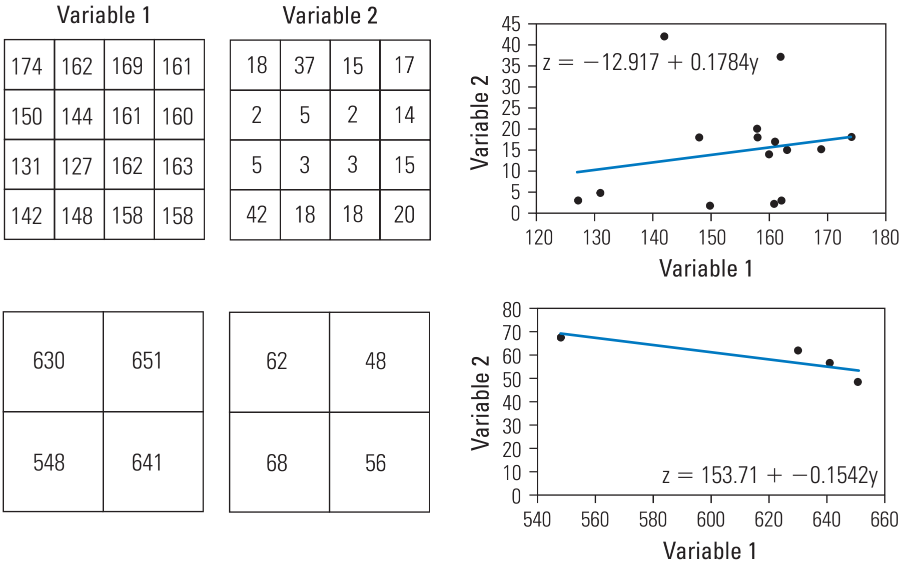
but if there is a housing estate just inside one edge of the zone then clearly generalizations about the population of the zone may be unsound. Geographical space can be divided in an infinite number of ways. In practice, it is often the case that data aggregated over only one set of areal units are available. It may be that no set of zones has intrinsic meaning about the underlying populations and that the units are 'modifiable'. This problem is sometimes termed the 'modifiable areal unit problem' (MAUP) (Openshaw and Taylor, 1979). The MAUP is composed of two parts:

The scale effect Statistical analyses based on data aggregated over areas of different sizes will produce different results.

The zoning effect Two sets of zones can have the same or similar areas but very different forms and analyses based on two such sets of zones may vary.

The scale effect and the zoning effect are illustrated in Figure 4.11. In this example, the values (representing numbers of people in each area) shown in Figure 4.10 have been aggregated into new larger zones.

Note that different terms are used in the literature to refer to the two component parts of the MAUP. For example, for what is termed here the 'scale effect' the term 'aggregation effect' is sometimes used (e.g. Atkinson and Tate, 2001; Waller and Gotway, 2004). To add to the confusion, the zoning effect is sometimes also referred to as the aggregation effect (e.g. Openshaw and Taylor, 1979). Other terms are also encountered for both components of the MAUP but, irrespective of this, the key distinction between the two components is that one deals with the size of zones (here called the scale effect) and the other with changes in their shape or position when the size of zones is the same or similar (here called the zoning effect).



Variable 1

174	162	169	161
150	144	161	160
131	127	162	163
142	148	158	158

Variable 2

18	37	15	17
2	5	2	14
5	3	3	15
42	18	18	20

630	651
548	641

62	48
68	56

Figure 4.12 Two scatter plots and fitted lines for different aggregations of the same values.

Wong (1997) explores changes in measures of residential segregation (i.e. the degree to which members of different groups live in different areas). Wong argues that, if the counts of the population group are negatively spatially autocorrelated (neighbouring zones are dissimilar), using zones of different sizes will result in different segregation measure values. Conversely, if the counts are positively spatially autocorrelated (neighbouring zones are similar), then using zones of different sizes will make little difference for zones smaller than the area (or scale) over which counts are positively autocorrelated. In general, the degree of spatial autocorrelation is important when considering the effect of changing the zonal system used.

Figure 4.12 gives an artificial example of the potential effects of altering the aggregation of values. In this case, two sets of variables are given for two different aggregations. Regression of one variable on the other is then conducted using the two sets of aggregations. Note that the sample size is *very* small and this example is used purely for illustrative purposes. Recall from Section 3.4 that assessing sample size is important when considering regression results. In the example, the slope changes from positive to negative as the values are aggregated over larger units. Even where the sign doesn't change, the effects of changing the size or form of zones may be highly significant. Openshaw and Taylor (1979) explore the issue using regression for many different zonal systems and they demonstrate large differences where the form of zones varies but their number is the same (the zoning effect) and where the number of zones varies (the scale effect). Note that r^2 tends to increase with increased aggregation, and this would clearly be the case given the example in Figure 4.12.

In summary, the potential impact of the size and shape of zones on results should be considered and it should be remembered that any pattern apparent in mapped areal data may be due as much to the zoning system used as to the underlying distribution

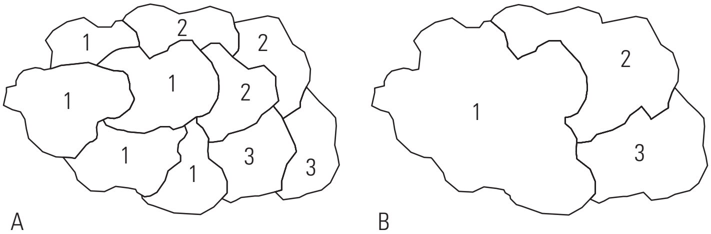


Figure 4.13 (A) Original polygons and (B) polygons with internal boundaries dissolved.

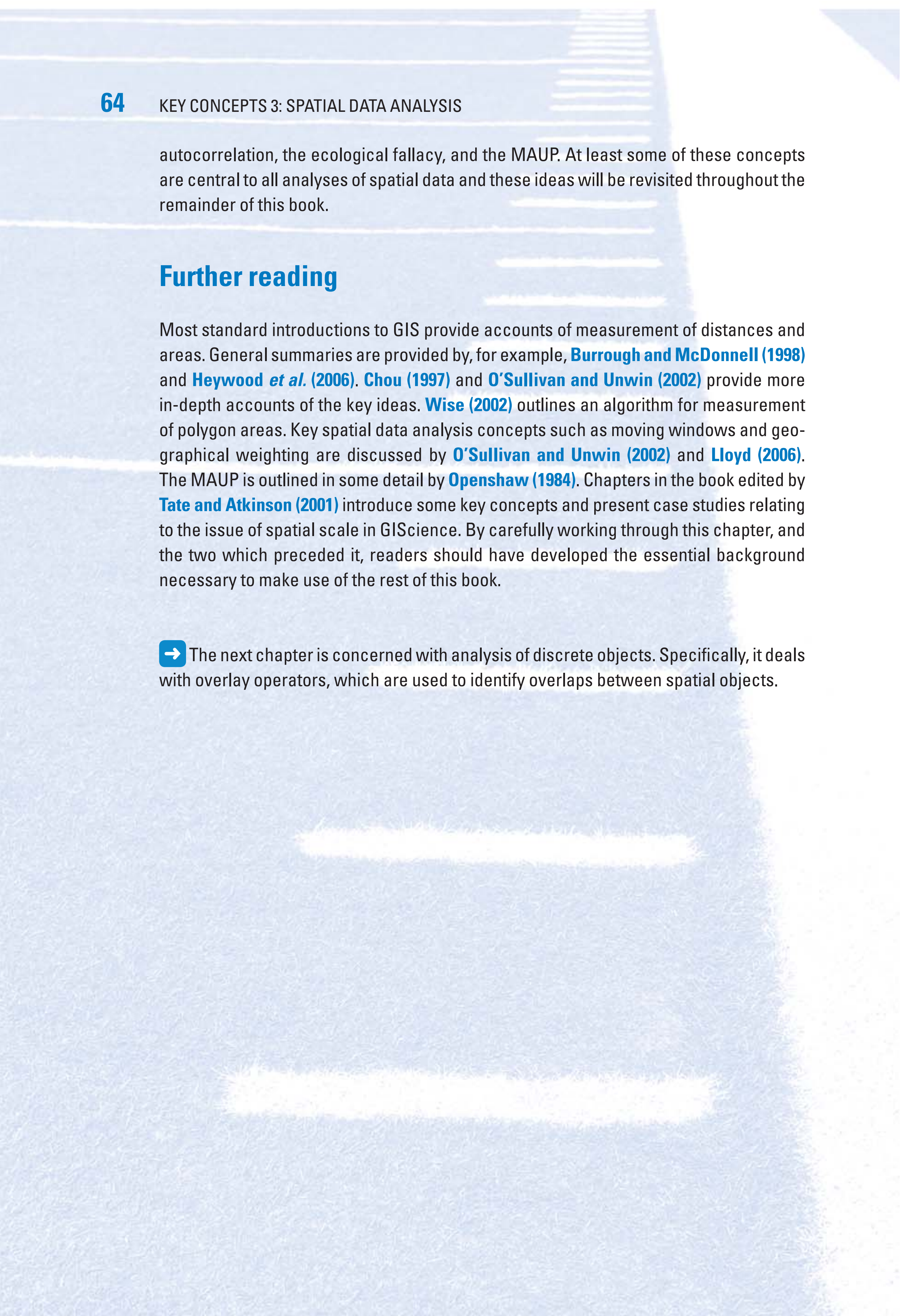
of the variable (Martin, 1996). Raster cells can also be conceptualized as zones and the spatial resolution of a raster will similarly determine results.

4.10 Merging polygons

Related to the previous theme, a common operation in GIS contexts is the merging of polygons with common attributes (this is sometimes termed the 'dissolve operator'). For example, if we have sets of areas all of which have the same area code then we may wish to merge those areas where their boundaries are adjacent. If there are several zones within one larger administrative area, all of which have the same zone identity, then if we dissolve the internal boundaries (i.e. the boundaries of the small zones) we are left with the boundaries of the larger administrative areas. Figure 4.13 gives an example.

Summary

This chapter was concerned with some very basic operations, but such methods form the core of many GIS-based analyses. Measurement of straight line (Euclidean) distances, an initial focus of this chapter, is sensible in many contexts; in others they may not be meaningful. For example, the movement of airborne pollutants is a function of various factors, such as wind direction, and simple distance may explain such a process only to a very limited degree. Also, buffers may only be useful in particular contexts. If we are concerned with accessibility of a particular location in a populated area then measuring the straight line distance of places to that location may not be very helpful. Instead, we may wish to measure distances (and perhaps calculate travel times) along a road network (perhaps using the approach detailed in Section 6.5). Other ways of accounting for the 'cost' of moving from one place to another include cost (or friction) surfaces and this idea is discussed in Section 10.6. Following the discussion about distances, areas, and buffers, some particular concepts in spatial analysis were introduced. These included moving windows, geographical weights, spatial dependence and spatial



autocorrelation, the ecological fallacy, and the MAUP. At least some of these concepts are central to all analyses of spatial data and these ideas will be revisited throughout the remainder of this book.

Further reading

Most standard introductions to GIS provide accounts of measurement of distances and areas. General summaries are provided by, for example, **Burrough and McDonnell (1998)** and **Heywood *et al.* (2006)**. **Chou (1997)** and **O'Sullivan and Unwin (2002)** provide more in-depth accounts of the key ideas. **Wise (2002)** outlines an algorithm for measurement of polygon areas. Key spatial data analysis concepts such as moving windows and geographical weighting are discussed by **O'Sullivan and Unwin (2002)** and **Lloyd (2006)**. The MAUP is outlined in some detail by **Openshaw (1984)**. Chapters in the book edited by **Tate and Atkinson (2001)** introduce some key concepts and present case studies relating to the issue of spatial scale in GIScience. By carefully working through this chapter, and the two which preceded it, readers should have developed the essential background necessary to make use of the rest of this book.

➔ The next chapter is concerned with analysis of discrete objects. Specifically, it deals with overlay operators, which are used to identify overlaps between spatial objects.